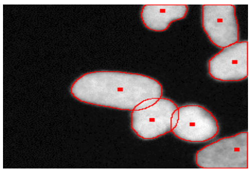


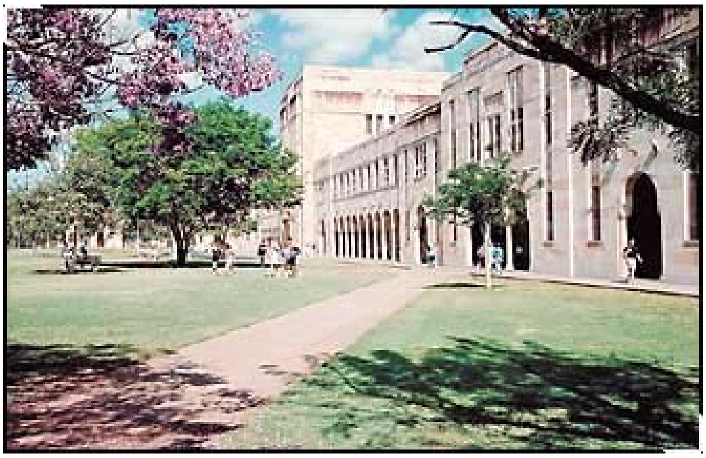
WDIC 2003

PROCEEDINGS OF THE 2003 APRS WORKSHOP ON DIGITAL IMAGE COMPUTING

Theme: Medical Applications of Image Analysis

7th February 2003

University of Queensland, St Lucia, Brisbane, Australia



Hosted by
The Australian Pattern Recognition Society
www.aprs.org.au

Member for Australia of the
International Association for Pattern Recognition
www.iapr.org

Edited by
Brian C. Lovell[1] and Anthony J. Maeder[2]
[1]School of Information Technology and Electrical Engineering
The University of Queensland
[2]School of Electrical and Electronic Systems Engineering
Queensland University of Technology

Proceedings of the 2003 APRS Workshop on Digital Image Computing (WDIC 2003)
Theme: Medical Applications of Image Analysis

Editors
Brian C. Lovell
Anthony J. Maeder

Published by
The University of Queensland
St Lucia, QLD 4072 Australia

First Published in February 2003

Printed in Brisbane

ISBN 0-9580255-2-5

PROCEEDINGS OF THE 2003 APRS WORKSHOP ON DIGITAL IMAGE COMPUTING (WDIC 2003)

THEME: MEDICAL APPLICATIONS OF IMAGE ANALYSIS

Organising Commitee

General Chair: Brian Lovell, UQ
Technical Chair: Anthony Maeder, QUT
Publicity: Clinton Fookes, QUT

Technical Committee

Ben Appleton, UQ
Pascal Bamford, CSSIP
Wageeh Boles, QUT
Andrew Bradley, CSSIP
Terry Caelli, U ALBERTA
Vinod Chandran, QUT
Vaughan Clarkson, UQ
Peter Kootsookos, UQ
Kurt Kubik, UQ
Geoff McLachlan, UQ
Binh Pham, QUT
Larry Spitz, DOCREC CORP
David Suter, MONASH U
Hugues Talbot, CSIRO CMS
Geoff West, CURTIN U
Gordon Wyeth, UQ

WELCOME FROM THE GENERAL CHAIR

Traditionally the APRS has organised a technical meeting every year since its inception in 1990 — a major 3-day refereed conference (DICTA) in odd numbered years and a workshop in even numbered years. This tradition was broken last year when we delayed DICTA by two months to run back to back with ACCV2002. This workshop is therefore what would have been the "even year" meeting, again slipped by a few months. Our plans are to run the next DICTA before the end of 2003, so that we are resynchronised.

After DICTA2002, a membership poll was conducted to determine whether members wanted the workshop to meet the same reviewing standards as DICTA, so that papers receive full academic credit. This motion was overwhelmingly supported, so WDIC2003 was run as an internationally peer reviewed conference with electronic submission, reviewing and publication.

The theme for the keynote address and first oral session is "Medical Applications of Image Analysis." Papers that are of general interest to the Pattern Recognition and Computer Vision Community appear in the other sessions.

We received a very large number of submissions despite the late advertising and registrations are also strong. As per APRS tradition, registrants at WDIC2003 are given one-year membership of the APRS which includes the newsletter and discounts on APRS and IAPR technical events. I look forward to an exciting technical program and to meeting you all at the workshop.

Finally, I would like to take this opportunity to thank Anthony Maeder and Clinton Fookes for their excellent support in organising this event. I would also like to express my gratitude to the members of the Technical Committee for their very speedy responses to my reviewing requests just before Christmas.

Hope you enjoy WDIC2003!

Brian Lovell
General Chair of WDIC2003
President of the Australian Pattern Recognition Society

FOREWORD FROM THE TECHNICAL CHAIR

WDIC2003 is the sixth in a series of bi-annual specialist workshops organised by APRS, intended to focus on current work-in-progress in topical areas of interest amongst our members. In keeping with the APRS aspirations for broad coverage of topics relevant to pattern recognition and engagement with a variety of application areas, WDIC2003 was established with a central theme of Medical Imaging, but submissions in other related areas were also invited. It is pleasing to note that the result is an event where the chosen theme is strong and yet is counterbalanced by a similar volume of coverage of other topics. In this way we hope that interests of most members can be addressed at the event.

The theme of Medical Imaging is one that is very popular in the international arena, with major annual meetings run by SPIE, ACR and many others. In Australia, APRS has been associated with previous workshops in this topic run in Ballarat (1998) and Gold Coast (1999), under ARC sponsorship. At these meetings, a network of contacts was established between researchers in university, government and health sectors, and much information was shared on current work and projects. We hoped that WDIC2003 provides an opportunity to maintain some of those links.

The acceptance of papers for WDIC2003 followed a rigorous process to comply with DEST category E1 requirements. All papers were submitted in full for independent review by 3 referees, chosen for their membership of the technical committee according to the relevance of their expertise to the workshop. They were drawn from a wide range of institutions, both within Australia and overseas, and are nationally recognised as qualified experts in their areas. An overall acceptance rate approaching 90% of submissions was achieved, and comments from reviewers were returned to authors where appropriate to allow improvements to be made to their papers for final publication. The resulting proceedings present a nationally significant body of work, with authors from 4 different states represented.

The decision to produce the full proceedings on CD only for this workshop is a new step for APRS, in line with intentions to boost our electronic media and web presence for further outreach to the academic and industry communities. The expedience of this process allowed a short timeline for review and publication to be achieved, and I must thank both the authors and the reviewers for their cooperation and patience in complying. I must also acknowledge the role of Clinton Fookes in assisting with the compilation of the final copy, alongside his active role as publicity coordinator.

Finally, I would like to pay tribute to the leadership of Brian Lovell in conceiving and promoting this event: his dedication to maintaining the active presence and impact of APRS in Australia is to be admired.

Anthony Maeder
Technical Chair of WDIC2003
Ordinary Member of APRS National Executive Committee

TABLE OF CONTENTS

Computer Vision and Pattern Recognition II *Page*

(This page left blank intentionally)

MEDICAL APPLICATIONS

(This page left blank intentionally)

Segmentation of Brain MR Images with Bias Field Correction

Seung-Gu Kim
Sangji University
Department of Applied Statistics
Wonju, Gangwon, 220-702, Korea
sgukim@mail.sangji.ac.kr

Shu-Kay Ng
University of Queensland
Department of Mathematics
Brisbane QLD 4072, Australia
skn@maths.uq.edu.au

Geoffrey J. McLachlan
University of Queensland
Department of Mathematics
Brisbane QLD 4072, Australia
gjm@maths.uq.edu.au

Deming Wang
University of Queensland
Centre for Magnetic Resonance
Brisbane QLD 4072, Australia
deming.wang@cmr.uq.edu.au

Abstract

We consider a statistical model-based approach to the segmentation of magnetic resonance (MR) images with bias field correction. The proposed method of penalized maximum likelihood is implemented via the expectation-conditional maximization (ECM) algorithm, using an approximation to the E-step based on a fractional weight version of the iterated conditional modes (ICM) algorithm. A Markov random field (MRF) is adopted to model the spatial dependence between neighbouring voxels. The approach is illustrated using some simulated and real MR data.

1. Introduction

Medical magnetic resonance imaging (MRI) has the advantages of being able to penetrate bony and air-filled structures with negligible attenuation and artifact. The modality has proven to be a very useful noninvasive medical imaging technique because of the ability to render high anatomical resolution of soft tissues with imaging in any arbitrary plane. Tissue-segmentation of magnetic resonance (MR) images of the human brain has a large potential to facilitate an imaging-based medical diagnosis, providing an aid to surgery and treatment planning [10]. Accurate estimation of the tissue parameters, including their volume sizes, will help to monitor changes in brain haemodynamics and metabolism resulting from neuronal activity [4], and so will assist in the diagnosis and treatment of neurogenerative disease such as Alzheimer's disease. MRI is also useful in providing anatomical information about the location of potential discontinuities in the Positron Emission Tomography (PET) image [17] and an opportunity to monitor the human brain activation effects to stimuli at relatively high spatial resolution [15]. Such tissue segmentation of MR images is often achieved by applying statistical classification techniques to the signal intensities [5, 21, 29], in conjunction with post-processing operations to remove acquisition artifacts [7, 12, 18]. A comprehensive review on MR image segmentation methods is provided by [6].

We consider here a statistical-based approach whereby the intensities on each voxel is modelled by a mixture of a finite number, say g, of normal distribution [19, 21]. In the latter, the expectation-maximization (EM) algorithm [8] is adopted to segment MR images and estimate the tissue parameters. An approximation to the E-step of the EM algorithm is employed based on a fractional weight version of Besag's iterated conditional modes (ICM) algorithm [2]. The prior (spatial) distribution of different tissue types is modelled by a hidden Markov Random Field (MRF) so as to incorporate spatial continuity constraints on the tissue segmentation. We refer to this model as GMM-HMRF (Gaussian Mixture Model with Hidden Markov Random Field). However, the intensity inhomogeneity of MR images due to acquisition equipments, severely degrades intensity-based segmentation of MR images [14, 30]. This low (spatial) frequency artifact known as the *bias field* arises from inhomogeneities in the radio-frequency (RF) field. Let $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$ and $\boldsymbol{Y}^* = (y_1^*, \ldots, y_n^*)^T$ be the observed and the ideal log-transformed intensities of a given image of n voxels, respectively, where the superscript T denotes vector transpose. The degradation effect of the bias field at the jth voxel can be expressed by an additive model as

$$y_j = y_j^* + b_j \qquad (j = 1, \ldots, n), \tag{1}$$

where b_j is the bias field at the jth voxel. It is noted that (1) implies that the observed MRI signal intensity is modelled as a product of the ideal intensity and a spatially varying factor (exponential of b_j).

In this paper, we extend the GMM-HMRF model by allowing the segmentation of MR images with bias field correction. Based on a penalized likelihood approach, we show how the estimation of the bias field and tissue parameters, and segmentation of the MR images can be obtained simultaneously via a partial version of an expectation-conditional maximization (ECM) algorithm [24].

2. Segmentation of MR Images for Gaussian Mixture Model with HMRF

Suppose that a continuous MR image is partitioned into a set of disjoint voxels labelled 1 to n, and that each voxel is assumed to belong to one of g distinct tissue types. This assumption is tenable because MR images have a spatial resolution at the range of the voxel size [19]. For notational convenience, we consider univariate intensity of each voxel, where the observed log-transformed intensities are denoted by a one-dimensional (1D) array $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$. Also, we let the g groups $G_1, \ldots, G_g$ represent the g possible tissue types. Further, we let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ denote the unobservable group-indicator vectors, where the ith element z_{ij} of $\boldsymbol{z}_j$ is taken to be one or zero according as to whether the jth voxel does or does not belong to the ith group. We put $\boldsymbol{z} = (\boldsymbol{z}_1^T, \ldots, \boldsymbol{z}_n^T)^T$.

A parametric mixture model approach [22] is adopted to represent the marginal distribution of $\boldsymbol{Y}$ of a given image of n voxel, where the bias field $b_1, \ldots, b_n$ are considered as unknown parameters. We assume that, unconditionally with respect to the group of origin, y_j $(j = 1, \ldots, n)$ has the finite mixture form of

$$p(y_j | b_j, \boldsymbol{\theta}, \beta) = \sum_{i=1}^{g} p(z_{ij} = 1 | \beta) p(y_j | z_{ij} = 1, b_j, \boldsymbol{\theta}) \tag{2}$$

where β is the parameter in the prior probability function of $\boldsymbol{Z}$ and $\boldsymbol{\theta}$ is the vector containing the tissue parameters. Suppose that the ideal log-transformed intensity of a voxel belonging to the ith group is normally distributed around a certain mean μ_i, with a variance σ_i^2. Then, we have

$$p(y_j | z_{ij} = 1, b_j, \boldsymbol{\theta}) = \phi_{\sigma_i}(y_j - \mu_i - b_j), \tag{3}$$

where $\phi_{\sigma_i}(\cdot)$ denotes a zero-mean normal distribution with variance σ_i^2 [28, 30] and $\boldsymbol{\theta} = (\mu_1, \sigma_1^2, \ldots, \mu_g, \sigma_g^2)^T$ is the vector containing the tissue parameters.

For the segmentation of MR images, the problem of inferring the vector $\boldsymbol{z}$ can be viewed as an incomplete-data problem, which can be approached by the application of the EM algorithm. This line of approach was undertaken by Kay and Titterington [16], who related some of the relaxation algorithms for image analysis with methods in the literature on the statistical analysis of incomplete data. The complete-data vector is given by $(\boldsymbol{y}^T, \boldsymbol{z}^T)^T$.

We consider a penalized complete-data log likelihood as

$$\begin{aligned} \log L_{Pc}(\boldsymbol{b}, \boldsymbol{\theta}, \beta) &= \log L_c(\boldsymbol{b}, \boldsymbol{\theta}, \beta) - \tfrac{1}{2} \boldsymbol{b}^T \Sigma_b^{-1} \boldsymbol{b} \\ &= \log p(\boldsymbol{y} | \boldsymbol{z}, \boldsymbol{b}, \boldsymbol{\theta}) + \log p(\boldsymbol{z} | \beta) \\ &\quad - \tfrac{1}{2} \boldsymbol{b}^T \Sigma_b^{-1} \boldsymbol{b}, \end{aligned} \tag{4}$$

where $\boldsymbol{b} = (b_1, \ldots, b_n)^T$, $\log L_c$ is the complete-data log likelihood, and $\Sigma_b = LL^T$, where L is a pre-defined low-pass filter [30]. In (4), The latter term can be viewed as a penalty term to stabilise the ML solution and to promote piecewise smoothness in the resulting segmentation and the bias field estimation. Thus, (4) can be regarded as a regularization method based on a penalized likelihood approach [13, 25]. It will be seen in Section 4 that this penalty term improves the segmentation and the bias field estimation.

Markov random fields are commonly employed in image processing problems to reflect the extent to which spatially neighbouring voxels belong to the same group [2, 11]. The incorporation of such spatial information on the images plays an important role in the estimation of $\boldsymbol{z}$ [5, 19]. The Hammersley-Clifford theorem states that the MRF prior can be specified using a Gibbs distribution [2, 11]

$$p(\boldsymbol{z} | \beta) = \exp\{-U(\boldsymbol{z} | \beta)\} / C(\beta), \tag{5}$$

where $C(\beta)$ is a normalizing constant known as the partition function and $U(\boldsymbol{z} | \beta)$ is the energy function specified by the neighbourhood system for the image. Because of the existence of the term $C(\beta)$ on the right-hand side of (5), there will be a stumbling block with the M-step with respect to β. Although the parameter β of a fairly general MRF can be estimated using Besag's pseudo-likelihood method [2], good estimates of β do not necessarily result in good segmentation [1]. We assume henceforth that β is specified *a priori*; see also the discussion in [3]. Besag [2] settled on $\beta = 1.5$ empirically. For the segmentation of MR images, the usage of such large value of β might, however, fail to detect small patches of voxels of one group surrounded by voxels of another group.

It will be seen that the E-step requires the calculation of $\tau_{ij}^{(k)} = E\{Z_{ij} \mid \boldsymbol{y}, \boldsymbol{\Psi}^{(k)}\}$, which unfortunately cannot be computed exactly under a HMRF mixture model [27]. McLachlan et al. [21] considered an approximation to the E-step based on a fractional weight version of the ICM algorithm. In the next section, we extend their GMM-HMRF model to simultaneously estimate the bias field and the tissue parameters, via an ECM algorithm.

3. An ECM Algorithm for Penalized ML Estimation

Concerning the probability density function of $\boldsymbol{Y}$ given $\boldsymbol{z}$ and $\boldsymbol{b}$, a common assumption in image analysis is to take $\boldsymbol{Y}_j$ to be independently distributed given the group membership and bias field. Thus, with (3), the first term of (4) can be expressed as

$$\log p(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{b},\boldsymbol{\theta}) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\log\phi_{\sigma_i}(y_j-\mu_i-b_j). \quad (6)$$

Let $\boldsymbol{\Psi}$ denote the vector of all the unknown parameters in the elements of $\boldsymbol{b}$ and $\boldsymbol{\theta}$. On the $(k+1)$th iteration of the EM algorithm, the E-step requires the calculation of

$$\begin{aligned} & Q_P(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)}) \\ & = E\{\log L_{Pc}(\boldsymbol{b},\boldsymbol{\theta},\beta) \mid \boldsymbol{y},\boldsymbol{\Psi}^{(k)}\} \\ & = \sum_{i=1}^{g}\sum_{j=1}^{n} E\{Z_{ij} \mid \boldsymbol{y},\boldsymbol{\Psi}^{(k)}\}\log\phi_{\sigma_i}(y_j-\mu_i-b_j) \\ & \quad + E\{\log p(\boldsymbol{z}|\beta) \mid \boldsymbol{y},\boldsymbol{\Psi}^{(k)}\} - \tfrac{1}{2}\boldsymbol{b}^T\Sigma_b^{-1}\boldsymbol{b}, \qquad (7) \end{aligned}$$

which is the conditional expectation of the penalized complete-data log likelihood given $\boldsymbol{Y}=\boldsymbol{y}$, using the current fit $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$. On the M-step of the $(k+1)$th iteration, the intent is to find the value of $\boldsymbol{\Psi}$ that maximizes $Q_P(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$, which gives $\boldsymbol{\Psi}^{(k+1)}$. The E- and M-steps are then alternated repeatedly until the penalized log likelihood changes by an arbitrary small amount, assuming convergence of the sequence of the penalized likelihood values.

E-step: We follow the approximation to the E-step considered in [21] by specifying the current conditional expectation of Z_{ij} given $\boldsymbol{y}$ and $\boldsymbol{\Psi}^{(k)}$ as

$$\begin{aligned} \tau_{ij}^{(k)} &= E\{Z_{ij} \mid \boldsymbol{y},\boldsymbol{\Psi}^{(k)}\} \\ &\approx E\{Z_{ij} \mid \boldsymbol{y},\boldsymbol{\Psi}^{(k)}, z_{N_j}=\hat{z}_{N_j}^{(k-1)}\} \\ &= \operatorname{pr}\{Z_{ij}=1 \mid \boldsymbol{y},\boldsymbol{\Psi}^{(k)}, z_{N_j}=\hat{z}_{N_j}^{(k-1)}\} \\ &= \frac{\pi_{ij}^{(k)}\phi_{\sigma_i^{(k)}}(y_j-\mu_i^{(k)}-b_j^{(k)})}{\sum_{h=1}^{g}\pi_{hj}^{(k)}\phi_{\sigma_h^{(k)}}(y_j-\mu_h^{(k)}-b_j^{(k)})}, \qquad (8) \end{aligned}$$

where N_j is some specified neighbourhood of the jth voxel, containing s voxels, labelled $j_1,\ldots,j_s$, and $z_{N_j}=(z_{j_1}^T,\ldots,z_{j_s}^T)^T$ is the vector containing the group labels of these s voxels in N_j. In (8), $\pi_{ij}^{(k)} = \operatorname{pr}\{Z_{ij}=1 \mid z_{N_j}=\hat{z}_{N_j}^{(k-1)}\}$ is the probability that the jth voxel belongs to the ith group G_i given the group membership of its specified neighbours as implied by $\hat{z}_{N_j}^{(k-1)}$. As in [21], we adopt the MRF model

$$\log\pi_{ij}^{(k)} \propto \beta(\gamma_1\sum_m \hat{z}_{im}^{(k-1)} + \gamma_2\sum_m \hat{z}_{im}^{(k-1)} + \gamma_3\sum_m \hat{z}_{im}^{(k-1)}), \quad (9)$$

where the summations in (9) are, respectively, over the prescribed first-, second-, and third-neighbours of the jth voxel. The parameters γ_1,γ_2, and γ_3 control the spatial relatedness between neighbouring voxels [5]. In the third-order model adopted by [19] for 3D MR images, γ_1,γ_2, and γ_3 are set equal to 1, $1/\sqrt{2}$, and $1/\sqrt{3}$, respectively. In the calculation of $\pi_{ij}^{(k)}$ in (9), we followed McLachlan et al. [21] and replaced $\hat{z}_{im}^{(k-1)}$, which is zero or one, by $\tau_{im}^{(k-1)}$. This modification avoids the discretization in counting the neighbours of the jth voxel and effectively avoids premature classification of the voxel with insufficient neighbourhood information. It can be viewed as a fractional weight version of the ICM algorithm [26]. Initially, we calculated $\log\pi_{ij}^{(0)}$ by using $\hat{z}_{im}^{(0)}$ in the right-hand side of (9).

M-step: The M-step involves the maximization of $Q_P(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$ with respect to $\boldsymbol{b}$ and $\boldsymbol{\theta}$. This maximization is implemented using a conditional approach, and the resulting algorithm can be viewed as an ECM algorithm. With the application of the ECM algorithm here, the M-step is replaced by two conditional maximization (CM) steps. The first involves the calculation of $\boldsymbol{b}^{(k+1)}$ by maximization (7) with $\boldsymbol{\theta}$ fixed at $\boldsymbol{\theta}^{(k)}$. The second CM step calculates $\boldsymbol{\theta}^{(k+1)}$ by maximization (7) with $\boldsymbol{b}$ fixed at $\boldsymbol{b}^{(k+1)}$.

On the $(k+1)$th iteration, the first CM-step yields

$$\begin{aligned} & \boldsymbol{b}^{(k+1)} \\ & = \arg\max_{\boldsymbol{b}} Q_P(\boldsymbol{b},\boldsymbol{\theta}^{(k)};\boldsymbol{b}^{(k)},\boldsymbol{\theta}^{(k)}) \\ & = \arg\max_{\boldsymbol{b}} \left\{\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_{ij}^{(k)}\log\phi_{\sigma_i^{(k)}}(y_j-\mu_i^{(k)}-b_j) \right. \\ & \qquad\qquad \left. -\tfrac{1}{2}\boldsymbol{b}^T\Sigma_b^{-1}\boldsymbol{b}\right\} \\ & = \arg\max_{\boldsymbol{b}} \left\{\boldsymbol{b}^T\boldsymbol{r}^{(k)} - \tfrac{1}{2}\boldsymbol{b}^T\boldsymbol{D}^{(k)}\boldsymbol{b} - \tfrac{1}{2}\boldsymbol{b}^T\Sigma_b^{-1}\boldsymbol{b}\right\}, \qquad (10) \end{aligned}$$

where, for $j=1,\ldots,n$,

$$r_j^{(k)} = (\boldsymbol{r}^{(k)})_j = \sum_{i=1}^{g}\tau_{ij}^{(k)}\left(\frac{y_j-\mu_i^{(k)}}{\sigma_i^{2(k)}}\right) \qquad (11)$$

and $\boldsymbol{D}^{(k)}$ is a diagonal matrix with the diagonal elements

$$d_j^{(k)} = (\boldsymbol{d}^{(k)})_j = \sum_{i=1}^{g}\tau_{ij}^{(k)}\left(1/\sigma_i^{2(k)}\right). \qquad (12)$$

Letting $\boldsymbol{I}$ denote the $n \times n$ identity matrix, it follows from (10) that we have

$$\boldsymbol{b}^{(k+1)} \;=\; (\boldsymbol{D}^{(k)}+\Sigma_b^{-1})^{-1}\boldsymbol{r}^{(k)}$$

$$\begin{aligned} &= (\Sigma_b \boldsymbol{D}^{(k)} + \boldsymbol{I})^{-1} \Sigma_b \boldsymbol{r}^{(k)} \\ &= (LL^T \boldsymbol{D}^{(k)} + \boldsymbol{I})^{-1} LL^T \boldsymbol{r}^{(k)}, \qquad (13) \end{aligned}$$

which coincides with result of Wells et al. [30]. To speed up the estimation of the bias field, we adopted in our implementation for (13) a computationally efficient lowpass filter proposed by Wells et al. [30], which is characterized as follows:

$$b_j^{(k+1)} = \frac{(L\boldsymbol{r}^{(k)})_j}{(L\boldsymbol{D}^{(k)}\boldsymbol{1})_j} = \frac{(L\boldsymbol{r}^{(k)})_j}{(L\boldsymbol{d}^{(k)})_j} \qquad (14)$$

for $j = 1, \ldots, n$, where $\boldsymbol{1} = (1, 1, \ldots, 1)^T$. The linear transformation $L\boldsymbol{d}$ in (14) is implemented by convolution of the point spread function (psf) ℓ and the 3D image matrix A corresponding to L and $\boldsymbol{d}$, respectively [14]. The convolution $\ell \odot A$ is operated repeatedly by 20–30 times in order to increase its lowpass filtering effect. An alternative filtering operation using Gaussian convolution may also be adopted [9].

With $\boldsymbol{b}$ fixed at $\boldsymbol{b}^{(k+1)}$ in (7), the second CM-step yields

$$\begin{aligned} &\boldsymbol{\theta}^{(k+1)} \\ &= \arg\max_{\boldsymbol{\theta}} Q_P(\boldsymbol{b}^{(k+1)}, \boldsymbol{\theta}; \boldsymbol{b}^{(k+1)}, \boldsymbol{\theta}^{(k)}) \\ &= \arg\max_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{g} \sum_{j=1}^{n} \tau_{ij}^{(k)} \log \phi_{\sigma_i}(y_j - \mu_i - b_j^{(k+1)}) \right\}, \quad (15) \end{aligned}$$

which can be carried out in closed form as

$$\mu_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)} (y_j - b_j^{(k+1)}) / \sum_{j=1}^{n} \tau_{ij}^{(k)} \qquad (16)$$

and

$$\sigma_i^{2(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)} (y_j - \mu_i^{(k+1)} - b_j^{(k+1)})^2 / \sum_{j=1}^{n} \tau_{ij}^{(k)} \quad (17)$$

for $i = 1, \ldots, g$.

The ECM algorithm preserves the appealing convergence properties of the EM algorithm. It thus has reliable global convergence in that it monotonely increases the penalized likelihood after each iteration, no matter what starting value is used. A detailed account of the convergence properties of the EM and ECM algorithms can be found in [20, 23]. In our proposed algorithm, we do not update the bias field (low frequency degrade) estimate $\boldsymbol{b}$ in every CM cycle. This approach speeds up the algorithm as the computational cost using an iterative lowpass filtering for a 3D MR image can be enormous. In addition, as the bias field is estimated using a more accurate estimate of $\boldsymbol{z}$ and $\boldsymbol{\theta}$ by more frequent update, it avoids the oscillations of the bias field estimate and hence improves the final segmentation and the bias field estimation, as demonstrated in Section 4. Our algorithm is summarized as follows:

1. *Obtain initial estimates of $\boldsymbol{b}^{(0)}$, $\boldsymbol{\theta}^{(0)}$, and $\boldsymbol{z}^{(0)}$.*
2. *E-step: Calculate the posterior probabilities $\tau_{ij}^{(k)}$ based on (8).*
3. *CM-Step 1: Estimate the bias field $\boldsymbol{b}^{(k+1)}$ based on (14), given $\tau_{ij}^{(k)}$ and $\boldsymbol{\theta}^{(k)}$.*
4. *Do T cycles with $\boldsymbol{b}$ fixed at $\boldsymbol{b}^{(k+1)}$:*

   4.1 E-step: Calculate $\tau_{ij}^{(k+\,t/T)}$ based on (8),

   4.2 CM-Step 2: Calculate $\boldsymbol{\theta}^{(k+\,t/T)}$ based on (16) and (17),

   where the $\tau_{ij}^{(k+\,t/T)}$ and $\boldsymbol{\theta}^{(k+\,t/T)}$ denote the posterior probabilities and the value of $\boldsymbol{\theta}$, respectively, after the tth cycle on the $(k+1)$th iteration ($t = 1, \ldots, T$).
5. *Repeat from 2 until parameter sequences converge.*

As not all the CM-steps were performed in every CM cycle, we refer to our algorithm as a “partial” ECM algorithm. As an E-step is performed before each CM-step, the algorithm corresponds to a “multi-cycle ECM”, where a cycle is defined by one E-step followed by one CM-step [24]. It can be seen that (10) and (15) imply

$$Q_P(\boldsymbol{b}^{(k+1)}, \boldsymbol{\theta}^{(k)}; \boldsymbol{b}^{(k)}, \boldsymbol{\theta}^{(k)}) \geq Q_P(\boldsymbol{b}^{(k)}, \boldsymbol{\theta}^{(k)}; \boldsymbol{b}^{(k)}, \boldsymbol{\theta}^{(k)})$$

and

$$\begin{aligned} &Q_P(\boldsymbol{b}^{(k+1)}, \boldsymbol{\theta}^{(k+1)}; \boldsymbol{b}^{(k+1)}, \boldsymbol{\theta}^{(k)}) \\ &\qquad \geq Q_P(\boldsymbol{b}^{(k+1)}, \boldsymbol{\theta}^{(k)}; \boldsymbol{b}^{(k+1)}, \boldsymbol{\theta}^{(k)}). \end{aligned}$$

It follows that the penalized log-likelihood

$$L(\boldsymbol{y}; \boldsymbol{b}, \boldsymbol{\theta}) - \tfrac{1}{2} \boldsymbol{b}^T \Sigma_b^{-1} \boldsymbol{b}$$

increases at each cycle and thus increases at each iteration monotonically if an exact E-step were used.

As in [14, 28], we set initially the bias field $\boldsymbol{b}^{(0)}$ to be zero. The initial estimates $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{z}^{(0)}$ can be obtained by performing the “noncontextual” segmentation of the MR image [21]. That is, the segmentation of the voxels is proceeded by ignoring all the spatial characteristics ($\beta = 0$) and the bias field estimation step (14). Van Leemput et al. [28] refer to this method as the independent model.

4. Experimental Results

We now demonstrate the use of the partial ECM algorithm for the fitting of the GMM-HMRF model with bias field correction. The first example is a simulated three-class image obtained by adding the true image $\boldsymbol{y}^*$ and true bias field as

$$y_j = (1 - d) y_j^* + d b_j,$$

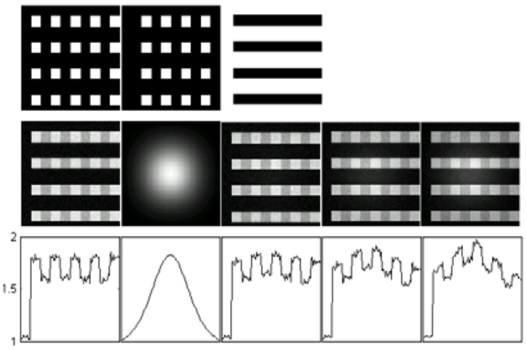


Figure 1. Simulated data: (From left to right) Top: the three groups; Middle: the true image, the true bias field, and the combined images with d=0.05, 0.12, and 0.20; Bottom: the intensity profile for each image

where $0 < d < 1$ is a constant governing the amount of bias field contamination. The true image simulates 256 gray-leveled brain MR image. The intensities for the white matter, gray matter, and the CSF are 150, 120, and 10, respectively. Gaussian noise with variance of 20, 14, and 12 is added to the three groups, respectively, before scaling by 256. The bias field simulates Gaussian-distributed function varied from zero (dark) to one (bright). Figure 1 presents the simulated data and the contaminated images. It can be seen that the image is heavily contaminated with bias field when $d = 0.20$. In this simulation experiment, we considered $g = 3$, $\beta = 0.9$, and $T = 3$. The results from fitting the GMM-HMRF model via our ECM algorithm are displayed in Figure 2 for $d = 0.20$.

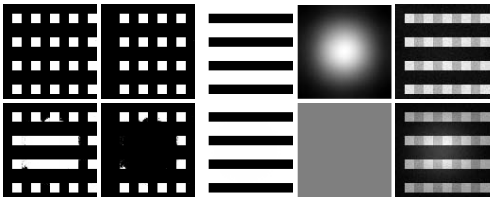


Figure 2. From left to right: the three groups segmented, estimated bias field, and the restored image. Top: GMM-HMRF model with bias field correction; Bottom: GMM-HMRF model without bias field correction

For comparison, we present also in Figure 2 the results obtained by the GMM-HMRF model without bias field correction. It can be seen that the GMM-HMRF algorithm fails to segment correctly. This result indicates how the segmentation is affected by the contamination of bias field. With the modification of GMM-HMRF model with bias field correction, it can be seen that almost perfect segmentation is obtained.

The second example is a real MR image of the human brain. The data set was a slice of a 3D T_1-weighted MR image, where the acquisition matrix was $256 \times 256 \times 256$. In this example, we considered $g = 3$ corresponding to the white matter, gray matter, and CSF. We adopted $\beta = 0.3$ and $T = 3$. The results are displayed in Figure 3. It can be seen that the three tissue types, brain-white matter, brain-gray matter, and CSF are well separated using the proposed GMM-HMRF model with bias field correction. For comparison, the segmented image from the GMM-HMRF algorithm without bias field correction is also given in Figure 3. It can be seen that white matter at the upper half of the brain is misclassified as gray matter, whereas the result is much better when the bias field correction is included in the algorithm.

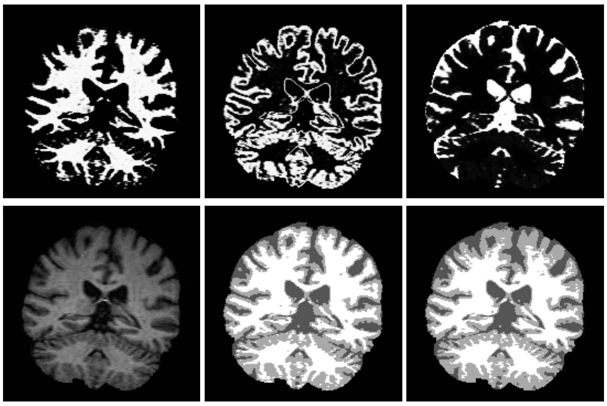


Figure 3. From left to right: Top (GMM-HMRF with bias field correction): Segmented white matter, Segmented gray matter, Segmented CSF; Bottom: Original image, Segmented image (with bias field correction), Segmented image (without bias field correction)

5. Discussion

For the segmentation of MR images with bias field correction, Wells et al. [30] have developed a mixture model-based approach via the EM algorithm to estimate the bias field and segment the images. Guillemaud and Brady [14] further refined this technique by introducing the extra tissue class “other” and initializing the EM algorithm automatically for a given number of classes. Both methods assume statistical independence of the voxel intensities (that is, the noise in the MR signal is spatially white). Moreover, the parameters of each tissue class are required to be pre-defined or estimated in advance of applying the algorithm. Recently, Zhang et al. [31] proposed a hidden MRF model to allow for the spatial continuity of image intensities and the bias field correction simultaneously. This model adopts the ICM algorithm [2] to sequentially update each z_{ij}, which are zero or one, by local minimization of the conditional posterior probability. A maximum *a posteriori* (MAP) approach is applied to estimate the bias field, and the tissue parameters are estimated by maximum likelihood. However, this algorithm fails to segment correctly when the bias field contamination is heavy.

The EM algorithm is a popular tool in statistics for carrying out ML estimation because of its simplicity of implementation and reliable global convergence [20]. Here we have been able to develop an extension of the EM algorithm which can handle penalized ML estimation for the present problem, while still preserving the desirable properties of the EM algorithm. The extension of the GMM-HMRF model with bias field correction is justified using some simulated and real MR data, as shown in Section 4.

References

[1] R. G. Aykroyd and P. J. Green. Global and local priors, and the location of lesions using gamma-camera imagery. *Philos. Trans. Roy. Soc. A*, 337(1647):323–342, 1991.

[2] J. E. Besag. On the statistical analysis of dirty pictures (with discussion). *J. Roy. Stat. Soc., ser. B*, 48(3):259–302, 1986.

[3] J. E. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. I. Stat. Math.*, 43(1):1–59, 1991.

[4] G. D. Cascino, C. R. Jack, J. E. Parisi, et al. Magnetic resonance imaging-based volume studies in temporal lobe epilepsy: Pathological correlations. *Ann. Neurol.*, 30(1):31–36, 1991.

[5] H. S. Choi, D. R. Haynor, and Y. Kim. Multivariate tissue classification of MRI images for 3D volume reconstruction – a statistical approach. *Proceedings SPIE Medical Imaging III: Image Processing*, 1092:183–193, 1989.

[6] L. P. Clarke, R. P. Velthuizen, M. A. Camacho, et al. MRI segmentation: Methods and applications. *Magn. Reson. Imag.*, 13(3):343–368, 1995.

[7] B. M. Dawant, A. P. Zijdenbos, and R. A. Margolin. Correction of intensity variations in MR images for computer-aided tissue classification. *IEEE Trans. Med. Imag.*, 12(4):770–781, 1993.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Stat. Soc., ser. B*, 39(1):1–38, 1977.

[9] R. Deriche. Recursively implementing the Gaussian and its derivatives. *Technical Report 1893*. INRIA Sophia-Antipolis, France, 1993.

[10] M. Freund, S. Hahnel, M. Thomsen, and K. Sartor. Treatment planning in severe scoliosis: the role of MRI. *Neuroradiology*, 43(6):481–484, 2001.

[11] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, 6(6):721–741, 1984.

[12] G. Gerig, O. Kübler, R. Kikinis, and F. A. Jolesz. Nonlinear anisotropic filtering of MRI data. *IEEE Trans. Med. Imag.*, 11(2):221–232, 1992.

[13] P. J. Green. On use of the EM algorithm for penalized likelihood estimation. *J. Roy. Stat. Soc., ser. B*, 52(3):443–452, 1990.

[14] R. Guillemaud and M. Brady. Estimating the bias field of MR images. *IEEE Trans. Med. Imag.*, 16(3):238–251, 1997.

[15] L. Jancke, T. W. Buchanan, K. Lutz, and N. J. Shah. Focused and nonfocused attention in verbal and emotional dichotic listening: An FMRI study. *Brain Lang.*, 78(3):349–363, 2001.

[16] J. Kay and D. M. Titterington. Image labelling and the statistical analysis of incomplete data. In *Proc. 2nd Int. Conf. Image Processing and Applications*. Institute of Electrical Engineers, London, 1986, pp. 44–48.

[17] R. Leahy and X. Yan. Incorporation of anatomical MR data for improved functional imaging with PET. *Lect. Notes Comput. Sc.*, 511:105–120, 1991.

[18] S. K. Lee and M. W. Vannier. Post-acquisition correction of MR inhomogeneities. *Magn. Reson. Med.*, 36(2):275–286, 1996.

[19] Z. Liang, J. R. MacFall, and D. P. Harrington. Parameter estimation and tissue segmentation from multispectral MR images. *IEEE Trans. Med. Imag.*, 13(3):441–449, 1994.

[20] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, New York, 1997.

[21] G. J. McLachlan, S. K. Ng, G. Galloway, and D. Wang. Clustering of magnetic resonance images. In *Proc. American Statistical Assoc. (Statistical Computing Section)*. American Statistical Assoc., Alexandria, Virginia, 1996, pp. 12–17.

[22] G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, New York, 2000.

[23] X. L. Meng. On the rate of convergence of the ECM algorithm. *Ann. Stat.*, 22(1):326–339, 1994.

[24] X. L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, 80(2):267–278, 1993.

[25] J. A. O'Sullivan. Roughness penalties on finite domains. *IEEE Trans. Image Processing*, 4(9):1258–1268, 1995.

[26] A. Owen. Contribution to the discussion of paper by B.D. Ripley. *Canad. J. Statist.*, 14(2):106–110, 1986.

[27] W. Qian and D. M. Titterington. Estimation of parameters in hidden Markov models. *Philos. Trans. Roy. Soc. A*, 337(1647):407–428, 1991.

[28] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens. Automated model-based tissue classification of MR images of the brain. *IEEE Trans. Med. Imag.*, 18(10):897–908, 1999.

[29] M. W. Vannier, R. L. Butterfield, D. Jordon, et al. Multispectral analysis of magnetic resonance images. *Radiology*, 154(1):221–224, 1985.

[30] W. M. Wells, III, W. E. L. Grimson, R. Kikinis, and F. A. Jolesz. Adaptive segmentation of MRI data. *IEEE Trans. Med. Imag.*, 15(4):429–442, 1996.

[31] Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imag.*, 20(1):45–57, 2001.

A New Deformable Model Using Dynamic Gradient Vector Flow and Adaptive Balloon Forces

Suhuai Luo, Rongxin Li, and Sébastien Ourselin
CSIRO Telecommunications & Industrial Physics
Medical Imaging Group
Cnr Vimiera & Pembroke Rd, Marsfield NSW 2122, Australia
suhuai.luo@csiro.au

Abstract

An extension of the gradient vector flow snake (GVF snake) is presented. The method is based on combining two other external forces. First, the adaptive balloon force has been developed to increase the GVF snake's capture range and convergence speed. Then, a dynamic GVF force is introduced to provide an efficient evolution-stop mechanism. In this way, we prevent the snake from breaking through the correct surface and locking to other salient feature points. Preliminary segmentation results demonstrate the potential of our approach in comparison with the original GVF snake method.

1. Introduction

There has been a substantial amount of research on segmenting images with deformable models in recent years [4]. Notably active contours, known as "snakes", have been widely studied and applied in medical image analysis, their applications including edge detection, segmentation of objects, shape modeling and motion tracking [7, 11]. Snakes were first introduced in 1987 by Kass *et al.* [5]. They generally represent an object boundary as a parameter curve or surface. An energy function is associated with the curve, so the problem of finding an object boundary is cast as an energy minimisation process. Typically, the curves are affected by both an internal force and external force. A snake can locate object contours well, once an appropriate initialisation is done. However, since the energy minimisation is carried out locally, the located contours can be trapped by a local minimum. A number of methods have been proposed to improve the snake's performance [8, 1]. A balloon model was introduced by Cohen *et al.* to enlarge the snake's capture range [2, 3]. Recently, Xu *et al.* have proposed a new deformable model called the "gradient vector flow snake" (GVF) [11, 12]. Instead of directly using image gradients as an external force, it uses a spatial diffusion of the gradient of an edge map of the image. GVF snake was proposed to address the traditional snake's problems of short capture range and inability to track at boundary concavity. But GVF still may not be able to capture object contours in some medical image segmentation. Efforts at improving the original GVF snake's performance have been published recently. Xu *et al.* combined GVF force with a constrained balloon force to segment gyri in the cortex [10]. Although this combination works well on this case, its requirement of an *a priori* knowledge of the region of interest may restrict its application. Yu *et al.* proposed to compute the GVF using a polar coordinate representation instead of cartesian coordinates [13]. In this way, the method can perform better than the original GVF snake in areas of long thin boundary concavities and boundary gaps. But the capture range of this improved GVF does not seem larger than the original method.

In our paper, after presenting the properties of the GVF snake, we propose a new approach to enhance the GVF snake performance on segmentation. The method consists of two major parts. First, an adaptive balloon force is incorporated into internal forces to increase capture range and speed-up evolution. Secondly, a dynamic GVF force is introduced to provide an evolution-stop mechanism. With this ability, the located contours are less sensitive to local minima.

This paper is organized as follows. First the mathematic foundation of active contour models, including conventional snakes and GVF snakes, are introduced in section 2. We detail in section 3 the different aspects of our improved GVF snake using an adaptive ballon model and a dynamic GVF. In section 4, we present some preliminary results of tumour segmentation on brain MRI and compare the performance of our approach with the GVF snake. Finally, we propose several avenues of research for future work in section 5.

2. Active Contour Models

In this section, we review the mathematic formulation of conventional snakes and GVF snakes. We also describe the strengths and weakness of each method.

2.1 Snakes

In 2D, a snake is a curve $\mathbf{C}(s) = (x(s), y(s))$ where $s \in [0, 1]$. The curve moves through the image domain to minimize a specified energy function. In traditional snakes, the energy is usually formed by internal forces and external forces as:

$$E_{snake} = E_{internal} + E_{external} \tag{1}$$

$E_{internal}$ tends to elastically hold the curve together (elasticity forces) and to keep it from bending too much (bending forces). This energy is defined in equation (2), where $\mathbf{C}_s$ and $\mathbf{C}_{ss}$ represent the first and second derivative respectively. We can control the snake's tension and rigidity by the coefficients α and β.

$$E_{internal} = \frac{1}{2}\int_s \alpha|\mathbf{C}_s|^2 ds + \frac{1}{2}\int_s \beta|\mathbf{C}_{ss}|^2 ds \tag{2}$$

$E_{external}$ intends to pull or push the curve towards the edges. Typically, the external forces consist of potential forces. This energy is defined in equation (3), where E_{image} represents the negative gradient of a potential function. This energy is generally the image force as defined in equation (4) where I denotes the image and $\mathbf{x} = \mathbf{x}(x, y) = [x\,y]^t$.

$$E_{external} = \int_s E_{image}(\mathbf{C}(s))ds \tag{3}$$

$$E_{image}(\mathbf{x}) = -|\nabla I(\mathbf{x})|^2 \tag{4}$$

Using variational calculus and the Euler-Lagrange differential equation, we can solve equation (1). Then, the solution of this force balance, as defined in equation (5), represents the snake final position. The differences in the ways the energy function is established will result in different snakes.

$$\alpha\mathbf{C}_{ss} - \beta\mathbf{C}_{ssss} - \nabla E_{image} = 0 \tag{5}$$

Although the traditional snakes have found many applications, they are intrinsically weak in three main aspects: First, they are very sensitive to parameters. Second they have small capture range and the convergence of the algortihm is mostly dependent of the initial position. Finally, they have difficulties in progressing into boundary concavities.

2.2 GVF snakes

Xu *et al.* have proposed a new GVF snake to achieve better object segmentation [12]. The basic idea of the GVF snake is to extend influence range of image force to a larger area by generating a GVF field. The GVF field is computed from the image. In detail, a GVF field is defined as a vector field $\mathbf{V} = \mathbf{V}(\mathbf{x})$ that minimizes the energy function

$$Q = \iint \mu\nabla^2\mathbf{V} + |\nabla f|^2|\mathbf{V} - \nabla f|^2 d\mathbf{x} \tag{6}$$

where f is the edge map which is derived by using an edge detector on the original image convoluted with a Gaussian kernel, and μ is a regularization parameter. Using variational calculus, the GVF field can be obtained by solving the corresponding Euler-Lagrange equations.

Similar to equation (5), the force balance equation of GVF snake can be expressed as

$$\alpha\mathbf{C}_{ss} - \beta\mathbf{C}_{sss} + \gamma\mathbf{V} = 0 \tag{7}$$

where γ is a proportional coefficient. GVF snake's larger capture range and concavity tracking ability are attributed to the diffusion operation shown in the above equation. When $|\nabla f|$ is small, the energy is dominated by the sum of the squares of the partial derivatives of the vector field, resulting a slowly varying yet large coverage field. Whereas when $|\nabla f|$ is large, the second term dominates the integral.

In applying the GVF snake on real data such as medical images, the capture range of the active contour did not seem as large as we expected. This is mainly because in the case of medical data, images often contain a lot of textures. Unfortunately, the GVF field is very sensitive to these variations and the active contour does not converge to the ideal solution. Another observation was that the GVF snake was sensitive to the shape irregularities. In these cases, the GVF force could not properly push the snake to the right contour.

To deal with these problems we have developed an improved GVF snake. This new method is presented in detail in the following section.

3 Improved GVF snake

The improvement we propose is to add new external forces, including an adaptive balloon force f_{pa} and a dynamic GVF force, defined as a vector field $\mathbf{V}_{dyn}$. Then, we propose a new scheme to integrate these external forces in the snake mathematic formulation.

3.1 Adaptive balloon force

In the balloon model proposed by Cohen *et al.*, a pressure force $f_p(s)$ is added to snake force as a second external force to push the curve outward or inward [2, 3]. In this way, the curve is considered as a balloon that has been inflated or deflated. Equation (8) represents the pressure force, where $\overrightarrow{n}(s)$ is the normal unit vector to the curve at point $\mathbf{C}(s)$.

$$f_p(s) = k.\overrightarrow{n}(s) \tag{8}$$

The balloon force is considered to increase the image potential force capture range. This is a proper consideration given that the snake can be set to start evolving inside the object. Unfortunately, balloon force introduces unpredictability to the performance of the active contour and make it more sensitive to the value of its different parameters.

To overcome the unpredictability problem introduced by the balloon force, this force is applied in an adaptive way. The main idea is to give the balloon force bigger weight compared to the GVF force at the early stage of the evolution, and to give the balloon force smaller weight at the later stage. In this way, the speed of the convergence is increased, and the snake can be correctly pushed toward the surface even if it starts far away with less chance of being over-pushed.

3.2 Dynamic GVF force

A dynamic GVF force is introduced to provide a unique evolution-stop mechanism as well as all the characteristics owned by the original GVF force.The evolution-stop mechanism is needed to prevent the snake from breaking through the correct contour and locking to other feature points. The breakage can happen in areas where two objects or organs are very close each other. The introduction of the dynamic GVF force is inspired by a property of the GVF field. That is, when the GVF field passes a contour, its direction will change. Figure 1 shows an ellipse and its corresponding gradient vector flow.

It can be easily observed that the field vector changes direction at the ellipse boundary. Therefore, a consistency degree is incorporated into the new dynamic GVF force. The force varies according to the consistency. If the evolution of the snake will cause the change of GVF force direction, it is said inconsistency has occured and the snake is not allowed to evolve to the new position.

3.3 New scheme

With these two novel inclusions, the proposed force balance equation can be expressed as:

$$\alpha\mathbf{C}_{ss} - \beta\mathbf{C}_{ssss} + \gamma V_{dyn} + \lambda f_{pa} = 0 \tag{9}$$

$\mathbf{V}_{dyn}$ is defined in equation (10) as a dynamic gradient vector flow force. Let $\mathbf{x_1}$ be a point on the current snake and $\mathbf{x_2}$ its possible next position in the evolution process. C_θ defines the consistency angle and is proportional to the angle between the GVF vectors at $\mathbf{x_1}$ and $\mathbf{x_2}$. T_θ represents the cut-off angle: based in our experiments, $T_\theta = 20^o$ is a good threshold.

$$\mathbf{V}_{dyn} = \begin{cases} \mathbf{V} & \text{if } C_\theta > T_\theta \\ \frac{-\alpha\mathbf{C}_{ss}+\beta\mathbf{C}_{ssss}-\lambda f_{pa}}{\gamma} & \text{otherwise} \end{cases} \tag{10}$$

The new dynamic gradient vector flow force will be the same as conventional GVF if the snake point moves towards the contour. But when the snake point tries to cross over an edge, the dynamic gradient vector flow force will stop the point from moving. The threshold T_θ will decide when this evolution-stop mechanism will be triggered.

4 Experiments

Brain Tumor Segmentation Database. In validating the performance of the proposed method, we used the SPL and NSG Brain Tumor Segmentation Database [6, 9]. The database consists of magnetic resonance images of several anonymous brain tumor patients, as well as segmentations of the brain and tumor from these scans (MR T1-weighted image in the sagittal plane, 256 x 256 x 124, 0.9375 x 0.9375 x 1.5 mm^3). Manual segmentations obtained by neurosurgeons and automated segmentations obtained by the method of [6] and [9]. In figure 3, we present the manual segmentation of a *glioma* on the MRI of the patient 4 in the database, the slice 35 is presented.

Validation criteria. Our validation criteria of tumor segmentation is based on both subjective and quantitative analysis. For subjective aspect, the contours drawn by experts (figure 3) and by automatic segmentation were compared (figure 4). For quantitative analysis, three validating parameters are defined. In defining the parameters, the accuracy of the snake results is checked against the manual segmentations done by four experts.

To evaluate the results, we propose to use three values which we will denote by C_r, FPN and FNN.

C_r is defined as the ratio between the area considered as tumor by both the snake and at least three experts and the area considered as tumor either by the snake or at least three experts. It is expressed as $C_r = \frac{A_\cap}{max(A_{expt}, A_s)}$, where A_s is the area confined by snake; A_{expt} is the area considered as tumor by at least three of the four experts and $A_\cap$ is the

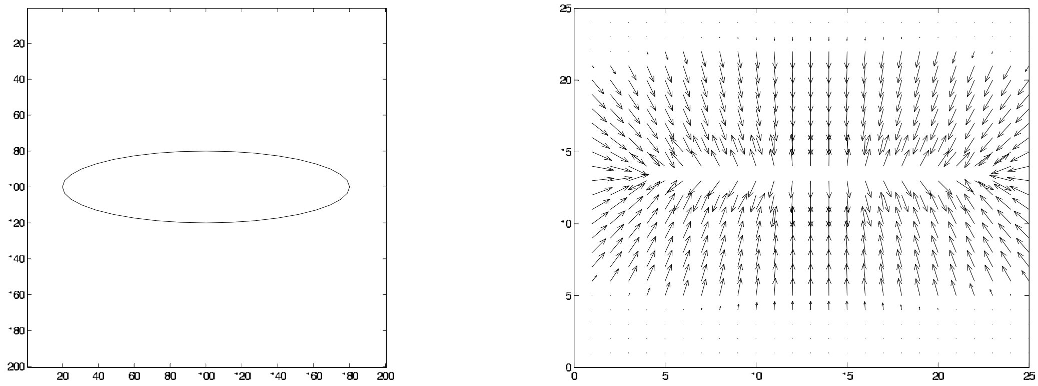


Figure 1. *An ellipse and its corresponding GVF forces.*

area overlapped between A_s and A_{expt}. We can note that C_r is a normalized value ($C_r \in [0, 1]$).

FPN is defined as the area considered as tumor by snake, but as non-tumor by at least 3 experts (False Positive Number).

FNN is defined as the area considered as non-tumor by snake, but as tumor by at least 3 experts (False Negative Number).

Results. To investigate the performance of our segmentation method and compare it with original GVF snake, we have designed four experiments using four different sets of parameters. For each experiment, both original GVF snake and our improved GVF snake have the same values of α, β and γ, we alse the initial position for the curve. Figure 2 shows the original image and the initial snake position drawn in white curve.

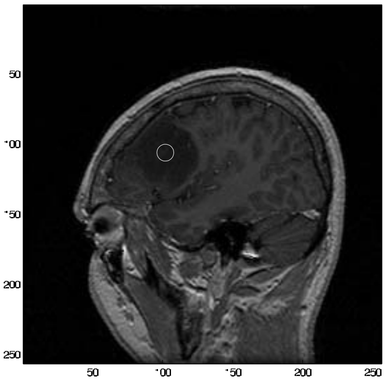


Figure 2. *The original MRI slice* 35 *of patient* 4 *and the initial snake position drawned in white.*

Table 1 summarizes the parameters set for the four experiments, we choose $\lambda = 0.2$ for our method. Table 2 and 3 presents the values of C_r, FPN and FNN of both the original GVF snake and our improved GVF snake.

parameters	α	β	γ
Test 1	0.05	0.1	0.3
Test 2	0.05	0.1	0.5
Test 3	0.05	0.1	0.1
Test 4	0.1	0.2	0.3

Table 1. *Parameters set for the four experiments.*

Analysis. Based on the figures and tables, two main points can be drawn as to the performance comparison. One point is that in terms of subjective criteria the original GVF snake's capture range is far from enough to locate the tumor and it easily became stuck on unwanted features and failed on most of the cases, whereas the proposed approach succeed in locating the *glioma* in most cases. The other point is that, according to quantitative analysis, our approach resulted more preferable results than the original GVF snake in most cases. One point we want to note is that in some particular cases the original GVF snake presented some better validating values, however, by analysis all the validating values for each case we can state that our method is still more preferable.

Based on figure 4, we can see that the original GVF can not correctly locate the *glioma* and is stuck at the top part of the tumor. For quantitative analysis, by observing the C_r values of the original GVF snake and our approach, we can see that the original GVF snake gives favorable value of 0.8127 comparing to 0.7783 of our approach. This does mean that in this case the area located by the original GVF snake is more likely as *glioma* than the area located by our approach. However, if we check the FPN and FNN values, we can see that our method presented much less false

Original GVF	Test 1	Test 2	Test 3	Test 4
C_r	0.7931	0.8127	0.1388	0.2055
FPN	129	235	0	37
FNN	307	278	1278	1179

Table 2. *Values of the three criteria using the original GVF snake.*

Improved GVF	Test 1	Test 2	Test 3	Test 4
C_r	0.8455	0.7783	0.6274	0.7871
FPN	259	407	2	389
FNN	67	55	553	46

Table 3. *Values of the three criteria using the improved GVF snake.*

negative number. This is ideal from the viewpoint of a neurosurgeon.

5 Conclusion

In this paper, we have presented a new method using Gradient Vector Flow and Balloons. We introduced an adaptive balloon force to increase GVF snake's capture range and speed up evolution. Then we proposed a dynamic GVF force to provide an efficient evolution-stop mechanism.

Based on experiments on segmenting a tumor in real brain MRI data, it has shown that the proposed method is robust to the variation in initial position and efficient in preventing the snake from breaking through correct contour and locking to other feature points.

A current limitation of the method is that GVF is computed independently slice by slice (2D version). As a consequence, we do not take into account the interslice spatial continuity of the gradient, or the possible anisotropy of the voxels. Such error of segmentation might be reduced by putting spatial constraints to the reconstructed structure.

Aknowledgments

The authors would like to thank Drs. Simon Warfield, Michael Kaus, Ron Kikinis, Peter Black and Ferenc Jolesz for sharing the tumor database.

References

[1] P. Bamford and B. Lovell. Unsupervised cell nucleus segmentation with active contours. *Signal Processing Special Issue: Deformable Models and Techniques for Image and Signal Processing*, 71(2):203–213, 1998.

[2] L. Cohen. On active contour models and balloons. *Computer Vision, Graphics, and Imasge Processing: Image Understanding*, 53(2):211–218, 1989.

[3] L. Cohen and I. Cohen. Finite-element methods for active contour models and balloons for 2-d and 3-d images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1146–1131, November 1993.

[4] A. Jain, Y. Zhong, and M. Dubuisson-Jolly. Deformable template models: A review. *Signal Processing*, 71:109–129, 1998.

[5] M. Kass, M. Witkin, and D. Terzopoulos. Snakes: active contour models. *International Journal of Vision*, 1:321–331, 1987.

[6] M. Kaus, S. Warfield, A. Nabavi, P. Black, F. Jolesz, and R. Kikinis. Automated segmentation of mri of brain tumors. *Radiology*, 218:586–591, 2001.

[7] T. McInerney and D. Terzopoulos. A Dynamic Finite Element Surface Model for Segmentation and Tracking in Multidimensional Medical Images with Applications to Cardiac 4D Image analysis. *Computerized Medical Imaging and Graphics*, 19(1):69–83, 1995.

[8] T. McInerney and D. Terzopoulos. Deformable models in medical image analysis: A survery. *Medical Image Analysis*, 1(2):91–108, 1996.

[9] M. Warfield, S.M. Kaus, F. Jolesz, and R. Kikinis. Adaptive, template moderated, spatially varying statistical classification. *Medical Image Analysis*, 4(1):43–55, 2000.

[10] C. Xu, D. Pham, M. Rettmann, D. Yu, and J. Prince. Reconstruction of the human cerebral cortex from magnetic resonance images. *IEEE Transactions on Medical Imaging*, 18(6):467–479, June 1999.

[11] C. Xu and J. Prince. Generalized gradient vector flow external forces for active contours. *Signal Processing*, 71:131–139, 1998.

[12] C. Xu and J. Prince. Snakes, shapes, and gradient vector flow. *IEEE Transactions on Images Processing*, 7(3):359–369, 1998.

[13] Z. Yu and C. Bajaj. Image Segmentation Using Gradient Vector Diffusion and Region Merging. In *ICPR'02*, pages 828–831, Quebec City, August 11-15 2001.

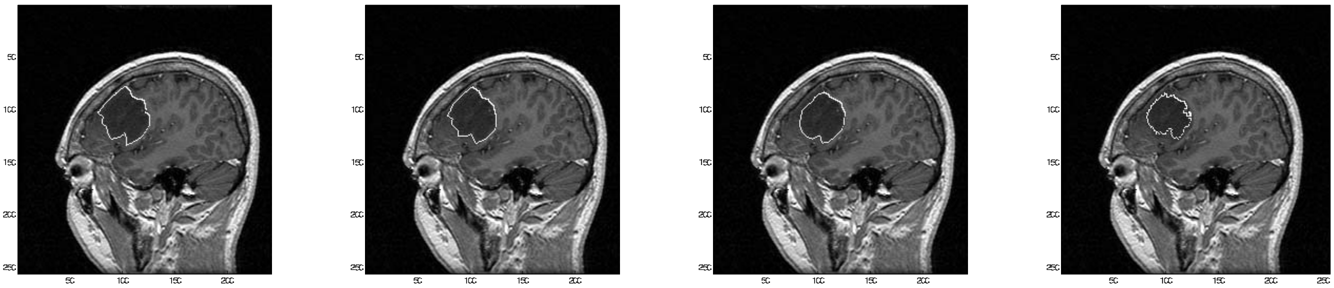


Figure 3. *Manual segmentation by four different experts of a glioma on the MRI.*

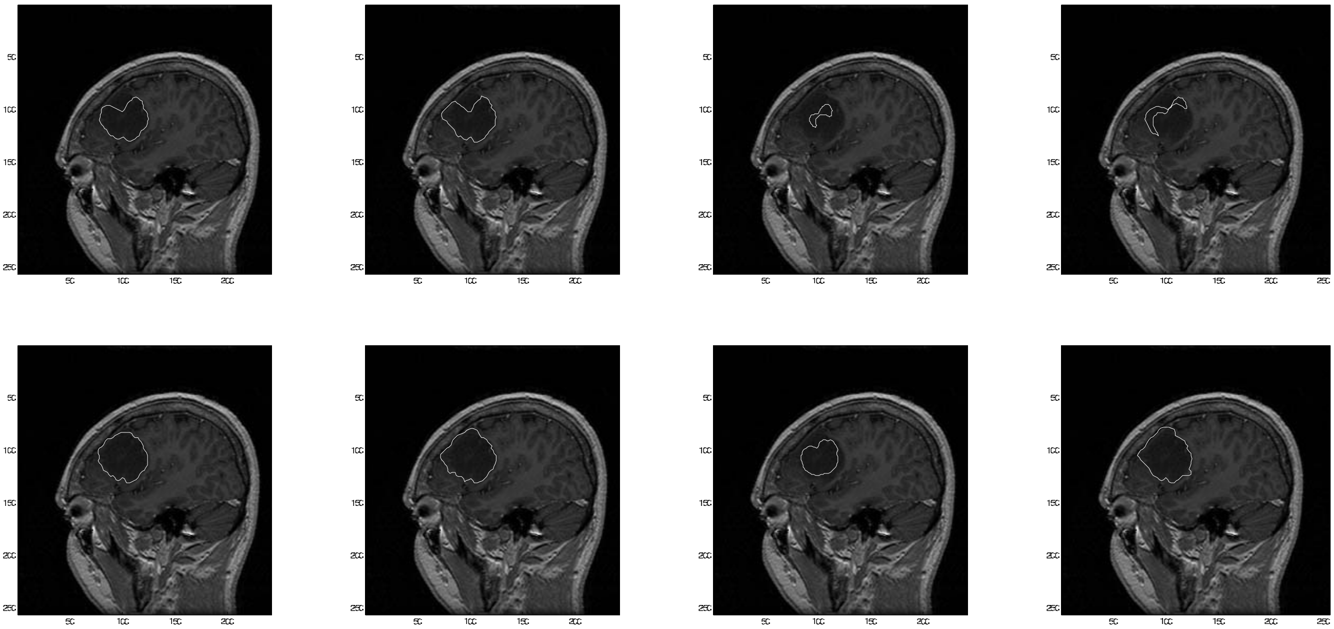


Figure 4. *Automatic Segmentation of the glioma, using two different techniques with four sets of parameters; Top: orignal GVF snake; Bottom: our improved GVF snake.*

Structural image texture and early detection of breast cancer

Shijia Lu and Murk J. Bottema
School of Informatics and Engineering
Flinders University
PO Box 2100, Adelaide SA 5001, Australia
and
Cooperative Research Centre for Sensor Signal and Information Processing
SPRI Building, Mawson Lakes Blvd, Mawson Lakes, SA 4095, Australia
lu@infoeng.flinders.edu.au
murkb@infoeng.flinders.edu.au

Abstract

Structural texture measures are used to address three aspects of early detection of breast cancer in screening mammograms: detection of microcalcification, detection and classification of clustered microcalcification as benign or malignant, and the detection of invasive lobular carcinoma. The use of structural texture features replaces the task of initial detection of complex and poorly modelled image structures such as masses or clustered microcalcifications by initial detection of primitive image structures.

Receiver operating characteristic (ROC) analysis yields high performance scores for detection of microcalcification ($A_z = 0.945$*) and for classification of clusters as benign or malignant (*$A_z = 0.858$*). In the case of invasive lobular carcinoma, cancer was detected in half of the images with no false positive detections. In these cases, no evidence of cancer could be found by visual inspection of the mammogram, thus demonstrating the potential of structural texture measures to encompass diagnostically useful information not accessible to the human expert observer.*

1. Introduction

In the last fifteen years, a plethora of papers have appeared in the literature on the use of computer algorithms for improving early detection of breast cancer in screening mammograms. A number of review papers have also appeared [2],[4], [12]. Currently, nearly all screening mammograms are film mammograms and, in order to implement computer algorithms, images must be digitised. Although commercial systems for computer assisted reading exist that accept film mammograms as input, wide spread use of computer aided screening awaits the impending switch to direct digital acquisition. The technology for direct digital mammography is just emerging and thus the motivation to further improve algorithms for detection is greater than ever before.

Algorithms reported in the literature for detection of breast cancer are designed to search for signs of cancer such as masses, clustered microcalcifications, and stellate patterns. These are the same signs of cancer used by radiologists to evaluate screening mammograms. Detection is difficult because the anomalies due to cancer are complex and are only subtly different from patterns arising from normal tissue. To overcome this problem, many algorithms work in two stages. In the first stage, regions of interest (ROI) are identified that resemble one of the signs of cancer and in the second stage, RIO are analysed more carefully to separate cancer from normal tissue. In the first stage, many false positive detections are accepted with the understanding that the second stage will separate these from true cancer. In the second stage, additional features are measured on the ROI. Some of these features reflect the experience of radiologists, and some, including image texture features, are not based on known manifestations of cancer in mammograms. Texture measures have been shown to contribute positively to discrimination between cancerous and normal tissue [6],[8],[9], [11],[12].

Although many good results have been reported, there are inherent weaknesses in this general approach.

1. Initial detection is focussed on complex signs of cancer.
2. Training studies for developing detection algorithms require accurate information regarding the location of anomalies (truthing). This information is often not available. In the case of detecting clustered microcal-

cifications, for example, nearly all algorithms initially detect individual calcifications. To train such an algorithm, the location in the image of every calcification must be known. Many calcifications are “obvious” but the greatest interest is in subtle calcifications. The error rate in assigning these is large and leaves open the possibility that information, not apparent to visual inspection, is omitted in the analysis.

3. Texture analysis is applied only to ROI, leaving open the possibility that better initial detection could be accomplished if texture information is used throughout the process. Measuring texture features over the entire image for the purpose of initial detection is often not feasible due to computational load.

1.1 Structural texture features

Structural texture refers to statistical distributions of image structures such as lines, edges, or bumps [5]. In contrast, statistical texture refers to statistical distributions of individual pixel values and includes measurements of the local mean, variance, higher order moments, run length statistics, and co-occurrence matrices.

Structural texture analysis provides a means to address the difficulties listed above.

1. Initial detection is focussed on simple structural primitives rather than complex manifestations of cancer.
2. Since statistics are computed over regions, only the disease state of regions (or the entire breast) is required for training.
3. Statistics are computed over image structures rather than pixels. The number of structures is usually at least two orders of magnitude smaller than the number of pixels in the image.

The structures considered in this paper are domains associated with local image intensity extrema. Their use is illustrated by three tasks in automatic detection of breast cancer: the detection of individual microcalcifications, detection and classification of clustered microcalcification, and the detection of invasive lobular carcinoma.

1.2 Clustered microcalcifications

Clustered microcalcifications appear in mammograms as groups of bright dots. Clustered microcalcifications are an important sign of cancer in mammograms, but their detection is not, in itself, a great achievement. The reason is that a very large percentage of normal breasts also show some signs of calcification. Although there are several forms of clusters associated with cancer and several forms associated with benign processes, there are some general differences between benign clusters and ones associated with cancer. Individual calcifications in clusters associated with cancer tend to be more irregular in shape and more varied in terms of contrast and size than those associated with normal tissue. Also, some benign clusters, particularly those associated with calcified blood vessels, can be distinguished based on the shape of the cluster itself [10].

In training algorithms for detecting clustered microcalcifications or for classifying clusters as benign or malignant, it is not possible to know with certainty, exactly which bright spots represent true calcification and which are manifestations of normal tissue such as crossing filaments. From histopathology reports, it is possible to know, with a great deal of certainty, the disease state of the tissue associated with the cluster. This provides motivation for using structural texture measures.

Our implementation of structural texture involves computing three features for every local image intensity maximum in the image: the radius of the bump associated with the local maximum, the height of the bump above background, and a measure of the symmetry of the bump. Detection of clusters is based on local distributions of these three features. In this way, the issue of identifying individual calcifications is bypassed. If there are very small calcifications or other disturbances in the tissue that are associated with cancer but cannot be discerned visually, this information is potentially incorporated into the detection process.

1.3 Invasive lobular carcinoma

The use of structural texture, and in particular the focus on local image intensity bumps, for detecting microcalcifications is natural given the visual appearance of microcalcifications in mammograms and guidance provided by the criteria used by radiologists during visual assessment. In contrast, there is no basis for using these structural textures detecting invasive lobular carcinoma, except by default.

Invasive lobular carcinoma is a form of breast cancer that is not visible at screening much more often than other forms of breast cancer. This form of cancer is not usually associated with the appearance of clustered microcalcifications and in many cases the growth pattern of the tumour is such that masses often are not seen in the mammogram.

Hence, there is no guide by which to select image features that are likely to lead to useful detection algorithms, except that visually obvious features need not be considered. This leaves spatially small and low contrast features as the only likely candidates. As a first attempt, image intensity minima were selected as candidate structure features.

2 Methods and materials

2.1 Data

Images used for these studies were digitised at 50 μ m spatial resolution and 12 bit depth using a Lumisys Lumiscan 150 laser digitiser. A number of image pre-processing steps were used to reduce the non-linear response of the film and digitiser system, to identify the breast region in the image, and to remove patient information, and subtract the background. These steps have been described elsewhere [3].

Images were included as examples of cancer only if the presence of cancer was confirmed by the pathology report and images were included as examples of normals only if no evidence of cancer was found in three years.

2.2 Feature extraction

In each image, local maxima (in the case of detecting microcalcification) and local minima (in the case of detecting invasive lobular carcinoma) were found by comparing pixel values with those of its 8-connected neighbours. For each local extrema point, p, the average pixel value on each concentric ring about p was found. In the case of image maxima, the average ring values form an initially decreasing function of the ring radius. In the case of image minima, the ring values form an initially increasing function. The radius at which these functions ceased to be strictly decreasing or increasing was taken to be the radius, R, associated with the local extremum. The region bounded by the ring of radius R was used to compute the net height, H, the volume, V, and the background level, B, associated with local extremum. The symmetry, S was taken to be the l^2 difference between the image values on the disk of radius R centred at p and the function obtained by revolving the average ring values about p. This latter function can be viewed as the ideally symmetric surface having the same height radius, volume, and average profile as the image intensity surface centred at p. Hence small values of S correspond to symmetric image features.

3 Experimental studies

3.1 Study 1. Detecting individual calcifications

Although the benefit of using structural texture features lies in detecting clusters without initial detection of individual microcalcifications, a study was conducted to determine the suitability of using the structural texture features described above to separate known microcalcifications from other image anomalies [3].

A total 107 individual microcalcifications were marked by a radiologist with experience in mammography. Using the features R, S, V, and H, and linear discriminant analysis, the detection rate was 90 percent at the operating point of 10 percent false positive rate. The area under the receiver operating characteristic (ROC) curve was $A_z = 0.945$.

This study does not confirm that structural texture features surpass other methods for detecting microcalcification. In fact, techniques for detecting microcalcifications are sufficiently mature, that there is very little room for improvement. At the recent International Workshop on Digital Mammography (IWDM, Bremen, June 2002) a panel of radiologists with experience in using commercial systems for automatic detection of microcalcifications agreed that current systems essentially do not miss clusters of diagnostic interest. However, many clusters not associated with caner are also detected by these systems. Hence, work on algorithms for classifying clusters as benign or malignant is still important.

This study does confirm that, in principle, structural texture features identify information relevant to distinguishing microcalcifications from other anomalies.

3.2 Study 2. Identification and classification of clusters

This study comprised 85 images, including 40 images containing clusters of microcalcification associated with benign clusters and 45 images containing clusters associated with cancer. The structural texture features R, S, V, and H were used for initial detection of clusters [7].

The algorithm identified many clusters not marked by the radiologist and assigned separate small clusters within general regions marked by the radiologist as a single cluster. This result cannot be used to measure the accuracy of the algorithm for identifying clusters for two reasons. First, radiologists usually note only the general region of the breast where a cluster or clusters occur and are not in the habit of delineating the spatial extent of individual clusters in detail since, in visual assessment of the mammogram, there is no benefit in doing so. Once a single cluster or region is found in an image that warrants calling the woman back for further tests, other regions are often not be noted even if further evidence of cancer is visible. Second, it is possible that clusters of subtle calcification are present that are not detectable by visual inspection.

The goal of the study was to assign each image as containing benign clusters only or as containing at least one malignant cluster. Accordingly, the performance of the entire system was measured in terms of the correct assignment of images as containing a malignant cluster or not. This process involved the initial detection of clusters as described above plus the extraction of statistical texture features from these regions. In this particular study, classification of the image as cancer or non-cancer was based on the statistical texture features measured on the clusters rather than the structural texture features used for initial detection of the clusters [7].

The classification of test data resulted in an area under the ROC curve of $A_z = 0.858$ if examples of ductal carcinoma in situ (DCIS) comedo type were excluded and $A_z = 0.701$ for all types together. The reason for the discrepancy is that DCIS comedo type calcification is not well characterised by the structural texture features considered here. This type of calcification forms large, linear, and branching structures (the calcium fills ducts and acquires their shape) rather than round spots. The failure of this algorithm to detect these clusters is not important since they are easily discovered by visual assessment of the mammogram or by a separate algorithm designed specifically for detecting DCIS comedo calcification [3].

3.3 Study 3. Detection of invasive lobular carcinoma

A preliminary study was conducted to test the feasibility of detecting invasive lobular carcinoma in screening mammograms [1]. For this project, 24 images with invasive lobular carcinoma were obtained from 12 women. In each of these cases, the screening mammogram had been judged to show no evidence of cancer, but cancer was discovered by other methods within 28 months. The sizes of the tumours ranged from 15 mm to 100 mm in diameter. The mean diameter was 35 mm. As part of normal policy, these mammographically occult cases were reviewed by radiologists with expertise in mammography. Each image was judged to show no evidence of cancer in retrospect. In addition, 24 normal images were included in the study.

In the case of detection and classification of microcalcifications, classification was based directly on the statistics of R, H, V, and S values in local neighbourhoods. In the case of invasive lobular carcinoma, the locations of the tumours in the training data were not well known and so it was necessary to classify entire images as containing evidence of cancer or not. The volume feature, V, was not used in this study, but the local background statistic, B, was used.

First, attention was restricted to two regions within the $RHSB$ feature space. Five images with cancer and five normal images were selected randomly. The H, R, S, B values for all local minima in these images were plotted as points in the four-dimensional feature space. Visual inspection showed that, the great majority of points from the two classes of images formed an indistinguishable glob in the feature space. This was expected since large portions of the images with cancer correspond to normal tissue, and noise characteristics of cancer tissue and normal tissue are the same. However, one portion of the feature space was occupied nearly entirely by points from cancer images only. This portion of the feature space was divided into two sets.

$$\Omega_1 = \{H > 19, R = 1, S \leq 150, B > 2100\}$$
$$\Omega_2 = \{H > 38, R = 2, S \leq 200, B > 2100\}$$

Second, two image features were defined by

$$N_1 = |\Omega_1|/N \qquad \text{and} \qquad N_2 = |\Omega_2|/N,$$

where N is the total number of local minima in the image. Normalisation by N was used to compensate for breast size. Classification was based on these two image features.

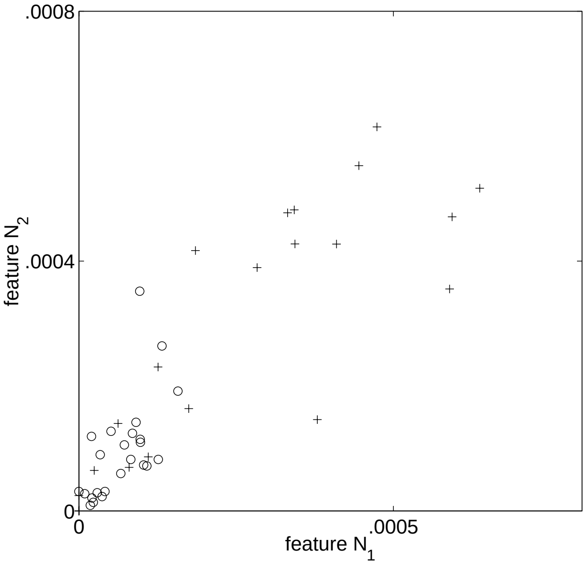


Figure 1. Scatter plot for detection of invasive lobular carcinoma. “o” - normal images, “+” - invasive lobular carcinoma. Approximately half of the carcinoma images cannot be distinguished from the normal images, but the other half are well separated from the normals.

An algorithm for detecting invasive lobular carcinoma must produce very few false positive detections. This is because the radiologist cannot verify a detection by visual inspection of the mammogram and the decision to call the woman back for further tests must be based on the algorithm’s finding alone. For this reason, the results for detection of invasive lobular carcinoma should be reported in terms of the number of detections at the operating point of zero false positives rather than in terms of A_z scores.

In this study 50 percent of the images with invasive lobular carcinoma were correctly detected with no false positive reports.

4 Discussion

Computer algorithms will not replace human interpretation of screening mammograms in the foreseeable future. Current commercial systems for computer assisted screening mammography focus on bringing suspicious regions of the mammogram to the attention of the radiologist and serve mostly to catch evidence of cancer that is missed by the radiologist due to inattention. These systems only detect evidence of cancer that could have been detected by the radiologist.

It is reasonable to conjecture that more information regarding the presence of cancer is present in mammograms than can be extracted by visual interpretation. A viable goal is to develop algorithms that complement and extend the information available to the radiologist rather than mimic this information.

Structural texture measures provide a class of features that are not based on the experience of radiologists but are well suited for embodying information of diagnostic interest. They offer a computational advantages since image segmentation is focused on simple structures, local image extrema in this study, instead of complex structures such as masses, stellate patterns, or clustered microcalcification.

The results on detection and classification of clustered microcalcification indicate that simple structural texture features are useful in identifying known signs of cancer in mammograms. Unfortunately, it has not been possible to compare our classification performance with other studies since there are no public databases available that are both of sufficient resolution and contain a repertoire of both benign and malignant clusters.

Invasive lobular carcinoma represents approximately 8 percent of breast cancer. At screening, roughly half of these cases can be detected by visual inspection. Thus 4 percent of breast cancer cases are not be detected at screening because they are examples of occult invasive lobular carcinoma. In our study, half of these occult cases were detected indicating a potential improved detection rate for breast cancer of 2 percent.

The structure features used in the detection of invasive lobular carcinoma were not motivated by models or by the experience of radiologists. The only guidance was that these carcinomas could not be seen in the mammograms by radiologists and thus large or high contrast features could be neglected. Local extrema were chosen mostly because algorithms for characterising these structures were available from previous studies. The results obtained were surprisingly good considering that these features were largely unmotivated and were the first ones tested for this purpose. Unless these choices were very fortuitous, this implies that other structure features may well supply equal or greater discriminatory power for detecting invasive lobular carcinoma.

The study on detection of invasive lobular carcinoma was based on only 48 images and should be regarded as a preliminary study.

References

[1] M. J. Bottema and J. P. Slavotinek. Automatic detection of invasive lobular carcinoma in screening mammograms. submitted.

[2] M. J. Bottema and J. P. Slavotinek. Computer aided screening mammography. In B. Pham, editor, *New Approaches in Medical Image Analysis*, volume 3747, pages 177–190. SPIE, 1999.

[3] M. J. Bottema and J. P. Slavotinek. Detection of microcalcifications associated with cancer. In M. J. Yaffe, editor, *IWDM 2000 5 th International Workshop on Digital Mammography*, pages 149–153, 2001.

[4] H.-P. Chan, N. Petrick, and B. Sahiner. Computer-aided breast cancer diagnosis. In A. Jain, A. Jain, S. Jain, and L. Jain, editors, *Artificial Intelligence techniques in breast cancer diagnosis and prognosis*, pages 179–264. World Scientific, 2000.

[5] R. C. Dubes and A. K. Jain. Random field models in image analysis. *Journal of Applied Statistics*, 16(2):131–164, 1989.

[6] H. D. Li, M. Kallergi, et al. Markov random field for tumor detection in digital mammography. *IEEE Transactions on Medical Imaging*, 14(3):565–576, 1995.

[7] S. Lu. Texture analysis for classification of microcalcification as benign or malignant in screening mammograms. PhD thesis, in preperation.

[8] N. R. Mudigonda, R. M. Rangayyan, and J. E. L. Desautels. Gradient and texture analysis for the classification of mammographic masses. *IEEE Transactions on Medical Imaging*, 19(10):1032–1043, 2000.

[9] B. Sahiner, H.-P. Chan, N. Petrick, M. A. Halvie, and M. M. Goodsitt. Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis. *Medical Physics*, 25(4):516–526, 1998.

[10] L. Tabar and P. B. Dean. *Teaching Atlas of Mammography*. Thieme, New York, 1985.

[11] G. M. te Brake, N. Karssemeijer, and J. H. C. L. Hendriks. An automatic method to discriminate malignant masses from normal tissue in digital mammograms. *Physics in Medicine and Biology*, 45:2843–2857, 2000.

[12] C. J. Vyborny, M. L. Giger, and R. M. Nishikawa. Computer-aided detection and diagnosis of breast cancer. *Radiology Clinics of North America*, 38(4):725–740, 2000.

(This page left blank intentionally)

Automating Cell Segmentation Evaluation with Annotated Examples

Pascal Bamford
Cooperative Research Centre for Sensor Signal and Information Processing,
Department of Information Technology and Electrical Engineering,
The University of Queensland
E-mail: P.Bamford@cssip.uq.edu.au

Abstract

Previously the development of a cell nucleus segmentation algorithm had been evaluated by eye by the author. This is an impractical method when attempting to evaluate and compare many algorithms and parameter sets on very large data sets. For this work, a dataset of 20,000 cell nucleus images was annotated by hand by three non-expert assistants. This paper concentrates on comparing the previous interactive approach to evaluating a segmentation algorithm to automated techniques using this annotated data.

1. Introduction

We have previously reported work on using a dynamic-programming algorithm for cell nucleus segmentation [1]. In that work, almost 20,000 cell images were segmented and the output judged as either a *pass* or a *fail* by the author (where any deviation from the perceived boundary was declared a fail). Then, an attempt was made to tune the algorithm parameter over a subset of the data using the same evaluation method.

This is clearly an extremely time consuming and subjective process. In fact, of the arguments encouraging more evaluation in computer vision [2], it seems that finding a better method than *eye-balling* results over the large datasets required to develop real systems is the most compelling! This is especially so if many algorithms and parameter sets are to be compared thoroughly.

The ultimate method of evaluating segmentation algorithms is to use the final outcome of the complete vision system as the performance metric [3]. Unfortunately this is very difficult except in the simplest of cases. This is due to the fact that there may be many processes, each with their own sources of variability and complex interactions, between the segmentation output and final measure [4], thus requiring very large datasets in order to perform a robust experiment. Attempts at evaluation therefore either evaluate individual components in isolation or consider the final outcome of the system [6].

Image segmentation is a module that is generally evaluated in isolation. This is either done by eye, via annotated examples or via some other goodness measure that does not rely on *ground truth* (e.g. inter-region contrast). This latter method is generally the only available option to those working in *general recovery* [2] work exemplified in [7]. Of the methods that employ annotated examples (Zhang's *empirical discrepancy methods* [8]), either the segmentation masks are pixel-wise compared or features extracted from those masks are compared. The latter method has been criticized as it is possible to obtain good agreement for a feature where the masks do not agree well [9] (trivial example: area). Also extracted features can be very sensitive to small differences in masks, complicating the detection of *significant* differences [5].

The difficulties associated with empirical discrepancy evaluation have been summarized to be [9]

- difficulties in defining measures/metrics,
- standardizing evaluation protocols, but mostly
- determining and acquiring ground truth data.

It is well known that image segmentation is a highly application-dependent task. Previous approaches to evaluation, which are briefly summarized in the following two sections, seem to show that this task is also more application dependent that one may expect. Selected techniques are then applied to the task of verifying results previously obtained by eye for cell segmentation [1].

2. Error Measures and Metrics

A framework for evaluating segmentation methods has recently been proposed where the measure was trained using examples of failure [10]. The error value measured edge-detection type errors (*bits* - false positive edges, and *holes* - false negative edges) that were assembled into patterns and then rated by human observers. The individual errors were weighted by a number of parameters and the measure trained to match the observers' score. This measure was classified as belonging to a group termed *low error models*, i.e. suited only to problems where the segmentation is already very near the final solution (and was tested upon synthetic images).

We investigated the failure modes for a number of algorithms in [1] and found that they generally either failed quite dramatically or performed an *acceptable* job (little or no perceived delineation error). Thus we are initially more interested in employing a measure that is capable of measuring *large* differences.

Zhang [8] reviewed a number of simple measures of which only two are applicable here: the number and position of misclassified (segmented) pixels. Also reviewed were methods for measuring over- and under-segmentation. These errors, and those of Roman-Roldin [10], are of less interest in this work as a well-formed mask is a pre-requisite to accepting the segmentation. Thus we assume that a mask for the object of interest is the final output of the segmentation stage (including all pre- and post-processing). A simple method of error checking, for this application, is then to evaluate the Euler number of the mask image. If it is not equal to one, then the mask is rejected outright and a failure assigned to that segmentation - the failure need no longer be quantitatively evaluated. This may be seen as first-step goodness measure that requires no ground truth.

More recently Chalana [9] employed the Hausdorff distance and average distance between human and computer boundaries. The Hausdorff distance is an attractive metric for this application as segmentation algorithms generally tend to fail in one localized position around the nuclear border - an error that may become masked when using normalized or average values [11]. The Hausdorff distance is defined to be the maximum of the set of shortest distances between corresponding points of two shapes.

3. Evaluation Protocols

Algorithms are generally evaluated using either the raw measures to establish how close competing algorithms (and different parameters sets) get to the annotated (observers') data or by thresholding the measures in order to obtain percentage success rates which are then compared. Chalana [9] used both methods and a number of observers' data to determine whether the computer boundary differed from the observers' boundaries as much as the observers' boundaries differed from one another. This was evaluated using a modified Williams' index and a *percent statistic*. The Williams' index, I', divides the average number of agreements (inverse disagreements, $D_{j,j'}$) between the computer ('observer' 0) and $n-1$ human observers (j) by the average number of agreements between human observers (eq. 1).

$$I' = \frac{\frac{1}{n}\sum_{j=1}^{n}\frac{1}{D_{0,j'}}}{\frac{2}{n(n-1)}\sum_{j}\sum_{j':j'\neq j}\frac{1}{D_{j,j'}}} \qquad (1)$$

If the upper value of the confidence interval of the result is greater than one, then it is concluded that the computer is a reliable member of the group of observers. The percent statistic measures the percentage of cases where the computer boundary lies within the inter-observer range.

Finally Everingham [12] has recently suggested an interesting method for combining large sets of results into an ROC type curve, where only the best performing results contribute to the final output.

4. Method

Although no *ground truth* exists for cell segmentation, the images do not necessarily require expert annotation. Figure 1 shows an example image and three corresponding non-expert annotations. This is quite a deviation from other imaging modalities where inter- and intra- observer variance can be very high, but valid (e.g. ultra-sound [9]). It is however desirable to obtain a number of interpretations so that inter-observer variability may be nonetheless investigated.

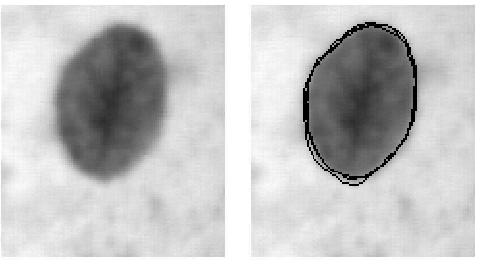


Figure 1. Example from dataset and corresponding observers' boundary

A Wacom PL400 pen-and-tablet was used to input the data. This device enabled almost immediate use by the observers to delineate the cell nuclei. The observers were instructed to draw a continuous line between the nucleus and background (cytoplasm), i.e. on the *transition region* [13] of the edge. Due to the low pass filtering effect of the optics used to capture the images, this covers a number of pixels and is readily identifiable in the majority of examples. However the exact area for delineation was not overly specified and left to the individual. Three observers were employed to annotate the entire 20,000 image dataset. The nucleus images are of the order of 128x128 pixels. The PL400 LCD screen has square pixels of pitch 0.264mm. By displaying the images at the native screen resolution therefore produced cell nuclei of approximately 1-2cm diameter on screen. This was found to be too fiddly and handshake became a problem. Thus the images were first upsampled to twice the original dimensions using a nearest-neighbour algorithm. The pen line thickness on the screen was made equal to one pixel at the original image resolution (i.e. four pixels on screen). This also assisted to reduce handshake.

There has been considerable work in improving the implementational performance of the Hausdorff metric for the more general problem of comparing shapes under transformation [14]. Here, we have implemented a rapid and simple routine to obtain the maximum distance between corresponding points, d_{MAX}, on two binary masks by

1. Obtaining the distance transforms, A_{DT} and B_{DT}, of the perimeter of the mask images, A and B.

2. Obtaining the pixel-wise maximum of A_{DT} and B_{DT} to produce AB_{DT}.

3. Obtaining the XOR of the mask images, AB_{XOR}

4. Using AB_{XOR} to mask AB_{DT} to produce AB_{MASK}

5. Obtaining d_{MAX} as the maximum value in AB_{MASK}.

These steps are illustrated in figure 2. The average distance between the masks was also computed. Chalana [9] used an iterative technique to evaluate the average distance between two curves, which yielded an average curve as a result. Here we implemented a rapid method of evaluating the average distance, d_{AV}, as the average value in AB_{MASK}. These two measures can be obtained very rapidly using operations for which implementations are widely available.

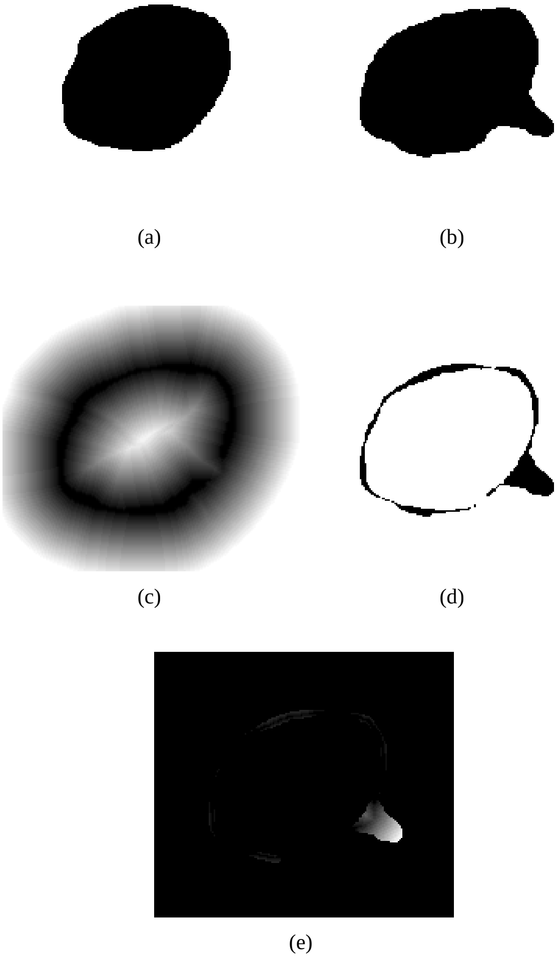

(a)
(b)
(c)
(d)
(e)

Figure 2. Steps for deriving the maximum distance between two binary masks. (a) and (b) are the two masks to be compared. (c) is the pixel-wise maximum of the distance transforms (DT) of the perimeters of (a) and (b). (d) is the result of the XOR operation on the two masks. (e) shows the final masked DT, the maximum value of which is the desired value, corresponding to the length of the protrusion in the mask (b).

The above data and measures were then used to confirm the results reported in [1]. This was done by comparing the algorithm performance against its (regularisation) parameter, λ. The Williams' index was first computed, using d_{MAX} as the discrepency measure $D_{j,j'}$, over half of the data over the full range of permissible [1] values of λ ($\in [0,1]$) at increments of 0.1. Figure 3 shows an example distribution of the Williams' index for $\lambda = 0.2$. This plot shows a group of (normally distributed) values near the observers' boundaries with a mean value near 1.0. In addition, there are a number of out-lying counts between 0.0 and roughly 0.5. These correspond to the failed segmentations. Thus rather than compare performance versus λ using summary statistics, as in [9], the Williams' index was thresholded in order to determine the percentage of correct segmentations at that threshold. However the selection of such a threshold is at this stage an arbitrary process - what constitutes a *correct* segmentation? Therefore algorithm performance was plotted over the natural range of the Williams' index, i.e. $\in [0,1]$. At a value of 0, the computer boundary is infinitely far from those of the observers. At a value of 1 the computer boundary is as close to the observers as they are to each other. Values above 1 are of less

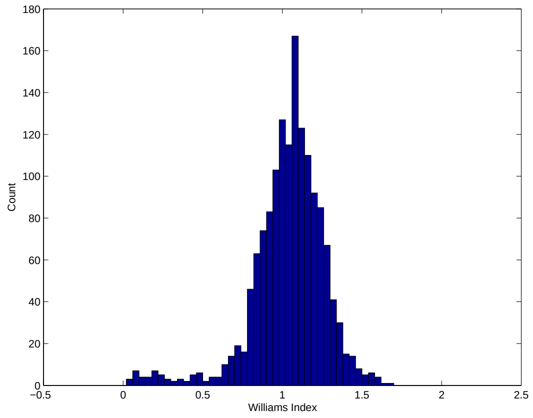


Figure 3. Distribution of the Williams index for $\lambda = 0.2$

interest - this can be understood to occur when one of the observers disagrees with the other two *more* than the computer does. Therefore all values greater than 1 are treated as a correct segmentation. Figure 4 is a convenient normalised and bounded representation to view segmentation algorithm performance against multiple observers in a single output. As a measure of overall performance at each value of λ, the area under this curve was computed and is represented in figure 5. The maximum value of these area values is 95.16% which occurs at $\lambda = 0.5$.

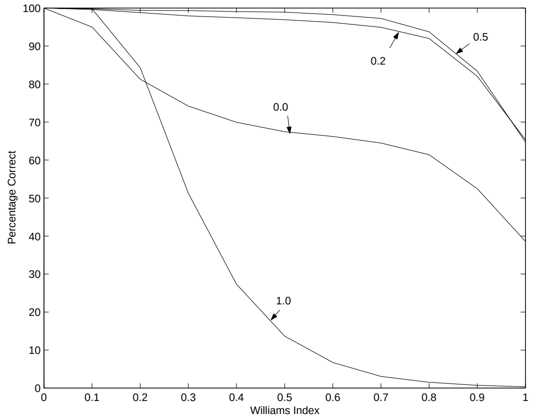


Figure 4. Plot of cumulative percentage success segmentation versus Williams Index for values of λ.

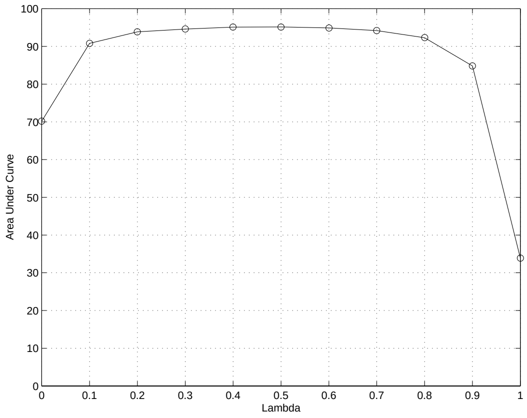


Figure 5. Plot of area under curves of figure 4 against λ.

5. Conclusion

We have considered methodologies for evaluating cell segmentation using annotated examples in an automated fashion. We found that recent approaches to segmentation evaluation have concentrated on low error models where the measures and metrics for segmentation error, in addition to the evaluation procedures, were unsuitable for this application. The shape of the graph in figure 5, the optimal value of algorithm parameter λ and the overall performance rate all agree well to the results reported in [1]. However, they were attained using a far more satisfactory and repeatable method.

This work represents a very early part of a greater project to thoroughly evaluate cell segmentation methods using annotated examples. The next stage is to attempt to compare other algorithms for this task. In addition, individual modules may be evaluated in isolation. For example marker extraction algorithms may be evaluated using the same data but measures that detect whether an inner marker is completely within the desired object. This work will then be expanded to include the original scene images from which the nucleus images were captured, representing a different segmentation task. Also, methods that use annotated data to obtain edge and region models, enabling the improvement or design of algorithms [15] will eventually be investigated. Finally, once a small number of discrepancies in the observers' data have been fixed (i.e. where the Williams' index is significantly greater than one!), this dataset will be made publicly available.

6. Acknowledgement

The author wishes to sincerely thank Wacom Co., Ltd for financially assisting the purchase of the PL400 pen-and-tablet.

References

[1] P. Bamford and B. Lovell, “Unsupervised cell nucleus segmentation with active contours,” *Signal Processing*, vol. 71, no. 2, pp. 203–13, 1998.

[2] R. C. Jain, T. O. Binford, M. A. Snyder, Y. Aloimonos, A. Rosenfeld, T. S. Huang, K. W. Bowyer, and J. P. Jones, “Ignorance, myopia, and naivete in computer vision systems,” *CVGIP: Image Understanding*, vol. 53, no. 1, pp. 112–28, 1991.

[3] B. McCane, “On the evaluation of image segmentation algorithms,” in *DICTA’97 and IVCNZ’97*, 1997, pp. 455–9.

[4] R. M. Haralick, “Performance characterization in computer vision,” in *Proceedings of 5th International Conference on Computer Analysis of Images and Patterns (CAIP’93)*, D. Chetverikov and W. G. Kropatsch, Eds. Washington Univ. Seattle WA USA, 1993, pp. 1–9.

[5] C. MacAulay, “Development, implementation and evaluation of segmentation algorithms for the automatic classification of cervical cells,” PhD, University of British Columbia, 1989.

[6] M. Greiffenhagen, D. Comaniciu, H. Niemann, and V. Ramesh, “Design, analysis, and engineering of video monitoring systems: an approach and a case study.” Visualization Dept. Siemens Corp. Res. Inc. Princeton NJ USA, 2001.

[7] K. Cho, P. Meer, and J. Cabrera, “Performance assessment through bootstrap,” *IEEE-Transactions-on-Pattern-Analysis-and-Machine-Intelligence*, vol. 19, no. 11, pp. 1185–98, 1997.

[8] Y. J. Zhang, “A survey on evaluation methods for image segmentation,” *Pattern-Recognition*, vol. 29, no. 8, pp. 1335–46, 1996.

[9] V. Chalana and Y. Kim, “A methodology for evaluation of boundary detection algorithms on medical images,” *IEEE Transactions on Medical Imaging*, vol. 16, no. 5, pp. 642–652, 1997.

[10] R. Roman-Roldan, J. F. Gomez-Lopera, C. Atae-Allah, J. Martinez-Aroza, and P. L. Luque-Escamilla, “A measure of quality for evaluating methods of segmentation and edge detection,” *Pattern Recognition*, vol. 34, no. 5, pp. 969–80, 2001.

[11] A. Hammoude, “An empirical parameter selection method for endocardial border identification algorithms,” *Computerized Medical Imaging and Graphics*, vol. 25, pp. 33–45, 2001.

[12] M. Everingham, H. Muller, and B. Thomas, “Evaluating image segmentation algorithms using the pareto front,” *Lecture Notes in Computer Science*, vol. 2353, pp. 34–48, 2002.

[13] J. J. Gerbrands, “Segmentation of noisy images,” PhD, Delft University, The Netherlands, 1988.

[14] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, “Comparing images using the hausdorff distance,” *IEEE-Transactions-on-Pattern-Analysis-and-Machine-Intelligence*, vol. 15, no. 9, pp. 850–63, 1993.

[15] M. Brejl and M. Sonka, “Object localization and border detection criteria design in edge-based image segmentation: automated learning from examples,” *IEEE Transactions on Medical Imaging*, vol. 19, no. 10, pp. 973–85, 2000.

(This page left blank intentionally)

Optimal Geodesic Active Contours: Application to Heart Segmentation

Ben Appleton
Intelligent Real-Time Imaging and Sensing Group, ITEE
The University of Queensland, Brisbane, Queensland 4072, Australia

Abstract

We develop a semiautomated segmentation method to assist in the analysis of functional pathologies of the left ventricle of the heart. The segmentation is performed using an optimal geodesic active contour with minimal structural knowledge to choose the most likely surfaces of the myocardium. The use of an optimal segmentation algorithm avoids the problems of contour leakage and false minima associated with variational active contour methods. The resulting surfaces may be analysed to obtain quantitative measures of the heart's function.

We have applied the proposed segmentation method to multislice MRI data. The results demonstrate the reliability and efficiency of this scheme as well as its robustness to noise and background clutter.

1 Introduction

The segmentation of organs in medical images is a challenging problem due to the high geometric variation common both between subjects and within a single subject. Local feature based methods are unable to account for geometric and structural properties and often yield unreliable segmentations. Rigid templates are rarely applicable however deformable templates and active contours have more prospect for success.

Classical active contours including snakes [11] and level sets [15, 1] have shown promise in a range of medical image segmentation problems. The relatively recent geodesic active contour framework has been shown to be simple, efficient and relatively accurate [7, 9]. These methods typically use a variational framework to obtain locally minimal contours by gradient descent of an energy functional. As a result the final segmentations are dependent upon their initialisation, requiring a simpler and less reliable segmentation as input. They also have a tendency to leak through gaps in object edges due to noise or indistinct boundaries and may become caught in irrelevant local minima.

Zhukov et. al. applied a 3D snake to segment the cavity of the left ventricle [18]. However their approach required a high level of user interaction to prevent contour leakage. To overcome the problem of contour leakage Ho et. al. used a competitive level set framework demonstrating significant improvement. They applied their approach to segment brain tumours in 3D magnetic resonance images [10].

A range of optimal active contour methods have been developed which avoid the problems of the variational framework. In the last decade numerous shortest path algorithms have been applied to locate curvilinear image features including road and valley detection from satellite images and crack detection on borehole cores [5, 6, 14].

In object segmentation the topology of the problem demands closed boundary curves. In [4] Bamford and Lovell considered the problem of segmenting nuclei in cell microscopy. They computed shortest paths across a polar trellis with the added restriction that the endpoints of the path met. The problem of shortest paths with connected endpoints was solved to optimality by Appleton and Sun [2] who used a branch and bound search to efficiently obtain the shortest closed path on trellises.

These approaches were recently unified by Appleton and Talbot to give an optimal form of geodesic active contour [3] . The resulting segmentations demonstrate superior quality to classic geodesic active contours for similar computational effort. This method is ideal for the segmentation of deformable objects characterised by homogeneous features making them attractive for medical image segmentation.

Our driving application is to assist the analysis of functional pathologies of the left ventricle in the human heart. The goal of this project is to provide quantitative measures of the heart's function, including the muscle thickness throughout the LV wall, the internal volume and the muscle volume, and the ejection fraction. Data is obtained in the form of multislice magnetic resonance images. Slices are acquired in a double oblique plane to the body along the long axis of the heart with $10\mathrm{mm}$ separation. The pixels composing each slice are $1.4\mathrm{mm} \times 1.4\mathrm{mm}$.

The aim of this paper is to present a segmentation method for this data modality which will robustly and accurately extract the epicardium and the endocardium of the left ventricle. The resulting surfaces may then be analysed to compute quantitative measures of the myocardial geometry. Due to the large separation between the slices and the lack of registration three-dimensional segmentation techniques are not considered. Instead we propose to independently segment each slice to form a layered collection of contour lines which may be suitably interpolated to approximate the myocardial surface.

2 Optimal Geodesic Active Contours

Our method utilises an existing algorithm for the segmentation of homogeneous objects under the geodesic active contour energy functional due to Appleton and Talbot [3]. This algorithm gives the simple closed curve of globally minimal energy which is required to contain a specified internal point p_{int}. This internal point selects the object of interest and may form the only input parameter to the algorithm, yielding a highly automated optimal object segmentation scheme.

The image to be segmented is represented as a Riemannian space S with a metric g induced by the image content. The metric quantifies the local homogeneity of the image. Using this space segmentation is achieved by locating the minimal closed geodesic containing p_{int}. A key feature of this approach is the separation of the segmentation problem into local feature extraction and global geometric optimisation.

The minimisation objective is the energy functional

$$E(C) = \int_C g(C(s))\, ds \tag{1}$$

where C is the segmentation contour and g is the metric. Locally minimal contours C are known as *geodesics*. For images composed of objects characterised by homogeneous intensity we choose a positive scalar metric of the form

$$g = \frac{1}{r}\left(\frac{1}{1 + |\nabla G_\sigma \star I|^p} + \varepsilon\right) \tag{2}$$

where $p = 1$ or 2 and r is the distance from p_{int}. Here G_σ is a Gaussian of variance σ^2 such that the denominator is a measure of the local intensity discontinuity at scale σ. $\varepsilon > 0$ is an arc-length penalty which implicitly smooths the contour. Here we alter the metric typically used in geodesic active contours [7, 9] by including an inverse radial weighting. This introduces scale invariance into the energy functional rendering it suitable for global minimisation.

Geodesics are computed on the discrete grid using Sethian's fast marching method [16, 8]. For the computation of closed geodesics we form a space S identical to the Riemann surface for the natural logarithm relation by augmenting the image plane $\mathbb{R}^2$. This space naturally embeds the information of whether a closed contour contains p_{int} without restricting the contour in any way.

Minimal closed geodesics are located efficiently using a best-first branch and bound search tree adapted from work by Appleton and Sun on circular shortest paths [2]. The optimal closed contour partitions the image into the object and the background such that the total similarity across the partition border is minimised.

This algorithm has been applied to the segmentation of microscope, x-ray, magnetic resonance and cDNA microarray images [3]. The resulting segmentations have been shown to be isotropic and demonstrate robustness to gaps in object boundaries as well as low sensitivity to the placement of the interior point p_{int}. They have been successfully applied to concave and convoluted boundaries, demonstrating the flexibility of this approach. The new approach compares quite favourably with the classic curve evolution approach of Caselles et. al. [7], achieving more reliable and accurate segmentations with very similar computational effort. However as opposed to classic geodesic active contours the segmentation is restricted to be a simple closed curve.

3 Adapting Optimal GAC

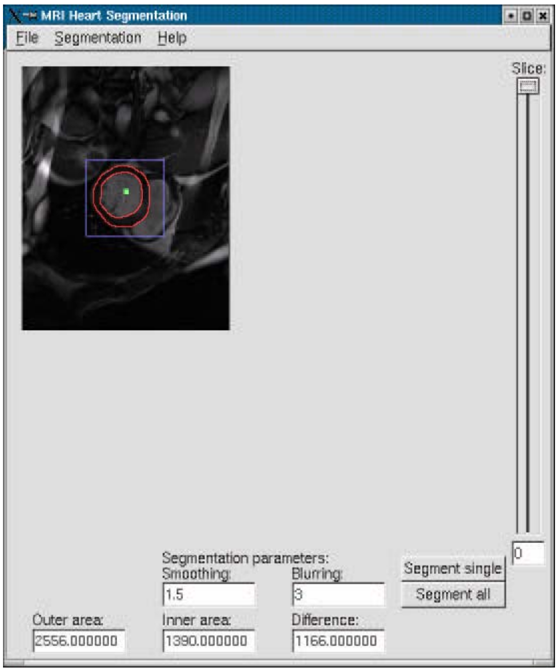
Due to the fact that the left ventricle is the major systemic pump for the cardiovascular system it has an approximately circular cross-section [13]. The internal cavity of the left ventricle appears brighter than the surrounding tissue as it is filled with blood. The myocardium appears darker than most of the surrounding tissue, with the exception of the cavity of the right lung. We therefore compute a radial gradient for each slice relative to the internal point, which is placed toward the centre of the LV. Strong negative gradients are likely to correspond to the intensity drop across the endocardium between the blood-filled LV cavity and the muscle of the myocardium, while strong positive gradients are likely to correspond to an intensity increase across the epicardium from the myocardial muscle to the generally brighter surrounding tissues.

From the radial gradient we generate a pair of metric images, g_{epi} and g_{endo}. Each metric should be low on points which are deemed likely from local features to lie on the epi- or endocardial surface respectively, and high on points which are unlikely to lie on the surface of interest. This inspires the following pair of metrics:

$$g_{epi} = \frac{1}{r}\left(\frac{1}{1 + \mathrm{u}(\nabla G_\sigma \star I)\,|\nabla G_\sigma \star I|} + \varepsilon\right) \tag{3}$$

$$g_{endo} = \frac{1}{r}\left(\frac{1}{1 + \mathrm{u}(-\nabla G_\sigma \star I)\,|\nabla G_\sigma \star I|} + \varepsilon\right) \tag{4}$$

where u is Heavyside's function [12].



Slice: 0

Segmentation parameters:
Smoothing: 1.5
Blurring: 3

Outer area: 2556.000000
Inner area: 1390.000000
Difference: 1166.000000

Figure 1. A snapshot of the user interface.

Segmentation is then performed independently on each metric image using the optimal geodesic contour algorithm of Appleton and Talbot[3].

4 User Interface

A prototypical user interface has been designed using the FLTK cross-platform GUI builder [17] in order to develop the segmentation scheme (Figure 1). The graphical interface allows users to view each slice in turn, alter segmentation parameters such as the scale of Gaussian blurring and the contour regularity, and specify a point inside the left ventricle for the segmentation. An optional cropping box may be used to restrict the range of the segmentation contour and increase the speed of the segmentation. Segmentations may be performed on the entire dataset with common parameters or on individual slices for greater control.

5 Results

Figure 2 depicts a segmentation of an 8-slice dataset with common parameters. The subject was part of a research project aimed at measuring myocardial viability and had had a heart attack within the last six months. No contrast agent was used.

Observe that this segmentation method does not require that the slices be registered; the fifth and sixth slices are clearly misaligned by 14mm in the vertical. Despite this perturbation to the placment of p_{int} the epicardium and endocardium are still correctly segmented.

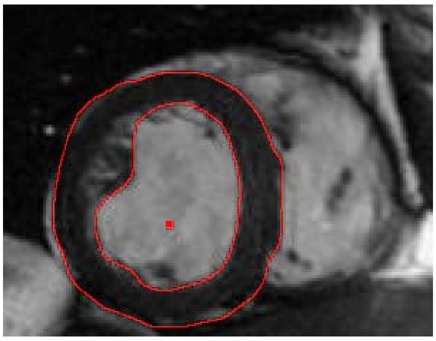


Figure 3. A single-slice segmentation at higher resolution (300×220). This slice is taken in the coronal plane and is rotated by a half circle with respect to the slices of Figure 2.

The epicardial surface is not easily distinguished by local edge strength alone due to the presence of the dark cavity in the right lung. This dark cavity produces weak edges along the heart-lung interface. Unlike classic geodesic active contours which are prone to leak through weakly defined edges the optimal geodesic active contour framework is able to correctly segment this interface in all slices.

The segmentation of the endocardial surface in the second, third and fourth slices shows a weakness of this segmentation technique. In these slices the segmentation contour follows a strong edge corresponding to a papillary muscle just interior to the superior surface of the LV cavity, instead of correctly tracking the less distinct edge of the endocardium. As a result the segmentation overestimates the thickness of the superior portion of the myocardium. This may be solved using prior structural knowledge of the heart or morphological preprocessing to remove small linear features such as papillary muscles.

All segmentations were performed on a 700MHz P-III Toshiba laptop with 192MB of RAM under the Linux operating system. The segmentation routines have been implemented in C and C++. Each slice of Figure 2 took 0.3 seconds to segment including all pre- and post-processing.

Figure 3 depicts the segmentation of a single slice at higher resolution. The increased resolution affords a better segmentation, however the computational load is greatly increased at 6.0 seconds per slice. A multiscale approach may be warranted to reduce the computational load at high resolutions.

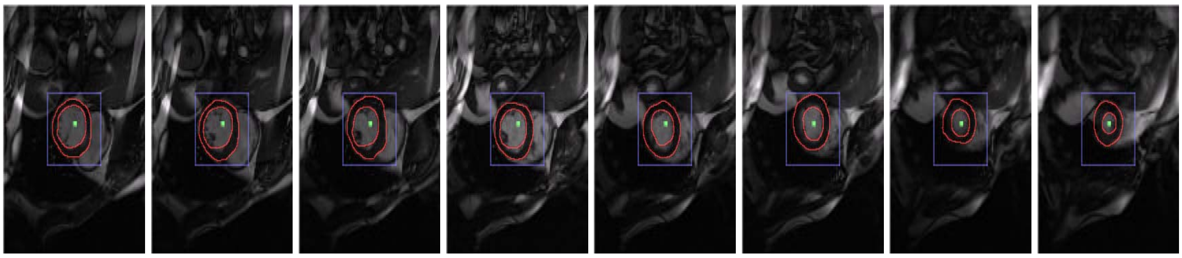


Figure 2. A multislice segmentation of the epi- and endomyocardial surfaces of the left ventricle. Segmentation is performed within an 80×75 crop box.

6 Conclusions

We have developed an automated method to assist in the analysis of functional pathologies of the left ventricle of the human heart. To do so we propose an active contour based method for segmenting multislice heart MR data. First we compute a pair of metric images which encapsulate local knowledge about the presence of the myocardial surface. Then we apply an optimal geodesic active contour algorithm to choose the most likely closed curves in each slice corresponding to the myocardium of the left ventricle. The resulting surfaces may be analysed to compute quantitative measures of the left ventricle's function such as myocardial thickness, internal and muscle volumes and ejection fractions. In addition we designed a graphical interface to facilitate the use of this segmentation scheme and the analysis of the resulting contours (see Figure 1).

We have applied the proposed segmentation method to a multislice dataset and a higher resolution single-slice image. The results have shown that this scheme is reliable and efficient, and that it performs well in the presence of indistinct boundaries and background clutter (see Figures 2 and 3). It was shown to be confused by the inclusion of papillary muscle fibers interior to the LV, suggesting the addition of prior structural knowledge or morphological preprocessing to avoid overestimating the thickness of the myocardium in these regions.

We are currently considering ways in which an expert user may interact with the segmentation. Soft spatial weightings may be used to bias the contour optimisation toward user-specified contour points without forcing the segmentation contour to pass through an inaccurately placed point, combining the knowledge of the expert with the precision of this segmentation method.

Acknowledgements

I would like to acknowledge Dr. Stephen Rose, Centre for Magnetic Resonance, University of Queensland and the Centre for Advanced MRI, The Wesley Hospital, Brisbane for making available the cardiac MR data presented in this paper.

References

[1] D. Adalsteinsson and J. A. Sethian. A fast level set method for propagating interfaces. *Journal of Computational Physics*, 118(2):269–277, 1995.

[2] Ben Appleton and Changming Sun. Circular shortest paths by branch and bound. *Pattern Recognition*, 2002. Submitted.

[3] Ben Appleton and Hugues Talbot. Globally optimal geodesic active contours. *Journal of Mathematical Imaging and Vision*, 2002. Submitted.

[4] Pascal Bamford and Brian Lovell. Unsupervised cell nucleus segmentation with active contours. *Signal Processing (Special Issue: Deformable models and techniques for image and signal processing)*, 71(2):203–213, 1988.

[5] M. Barzohar and D. B. Cooper. Automatic finding of main roads in aerial images by using geometric-stochastic models and estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(7):707–721, 1996.

[6] M. Buckley and J. Yang. Regularised shortest-path extraction. *Pattern Recognition Letters*, 18(7):621–629, 1997.

[7] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *IJCV*, 22(1):61–79, 1997.

[8] Laurent D. Cohen and Ron Kimmel. Global minimum for active contour models: A minimal path approach. *International Journal of Computer Vision*, 24(1):57–78, August 1997.

[9] R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky. Fast geodesic active contours. *IEEE Trans. On Image Processing*, 10(10):1467–1475, 2001.

[10] Sean Ho, Elizabeth Bullitt, and Guido Gerig. Level-set evolution with region competition: Automatic 3-D segmentation of brain tumors. In *Proceedings of the 16th International Conference on Pattern Recognition*, August 2002.

[11] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1998.

[12] E. Kreyszig. *Advanced Engineering Mathematics*. W. Anderson, seventh edition, 1993.

[13] Elaine N. Marieb. *Human Anatomy and Physiology*. The Benjamin/Cummings Publishing Company, Inc., fourth edition, 1998.

[14] N. Merlet and J. Zerubia. New prospects in line detection by dynamic programming. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(4):426–431, April 1996.

[15] Stanley Osher and James A Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. *Journal of Computational Physics*, 79:12–49, 1988.

[16] J. Sethian. A fast marching level set method for monotonically advancing fronts. In *Proceedings of the National Academy of Sciences*, volume 93(4), pages 1591–1595, 1996.

[17] Bill Spitzak et al. The fast light toolkit home page. http://www.fltk.org.

[18] L. Zhukov, Z. Bao, I. Guskov, J. Wood, and D. Breen. Dynamic deformable models for 3D MRI heart segmentation. In *Proceedings of SPIE Medical Imaging 2002*, pages 1398–1405, February 2002.

(This page left blank intentionally)

An overview of the Polartechnics SolarScan melanoma diagnosis algorithms

Hugues Talbot and Leanne Bischof
CSIRO – Mathematical and Information Sciences
Locked Bag 17, Building E6B, Macquarie University
North Ryde NSW 1670 Australia
{Hugues.Talbot,Leanne.Bischof}@csiro.au

Abstract

The Polartechnics SolarScan is an Australian medical instrument designed to help physicians diagnose melanoma. It has been in development in Sydney since 1994 and has now entered its commercial phase. Because of IP issues many aspects of the image analysis subsystems could not be discussed openly until recently. In this paper we describe some of the algorithm designs that makes this instrument powerful and reliable.

1. Introduction

Melanoma is the most deadly form of skin cancer. About 1000 Australians die each year as a result of melanoma. As melanoma can readily spread through the whole body it must be detected and treated early for best survival chances. Most Australians go to their GPs for skin checkups, however not all GPs can be trained to diagnose early melanoma, and as it is a relatively rare skin condition, most GPs will only see a handful of melanomas in their entire career. To err on the side of caution, a lot of benign skin lesions are also needlessly excised.

The idea of the Solarscan instrument is to allow clinics to have access to a reliable diagnostic aid that will lower error rates. This instrument needs to be relatively cheap, fast, accurate, give reproducible results, be able to follow up patients and above all have a low error rate.

The result of a collaboration between Polartechnics Ltd, The Sydney Melanoma Unit at Royal Prince Alfred Hospital and CSIRO-MIS, the Solarscan instrument is one of a few such instruments that could fit the above description.

In this paper we present an overview of the image analysis algorithms that makes this instrument useful.

2. Instrument design

The Solarscan instrument consists of a video capture device connected to a standard PC running Microsoft Windows NT. The video capture device is a special-purpose designed camera with a 3-CCD 760×560 captor, self-contained flat illumination and surface microscopy (epiluminescence) optics. The front cover of the camera is changeable and can be replaced to take standard non-microscopy pictures, for example to save the location of moles on the body in the patient database for future reference.

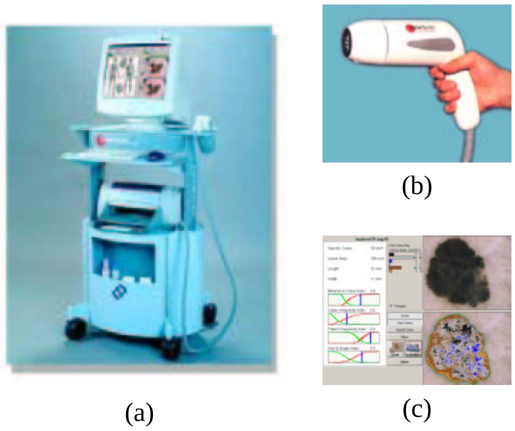

(a) (b) (c)

Figure 1. The SolarScan instrument: (a) global appearance, (b) camera, (c) user interface.

All captured images have 4 colour swatches in each corners for colour calibration.

The diagnosis system has two main components:

- The image analysis pipeline. Its aim is to segment the lesion from the image and to provide feature extraction.

- The diagnosis model. Its aim is to provide diagnosis aid to the physician. It can simply be a probability of a lesion to be a melanoma in the case of a full diagnosis system, or it can provide significant numbers that are in tune with the physician's training (ABCD rules for example [5, 8]).

3. Image analysis pipeline

In this section we describe in more detail the image analysis pipeline leading to feature extraction.

3.1. Image calibration

All images taken with the SolarScan instrument contain four swatches of known reflectance which can be used to calibrate images between session and between instruments. The aim is to be able to refer all pixels to a known fixed colour space such as CIE XYZ.

A standard uniform gray background is also imaged from time to time to provide background illumination correction.

3.2. Artifact removal

Surface microscopy involves a layer of oil between the lesion and the optics. Air bubbles can be trapped in this layer. Additionally, hair is frequently present on images of skin. Because of the danger to the lesion it is not usually shaved and must be detected and masked out.

The procedure to detect air bubbles involves taking a top-hat filter on the lightness component of the calibrated image, together with strong edges detected with a morphological gradient [11].

Hairs are detected using a more complex procedure involving the algebraic closing consisting of the intersection of a series of closings by linear structuring elements [13]. The artifact detection is illustrated in Fig. 2

Obviously the calibration patches within the field of view are also masked out.

3.3. Lesion segmentation

An accurate lesion segmentation is critical to the whole procedure and has been subject to a significant body of research [15, 9]. The procedure is difficult because of the large variation in skin colour in the population and the equally large variation of lesion appearance.

For reliability an entirely automated procedure based on seeded region growing [1] is first run. If the result is judged unsatisfactory by the operator a semi-automated procedure is tried next, based on colour clustering.

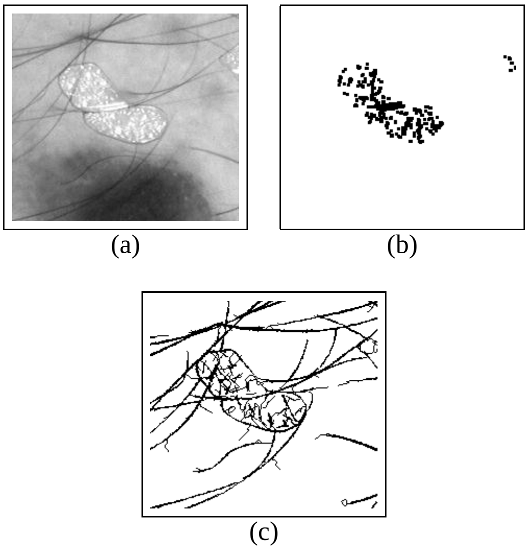


Figure 2. Artifact removal on subset of lesion image: (a) lightness component, (b) air bubbles, (c) lesion hair.

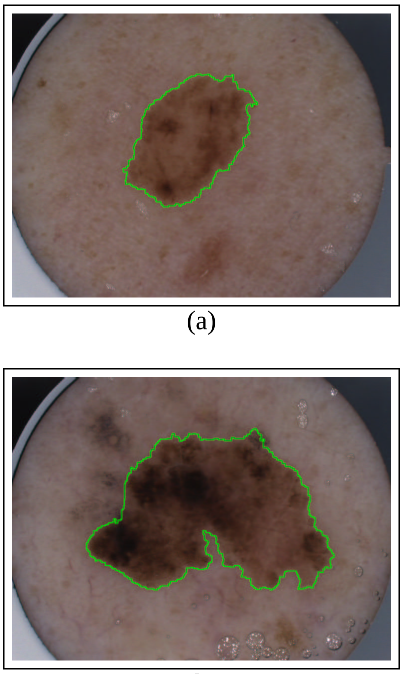
In both approaches, a principal component analysis of a representative sample of calibrated images is peformed. PC1 and PC2 are the first and second principal components, respectively.

3.3.1. SRG-based procedure

The main difficulty in an SRG-based procedure is to find the seeds. In this case the lesion seed is obtained from the PCA analysis. PC1 is roughly equivalent to lightness,the lower 20% of which are used as a seed for the lesion, assumed to be darker than the rest of the image. The brightest 20% are assumed to be skin. A standard SRG algorithm is run on the calibrated data with these seeds.

3.3.2. Colour clustering

This approach is based on clustering the bivariate histogram PC1 vs. PC2 of the PCA transform of the calibrated image. The clustering technique is based on finding peaks in the bivariate histogram and segmenting by watershed with a technique similar to that described in [12]. The colour clusters are then ordered on the basis of increasing lightness, which yields an ordered series of image segmentations, from darkest to lightest. To limit the number of boundaries to a reasonable number, skin and lesion statistics are used from the previous SRG procedure.



(a)

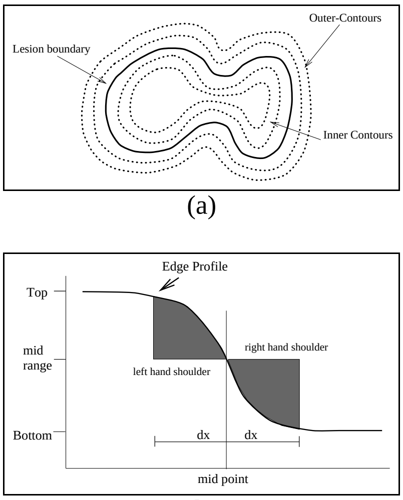
(b)

Figure 3. Example of lesion segmentation.

3.4. Lesion border analysis

Clinical analysis of the border of a lesion has been shown to be significant [7]. A number of researchers have published automated methods of boundary analysis [2, 10]. In this project we systematically reused all available published methods including fractal analysis [4], notch analysis, symmetry analysis and more. One of the most significant variables in subsequent statistical analysis turned out to be the lowly area measurement (i.e: larger lesions were more likely to be melanomas than smaller ones).

Novel methods were also developed. As an example we present our method for determining edge abruptness. Clinical studies have shown that the sharpness, or abruptness of the transition between the lesion and the surrounding skin is diagnostic: melanomas tends to have sharper transitions than benign lesions. To extract the edge profile around the boundary, a distance function (DF) is computed from the edge of the lesion (in both directions, inside and outside). From each point on the boundary a profile is run towards the inside of the lesion and towards the outside, following the upstream [6] of the DF, recording the lesion and skin intensities respectively. This allows the reconstruction of a reliable surrogate for the gradient of the border, as shown on Fig. 4.



(a)

(b)

Figure 4. Edge abruptness measure.

3.5. Colour segmentation

Colour analysis is also significant for the melanoma diagnosis. Melanomas tend to be more colourful than benign lesions. Both absolute colour (i.e: calibrated colours) and relative colour (variation in colours within a given image) measures were derived.

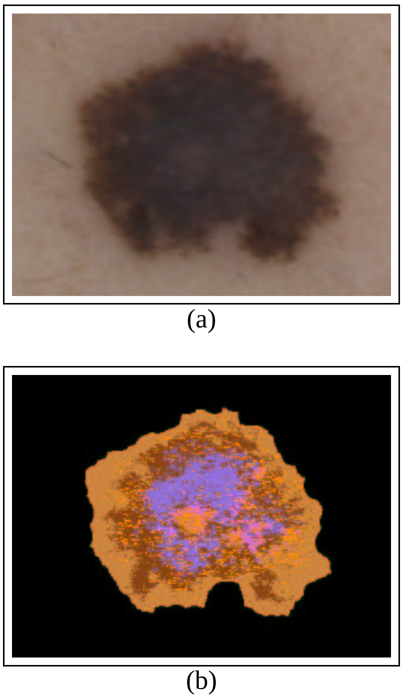
3.5.1. Absolute colours

An RGB cube of calibrated colours derived from a significant subset of images was classified into various clusters. Several methods were used for this, including a complete watershed-based segmentation of the cube, which resulted in a fixed 3D look-up table. Applying this LUT to any calibrated colour skin image results in a pixel-wise classification. From the statistics of the classified regions various measures were derived, such as relative areas of colours, presence or absence of particular colours, etc.

3.5.2. Relative colours

For this an automated clustering of the first two PCA components of individual lesions was performed, resulting again in pixel classification. Because these regions do not correspond to known colours, the resulting regions were used as input for a categorical symmetry analysis (symmetry analysis between regions) based on simple Euclidean distances between clusters. Various other inputs were used for categorical symmetry analysis, such as 1-D histogram segmentation (in all 3 channels) and the absolute colour classification.

Figure 5 shows a sample absolute colour segmentation.



(a)

(b)

Figure 5. Lesion absolute colour segmentation. (a) original image, (b) colour segmentation.

3.6. Texture analysis

Some malignant lesions are not colourful, but can still be distinguished from benign lesions by texture analysis. More precisely melanomas tend to have less regular textures and less symmetry than benign lesions. However defining regularity and symmetry in skin lesions is challenging [14]. To capture the notion many different features were extracted. We present a few examples:

3.6.1. Simple measures

Relatively obvious measures were obtained, such as the total intensity variance in lightness or in each of the colour channels. Some extension of symmetry measurements applied to the binary mask of the lesion were also extended to grey-level and colour, such as the flip and rotation measures. These consist of finding the axes of symmetry of the lesion (by moment-based methods), and then measuring the amount of correspondence between parts of the lesions on each side of the axes. This can be done by flipping the lesion about each axis and measuring the average absolute pixelwise difference over the overlap.

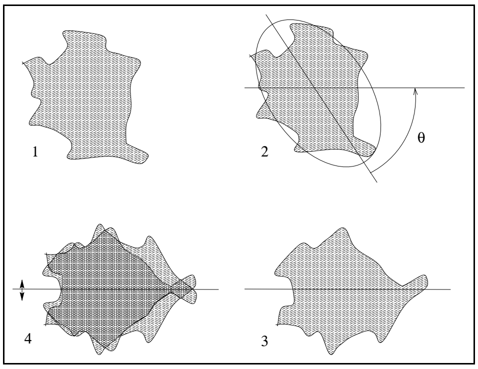


Figure 6. The flip symmetry measures.

3.6.2. More complex measures

Standard unsupervised texture classification methods (Wavelet, Gabor, etc) were too slow and did not appear to be suitable to the particular kind of biological texture found in skin lesions. However we did persist with some FFT-based texture measures with mixed results.

A number of ad-hoc texture measures were also developed. These measures were designed to match patterns often found in lesions, often labelled as “network”. The associated measures correspond to a coverage percentage of a lesion. The idea being that if a single particular texture covers a large area portion of a lesion, it is probably regular.

3.6.3. Specific features

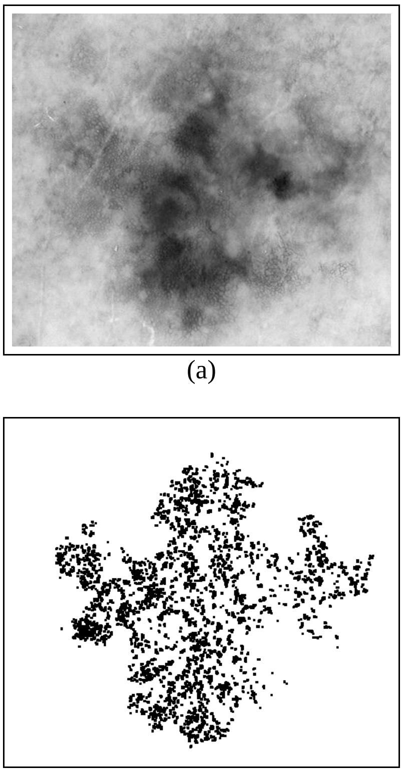
Some malignant lesions can only be diagnosed as such by looking at some specific visual features that need to be looked for. A list of such significant features can be found in [7].

An example of those are the border spots, which are small dark spots on the boundary of a lesion corresponding to melanocytes in malignant stage. Such features can be extracted by black top-hat restricted to the boundary of the lesion, as seen on Fig. 8

4. Feature extraction and selection

Many more features were extracted and measures derived than can be presented here. As an overview, the features not already present in the literature (the vast majority) were designed in collaboration with an expert specialist clinician. The idea of the design phase was not necessarily to come up with features that by themselves were correlated to malignancy, but which reasonably matched something that the clinician could see *and* was in some way related to the highly visual human diagnosis procedure.

The feature selection was made by letting the data itself drive the selection. The set of all measures was fed to a



(a)

(b)

Figure 7. Example of network extraction: (a) lesion with network, (b) network coverage.

series of classification methods, from logistic regression to regression trees. After culling a number of models were built using cross-correlation.

The database of images used for building these models grew during the length of the project. It started relatively small with about 30 melanomas and a few hundred atypical benign lesions but now contains several hundred melanomas and several thousand both typical and atypical benign lesions.

At some point in the research the number of extracted features grew to over 600, but was later reduced to about 80. Fewer than a dozen are typically needed for a model.

4.1. Diagnostic features

Not much detail can be given here, but it has become apparent that relatively simple measures are often more valuable than complex one. Simple measurements such as lesion area often appear at the root of the model. These measures have the advantage of being robust and highly repeatable whereas more complex one can give signficantly different answers under relatively slight changes in image capture conditions.

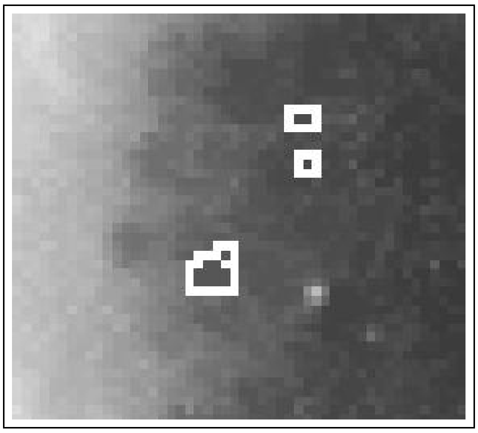


Figure 8. Example of border spot segmentation.

However we have found it impossible to build an accurate model without some of the more advanced measures, especially those that look for specific features in lesions.

4.2. Accuracy

The SolarScan instrument was originally designed to be a diagnostic instrument, however it is now being marketed simply as a diagnostic aid which gives several indications to the operator rather than operate as a black box giving a single probablity or yes/no answer. As such it is not possible to give a concise accuracy figure for this instrument.

However early in the development of the instrument, with only some of the features now being collected and a much smaller image database, this instrument obtained a cross-validated sensitivity (ability to diagnose a melanoma correctly) of 92% and specificity (ability to diagnose a non-melanoma) of 62% [3], which is comparable to an inexperienced skin specialist and better than most GPs. It is presumed that the numbers for the present instrument are higher.

5. Conclusion

The Polartechnics SolarScan melanoma diagnosis assistance system is an innovative Australian instrument which has followed the difficult road from research to commercialisation thanks to a unique collaboration between research organizations and a private company. Image analysis is a key component of this complex system. In this paper we were able to describe parts of the design of the image analysis subsystem.

The proof of the validity of this research will be if the instrument becomes successful as a commercial product, keeping in mind that the final objective of such instruments is really to save lives.

References

[1] R. Adams and L. Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):641–647, 1994.

[2] M. Binder, H. Kittler, A. Seeber, A. Steiner, H. Pehamberger, and K. Wolff. Epiluminescence microscopy-based classification of pigmented skin lesions using computerized image analysis and and artificial neural network. *Melanoma Research*, 8:261–266, 1998.

[3] L. Bischof, H. Talbot, E. Breen, D. Lovell, D. Chan, G. Stone, S. Menzies, A. Gutenev, and R. Caffin. An automated melanoma diagnosis system. Ballarat, Vic., Australia, 1998. Submitted.

[4] E. Claridge, P. Hall, M. Keefe, and J. Allen. Shape analysis for classification of malignant melanoma. *Journal of Biomedical Engineering*, 14:229–234, May 1992.

[5] NIH Consensus Conference. Diagnosis and treatment of early melanoma. *Journal of the American Medical Association*, 268(10):1314–1319, 1992.

[6] G. Matheron. Example of topological properties of skeletons. In J. Serra, editor, *Image Analysis and Mathematical Morphology*, volume 2, Theoretical Advances, pages 217–238. Academic Press, London, 1988.

[7] S. W. Menzies, K. A. Crotty, C. Ingvar, and W. H. McCarthy. *An Atlas of Surface Microscopy of Pigmented Skin Lesions*. McGraw-Hill, Roseville, Australia, 1996. ISBN 0 07 470206 8.

[8] F. Nachbar, W. Stolz, T. Merkle, and et al. The ABCD rule of dermatoscopy. *J Am Acad Dermatol*, 30:551–559, 1994.

[9] T. Schindewolf, W. Stolz, R. Albert, W. Abmayr, and H. Harms. Comparison of classification rates for conventional and dematoscopic images of malignant and benign melanocytic lesions using computerized colour image analysis. *European Journal of Dermatology*, 3:299–303, 1993.

[10] S. Seidenari, G. Pellacani, and A. Giannetti. Digital videomicroscopy and image analysis with automatic classification for detection of thin melanomas. *Melanoma Research*, 9:163–171, 1999.

[11] J. Serra. *Image Analysis and Mathematical Morphology - Volume II : Theoretical Advances*. Academic Press, London, 1988.

[12] P. Soille. *Morphological Image Analysis, principles and applications*. Springer, 1999.

[13] P. Soille and H. Talbot. Directional morphological filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1313–1329, 2000.

[14] W. Stoecker, C.-S. Chiang, and R. Moss. Texture in skin images: comparison of three methods to determine smoothness. *Comp. Med. Imag. Graph.*, 16(3):179–190, 1992.

[15] L. Xu, M. Jackowski, A. Goshtasby, D. Roseman, S. Bines, C. Yu, A. Dhawan, and A. Huntley. Segmentation of skin cancer images. *Image and Vision Computing*, 17:65–74, 1999.

COMPUTER VISION AND PATTERN RECOGNITION I

(This page left blank intentionally)

Can Region of Interest Coding Improve Overall Perceived Image Quality?

Andrew P. Bradley
Centre for Sensor Signal and Information Processing (CSSIP),
School of Information Technology and Electrical Engineering,
The University of Queensland,
St Lucia, QLD 4072, Australia.
a.bradley@cssip.uq.edu.au

Abstract

In this paper we investigate the possibility of improving the overall perception of image quality by preferentially coding certain regions of interest (ROI) in an image. Experiments are conducted utilising an automated algorithm for visual attention (VA) to detect the primary ROI(s) in an image, and then encoding the image using the maxshift algorithm of JPEG 2000. The results from a 2 alternative forced choice (2AFC) visual trial show that, while there is no overall preference for the ROI encoded images, there is an improvement in perceived image quality at low bit rates (below 0.25 bits per pixel). It is concluded that a perceived increase in overall image quality only occurs when the increase in quality of the ROI more than compensates for the corresponding decrease in quality of the image background (i.e., non ROI).

Keywords

JPEG 2000; Region of Interest Coding; Image Quality.

Introduction

With the introduction of third generation (3G) mobile devices there will be an increasing demand for the efficient transmission of multi-media data, such as speech, audio, text, images, and video. Of these multi-media data types image and especially video data will provide the toughest challenges because of their high bandwidth and user expectations in terms of high quality of service. Therefore, to enable the successful adoption of the multitude of 3G applications, the transmission of multi-media data must be at high compression ratios and be of a perceptually high quality. In this paper we shall investigate the suitability of the region of interest (ROI) coding feature provided by JPEG 2000 (JP2K) to improve the perceptual quality of compressed images, where the ROI has been automatically extracted from the image by using an algorithm that simulates visual attention (VA).

JPEG 2000 (JP2K) is the emerging image and video compression standard developed by the International Organisation for Standardisation/International Telecommunications Union (ISO/ITU-T). JP2K has been designed to complement the current JPEG standard by providing improved compression performance and a rich set of new functionalities [7]. JPEG 2000 Part I, the core-coding algorithm, became an international standard in December 2000 and further work is ongoing to tailor the standard to specific applications, such as medical imaging and video coding. JP2K provides, in a single bit-stream, a broad set of functionalities, such as: progressive transmission by resolution, quality, component, or location; random access; lossless to lossy compression; and error tolerance [3]. The specific functionality investigated in this paper is the ability of JP2K to encode a region of interest (ROI) in an image with more detail than the background. In this paper JP2K ROI coding is used in combination with an algorithm for visual attention (VA) [5, 6], to provide a progressive bit-stream where the regions highlighted by the VA algorithm are presented first in the bit-stream. This results in an *interest ordered* bit-stream where any valid bit-stream termination results in an image where the ROI is coded to a higher quality than the background. The efficacy of this technique is then evaluated using a visual trial to determine under what conditions it provides an increase in overall perceived image quality compared to conventional JP2K at the same bit-rate.

As has been demonstrated in previous rate distortion experiments [1], it is important to reduce the overhead associated with ROI coding in order to ensure maximum coding efficiency. Briefly, this can be achieved by:

1. Ensuring the ROI is < ¼ of the area of the whole image;
2. Reducing the number of regions of interest (to two or less); and
3. Ensuring region boundaries are reasonably regular (smooth).

The first constraint ensures that there are enough available bits in the background to be able to preferentially encode the foreground ROI. Whilst the last two constraints ensure that the overhead associated with ROI coding is minimised, e.g., for maxshift coding it minimises the number of code-blocks that contain coefficients from both the ROI and the background. The algorithm for processing the VA map produced by [5, 6] to meet the above constraints is detailed in [2]. An example image, with the ROI selected by the VA algorithm highlighted, is shown in Figure 1.

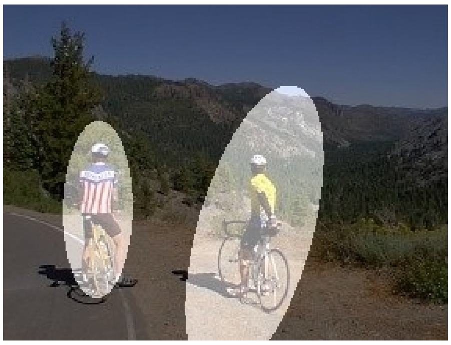


Figure 1. Cycles image with VA ROI highlighted.

Experimental Methodology

As discussed in [1] selection of the most appropriate JP2K ROI encoding methodology for a particular application is dependent upon a number of factors: the desired bitrate; relative ROI/background importance; the shape and size of the ROI; and whether the ROI is fixed or is to be selected by the user. For client/server applications it is essential to be able to extract any ROI from an encoded image, in which case code-block selection is the best method to use [4]. However, in the type of applications we are considering the ROI can be calculated directly from the image and is fixed. Therefore, it is desirable to have the ROI embedded in the bit-stream using coefficient scaling [1]. In addition, the ROI is assumed to be of primary importance and so we desire to receive it as early as possible in the bitstream. Therefore, we shall use the method of coefficient scaling provided in Part I of the JP2K standard, the maxshift algorithm. In addition, we shall tailor the JP2K ROI coding to our particular requirements by using: small (16x16) code-blocks for fast ROI refinement; a 5 level irreversible (bi-orthogonal spline 9/7) wavelet transform for high compression (lossy) efficiency with the lowest level of the DWT defined to be part of the ROI; and an increased quantisation step size (of 0.03125 which is four times the default) to prevent ROI over-coding.

The visual trial was based upon six images, namely: boat, cycles, beach, helicopter, land, and road sign. These images were chosen to have a reasonably varied content, whilst still containing one or two primary objects that could be considered to be more important (visually interesting) than the background. The images selected for the visual trial are not intended to be representative of any particular potential application, but were chosen solely to judge the efficacy of ROI coding in JP2K.

The purpose of the visual trial was to directly compare images encoded to a specified bit-rate using standard JP2K and JP2K ROI coding, where the ROI is determined using the VA algorithm [2]. The comparisons were made at four logarithmically spaced bit-rates (and hence varying image qualities) of 0.125, 0.25, 0.5, and 1 bits per pixel (bpp). A two alternative forced choice (2AFC) methodology was selected because of its simplicity, i.e., the observer views the two images and then selects the one preferred, and so there are no issues with scaling opinion scores between different observers. There were ten observers (8 male and 2 female) all with good, or corrected, vision and all observers were non-experts in image compression. The viewing distance was approximately 40cm (i.e., a normal PC viewing distance) and the image pairs were viewed one at a time in random order. The observer was free to view the images multiple times before making a decision, however a buzzer sounds after 20 seconds to indicate that they should make a decision. In addition, a blank mid-grey image is shown between each image (for 2 seconds) to prevent observers switching between the two images to find insignificant differences. Each image pair was viewed twice, giving (6×4×2) 48 comparisons, which means that each observer takes approximately 10 minutes to view all of the images. Images were viewed on a 12.1” Thin Film Transistor (TFT) display, in a darkened room (i.e., daylight with drawn curtains). The test images were displayed on a mid-grey background to a maximum size of 410×600 pixels. Prior to the start of the visual trial all observers were given a short period of training on the usage of the visual trial software and they were told to select they image they preferred assuming that it had been downloaded over the internet or wireless network.

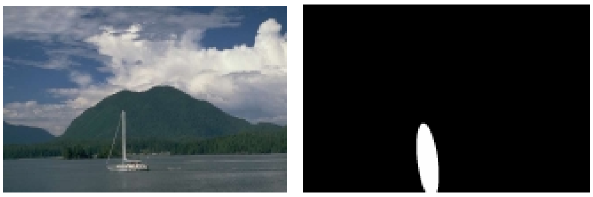


Figure 2. Boat image and VA ROI mask.

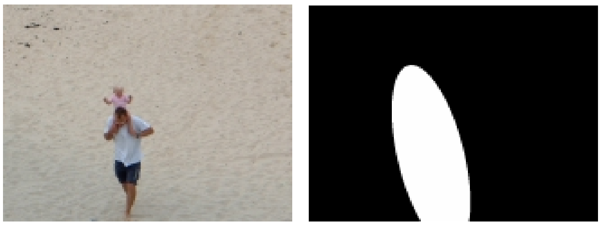


Figure 3. Beach image and VA ROI mask.

Results

Table 1 shows the overall preferences, i.e., independent of (summed over) image and bit-rate, for standard JP2K and JP2K ROI coding with the ROI determined using the VA algorithm. Table 1 also shows the standard errors associated with the preferences assuming a Gaussian approximation to the Binomial distribution. From Table 1 it can be seen that standard JP2K is preferred over ROI coding approximately 65% of the time. This shows that standard JP2K produces good quality images over a wide range of bit-rates and indicates that ROI coding may not be suitable as a general-purpose image coding technique. Therefore, we will have to examine the results in more detail to identify the conditions to which the ROI JP2K coder is best suited.

Table 1. Overall preferences (independent of image and bit-rate)

Compression Method	Number of Preferences	Standard Error
JP2K	311	± 12.3
JP2K ROI	169	± 12.3

Figure 4 shows that preferences vary both across the image set (independent of bit-rate) and with bit-rate (independent of image). Standard JP2K is shown with red standard error bars (on the left) whilst JP2K ROI coding is shown with blue standard error bars (on the right). From Figure 4 it can be seen that there is a large variation in preferences across each of the images in the test set. For example, standard JP2K is preferred at every bit-rate on the boat image, whilst the two methods are equivalent on the cycles and beach images (within 1 standard error). However, the second, and more important, trend that can be observed in Figure 4 is an increase in preferences for ROI coding as the bit-rate decreases. At the lowest bit-rate tested (0.125 bpp) the preferences for ROI coding are 68, with a standard error of ± 5.8, and 52 (± 5.8) for standard JP2K. This indicates a clear preference (i.e., statistically significant) for the JP2K with ROI coding at this bit-rate.

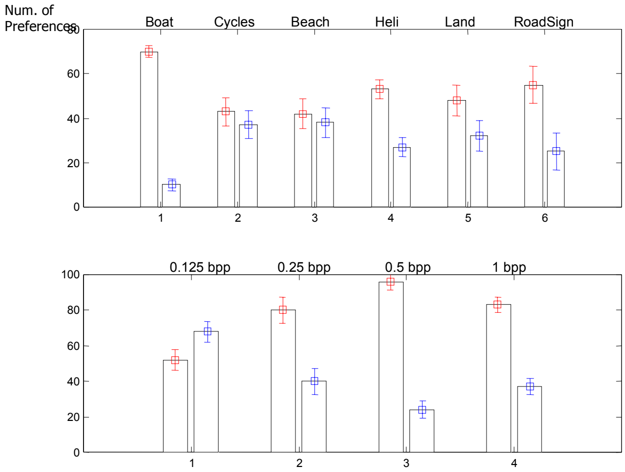


Figure 4. JP2K (left) and ROI JP2K (right) preferences for each image (independent of bit-rate) and preferences at each bit-rate (independent of image).

Note: the decrease in preferences for standard JP2K at 1 bpp in Figure 4 is due to the two methods producing image that look increasingly similar. Therefore, preferences between the two methods will tend to random (i.e., 50/50) selection.

Discussion

As illustrated in Figure 4, there are two main sources of variation that can explain the differences in preferences: variation with image and variation with bit-rate (variation with observer being indicated by the standard errors in the results). The variation across the images in the test set shows that for an image that has an ROI and a background of little importance, such as the beach image (see Figure 3), the ROI coding works well. However, for an image that has an ROI and also some visually important contextual details in the background, such as the boat image (see Figure 2), the ROI coding works less well. The increase in preferences for the ROI coding as bit-rate decreases, illustrated in Figure 4, is undoubtedly the most significant and consistent effect observed in the visual trial (being apparent in 5 of the 6 images in the test set).

It is worthwhile noting that the performance of ROI coding on the boat image was degraded by the fact that the ROI found did not enclose the whole region of primary interest. However, in general it is probable that most images that have a primary ROI will also have some important contextual details in the background and so ROI coding is unlikely to provide an overall improvement in image quality at all bit-rates.

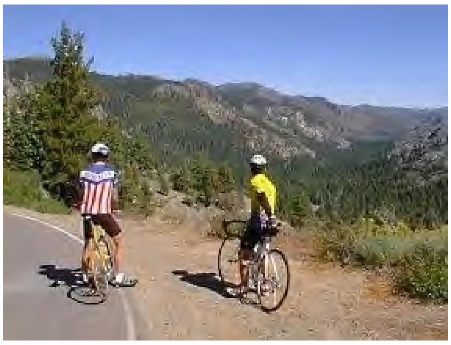


Figure 5. Cycles image JP2K (1bpp).

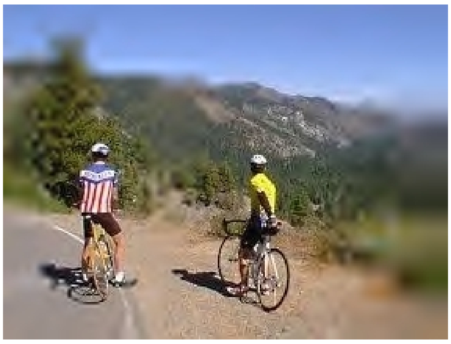


Figure 6. Cycles image JP2K ROI (1bpp).

At low bit-rates (< 0.25 bpp), having the ROI encoded first in the bit-stream can significantly improve the visual quality of the ROI compared to standard JP2K. In addition, the background (non-ROI) areas are not of significantly poorer visual quality and are often of preferable visual quality as they contain less (wavelet) compression artefacts. At the low bit-rates the background regions tend to contain only coefficients from the lowest level of the wavelet transform rather than sporadic coefficients from higher levels of the DWT (as in standard JP2K). This results in a background that is uniformly blurred, which is often preferable to a less blurred background that also has wavelet artefacts.

At the higher bit-rates (> 0.25 bpp) the ROI is often not of significantly better visual quality than standard JP2K. This combined with the fact that the background areas are often more blurred and pixelated than standard JP2K results in lower preferences (see Figures 5 and 6). This effect should come as no surprise as once the ROI is coded to a visually acceptable level it takes a significant number of bit refinements (of the high entropy least significant bits) to get a visible improvement in image quality. In addition, because small code-blocks were used at all bit-rates for the ROI coding (16×16 compared to 64×64) there is a reduced compression efficiency, especially with the entropy coding of the code-blocks. This reduction in compression efficiency is particularly apparent at the higher bit-rates due to the increased number of significant coefficients [1]. However, using small code-blocks reduces the ROI coding overhead and therefore ensures that the complete ROI appears as early as possible in the bit-stream.

Another reason for the reduction in preferences at bit-rates > 0.25 bpp is due to the inherently uneven image quality in the majority of ROI coded images. This results in images that do not appear *natural* as the ROI is in sharp focus, whilst the background appears blurred. A more gradual change in image quality between ROI and background would, however, increases the size of the ROI, which has a negative impact on ROI coding efficiency.

There is an anecdotal explanation for the reduction in preferences at bit-rates > 0.25 bpp by considering the rule of thumb that to observe a significant increase in the visual quality of an image you have to (approximately) double the bit-rate. This means that the ROI has to be coded to twice the bit-rate of the background to observe a significant improvement in perceived visual quality. Therefore, if we assume the ROI is ¼ of the image area, then to code an image to the same (target) bit-rate as standard JP2K, the background can only be coded to half the target rate to allow the ROI to be coded at twice the target rate. For example, if the target bit-rate is 0.5 bpp, then we can either code to this bit-rate using standard JP2K, or code the ROI to 1 bpp and the background to 0.25 bpp using JP2K ROI coding. Therefore, ROI coded images will invariably have an ROI that looks better, but a background that looks worse, than standard JP2K images coded to the same bit-rate. Results observed in this visual trial indicate that the ROI encoded images only score an overall improvement in image quality at target bit-rates less than 0.25 bpp (i.e. 0.125 bpp). At bit-rates greater than 0.25 bpp the increase (if any) in quality of the ROI does not compensate for the decrease in background quality when observers judge overall image quality.

Conclusions

Results from the visual trial indicate an overall preference for standard JP2K independent of image and bit-rate. However, the proposed VA ROI JP2K coding method was clearly preferred at the lowest bit-rate tested (0.125 bpp). This indicates that, when observers judge overall image quality, it is only at this bit-rate that the visible increase in quality of the ROI more than compensates for the decrease in quality of the background. Therefore, it can be concluded that ROI coding in JP2K will only produce an overall increase in perceived image quality when: the image contains a small number (≤ 2) of regions of interest; these regions are relatively small (< ¼ of the total image area); and the bit-rate is low enough to produce visible compression artefacts (< 0.25 bpp).

Acknowledgements

This work was undertaken whilst the author was a visiting research fellow at Btexact Technologies, Adastral Park, Martlesham Heath, Ipswich, UK. All experiments were performed with JJ2000, version 4.1 (the current version is available from http://jj2000.epfl.ch).

References

[1] A. P. Bradley and F. W. M. Stentiford, JPEG 2000 and Region of Interest Coding, *Digital Image Computing Techniques and Applications* (DICTA), Melbourne, Australia, January 2002

[2] A. P Bradley and F. W. M. Stentiford, Visual Attention for Region of Interest Coding In JPEG 2000, Submitted to *Journal of Visual Communication and Image Representation*, January 2002.

[3] C. Christopoulos, A. Skodras, and T. Ebrahimi, The JPEG 2000 Still Image Coding System: An Overview, *IEEE Transactions on Consumer Electronics*, Vol. 46, No. 4, pp. 1103-1127, November 2000.

[4] D. Santa-Cruz, T. Ebrahimi, M. Larsson , J. Askelof, and C. A. Christopoulos, Region of Interest Coding in JPEG 2000 for Interactive Client/Server Applications, *IEEE 3rd Workshop on Multi-media Signal Processing* (MMSP), pp.389-394, September 1999.

[5] F. W. M. Stentiford, An Evolutionary Programming Approach to Simulation of Visual Attention, *Congress on Evolutionary Computation*, Seoul, Korea, pp. 851-858, May 2001.

[6] F. W. M. Stentiford, An Estimator for Visual Attention Through Competitive Novelty with Application to Image Compression, *Picture Coding Symposium*, Seoul, Korea, pp. 101-104, April 2001.

[7] D. S. Taubman and M. W. Marcellin, *JPEG 2000: Image Compression Fundamentals, Standards, and Practice*, Kluwer Academic Press, 2001.

Towards a Maximum Entropy Method for Estimating HMM Parameters

Christian J. Walder, Peter J. Kootsookos and Brian C. Lovell
Intelligent Real-Time Imaging and Sensing Group,
School of Information Technology and Electrical Engineering,
The University of Queensland, Brisbane, Queensland 4072, Australia.
{walder, lovell}@itee.uq.edu.au

Abstract

Training a Hidden Markov Model (HMM) to maximise the probability of a given sequence can result in over-fitting. That is, the model represents the training sequence well, but fails to generalise. In this paper, we present a possible solution to this problem, which is to maximise a linear combination of the likelihood of the training data, and the entropy of the model. We derive the necessary equations for gradient based maximisation of this combined term. The performance of the system is then evaluated in comparison with three other algorithms, on a classification task using synthetic data. The results indicate that the method is potentially useful. The main problem with the method is the computational intractability of the entropy calculation.

1 Introduction

In recent years, the HMM has become one of the main tools for spatio-temporal pattern recognition, especially in the area of speech recognition [8]. In 1983, Levinson, Rabiner and Sondhi described a method of estimating HMM parameters from multiple training sequences in the maximum likelihood sense, via a special case of the expectation-maximisation algorithm. This method, known as the Baum-Welch algorithm, has been widely used, however it is well known that it is susceptible to the problem of "over fitting".

Several attempts have been made to deal with the over-fitting problem. In 1998, Brand described an effective method involving maximum likelihood parameter estimation, but with the additional constraint of an "entropic prior" [1]. That is, an a priori assumption was made regarding the probability distribution of the HMM parameters themselves.

Recently, Davis et al have explored [2] the possibility of using parameter averaging, as suggested by Mackay in 1997 [6]. This method involves training a separate HMM for each training sequence, and then averaging the parameters of the resulting HMMs. The reported results indicate that the averaging method offers an improvement over the basic Baum-Welch algorithm.

This paper presents another method of HMM parameter estimation which is intended to overcome the over-fitting problem. The method herein was mentioned by Brand in 1998, with reference to combinatorial optimisation problems [1], however the approach does not appear to have been investigated for the task of HMM parameter estimation.

2 Background Theory

2.1 HMM Preliminaries and Notation

An HMM can be described as a probabilistic function of a Markov Chain. For the case of HMMs with discrete outputs and discrete states, we can assume that the underlying Markov Chain has N states, $q_1, q_2, \ldots, q_N$. Such a Markov chain can be specified in terms of an initial state distribution vector, $\pi = (\pi_1, \pi_2, \ldots, \pi_N)$, and a state transition probability matrix, $A = [a_{ij}], 1 \leq i, j \leq N$. Here, π_i is the probability of q_i at time time $t = 0$, and a_{ij} is the probability of transiting to state q_j given that the current state is q_i, that is $a_{ij} = p(q_j$ at time t + 1 $|q_i$ at time t$)$. In the previous expression and the remainder of the paper, $p(x)$ is to be taken as the probability of occurrence of event x.

Each of the Markov states have an associated random process which provides a probabilistic mapping to the output of the HMM, which is drawn from an alphabet V of M possible outputs, $v_1, v_2, \ldots, v_M$. These probabilistic mappings from hidden state to observed output can be collectively specified by another stochastic matrix $B = [b_{jk}]$ (the "observer matrix") in which for $1 \leq j \leq N$ and $1 \leq k \leq M$, b_{jk} is the probability of observing symbol v_k given that the current state is q_j, that is, $b_{jk} = p(v_k$ at time t $|q_j$ at time t$)$.

2.2 Entropy of a Random Variable

Consider a random variable X, with N discrete outcomes, $x_1, x_2, \ldots, x_N$. The "information" of outcome x_i is [4]:

$$I(x_i) \triangleq -\log p(X = x_i)$$

The entropy of X is the expected information [4]:

$$H(X) \triangleq -\sum_{i=1}^{N} p(X = x_i) \log(p(X = x_i))$$

2.3 Entropy of an HMM

From the equation of Section 2.2, the entropy of a sequence of length T produced by HMM λ can be written as:

$$H(\lambda, T) = -\sum_{\forall O \in \tilde{O}_T} p(O|\lambda) \log p(O|\lambda) \qquad (1)$$

Where $\tilde{O}_T$ is the set of all sequences of length T that can be produced by λ. For a symbol alphabet size M, $|\tilde{O}_T| = M^T$, so the computation in equation 1 is intractable for large T.

2.4 Maximum Entropy Parameter Estimation

Let X be a random variable taking on values $x_1, x_2, \ldots, x_K$, with an unknown probability mass function (*pmf*), $p_k = p(X = x_k)$. Suppose we would like to estimate the *pmf* of X given only the expected value of some function $g(X)$ of X:

$$\sum_{k=1}^{K} g(x_k) p_k = c \qquad (2)$$

For example if $g(X) = X$ then c is the mean of X. Since Equation 2 does not, in general, specify the *pmf* of X uniquely, we must apply further constraints in order to solve for the p_k. One additional constraint that is commonly applied [4, 3] is that of "maximum entropy". That is, we seek the *pmf* that maximises the entropy subject to the constraint in Equation 2. This is intuitively appealing, since the maximum entropy solution is that which satisfies our known constraints, while asserting as little as possible about the nature of the underlying *pmf*. The maximum entropy parameter estimation can be set up as an optimisation problem and in some cases solved using classical methods such as Lagrange multipliers [4, 3].

3 Maximum Entropy HMM Parameter Estimation

In its most fundamental form, HMM parameter estimation proceeds as follows [5]. First of all, the source which is to be modelled is sampled one or more times, to provide "training data" for the parameter estimation. An HMM topology is then chosen, and an HMM is initialised randomly within the chosen topology. The HMM parameters are then adjusted so as to maximise the likelihood of the HMM producing the training sequence(s). This is known as "maximum likelihood" parameter estimation. For the case of HMMs, maximum likelihood parameter estimation is in itself a difficult problem: in general only locally optimal solutions can been found. A well known problem with the maximum likelihood approach is that of "overfitting" to the training data. That is, the model fits the training sequences *too* well, thereby failing to generalise.

In 1998, Brand proposed a means of dealing with the overfitting problem [1]. The method is essentially Bayesian inference with an "entropic prior". That is, maximum a posteriori (MAP) parameter estimation using an a priori distribution over parameter space. Formally, the method seeks the parameter set θ which maximises the posterior

$$p_e(\theta|x) \propto p(x|\theta) p_e(\theta) \qquad (3)$$

where x is the observed (training) data, and $p_e(\theta)$ is the entropic prior:

$$p_e(\theta) \propto e^{-H(\theta)} \qquad (4)$$

where $H(\theta)$ is the entropy of the model. A detailed explanation of the method can be found in [1].

The method proposed in this paper is similar to that of Brand [1], in that the generality of the model (as measured by its entropy) is accounted for during the training process, however the knowledge of model entropy is used in in a different way. Following the same approach as the classical maximum entropy method, we would like the model to have high entropy as well as to match our knowledge of the data. This leads to the following idea: rather than maximising the probability of the training sequences, maximise a linear combination of the likelihood and the model entropy. Formally, we seek to maximise the following "objective function":

$$C = b \log p(O|\lambda) + (1 - b) H(\lambda, T) \qquad (5)$$

Where $b \in [0, 1]$, the "balancing parameter", is the free parameter that sets the desired "generality" of the model, and O is our training sequence. For example, $b = 1$ results in normal maximum likelihood learning, whereas $b = 0$ ignores the training data and maximises the entropy of the model.

In equation 5, $\log p(O|\lambda)$ is maximised rather than $p(O|\lambda)$ to ensure that we are comparing equivalent units of likelihood and entropy – log of probability is information, and entropy is expected information, so the units are comparable. If the $\log$ is to base 2, then the units are "bits", while a natural log has units "nats".

In the next section we begin describing how the model can be optimised according to the objective function above. Before proceeding, however, it is worth making a few comments regarding the b parameter of Equation 5. The inclusion of the parameter can be justified by the following argument. In an extremely "data poor" training problem in which we have only one training sequence, it may be possible to find a deterministic (zero entropy) HMM that fits the data perfectly in the maximum likelihood sense, however this would obviously be of no value for either regression or classification tasks. By applying domain knowledge, it may be possible to sensibly choose b such that a useful model is obtained. The necessity for the free parameter is a symptom of the inherent difficulties of all inductive inference tasks – as is well known, logical induction is a flawed process, and one that requires the assumption of some prior knowledge in order to reach a conclusion [7].

4 Gradient Descent Equations

To maximise the objective function in equation 5, we can perform gradient descent w.r.t. C. To do this we need the partial derivative of C w.r.t. an arbitrary HMM parameter, θ. This follows directly from equation 5:

$$\frac{\partial C}{\partial \theta} = \frac{b}{p(O|\lambda)} \frac{\partial p(O|\lambda)}{\partial \theta} + (1-b)\frac{\partial H(\lambda, T)}{\partial \theta}$$

We now proceed, in a top-down approach, to relate the above expression back to all of the specific HMM model parameters.

4.1 Partial Derivatives of HMM Entropy with respect to Model Parameters

Taking the partial derivatives of equation 1 with respect to an arbitrary parameter θ we get,

$$\frac{\partial H(\lambda, T)}{\partial \theta} = -\sum_{\forall O \in \tilde{O}_T} \frac{\partial p(O|\lambda)}{\partial \theta}(1 + \log p(O|\lambda)) \tag{6}$$

Next we need the expressions for $\frac{\partial p(O|\lambda)}{\partial \theta}$.

4.2 Partial Derivatives of Likelihood with respect to HMM Parameters

This is our own derivation of the partial derivatives of likelihood with respect to HMM parameters. An alternative derivation is available in [5]. From [5] we have the probability in terms of the forward variable, $\alpha_t(n)$:

$$p(O|\lambda) = \sum_{n=1}^{N} \alpha_T(n) \tag{7}$$

$$\alpha_{t+1}(n) = \sum_{m=1}^{N} \alpha_t(m) a_{mn} b_n(O_{t+1}) \tag{8}$$

$$\alpha_1(n) = \pi_n b_n(O_1) \tag{9}$$

Where O_t is the t-th observation symbol in our training sequence, O. From equation 7 we get:

$$\frac{\partial p(O|\lambda)}{\partial a_{ij}} = \sum_{n=1}^{N} \frac{\partial \alpha_T(n)}{\partial a_{ij}}$$

$$\frac{\partial p(O|\lambda)}{\partial b_j(k)} = \sum_{n=1}^{N} \frac{\partial \alpha_T(n)}{\partial b_j(k)}$$

$$\frac{\partial p(O|\lambda)}{\partial \pi_i} = \sum_{n=1}^{N} \frac{\partial \alpha_T(n)}{\partial \pi_i}$$

From equations 8 and 9 we get, for a_{ij}:

$$\frac{\partial \alpha_{t+1}(n)}{\partial a_{ij}} = \alpha_t(i) b_j(O_{t+1}) + \sum_{m=1}^{N} a_{mn} b_n(O_{t+1}) \frac{\partial \alpha_t(m)}{\partial a_{ij}}$$

$$\frac{\partial \alpha_1(n)}{\partial a_{ij}} = 0$$

for $b_j(k)$,

$$\frac{\partial \alpha_{t+1}(n)}{\partial b_j(k)} = \sum_{m=1}^{N} (\delta(v_k, O_{t+1})\delta(j,n) a_{mn} \alpha_t(m) +$$

$$b_n(O_{t+1}) a_{mn} \frac{\partial \alpha_t(m)}{\partial b_j(k)})$$

$$\frac{\partial \alpha_1(n)}{\partial b_j(k)} = \pi_n \delta(O_1, v_k)\delta(j,n)$$

and for π_i:

$$\frac{\partial \alpha_{t+1}(n)}{\partial \pi_i} = \sum_{m=1}^{N} a_{mn} b_n(O_{t+1}) \frac{\partial \alpha_t(m)}{\partial \pi_i}$$

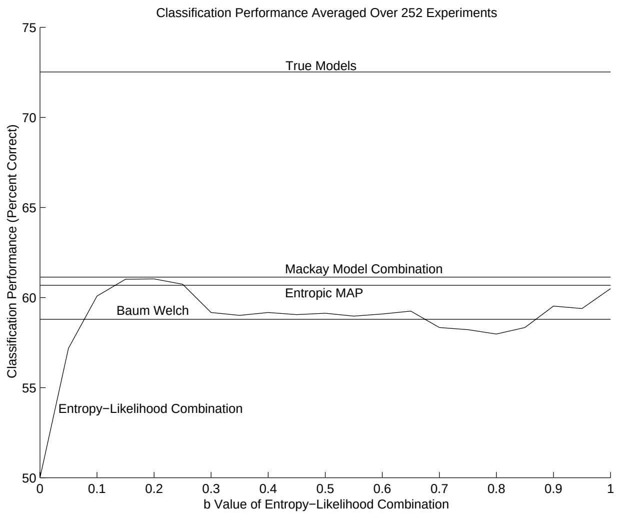


Figure 1. Mean classification performance for various model pairs.

$$\frac{\partial \alpha_1(n)}{\partial \pi_i} = b_i(O_1)\delta(i,n)$$

In the above, $\delta(x,y)$ is the Kronecker delta function:

$$\delta(x,y) = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

The equations above provide a recursive means of computing the partial derivatives of probability, w.r.t. model parameters, for a given observation sequence. The time complexity of the calculation is linear w.r.t. T. This and the results of section 4.1 allow us to calculate $\frac{\partial H(\lambda,T)}{\partial \theta}$ for all parameters θ, ie. a_{ij}, $b_j(k)$ and π_i. Unfortunately the complexity is then exponential w.r.t. T, that is, the operation has time complexity of order $O(M^T)$.

Since we now have the partial derivatives of both entropy and likelihood with respect to all of the HMM parameters, we can calculate $\frac{\partial C}{\partial \theta}$ using Equation 4, and so we can maximise C using standard hill climbing/gradient descent based numerical optimisation.

5 Results

The performance of the method has been tested in a classification task with the following methodology. An HMM topology of two hidden states and two observation symbols was chosen, with a “feed forward” structure (upper triangular in the transition matrix). Two HMMs were then randomly initialised subject to the topology and structure constraints above, and a bias was placed on the long diagonal of the transition matrix by choosing uniformly random (in the range $[0, 1]$) transition probabilities, then adding 3 to the long diagonal and normalising each *pmf* to satisfy the stochasticity constraint. Structure was also added to the observer matrix by similarly biasing a single randomly chosen probability from the observation *pmf* of each state. From each of these two “true” or “generating” models, 5 training and 500 testing sequences of length 4 were randomly generated. The sequences were chosen to be so short, and the number of states so few, due to the exponential time complexity the entropy calculation (see Section 2.3).

The training set of each model was then used to estimate an HMM with the same topology as that of the initial models, using the following training algorithms: Baum-Welch maximum likelihood [5], Mackay model averaging [2], Brand’s entropic prior [1], and finally the maximum entropy method presented in this paper. For the maximum entropy method, the b value of Equation 5 was varied from 0 to 1 with a step size of 0.05. Following this, the test set was then classified by all of the pairs of learnt models, and also by the “true” models. This entire procedure was repeated 252 times with different random seeds. The mean classification performance over the 252 trials is shown for each model pair in Figure 1.

6 Discussion

Before considering the information presented in Figure 1, it should first be noted that the amount of data used to construct the curves is somewhat insufficient. For example, the sign test shows that the hypothesis “Mackay Model Combination is no better or worse than Brand’s entropic MAP” is correct to the significance level $p \leq 0.796$, however the true models are significantly better than all others, and the entropic MAP and Mackay model averaging methods are better than Baum Welch at the 90% significance level. With these considerations of statistical significance in mind, we proceed to discuss those features Figure 1 that are likely to be significant.

The first thing to notice is that there are indeed improvements to be made over the Baum Welch algorithm. Next, we see that the “Entropy-Likelihood Combination” method is no better or worse than random for $b = 0$ – this is to be expected since $b = 0$ corresponds to pure entropy maximisation, which gives an HMM equivalent to the independent sampling of a random variable with uniform *pmf*. As b increases, so does the performance of the maximum entropy models, until $b = 0.2$. This may seem to be a surprisingly small value for optimum b, but this is partly due to the fact that in our implementation of the training method, the log-likelihood term of Equation 4 is in fact the sum of the log-likelihood for each of the five training sequences used in the test. This results in the likelihood term effectively being increased in magnitude by a factor of five. Our conjecture, however, is that the optimal value for b is a function of the entropy of the generating HMM (or more generally, the entropy of the generating source, which in most practical applications will not be an HMM). If this conjecture is correct, then it may well be possible to determine the correct value of b for a given application, based on the statistics of known sequences.

The main problem with the algorithm in its current form is the computational intractability of the entropy calculation. Unfortunately, it is unclear whether an efficient calculation exists. It may be possible to use an ad-hoc function that is similar to entropy, however it is unclear whether this will be effective. To illustrate some of the difficulties, imagine that our ad-hoc "approximation" of $H(\lambda)$ is $H_A(\lambda) + H_B(\lambda)$, where $H_A(\lambda)$ is the entropy of the states given the transition matrix and initial state *pmf*, and $H_B(\lambda)$ is the sum of the entropies of the observer matrix *pmf*s. Now consider the pathological case in which all of the observer *pmf*s have zero entropy except one, then by varying only the transition matrix, the maximum entropy HMM is obtained when the transition matrix always transitions (with probability 1) to the state with the non-zero observation *pmf* – that is, when $H_A(\lambda) = 0$! Nonetheless, there may exist a function that performs well, for example the "variance" of an HMM as defined in [9].

7 Future Work

Some possibilities for the continuation of the work are the following:

- Attempt to find an efficient calculation for $H(\lambda)$. Failing that, prove the hardness of the problem.
- Examine the performance of the system using various easily calculated ad-hoc alternatives to $H(\lambda)$.
- Investigate the relationship between the optimal b value (of Equation 5) and the entropy of the generating source.

References

[1] Matthew Brand. Structure learning in conditional probability models via an entropic prior and parameter extinction. *Neural Computation*, 11(5):1155–1182, 1999.

[2] Richard I. A. Davis, Brian C. Lovell, and Terry Caelli. Improved estimation of hidden markov model parameters from multiple observation sequences. *Proceedings International Conference on Pattern Recognition*, pages 168–171, August 2002.

[3] E.T. Jaynes. *Papers on Probability, Statistics and Statistical Physics*. Kluwer Academic Publishers, 1989.

[4] Alberto Leon-Garcia. *Probability and Random Processes for Electrical Engineering*. Addison Wesley, 2nd edition, May 1994.

[5] S.E. Levinson, L.R. Rabiner, and M.M. Sondhi. An introducton to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. *Bell System Technical Journal*, 62:1035–1074, April 1983.

[6] D.J.C. Mackay. Ensemble learning for hidden markov models. *Technical Report, Cavendish Laboratory, University of Cambridge*, 1997.

[7] K.R. Popper. *The Logic of Scientific Discovery*. Hutchinson, 1968.

[8] L.R. Rabiner and B.H.Juang. *Fundamentals of Speech Recognition*. Prentice Hall, New Jersey, 1993.

[9] Roy L. Streit. The moments of matched and mismatched hidden markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(4), April 1990.

(This page left blank intentionally)

Inherent Visual Information for Low Quality Image Presentation

Justin Boyle, Anthony Maeder, Wageeh Boles
School of Electrical and Electronic Systems Engineering
Queensland University of Technology, Brisbane, Australia
Email: jr.boyle@qut.edu.au, a.maeder@qut.edu.au, w.boles@qut.edu.au

Abstract

This paper describes our experiments to quantify the inherent information content in images as a means to optimally present images where display pixels are limited. Such low image quality applications include visual prostheses, or "bionic eyes", where implant electrodes are limited in number. Results from subjective tests with 225 normally sighted viewers are compared to predictions made with a metric for information content.

Keywords

Information content, image presentation, low quality images

INTRODUCTION

Ideally image displays should be of highest spatial resolution for adequate human perception. However in cases where size or manufacturing constraints limit the number of display pixels possible, intelligible perception is still desired from these low quality coarse images. One example is the developing area of visual prostheses, or "bionic eyes", where implanted electrodes in contact with nerve cells in the visual pathway are stimulated by electric pulses. Electrode array sizes of current prototypes include 25x25 [1] and 10x10 [2] which result in significant information loss.

We are reviewing image processing methods to efficiently use limited display pixels. Previously we have determined the impact of several novel image processing techniques on object recognition [3]. The concept of importance mapping was found to improve recognition of low quality images. Importance mapping aims to predict where the human eye will fixate in an image, ie. what are the salient areas or regions of interest in an image.

In this paper we propose an improved model which maximizes the information content in the resulting final saliency/importance map. We describe the Importance Map concept and then introduce the concept of Visual Information Content. Our psychophysical experiments to quantify this term are outlined along with the development of our metric for information content in images. Results of subjective visual information are compared to metric predictions. Finally we show that subjective information content is correlated with object recognition and is thus a suitable measure to use to optimise image presentation.

IMPORTANCE

Several region-of-interest algorithms which predict where the human eye fixates on an image are reported in the literature (eg. [4-7]). When compared against subjective tests using eye-tracking machines or similar attention-recording devices, the region-of-interest algorithms correlate highly. An extension of these algorithms is the concept of assigning an importance score or weighting to each area in an image to generate an "importance map" [5,7]. This importance ranking has previously been applied in visually lossless compression, where improved compression ratios have been achieved with high perceived image quality.

Several image features are known to influence attention in the human viewer, including motion, location, contrast, size and shape. Feature maps/images are developed representing each feature and then combined to form an overall importance map. Several combination strategies have been attempted, ranging from linear summation of features (all weighted equally) [4,5] to weights selected in accordance with eye-tracker data [7]. We propose a new method where feature map weights are selected iteratively to maximise the information content in the resulting importance map. Thus there is a need for defining the concept of information content.

QUANTIFYING INFORMATION CONTENT

We have conducted experiments to attempt to quantify the amount of inherent visual information in images. In the experiments images were compared with each other to obtain a ranking from most to least visually informative.

There were 9 image quality classes tested. Original images were 256x256 pixels representing a range of scene types. A decreasing image quality scale was presented using spatial resolutions typical of visual prosthesis designs (25x25, 16x16, 10x10) and reducing the grey levels from full greyscale to binary. It was also of interest to expose the structure of an image by presenting image edges.

A Metric for Information Content

Reduced quality image sets were prepared for each of the images shown in Figure 1. A visual information metric was developed from analyzing the subjective scores of subjects ranking the images.

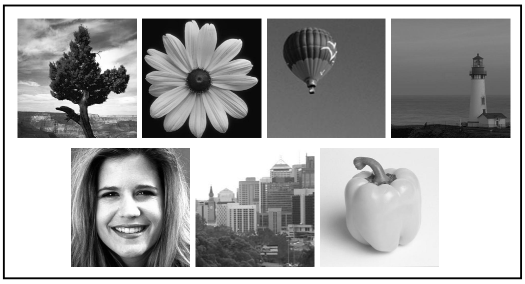


Figure 1 – Visual Information Metric obtained from 7 images representing a range of scene types.

Participation was on a voluntary basis and comprised 271 Year 11 students and 11 mature age respondents. 57 questionnaires (21%) were rejected due to invalid data. Thus the final sample size was 225, representing sample sizes of 25 for each of the 9 image quality classes.

Participants had no prior knowledge of the images. Booklet instructions stated that a range of high quality and low quality images could be expected, and although the low quality images might just appear as a range of blocks, they may be similar to what a blind person might see with a bionic eye. Viewing conditions for the experiment were not controlled.

Two questionnaire-based methods were used:

1) Seven images presented all on one page

The following instruction was presented with the images:

> *Rank the images shown on each page for visual information. Place a number in each box beside the image.*
>
> *Rank the images for how much visual information they contain:*
>
> *1 = contains most visual information*
>
> *7 = contains least visual information*

2) Paired comparison (binary decision) questionnaire test

Subjects were presented with an image pair and the instruction:

> *WHICH IMAGE APPEARS TO CONTAIN MORE INFORMATION?*
>
> *Which image could you answer the most questions about? (eg. What is the scene? How many objects?)*
>
> *If you had to rely on only one of the images to perform a task which would it be?*

Subject response was measured on a 5-point scale:

Box 1 = left image has much more information than right image

Box 2 = left image has slightly more information than right image

Box 3 = images have same amount of visual information

Box 4 = right image has slightly more information than left image

Box 5 = right image has much more information than left image

Both methods gave similar results for ranking of subjective information content. For example, when considering the ranking for all quality classes (n=225) both methods gave the following near identical ranking order:

Face > Flower > Tree > Buildings > Lighthouse/Capsicum > Balloon.

Subjective rankings have been used to propose a metric to quantify visual information that is stable across all image quality classes (not just the ones used in these tests). 15 image attributes were considered for the visual information metric:

1. file size
2. standard deviation
3. maximum standard deviation in 4 image quadrants
4. variance
5. maximum variance in 4 image quadrants
6. entropy
7. number of edges
8. number of segments
9. fractal dimension
10. 11. 12. image internal similarity measures
13. 14. 15. image symmetry measures

Three measures were used for image internal similarity (exact match across x and y axes) and image symmetry (mirror match across x and y axes):

- exact pixel match - no sub-block analysis (same result operating on big or small block)
- shaded pixel difference between blocks - 5 level sub-block analysis (objects might be in a different position within a block)
- average pixel value - 5 level sub-block analysis

Stepwise regression was used to search for the optimum subset of variables. The procedure was based on sequentially introducing variables into the model one at a time and testing the significance of all variables at each stage. The most stable performance was found to be from a metric consisting of the number of image edges alone.

ie. Information Content = *f*(edges)

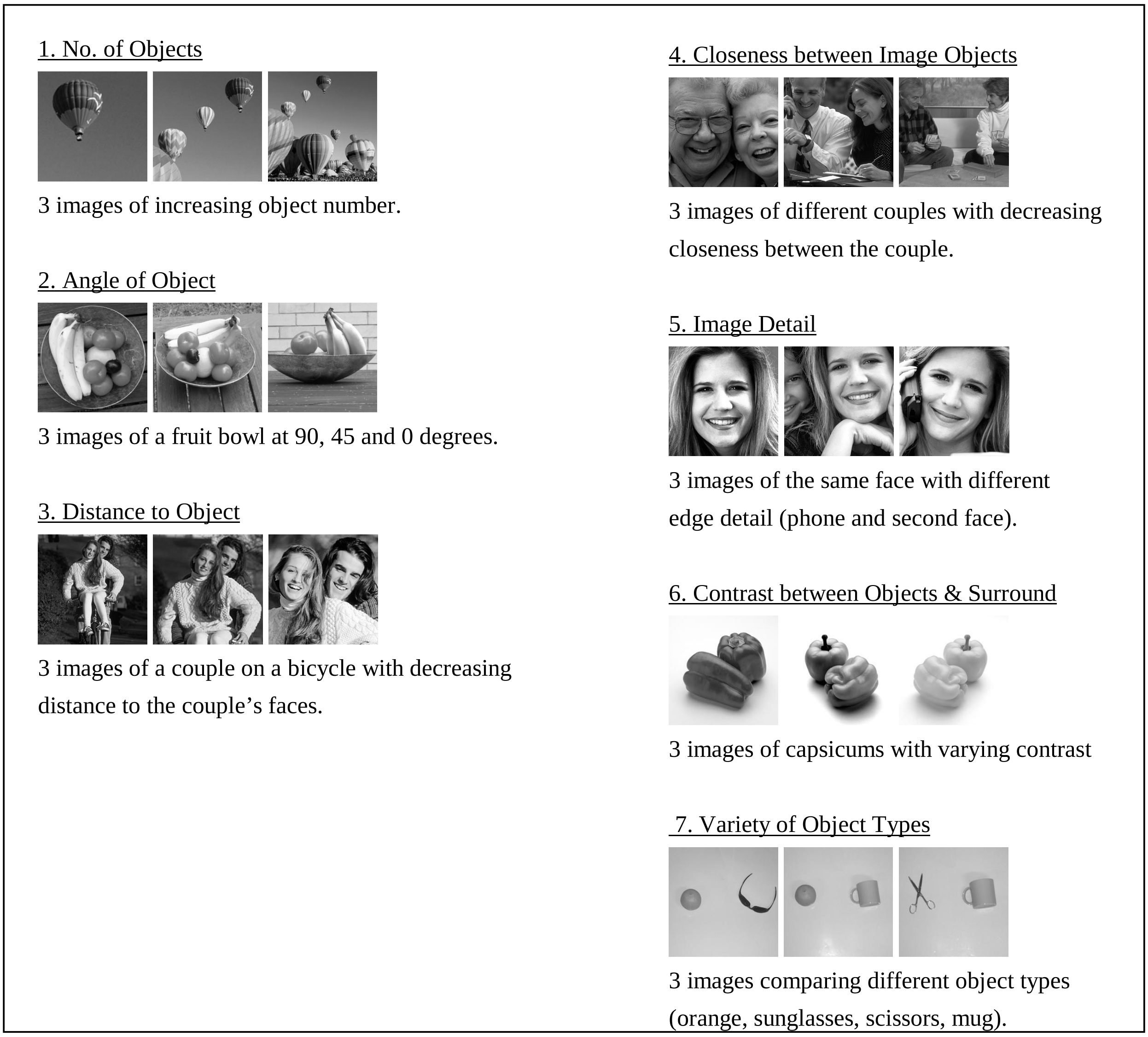
This is an interesting result considering Marr's emphasis of zero crossing (edge) detection in producing images of the external world [8]. This includes their role in the formation of a primal sketch to derive shape information from images, and biological mechanisms for detecting oriented zero-crossing segments in retinal ganglion cells.

This metric is now validated against additional data collected in the experiment.

Validating the Information Content Metric

A number of additional aspects/dimensions were explored to provide data to validate the metric and also determine what impact (if any) they had on perceived information content. These issues were assessed by comparing sets of 3 images against each other.

Predictive performance of the information content metric is tested against these results. The additional dimensions explored are shown below in Figure 2. Low quality image sets were developed for the images and subjects were asked to rank the images for the amount of visual information they contain.



1. No. of Objects

3 images of increasing object number.

2. Angle of Object

3 images of a fruit bowl at 90, 45 and 0 degrees.

3. Distance to Object

3 images of a couple on a bicycle with decreasing distance to the couple's faces.

4. Closeness between Image Objects

3 images of different couples with decreasing closeness between the couple.

5. Image Detail

3 images of the same face with different edge detail (phone and second face).

6. Contrast between Objects & Surround

3 images of capsicums with varying contrast

7. Variety of Object Types

3 images comparing different object types (orange, sunglasses, scissors, mug).

Figure 2 – Visual Dimensions used to validate the metric

Results

63 visual information rankings were obtained (7 factors/dimensions x 9 image quality classes). Dominant patterns (ie. the most frequently specified ordering in terms of perceived information content) were identified for each case. The strength of the dominant patterns (ie. the frequency with which that pattern was specified by observers) ranged from 96% (24 of 25 respondents ranking images in that order) to 28% (only 7 of 25 respondents). The number of cases for each ten percentile class is given in the first column of Table 1.

Table 1 – Metric Performance

STRONG

Dominant patterns

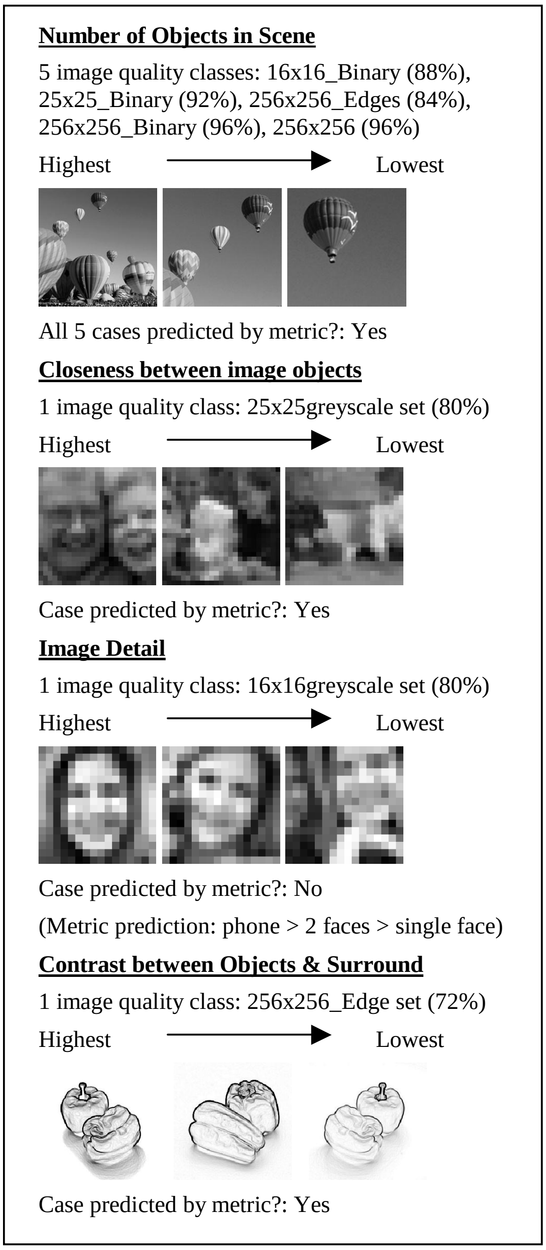
WEAK

Strength and number of cases for dominant viewer patterns (63 in total)	Frequency of image with highest info content being predicted by metric	Frequency of exact ranking being predicted by metric
90-100%: 3	100%	100%
80-89%: 4	75%	75%
70-79%: 1	100%	100%
60-69%: 12	67%	25%
50-59%: 6	83%	50%
40-49%: 16	38%	19%
30:39%: 19	32%	21%
20-29%: 2	100%	100%
10-19%: 0	-	-
0-9%: 0	-	-

The performance of the Information Content metric in predicting subjective dominant viewer patterns is also shown in Table 1. Out of the 63 test cases examined, three cases had 90% or above consensus from subjects viewing the sample set. For each of these cases, the metric successfully predicted not only which of the 3 images had the highest information content (2nd column above) but also the ranking order chosen by subjects (3rd column above). Metric performance at weaker subject consensus levels are also shown.

It was of interest to further examine strong dominant viewer patterns in the data. Eight of the 63 rankings had 70% or above consensus among viewers. Five of these related to the number of objects in the scene.

Strong viewer preferences are shown in Figure 3.



Number of Objects in Scene

5 image quality classes: 16x16_Binary (88%), 25x25_Binary (92%), 256x256_Edges (84%), 256x256_Binary (96%), 256x256 (96%)

Highest → Lowest

All 5 cases predicted by metric?: Yes

Closeness between image objects

1 image quality class: 25x25greyscale set (80%)

Highest → Lowest

Case predicted by metric?: Yes

Image Detail

1 image quality class: 16x16greyscale set (80%)

Highest → Lowest

Case predicted by metric?: No

(Metric prediction: phone > 2 faces > single face)

Contrast between Objects & Surround

1 image quality class: 256x256_Edge set (72%)

Highest → Lowest

Case predicted by metric?: Yes

Figure 3 – Strong viewer preferences (70% or above consensus among viewers) showing images ranked from highest to lowest perceived information content

Four conclusions can be drawn from Figure 3:

1. the more objects in the scene, the higher the visual information
2. the closer the objects in the scene, the higher the visual information
3. a simple face with no surrounding clutter was most visually informative at low resolution levels
4. strong edges, arising from high intensity contrast, correspond with high perceived information content

The visual information metric predicted 7 of the 8 strong viewer preferences (70% or above consensus level). Viewers of the 16x16 greyscale Image Detail set ranked a simple face as containing most visual information, while the metric ranked the image of the phone and two faces ahead of the single face. The familiarity and strong recognition of the human face at low levels of image quality may cause viewers to select it over others containing unrecogniseable blobs.

The metric was found to work best with binary images, which are expected from at least early prototype designs. (Limited greyscale may be possible by modulating stimulus amplitude, frequency and pulse duration [9]). The number of ranking cases where the metric was able to predict the image with the highest information content is shown in Table 2 below. There are a total of seven ranking cases for each image quality class, corresponding to each visual dimension explored.

Table 2 – The number of correct metric predictions of images with the highest information content

10x10 Binary set - 4/7	10x10 Greyscale set – 1/7
16x16 Binary set - 6/7	16x16 Greyscale set – 1/7
25x25 Binary set - 4/7	25x25 Greyscale set – 3/7
256x256 Binary set - 6/7	256x256 Greyscale set – 3/7
256x256 Edge set - 6/7	

This may be another reason why the metric prediction for the 10x10 greyscale Image Detail set did not agree with the ranking chosen by 80% of viewers. Table 2 shows that for 16x16 greyscale images, the metric was successful in predicting the image with the highest information content in only 1 out of 7 cases. However for 16x16 binary images, the metric prediction was correct for 6 out of 7 cases. It should be remembered that the strength of dominant patterns on which metric performance is assessed range from 96% to 28%. At high levels of viewer consensus, the metric is accurate in predicting images with the highest information content, and is thus considered acceptable for this application.

It is useful to now show that this measure for information content is an adequate pointer to how well an image might be recognised.

CORRELATIONS BETWEEN RECOGNITION RATE AND PERCEIVED INFORMATION CONTENT

We wished to determine if there was any relationship between recognition rates and the amount of visual information as perceived by viewers.

The experiment also included a component where recognition ability was assessed. Subjects were presented with images shown in Figure 1 and the following instruction:

CAN YOU TELL WHAT IS SHOWN IN EACH IMAGE.

> *Write a word under each image to describe the main object or content of the scene.*
>
> *Put a circle around the images that you are confident about.*

There were no clues provided as to the context of the image (ie. an open-ended guess). Relationships between correct object recognition and subjective information content scores were obtained for each image quality class (for example, Figure 4 shows the relationship for 25x25 binary Paired Comparison experiments).

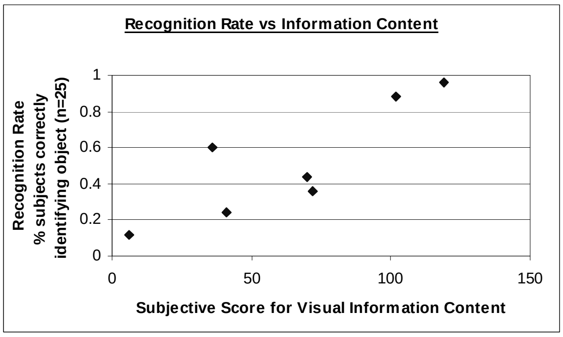


Figure 4 – Example relationship between recognition and information content (25x25 Binary Paired Comparison data)

We then assessed the significance of these relationships. Linear regression models for each quality class were developed for two series of data:

1. where images were presented at one time
2. paired comparison data

The significance of the models and correlation coefficients appears in Table 3 over.

There was some evidence for correlation between ranked information content and recognition rates with significance levels ranging from P=0.05 to P=0.1 for all but the 10x10 greyscale image set. Thus the concept of visual information content can be considered an adequate measure to optimise in importance map generation to enhance recognition.

Table 3 – Correlation coefficients between recognition rate and perceived information content

	Images presented at one time		Paired Comparison	
Image Quality Class	R Correlation Coeff.	Significance $F_{(1,7-1-1)}$ test	R Correlation Coeff.	Significance $F_{(1,7-1-1)}$ test
10x10 Binary	0.76	0.05	0.76	0.05
10x10 G.S	0.54	0.21	0.36	0.43
16x16 Binary	0.69	0.09	0.71	0.07
16x16 G.S	0.70	0.08	0.61	0.15
25x25 Binary	0.90	0.01	0.85	0.01
25x25 G.S	0.75	0.05	0.81	0.03
256x256Edge	0.70	0.08	0.66	0.10
256x256 Bin	0.69	0.09	0.73	0.06

CONCLUSIONS

In the field of low quality vision, there is a need for delivering maximum scene information to a limited number of display electrodes/pixels. In this paper we have proposed a method to enhance recognition using importance maps weighted to maximise the “information content” in the resulting importance map. We have described our experiments to quantify this term. The number of edges in an image was found to be the best statistic out of a 15-variable multiple regression analysis, to correlate with subjective rankings of visual information. The metric was tested on additional data and found to be appropriate in assessing information content. Finally we showed that subjective information content was significantly related to object recognition. We are applying this now to generating improved importance maps which will be compared to other predictive algorithms and eye-tracker data.

REFERENCES

[1] Normann R, Maynard E, Rousche P, Warren D, A neural interface for a cortical vision prosthesis, Vision Research 39(15), pp. 2577-2587, 1999

[2] Suaning G, Lovell N, CMOS Neurostimulation System with 100 Electrodes and Radio Frequency Telemetry, Inaugural Conference of the IEEE EMBS (Vic), Melbourne, pp.37-40, Feb 1999

[3] Boyle J, Maeder A, Boles W, *Image Enhancement for Electronic Visual Prostheses*, Australasian Physical & Engineering Sciences in Medicine Journal 25(2), pp.81-86, 2002

[4] Itti L, Koch C, Feature combination strategies for saliency-based visual attention systems, Journal of Electronic Imaging, 10(1), pp.161-169, 2001

[5] Osberger W, Maeder A, “Automatic identification of perceptually important regions in an image using a model of the human vision system”, *14th International Conference on Pattern Recognition, Brisbane, Australia*, pp. 701-704, 1998

[6] Privitera C, Stark L, “Focused JPEG encoding based upon automatic pre-identified regions-of-interest,” in *Human Vision and Electronic Imaging IV*, Rogowitz T, Pappas T, Editors, Proceedings of SPIE Vol. 3644, pp. 552-558, 1999

[7] Osberger W, Rohaly A, “Automatic detection of regions of interest in complex video sequences,” in *Human Vision and Electronic Imaging VI*, Rogowitz T, Pappas T, Editors, Proceedings of SPIE Vol. 4299, pp.361-372, 2001

[8] Marr D, Vision, W.H. Freeman & Co, New York, pp.54-79, 1982

[9] Suaning G, Lovell N, Schindhelm K, Coroneo M, *The bionic eye (electronic visual prosthesis): A review*, Australian and New Zealand Journal of Ophthalmology 26, 195-202, 1998

Comparison of Popular Non-Rigid Image Registration Techniques and a New Hybrid Mutual Information-Based Fluid Algorithm

C. Fookes and A. Maeder
Research Concentration for Computer Vision and Automation
Queensland University of Technology,
GPO Box 2434 Brisbane, 4001 QLD Australia
c.fookes@qut.edu.au

Abstract

Recently there has emerged a need to compute multimodal non-rigid registrations in a lot of clinical applications. To date, the viscous fluid algorithm is perhaps the most adept method at recovering large local mis-registrations that exist between two images. However, this model can only be used on images from the same modality as it assumes similar intensity values between images. This paper presents a solution to this problem by proposing a hybrid non-rigid registration using the viscous fluid algorithm and mutual information (MI). The MI is incorporated via the use of a block matching procedure to generate a sparse deformation field which drives the viscous fluid algorithm. This algorithm is compared to two other popular local registration approaches, namely Gaussian convolution and the thin-plate spline warp. Results show that the thin-plate spline warp and the MI-Fluid approach produce comparable results. However, Gaussian convolution is the superior choice, especially in controlled environments.

1. Introduction

Non-rigid image registration is an essential tool required for overcoming the inherent local anatomical variations that exist between images acquired from different individuals or atlases. The majority of these non-rigid algorithms assume the existence of similar intensities between images, restricting their use to intra- or mono-modality registrations. Recently, however, there has emerged a need to compute multimodal non-rigid registrations in a lot of clinical applications. The most prominent application of this is in the registration of pre-operative and intra-operative images. This allows the display of pre-operative anatomical and pathological tissue discrimination in the interventional field [7].

An important concept that arouse in the computer vision field during the mid 1990's was an entropy-based measure known as mutual information (MI). This measure has its roots in information theory and has demonstrated its power and robustness for use in multimodality registration in the rigid domain repeatedly. The strength of this measure lies in its simplicity as it does not assume the existence of any particular relationship between image intensities. It only assumes a statistical dependence.

MI has been incorporated into a non-rigid registration by several researchers. The main distinction between the proposed methods lie in the way the MI is calculated. This is accomplished either globally or locally [4]. However, to date MI has never been incorporated with a physical continuum model, (such as the elastic or viscous fluid algorithm). The viscous fluid algorithm is a popular approach which is capable of recovering large local mis-registrations. It also ensures that the deformation field is physically smooth. However, like most other non-rigid registrations, it assumes similar intensity values between images.

This paper proposes a novel hybrid non-rigid registration using the viscous fluid algorithm and MI. This new technique is also compared to two other popular non-rigid registration approaches, namely Gaussian convolution and the thin-plate spline warp. All three methods rely on the execution of a block matching procedure to generate an initial sparse deformation field. However, the way in which this sparse deformation field is propagated to the rest of the image depends on the technique utilised.

The outline of the paper is as follows. Some MI preliminaries are outlined in Section 2. Section 3 introduces non-rigid image registration in general, while Section 4 describes the techniques examined by this paper. This includes a general block matching approach, Gaussian convolution, thin-plate spline warps, and the new hybrid algorithm incorporating MI and the viscous fluid algorithm. Results are presented in Section 5 and conclusions are drawn in Section 6.

2. Mutual Information Preliminaries

MI is an information theoretic measure and was proposed for use in image registration by two independent groups, Viola et al. [10] and Collignon et al. [3], in 1995. The basic concept behind the use of this measure is to find a transformation, which when applied to an image, will maximise the MI between the two images.

The MI formulation used by the techniques described in this paper is based on the Kullback-Leibler measure [9] and is given by,

$$I(X,Y) = \sum_{x,y} p_{X,Y}(x,y) \log \left(\frac{p_{X,Y}(x,y)}{p_X(x) P_Y(y)} \right) \qquad (1)$$

where the densities are estimated by normalisation of the 2D frequency histograms. MI is a measure of the degree of dependence of the random variables X and Y. When formulated using the Kullback-Leibler measure in Equation 1, the MI measures the distance between the joint distribution $p_{X,Y}(x,y)$ and the distribution associated with complete independence, i.e. $p_X(x).p_Y(y)$ [8]. This measure is bounded below by complete independence and bounded above by one-to-one mappings.

3. Non-Rigid Image Registration

A rigid registration is composed solely of a rotation and translation and literally preserves the ‘rigid’ body constraint, i.e. a body is rigid and must not undergo any local variations during the transformation. This type of registration is distance preserving and is adequate for many applications in medical imaging including multimodality and intrapatient registration. However, for inter-patient registration or patient-atlas matching, non-rigid algorithms are required. In a non-rigid approach, the ‘rigid’ body constraint is no longer acceptable as it does not account for the non-linear morphometric variability between subjects [6], i.e. there exists inherent anatomical variations between different individuals resulting in brain structures that vary in both size and shape. These non-rigid algorithms allow one image to deform to match another image, thus overcoming any local variations.

A non-rigid registration defines a deformation field that gives a translation or mapping for every pixel in the image. This is generally described by the following relationship.

$$I_f \circ T(\mathbf{x}) = I_f(\mathbf{x} - \mathbf{u}(\mathbf{x})) = I_r \qquad (2)$$

In the above expression, I_f is referred to as the floating image that is undergoing the deformation while I_r is the reference image. T denotes the non-rigid transformation which equates to a translation of every pixel $\mathbf{x}$ in the floating image by a certain displacement defined by the displacement field $\mathbf{u}(\mathbf{x})$.

4. Description of Techniques

There are many ways of estimating the required displacement field $\mathbf{u}(\mathbf{x})$ in Equation 2. This includes deformable models, optical flow, elastic and viscous fluid models, spline warps, truncated basis function expansion methods, and also local registration approaches [4]. The type of method employed will also determine what constraints are imposed on the deformation field. Generally speaking, the constraints are used to ensure the existence of a smooth and continuous deformation field.

The techniques that will be described here however, are all based on a local registration approach referred to as block matching. This method is quite popular as it easily allows the incorporation of the MI measure into the non-rigid registration. This approach is described below, along with the three techniques which are used to propagate the sparse deformation field to the entire image. They are Gaussian convolution, the thin-plate spline warp, and a new hybrid algorithm incorporating MI and the viscous fluid algorithm.

4.1. Block Matching

Non-rigid registration can be made possible through local registration approaches and several methods exist to accomplish this. One common method, known as block matching, is where a grid of control points are defined on an image which are each taken as the centre of a small window. These windows, which usually overlap their neighbours, are then translated to maximise a local similarity criterion. MI is used as the similarity measure in order to obtain a robust multimodality non-rigid registration.

The location of the maximum can then be found through an exhaustive search or with the use of local optimisation strategies. The location of the maximum then represents the existence of a corresponding window in the second image, the centre of which being the homologue point of the corresponding grid point defined in the first image. Thus, this block matching approach can be used to generate two corresponding sets of control points (or landmark points) between two images. This information can then be used to generate a sparse deformation field with the translations known at each of these grid points. An example of a sparse field generated using block matching procedures is shown in Figure 1.

4.2. Gaussian Convolution

As described above, the execution of a block matching procedure results in the generation of two corresponding sets of control points. By using these control points with known deformations in a non-rigid registration, constraints are being imposed on the space of possible deformations.

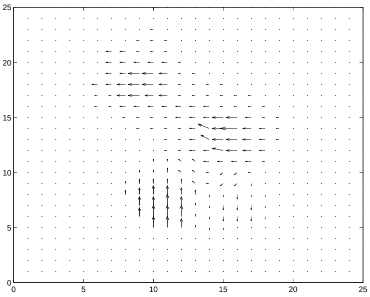


Figure 1. Sparse deformation field calculated using a block matching procedure.

This has been described as a static constraint problem [5], or an interpolation issue as the problem then becomes one of how to interpolate the deformations at these known locations to the rest of the image. Several techniques exist to accomplish this.

One of the simplest approaches is to convolve this sparse deformation field with a 2D Gaussian kernel (Gaussian smoothing), to propagate the deformations to the rest of the image. It has been described in [6] that Gaussian smoothing is equivalent to solving a heat or diffusion equation. Thus, this approach equates to an oversimplified version of a physical model-based algorithm (such as the elastic or viscous-fluid model). As model-based techniques are solved in an iterative process, the two choices essentially become whether to perform Gaussian smoothing on either the final or incremental deformation field. The first choice equates to an oversimplified elastic transformation while the second choice equates to an oversimplified viscous fluid transformation [6].

4.3. Thin-Plate Spline Warp

Another popular approach very suited to the propagation of a sparse deformation field is the thin-plate spline warp. In this method, an image is represented as a thin metal plate which undergoes certain deformations at selected points, defined by the sparse deformation field. The thin-plate spline has an elegant algebra that expresses the dependence of the physical bending energy of the thin metal plate to these point constraints [1].

For 2D image registration, two 2D thin-plate spline warps are used to describe an interpolation map from R^2 to R^2 relating two sets of landmark points, (one for the deformation in the x and y-directions respectively). The fundamental basis function used by the thin-plate spline is given by the following expression,

$$z(x,y) = -U = -r^2 \log r^2 \tag{3}$$

where r is the distance $\sqrt{(x^2+y^2)}$ from the Cartesian origin. The function $U(r)$ also satisfies the following equation.

$$\delta^2 U = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^2 U \propto \delta_{0,0} \tag{4}$$

Thus, U is a fundamental solution of the biharmonic equation $\Delta^2 U = 0$, the equation for the shape of a thin metal plate vertically displaced as a function $z(x,y)$ above the (x,y)-plane. Note that this basis function is the natural generalisation to two dimensions of the function $|x|^3$ which describes the common 1D cubic spline [1].

A thin metal plate which is subjected to vertical displacements at selected points with any arbitrary spacing will minimise the 2D bending energy of the metal plate. This is equivalent to minimising the following expression.

$$\int\int_{R^2} \left(\left(\frac{\partial^2 z}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 z}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 z}{\partial y^2}\right)^2\right) \mathrm{d}x\mathrm{d}y \tag{5}$$

The minimisation of this energy represents a smoothness criterion which imposes constraints on the deformation field, ensuring that the deformation in between the known landmark points varies smoothly. Note that this process is repeated twice - for the deformation in the x and y directions respectively.

4.4. A New Hybrid MI-Based Fluid Algorithm

To date, the viscous fluid registration algorithm is perhaps the most adept method at recovering large local misregistrations that exist between two images. This is due to the internal restoring forces which relax as the image deforms over time. This method ensures that the deformation field is physically smooth. However, like the elastic model, the viscous fluid model can only be used on images from the same modality as it assumes similar intensity values between images.

In the viscous fluid model, the instantaneous velocity field $\mathbf{v}(\mathbf{x},t)$ is linked to external forces by the Navier-Stokes viscous fluid partial differential equation which is shown below [2],

$$\alpha\nabla^2\mathbf{v}(\mathbf{x},t) + (\alpha+\beta)\nabla(\nabla^T \cdot \mathbf{v}(\mathbf{x},t)) + \mathbf{b}(\mathbf{x},\mathbf{u}(\mathbf{x},t)) = 0 \tag{6}$$

where $\mathbf{v}(\mathbf{x},t)$ is the instantaneous velocity of the displacement field $\mathbf{u}(\mathbf{x},t)$ at time t. The term $\mathbf{b}(\mathbf{x},\mathbf{u}(\mathbf{x},t))$ represents the applied forces and the parameters α and β are the viscous fluid coefficients. This equation is solved at each time step and the driving forces are derived from image differences and intensity gradients.

The main motivation behind the creation of a hybrid algorithm was to incorporate the strengths of both the viscous fluid algorithm and an information theoretic measure such as MI. This would allow the execution of a fluid registration on multimodal images. In the original viscous fluid algorithm described above, the driving forces are formulated in the most possible local manner, i.e. the force acting at a particular voxel is derived from the intensity difference and gradients of a point, not a region. However, in the approach of the hybrid algorithm, these driving forces are replaced with those derived from the MI block matching scheme. As mentioned in Section 4.1, the block matching is used to produce two sets of corresponding point sets with known deformations at each point. MI is the similarity criterion used in order to allow for a multimodal registration. The MI is also formulated using the frequency histogram approach and Equation 1.

The forces derived from the sparse deformation field are then fed into the viscous fluid algorithm which are used as the driving potentials instead of the original image differences and gradients. The significant difference however, lies in the manner in which the forces were calculated. Instead of utilising the intensity difference and gradients of a point, the block matching approach estimates the displacement field and hence the driving forces of a point by incorporating information that is contained in a small region around the point.

5. Results

The three local registration approaches were tested on a pair of simulated multimodal images with known deformations. The simulated multimodal images were generated from a single MR image and a deformed version of itself. The intensities of the reference image were also transformed using $I^* = \sin(I \times \frac{\pi}{255})$ to simulate images from different imaging modalities. These images are shown in Figure 2, along with a rescaled difference image. This difference image however, was computed without the intensity transformation in order to display more meaningful results. This is also the case for other difference images displayed later.

The results of the registration are shown in Figure 3. The letters (a), (b), and (c) are used to represent results computed with Gaussian convolution, the thin-plate spline warp, and the MI-Fluid algorithm respectively. The numbers (1), (2), and (3) are used to represent the final image after registration, the rescaled difference image, and a histogram of the intensity differences respectively. Quantitative results are shown in Table 1. This includes the SSD (sum of square differences) and SAD (sum of absolute differences) measures, and the mean μ and standard deviation σ of the error. These results are also shown for the two images before registration described by the term ‘pre-reg’.

From the rescaled difference images shown in Figure 3, it appears that all three algorithms have reduced local anatomical differences quite considerably when compared to the differences before registration. However, these rescaled difference images can be a little misleading as the intensities representing the differences are scaled to fit into the range $\{0 - 255\}$, no matter how large the actual difference. Thus, the histogram of intensity differences is also presented as another helpful avenue for evaluating the results.

Method	SSD	SAD	Error μ	Error σ
Pre-Reg	1.21×10^7	2.19×10^5	0.389	185.21
Gauss Conv	4.16×10^4	1.17×10^4	-0.003	0.64
TPS	2.52×10^6	9.43×10^4	0.099	38.48
MI-Fluid	2.82×10^6	6.49×10^4	0.109	42.98

Table 1. Quantitative error measures of registration results.

From the histograms, it can be seen that all methods have errors concentrated around the origin. However, Gaussian convolution has a much lesser spread of its errors than the other two approaches. This is also illustrated in Table 1. The Gaussian convolution method also has significantly lower SSD and SAD scores, as well as a mean error closer to the origin, and a much smaller error standard deviation. The MI-Fluid algorithm has a larger SSD score than the thin-plate spline warp, yet it has a smaller SAD score. This suggests that overall, the MI-Fluid approach produces less errors than the thin-plate spline warp. However, MI-Fluid has more errors in the outer regions which carry more weighting in the SSD measure, resulting in a higher SSD score than the thin-plate spline warp. Other results between these two methods are comparable.

6. Conclusion

This paper has proposed a hybrid non-rigid registration algorithm using MI and the viscous fluid algorithm. The MI is incorporated via the use of a block matching procedure to generate a sparse deformation field which drives the viscous fluid algorithm. Results show that the hybrid approach is successful in recovering local deformations between multimodal images. However, it is susceptible to interpolation artifacts which prevent the estimation of sub-pixel translations. Thus, the estimated deformation field will not vary smoothly, instead it will vary with integer valued steps.

This algorithm was also compared to two other popular local registration approaches, namely Gaussian convolution and the thin-plate spline warp. Results showed that the thin-plate spline warp and the MI-Fluid approach produced comparable results. However overall, simple Gaussian convolution was significantly superior. The main drawback of using Gaussian convolution is that appropriately sized variances

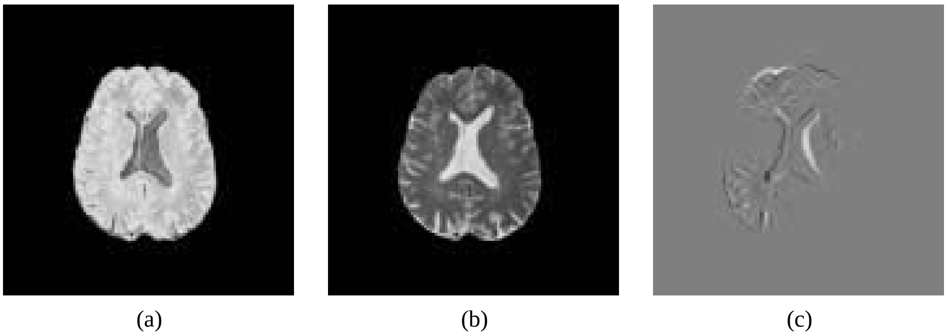


Figure 2. Manually deformed simulated multimodal images. (a) Reference image with intensity transformation $I^{*} = \sin(I \times \frac{\pi}{255})$, (b) Deformed floating Image, (c) Rescaled difference image.

and window dimensions must be selected for the Gaussian smoothing functions. The size of the variance will determine the extent of the deformation and its region of influence. In controlled environments, these variances can be manually selected for good results, as was the case in this paper. However, for situations where the amount of deformation involved and the spacing of control points in the block matching are unknown, then variance selection can have a much greater impact on final results.

From the results it is concluded that the thin-plate spline and hybrid MI-Fluid approach would be appropriate for multimodal applications that require a coarse-to-medium registration. However, these two methods do not rely so heavily on parameter selection. This suggests that these two methods may be the optimal choice in unknown situations. Overall though, if conditions are known, then Gaussian convolution is the better selection as it can be tailored to a situation, is simpler and also computationally faster than both the thin-plate spline and the MI-Fluid approaches.

References

[1] F. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, June 1989.

[2] G. Christensen. *Deformable Shape Models for Anatomy*. PhD thesis, Washington University, August 1994.

[3] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, and G. Marchal. Automated mutli-modality image registration based on information theory. In Y. B. et. al., editor, *Information Processing in Medical Imaging*, pages 263–274, The Netherlands, 1995. Kluwer Academic Publishers.

[4] C. Fookes and M. Bennamoun. Rigid and non-rigid image registration and its association with mutual information: A review. Technical Report ISBN: 1 86435 569 7, RCCVA, QUT, Brisbane, Australia, May 2002.

[5] T. Gaens, F. Maes, D. Vandermeulen, and P. Suetens. Non-rigid multimodal image registration using mutual information. In *Proceedings of the First International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 1099–1106, 1998.

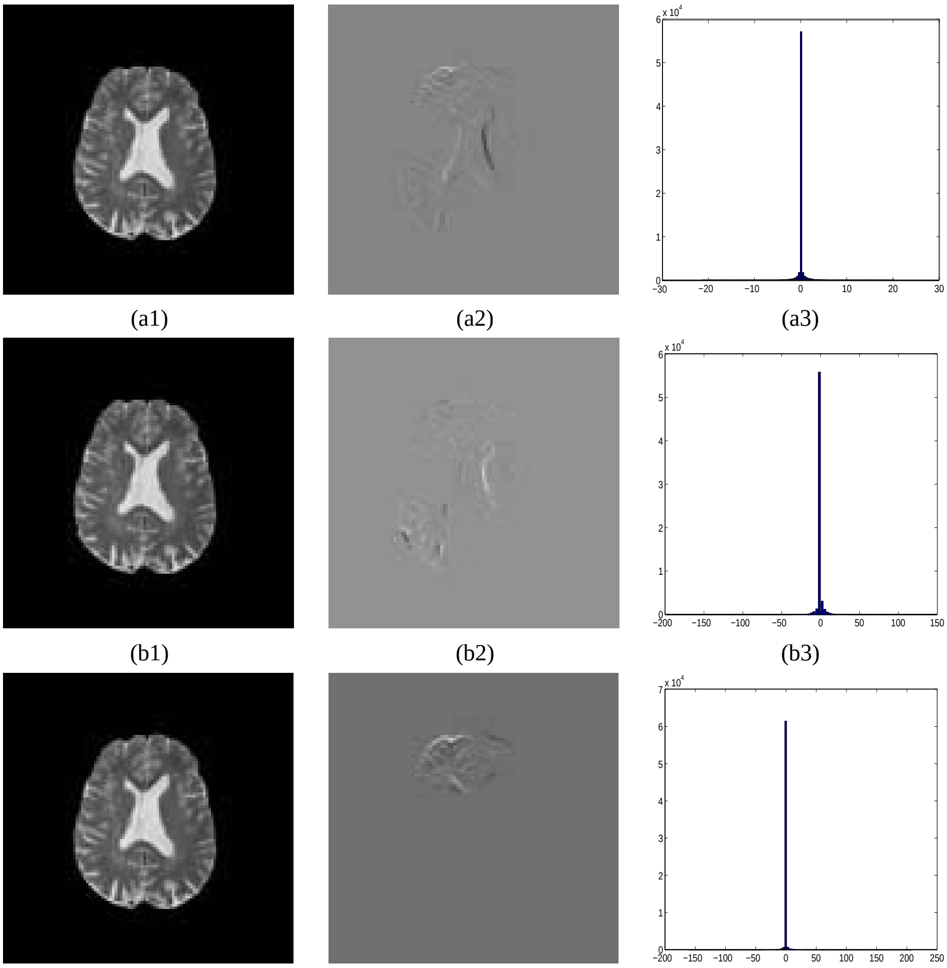
[6] T. Gaens, S. Sagaert, and D. Vandermeulen. Non-rigid registration using mutual information. Technical Report 97/10, Katholieke Universiteit Leuven, Medical Image Computing, Leuven, Belgium, 1997.

[7] N. Hata, T. Dohi, S. Warfield, W. Wells, R. Kikinis, and F. Jolesz. Multimodality deformable registration of pre- and intraoperative images for mri-guided brain surgery. In *Proceedings of the First International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 1067–1074, 1998.

[8] F. Maes, A. Collignon, D. Vandermeulen, and G. M. P. Suetens. Multimodality image registration by maximisation of mutual information. *IEEE Transactions on Medical Imaging*, 16(2):187–198, April 1997.

[9] I. Vajda. *Theory of statistical inference and information*. Kluwer, Dordrecht, The Netherlands, 1989.

[10] P. Viola and W. Wells. Alignment by maximisation of mutual information. *5th International Conference on Computer Vision*, pages 16–23, 1995.



(a1) (a2) (a3)

(b1) (b2) (b3)

(c1) (c2) (c3)

Figure 3. Registration results. Letters (a), (b), and (c) represent results computed with Gaussian convolution, thin-plate spline warp, and MI-Fluid algorithm respectively. Numbers (1), (2), and (3) represent the final image after registration, the rescaled difference image, and a histogram of the intensity differences.

Adaptive Arc Fitting for Ball Detection in RoboCup

Genevieve Coath
University of Melbourne
Dept. of Mechanical & Manufacturing Eng.
Melbourne, Australia
gcoath@ieee.org

Phillip Musumeci
James Cook University
School of Information Technology
Cairns, Australia
p.musumeci@ieee.org

Abstract

This paper presents an edge-based ball detection system for use in RoboCup robots. The algorithm can be economically coded for real-time operation in integer maths digital signal processing units. It allows colour dependencies in existing ball detection algorithms to be relaxed. This work is undertaken as part of the world-wide RoboCup project which aims to stimulate robot development by investigating difficult test problems.

1 Introduction

The international RoboCup competitions stimulate development in mobile robots that function collectively and incorporate a team approach to problem solving. While the RoboCup challenge is only part way towards its stated 2050 goal of achieving autonomous humanoid robots that can defeat humans in soccer, the existing systems are relatively complex and display team cooperation. A hierarchical view of the RMIT University robot system shows a top-level strategy module which coordinates the operation of robot team members using sensory inputs such as vision and touch, and implements strategies via actuator outputs including motor control, steering, and ball kicking systems.

This paper presents research done at RMIT University to enhance the vision system of its middle league (F2000) robots. In particular, the strict colour dependency of previous ball detection systems has been relaxed by developing an adaptive arc identification and location system that processes image data containing edge information. Analysis of arc parameters such as radius, location, and recent trajectory is then used in revised ball identification schemes.

Examples are presented for a number of edge extraction techniques applied to colour image field data, and the adaptive arc detection scheme is shown applied to a ball in full view and also a ball obscured by a (worst case) curved object.

This paper provides a brief description of the existing image processing system in order to set the context for the arc-based scheme. Some preprocessing algorithms used to generate edges are described, including schemes with simplified computation to assist in real-time application, and finally the edge tracking algorithm is presented with field images.

2 Visual Environment

In current competitions, the robot environment is tightly controlled so that robots can interact with a world that can be understood with very limited knowledge. The field lighting levels are specified so that robots do not need to handle significant dynamic lighting effects. In addition, the ball and each goal region may be uniquely coloured to simplify robot orientation and target acquisition. The dependency of current robots on the use of colour in identifying objects is highlighted by the need, in some tournaments, to remove coloured items from spectators because they can confuse the vision systems.

Following the Melbourne 2000 event, a RoboCup rules debate suggested the possibility of increasing the complexity of the environment in which the robots must compete, through such changes as removal of field walls and unique colour schemes. This debate has lead to the current work which attempts to incorporate additional shape information into ball identification while reducing the relative importance of colour.

2.1 Overview of Existing System

The existing vision system [UVD,SBK] uses a Pulnix 1/3" CCD colour analogue camera attached to a custom interface board that connects a Brooktree video capture device to a Texas Instruments DSP (TMS320C6211).

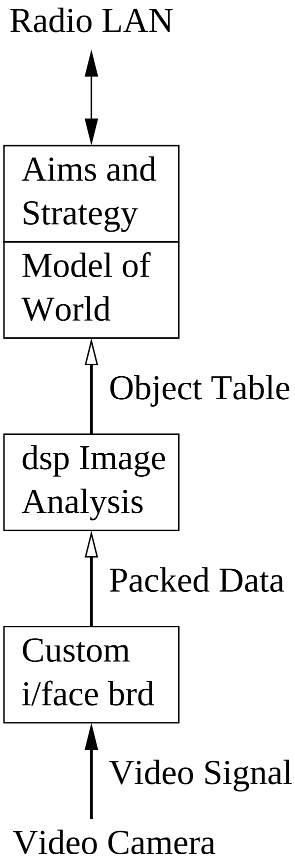


Figure 1. System Information Flows

Figure 1 shows the information flows in the three processing stages. The first stage accepts a video signal from the camera, digitises it, and reduces the video resolution by coding the digital data to 16 bits/pixel to allow more efficient pixel data packing in DSP memory. In the second stage, the enhanced direct memory access (EDMA) controller provided by the DSP is programmed to be a state machine which retrieves image frames from the video capture device at a rate of 25 frames/second and an image size of 450x450 pixels with little CPU intervention.

The DSP analyses the image data and creates an object table with attributes of type (line, solid, etc.), colour, and image coordinates to describe each element. Field information is based on straight line edges detected by a modified Hough transform using integer maths. Information relating to other robots and the ball is derived from colour information.

The ball detection system analyses image scans and detects colour transitions. Regions that match the colour of the ball are then analysed and an estimate of center and diameter are generated. If these parameters are within suitable limits, a ball detection occurs.

To reduce computation, the estimated diameter is obtained from the widest horizontal scan segment with the desired ball colour. If a closer object obscures either side of a ball, then the estimate of center and diameter will be incorrect. Ideally, both the largest vertical and largest horizontal dimension should be used to derive diameter estimates. However, the use of overhead lighting systems in competition venues may cause shadows which make the detection of the bottom edge of the ball unreliable.

The final stage is implemented by interfacing the DSP output to a robot-mounted laptop which processes aims and strategies, and estimates range information based on the vertical position of an object within an image. The laptop also handles the other sensor inputs such as proximity detectors, and outputs such as motor controller commands.

2.2 Deployment of Image Analysis Algorithms

The system was developed by incorporating a shape-based ball identification algorithm into the existing colour-based ball detection scheme. The new algorithm initially provides an additional test of results from the previous ball detection system.

3 Edge-based Arc Detection

Arc detection is used to determine the ball center and diameter as a full circle is not available when the ball is partially obscured. Because there are potentially many other arc sources in a field, additional information such as dimensional constraints, relationship to previous ball sightings, and colour is used to filter out spurious responses.

To detect arcs, edges in the image are enhanced and a contiguous set of straight line segments is constructed between adjacent pixels that may form an edge. By interpreting the straight line segments as chords of an arc, perpendicular projections from each chord are used to identify the centers of these potential arcs. A cluster of centers will then confirm and identify an arc.

3.1 Edge Detection

The edge information is obtained by applying a spatial differentiation operator to the image — see for example [SBA]. For colour independence, this operator can be fed a preprocessed image such as image energy i.e. an L_2 norm is applied to the raw image, or the absolute value of the pixel i.e. an L_1 norm is applied to the raw image. An advantage of preprocessing with an L_1 operator before applying the spatial differentiation operator is a reduction of computation. The application of these operators is now detailed.

Figure 2 shows an ideal ball image sampled horizontally, an image energy curve, its first derivative, and its second derivative.

Any rapid changes in image energy I across a small region give rise to higher peak values in the differen-

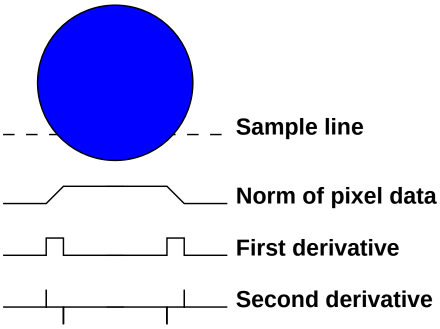


Figure 2. Ideal Ball

tiated image. An approximation to a one-dimensional differential operator at position $x = i\Delta$ is $\frac{dI}{dx} \approx \frac{I(i+1)-I(i)}{\Delta}$, where Δ is the spatial sample interval in the x direction. Extending this to two dimensions gives

$$\begin{aligned}\frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} &= k\left(I(i+1,j) + I(i-1,j) + I(i,j+1)\right.\\ &\quad \left.+I(i,j-1) - 4I(i,j)\right); \mathrm{k} = \mathrm{constant.}\end{aligned}$$

Applying this operator to the typical field view of a robot soccer field as shown in Figure 3(a) gives the result shown in Figure 3(b). Previous research in [ERD,UVD] shows that binarisation of RGB components of the image, in an image preprocessing stage prior to differentiation, leads to an enhanced edge display as shown in Figure 4(a). This provides the added benefit of data reduction.

A further improvement in edge definition may be possible for a typical field view if the preprocessing includes: a band-pass filter centered on the target colour; and a norm operator to generate monochrome image data where the maximum output corresponds to the target object's colour. Preprocessing is then completed with binarisation and differentiation to yield the image in Figure 4(b). In effect, this recognises that *typical* field objects have a colour that is clearly different to the target. Note that if the target has a colour that matches one of the dimensions used to represent colour, e.g. red, green or blue in an RGB format, then further computation reductions present themselves.

3.2 Radius Projection

Given a region of interest, our algorithm chooses as a reference a pixel in the edge enhanced image with significant energy[1] i.e. point p0 in Figure 5. Then, a nearby pixel p1 with significant energy is located such that the two points form a chord of length $> l$. A perpendicular projection through the mid-point of chord p0-p1 indicates potential locations for the center of the arc p0-p1. This process is then repeated for a sequence of pixels pairs located on the potential arc, and the intersections between the sequential projections leads to a cluster of potential center points. This example shows two chords p0-p1 and p1-p2 with projections intersecting at a potential arc center.

The minimum length constraint l is imposed on each chord because the location of each point is spatially discretised based on the pixels in the sampled image, and so very small length chords would introduce excessive error in the calculation of centers.

When this algorithm is applied to an unobscured ball image, the result shown in Figure 6(a) is obtained. In this example, the algorithm has been run a number of times with slightly different starting points, so that a larger number of projections have been produced. The clustering of intersections between successive projections gives the location of the center of the arc (or in this case, circle).

Note that, while we have highlighted the projections in this image, the actual system is only concerned with the intersection point of pairs of equations that are illustrated by the projections from pairs of adjacent chords.

An advantage of this algorithm is that it can function on obscured balls. In Figure 6(b), a number of irregular objects are covering parts of the right hand side of the ball which is also slightly distorted vertically. A number of potential centers are identified, in rough proportion to the size of the corresponding arc. In this example, an arc associated with the ball is a dominant image feature so the ball center is correctly found. If the ball is mostly obscured, then the result is unpredictable without additional information. The algorithm is now described in detail.

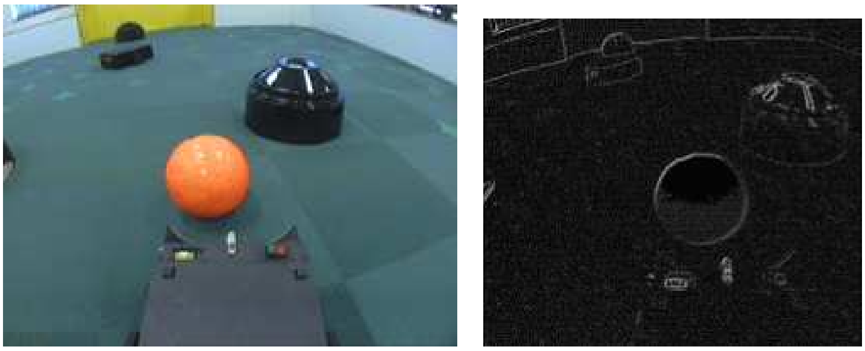
3.3 Arc Location Algorithm

As mentioned previously, the points located on the potential arc or circle have a minimum spacing of l enforced. The algorithm starts from the initial point identified for the largest edge image pixel, and then generates a sequence of points along the potential arc or circle.

1. Obtain edge image via application of two-dimensional differentiation operator or similar transform;

2. Find the magnitude of the maximum pixel e_{max}

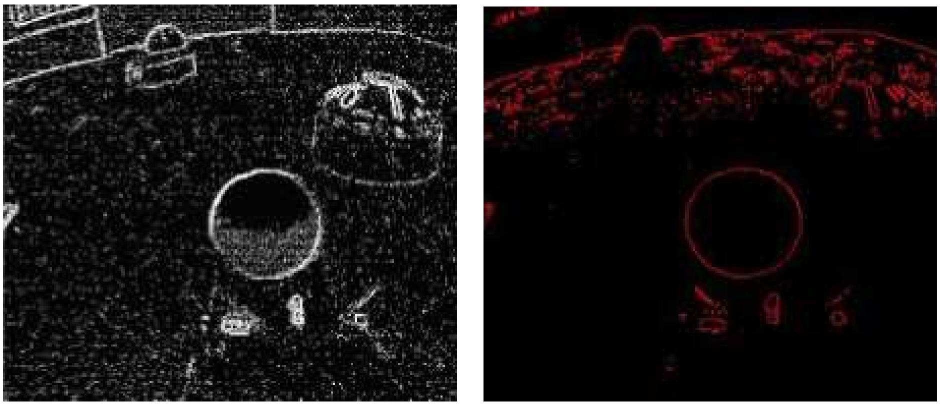
[1] In off-line testing, the initial pixel may be chosen as the largest value in the edge enhanced image for comparison purposes.



(a) Typical RoboCup Field

(b) Differentiated Image (smaller after processing)

Figure 3. Field Views



(a) Binarised & differentiated

(b) 1-colour, binarised & differentiated

Figure 4. Field Views with Preprocessing

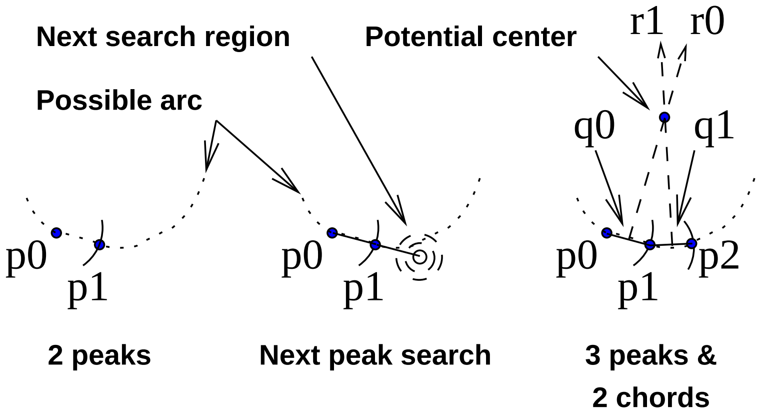
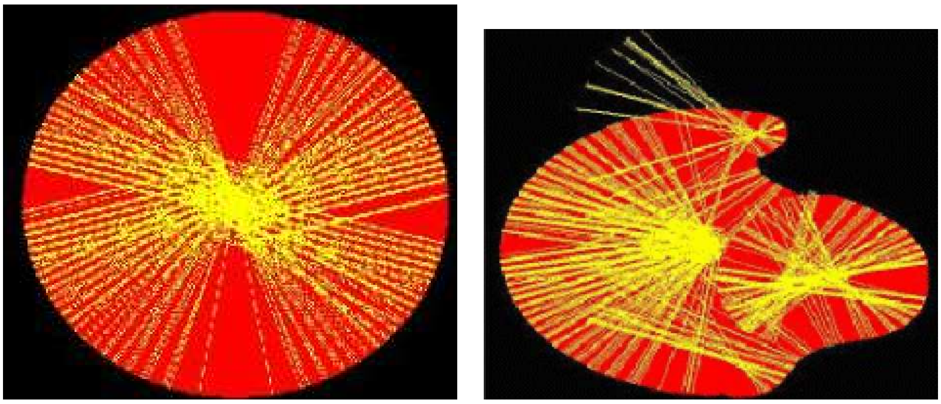


Figure 5. Edge Tracking (3 stages shown)



(a) Test Circle Detection

(b) Obscured Circle

Figure 6. Projections for Center Location

in edge image, and choose α such that threshold $e_t = \alpha e_{max}$ can be used as a lower bound to identify significant peak values $e \geq e_t$. In our work, we typically use $\alpha = 0.9$;

3. Designate the first point $p_0 = \{x_0, y_0\}$ where $e(p_0) = e_{max}$;

4. Find the next point, $p_1 = \{x_1, y_1\}$, where $e(p_1) \geq e_t$ and $\|p_1 - p_0\|^2 > l^2$, by conducting a localised horizontal and then vertical scan for a pixel with significant energy in the next search region found by extending the current chord (illustrated in the 2^{nd} stage of Figure 5). Note that calculation of a square root to ensure adequate spacing of points p_i is avoided by testing distance squared;

5. Identify the chord mid-point $q_0 = \{(x_1 + x_0)/2, (y_1 + y_0)/2\}$. The perpendicular projection from point q_0 to r_0 is defined in terms of a vector dot product i.e. $\overline{r_0 - q_0}.\overline{p_1 - p_0} = 0$;

6. Repeat the previous two steps to generate the point sequences $\{p_0, p_1, p_2, \ldots\}$, $\{q_0, q_1, q_2, \ldots\}$, and a set of equations for $\{r_0, r_1, r_2, \ldots\}$;

7. Choose to solve for intersection equations for $\{r_i, r_{i+1}\}$ i.e. sequential projections, to generate potential center locations c_i. Solving for intersection of adjacent projections also minimises storage of point data;

8. Finally, an average of the coordinates of potential center locations c_i gives an estimate of the arc (or ball) center. For more complex field views, the image may be segmented and the average c_i in a segment suggests a localised arc center.

The potential center locations c_i can be interpreted as a vote for a potential arc center and therefore a vote for the presence of a ball. An additional filtering stage under development is to reject c_i values that are located more than a specified distance from a potential center, in order to obtain a more tightly bound cluster that is not influenced by outliers. For the current robots, an ad-hoc filtering stage rejects a potential center that fails to have 60% of its center estimates within ±20 pixels of the horizontal center or 60% of its center estimates within ±30 pixels of the vertical center. However, this scheme needs to be recast in relation to the object size.

Where ambiguous results are obtained, additional information must be used to clarify the location of a single ball. For the RMIT University systems, this arc detection algorithm is being used to enhance reliability of other ball location schemes.

3.4 Algorithm Tuning

The size of l can be found by testing the algorithm on the top and bottom arcs of a test ball, and then on the far left and far right arcs of a test ball. The smaller l is, the better tracking is obtained for the edge. However, projection errors due to the spatial discretisation due to horizontal and vertical (scan) image sampling become significant if l is made too small. For the range of ball sizes encountered when viewing a ball at a range of distances, l is chosen so that the chord subtending each arc is approximately 10 or more pixels long.

The threshold selection parameter α is chosen so that a suitable number of preprocessed edge peaks are generated — if α is too small, the edge tracker can be confused by noise while if α is too high, any arcs may be undersampled. It is possible to add a feedback loop that adapts α up or down in order to achieve a reasonable number of peaks. However, the effect of α is not unduly sensitive in typical field lighting conditions due to the use of preprocessing stages (binarisation and differentiation).

4 Conclusion

We have presented an edge tracking algorithm and arc location scheme that has been successfully applied to RoboCup field images. A number of image data preprocessing strategies have been discussed, with features such as data reduction and colour-centered edge enhancement. The edge tracking algorithm provides a means of locating arcs and circles to identify the soccer ball. In the implementation on the TMS320C6211 DSP, the actual chord and associated perpendicular projection equations were implemented with minimal transcendental function calls because this DSP does not have a floating point unit.

The availability of alternative ball detection algorithms allows comparison testing and also focusses attention on benchmarking. Previous RMIT RoboCup systems have concentrated on getting a single algorithm working for each critical stage so comparison testing was not possible. In the current robot architecture, target tracking is performed at higher levels so comparisons will apply to the processing of individual image frames with the assumption that a recently acquired target allows analysis to be restricted to smaller parts of an image over a short time interval.

As the whole system must operate in real-time, it is desirable that partial results obtained in the sample time be of use — the algorithm proposed here can identify a ball before all of the visible circumference has been traversed if a suitable number of potential centers has been found i.e. an early result is possible without identifying all chords or processing all significant edge pixels.

We acknowledge the RMIT University RoboCup teams of 1999 and 2000, and specifically thank Mark Makies, Duy Khuong Do, Sy Quang Dinh, Karl Fantone, Alan Harvey, and Shereene Jesurajah. We thank Colin Lemmon (JCU) and the anonymous reviewers for improving this paper.

5 References

ERD “Machine Vision - Theory, Algorithms, Practicalities”, E.R. Davies, 2nd edition, Academic Press, 1997.

SBA “Diffuse edge fitting and following: a location-adaptive approach”, A.L. Shipman, R.R. Bitmead and G.H. Allen, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-6, no. 1, January 1984, pp. 96-102.

UVD “Field Understanding System and Vision System”, RMIT United Vision & DSP Group, 2000, Internal Report.

SBK “RoboCup 2000: Robot Soccer World Cup IV”, Peter Stone, Tucker R. Balch, Gerhard K. Kraetzschmar (editors), Lecture Notes in Computer Science 2019, Springer 2001, ISBN 3-540-42185-8.

Object Tracking in Image Sequences using Point Features

P. Tissainayagam and D. Suter
raj@tns.nec.com.au and d.suter@eng.monash.edu.au
Dept. of Electrical and Computer Systems Engineering
Monash University, Clayton, Vic. 3168, Australia

Abstract

This paper presents an object tracking technique based on the Bayesian Multiple Hypothesis Tracking (MHT) approach. Two algorithms, both based on the MHT technique are combined to generate an object tracker. The first MHT algorithm is employed for contour segmentation (based on an edge map). The second MHT algorithm is used in the temporal tracking of a selected object from the initial frame. An object is represented by key feature points that are extracted from it. The key points (mostly corner points) are detected using information obtained from the edge map. These key points are then tracked through the sequence. To confirm the correctness of the tracked key points, the location of the key points on the trajectory are verified against the segmented object identified in each frame. The results show that the tracker proposed can successfully track simple identifiable objects through an image sequence.

Key words: Object tracking, Key points, Multiple Hypothesis Tracking, Contour segmentation, Edge grouping.

1 Introduction

The primary purpose of this paper is to track a selected object (as opposed to a single point feature) from the initial frame through the image sequence. The process is an attempt to extend the point feature tracking introduced in [13, 14] to object tracking. In this case, key points from the object are selected using a curvature scale space technique [11] to represent that object. The key points are temporally tracked and are validated against the object contour (obtained by grouping edge segments) in each frame. The tracking technique involves applying the MHT algorithm in two stages: The first stage is for contour grouping (object identification based on segmented edges) and the second stage is for temporal tracking of key features (from the object of interest). For the contour grouping process, we employed the algorithm developed by Cox et al [6], and for the key point tracking procedure we used the tracker introduced by the authors in [13]. Both algorithms combine to provide an object tracker.

The set of image contours produced by objects in a scene, encode important information about their shape, position, and orientation. Image contours arise from discontinuities in the underlying intensity pattern, due to the interaction of surface geometry and illumination. A large body of work, from such areas as model-based object recognition and contour motion flow, depend critically on the reliable extraction of image contours. Reliable image contours are necessary to identify an object with certainty, which in turn is necessary for tracking the object over a period of time in a sequence of images. We use the term ‘object’ for a group of edge segments that form a recognisable object (identified as belonging to the same object). The object will be identified by an enclosed (*or* near-enclosed) contour.

This paper is organised as follows: Section 2 gives a brief description of the Multiple Hypothesis Tracking (MHT) approach relating to edge segmentation. Section 3 shows how the multiple hypothesis approach can be used for object recognition. In section 4 we briefly show the process to extract key points from an object, and the MHT approach for tracking key point features through an image sequence. Section 5 provides the object-tracking framework employed using methods described in section 3 and 4. Section 6 gives results obtained from experiments. Section 7 gives a general discussion, and finally section 8 provides the conclusion.

2 Multiple Hypothesis Framework for Contour Grouping

This section briefly describes the multiple hypothesis approach in relation to contour segmentation. The details of which are discussed in [6-8].

Fig. 1 outlines the basic operation of the MHT algorithm for contour grouping. At each iteration, there are a set of hypotheses (initially null), each one representing a different interpretation of the edge points. Each hypothesis is a collection of contours, and at each iteration each contour predicts the location of the next edgel as the algorithm follows the contour in unit increments of arc length. An adaptive search region is created about each of these predicted locations as shown in Figure 2 [6]. Measurements are extracted from these surveillance regions and matched to predictions based on the statistical Mahalanobis distance. This matching process reveals ambiguities in the assignment of measurements to contours. This procedure provides an associated ambiguity matrix (Ω) for each global hypothesis from which it is necessary to generate a set of legal assignments. As a result, the hypothesis tree grow another level in depth, a parent hypothesis generating a series of hypotheses each being a possible interpretation of the measurements. The probability of each new hypothesis is calculated based on assumptions described in [6, 8]. Finally, a pruning stage is invoked to constrain the exponentially growing hypothesis tree. This completes one iteration of the algorithm.

In the following sections we briefly describe the contour-grouping algorithm employed, and the key point selection and tracking process used. Both these methods are based on the multiple hypothesis approach.

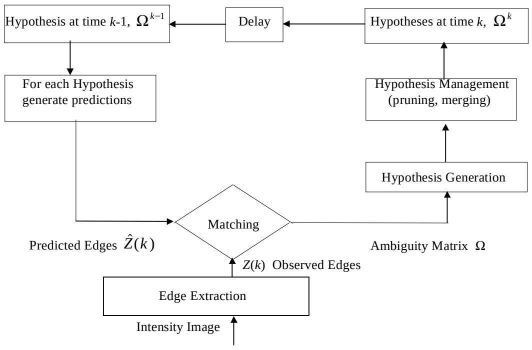

Hypothesis at time k-1, Ω^{k-1}
Delay
Hypotheses at time k, Ω^{k}
For each Hypothesis generate predictions
Hypothesis Management (pruning, merging)
Hypothesis Generation
Matching
Predicted Edges $\hat{Z}(k)$
Ambiguity Matrix Ω
$Z(k)$ Observed Edges
Edge Extraction
Intensity Image

Figure 1: Outline of the multiple hypothesis algorithm for edge grouping

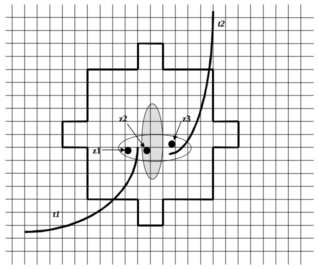


Figure 2: Predicted contour locations, a surveillance region and statistical Mahalanobis (elliptical) regions for a situation with two known contours (t1 and t2) and three new measurements (z1, z2 and z3).

3 Object Recognition

3.1 Contour Segmentation

The contour grouping problem examined in this paper, involves assigning edge pixels produced by an edge detector [4, 5] to a set of continuous curves. Associating edge points with contours is difficult because the input data (from edge detectors) is noisy; there is uncertainty in the position of the edge, there may be false and / or missing points, and contours may intersect and interfere with one another. There are four basic requirements for a successful contour segmentation algorithm. First, there must be a mechanism for integrating information in the neighbourhood of an edgel to avoid making irrevocable grouping decisions based on insufficient data. Second, there must be a prior model for the smoothness of the curve to base grouping decisions on. This model must have an intuitive parameterisation and sufficient generality to describe arbitrary curves of interest. Third, it must incorporate noise models for the edge detector, to optimally incorporate noisy measurements, and detect and remove spurious edges. And finally, since intersecting curves are common, the algorithm must be able to handle these as well. The algorithm due to Cox et.al. [6-8] is one which has a unified framework that incorporates these four requirements, and we will use this algorithm for contour segmentation.

The contour grouping is formulated as a Bayesian multiple hypothesis 'tracking' problem (as in [12]). The algorithm has 3 main components. A dynamic contour model that encodes the smoothness prior, a measurement model that incorporates edge-detector noise characteristics, and a Bayesian hypothesis tree that encodes the likelihood of each possible edge assignment and permits multiple hypothesis to develop in parallel until sufficient information is available to make a decision.

A key step in assigning probabilities to segmentation hypothesis is the computation of the likelihood that a given measurement originated from a certain contour. This likelihood computation depends on two things: a dynamic model that describes the evolution of the curve in the image, and a measurement model that describes how curves produce edgels. In this formulation, the curve state vector is $[x \;\; \dot{x} \;\; y \;\; \dot{y}]^T$ (where (x, y) are the position in a Cartesian coordinate) and its dynamics are described by a linear noise-driven acceleration model common in the tracking literature [1, 2]. The autocorrelation of the white Gaussian acceleration noise can be varied to model curves of arbitrary smoothness. Thus the tip (end point) of the contour as a function of arc length, u, is $(x(u), y(u))$ and has tangent $(\dot{x}(u), \dot{y}(u))$. Since many edge detectors provide gradient information, it is assumed that the entire state vector is available for measurement (a good edge detector such as Canny [5], Boie-Cox algorithm [4] etc. which provide both position and coarse gradient information (horizontal, vertical, and two diagonals) is employable for this application). A Kalman filter is then employed to estimate curve state and predict the location of edgels. These predictions are combined with actual measurements to produce likelihoods.

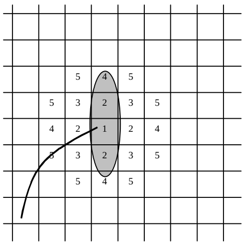


Figure 3: A contour, its surveillance region (labelled 1 – 5) and its validation region.

Once the location of a given curve has been predicted by the Kalman filter and discretized to image coordinates, a surveillance region is employed to extract measurements. A surveillance region is an adaptive variable sized window that travels with the tip of the contour and is used to extract measurements from the edge map. Every iteration, each contour searches for edge points in a series of circles of increasing diameter centred at the predicted contour endpoint. The search halts as soon as at least one measurement is found, or the maximum search radius is reached. The size of the surveillance region determines the distance the curve must travel in that time period, and is reflected in the step size for the curve. The use of a set of windows of increasing size ensures that no more than one measurement from the given contour will be found in a single time period.

The search for measurements takes place after the prediction phase of the state estimator generates an extrapolated endpoint location, (x, y), for the contour. This location determines the discrete image coordinates,

(x_i, y_j), at which the surveillance region is centred. If there is no edge at the predicted location, concentric circles (see Fig. 3), of radius $1,\sqrt{2},2,\sqrt{5}$, are searched for edgels (the radii define discrete pixel neighbourhoods). These surveillance regions are labelled 1 to 5 in Fig. 3 (It should be noted that the surveillance region of a contour is not equivalent to its validation region, which is defined by the Mahalanobis distance and is depicted in Fig. 3 as an ellipse). It is these measurements that form segmentation hypothesis whose probabilities are computed. See [6] for details.

3.2 Contour Merging

The grouped contours resulting from the above mentioned process still might have breaks and gaps between segments of the same object. A further refinement process can be employed to merge segments to form identifiable objects. A merging technique is employed by using a distance test (eg: Mahalanobis distance) applied to the end points of contours (assuming a non-closed contour). In this case the multiple contours can be merged to recover the correct segmentation, compensating for the incorrect initial conditions. Two contours with state estimates $\hat{x}_i$ and $\hat{x}_j$ at common boundary are merged if $dx'_{ij}\,\mathrm{T}^{-1}_{ij}\,dx_{ij} \le \delta$, where $dx_{ij} = \hat{x}_i - \hat{x}_j$. T_{ij} is the covariance, and δ is obtained from χ^2 tables or set appropriately as a threshold. This test is applied after the algorithm produces an initial segmentation. The procedure resolves many ambiguities left by the contour segmentation algorithm. A simpler algorithm can also be used simply by using the end-point positions and derivatives of the end-point positions of each curve (produced by the edge detector) which can be quicker.

4 Temporal Tracking of Key Feature Points

In this section we discuss the process to extract key points from the object of interest and we also discuss the procedure to track them temporally.

4.1 Extracting Key Feature Points from Objects

In order to temporally track the object of interest, key points from the object are extracted to represent the object. The key point extraction method should ensure that only true corner points (or any clearly identifiable and definable points) are extracted. Extraction of multiple points within a small region should be avoided (eg: in a curved object, ideally only 1 point should be selected from the curved portion) for good tracking. Since contour grouping (discussed in the previous section) is based on an edge-map, it is desirable that key points should also be selected from the same edge map. Such a process will be efficient and will eliminate the requirement to employ a separate corner detection algorithm. Because of these limitations, we cannot effectively use any of the standard corner extraction algorithms [11] (these calculate corner values directly from the raw image). Instead we have employed a method called the curvature scale space technique [11], which selects key points directly from an edge map efficiently. In the next section the curvature scale space technique is discussed in brief.

4.2 The Curvature Scale Space Algorithm (CSS)

The CSS technique is suitable for recovering invariant geometric features (curvature zero-crossing points and / or extrema) of a planar curve at multiple scales. To compute it, a curve Γ is first parameterised by the arc length parameter *u*:

$$\Gamma(u) = (x(u), y(u))$$

An evolved version Γ_σ of Γ can then be computed. Γ_σ is defined by:

$$\Gamma_\sigma(u) = (\chi(u,\sigma), \gamma(u,\sigma)),$$

where

$$\chi(u,\sigma) = x(u) \otimes g(u,\sigma) \qquad \gamma(u,\sigma) = y(u) \otimes g(u,\sigma),$$

where $\otimes$ is the convolution operator and $g(u,\sigma)$ denotes a Gaussian of width σ (σ is also referred to as the *scale* parameter). The process of generating evolved versions of Γ as σ increases from zero to infinity is referred to as the evolution of Γ. This technique is suitable for removing noise from, and smoothing a planar curve as well as gradual simplification of its shape. In order to find curvature zero-crossings or extrema from evolved versions of the input curve, one needs to compute the curvature accurately and directly on an evolved version Γ_σ. Curvature κ on Γ_σ is given by [11]:

$$\kappa(u,\sigma) = \frac{\chi_u(u,\sigma)\gamma_{uu}(u,\sigma) - \chi_{uu}(u,\sigma)\gamma(u,\sigma)}{\left(\chi_u(u,\sigma)^2 + \gamma(u,\sigma)^2\right)^{1/2}}$$

where

$$\chi_u(u,\sigma) = x(u) \otimes g_u(u,\sigma) \qquad \chi_{uu}(u,\sigma) = x(u) \otimes g_{uu}(u,\sigma)$$

$$\gamma_u(u,\sigma) = y(u) \otimes g_u(u,\sigma) \qquad \gamma_{uu}(u,\sigma) = y(u) \otimes g_{uu}(u,\sigma)$$

4.3 CSS Key Point Detection Method

4.3.1 Brief Overview

The corners (key points) are defined as the local maxima of the absolute value of curvature. At a very fine scale, there exist many such maxima due to noise on the digital contour. As the scale is increased, the noise is smoothed away and only the maxima corresponding to the real corners remain. The CSS detection method finds the corners at these local maxima.

As the contour evolves, the actual locations of the corners change. If the detection is achieved at a large scale the localisation of the corners may be poor. To overcome this problem, local tracking is introduced in the detection. The corners are located at a high scale σ_{high}, assuring that the corner detection is not affected by noise. σ is then reduced and the same corner points are examined at lower scales. As a result, location of corners may be updated. This is continued until the scale is very low and the operation is very local. This improves localisation and the computational cost is low, as curvature values at scales lower than σ_{high} do not need to be computed at every contour point but only in a small neighbourhood of the detected corners.

There are local maxima on the evolved contours due to rounded corners or noise. These can be removed by introducing a threshold value *t*. The curvature of a sharp corner is higher than that of a rounded corner. The final stage to the candidate corner declaration is that each local maximum of the curvature is compared to its two neighbouring local minima. The curvature of a corner point should be double the curvature of a neighbouring extremum. This is necessary since if the contour is continuous and round, the curvature values can be well above the threshold value *t* and false corners may be declared.

4.3.2 CSS Detection Process

The CSS key point detection process can be given by the following steps:

1. Utilise an Edge detector (such as Canny [5] or Boie-Cox [4] etc.) to extract edges from the original image.
2. Extract the edge contours from the edge image:
   - Fill gaps in the edge contours
   - Find the T-junctions and mark them as T-corners
3. Compute the curvature at highest scale σ_{high} and determine the corner candidates by comparing the maxima of curvature to the threshold *t* and the neighbouring minima.
4. Track the corners to the lowest scale to improve localisation.
5. Compare the T-corners to the corners found using the curvature procedure, and remove corners which are very close.

The details of the CSS process can be found in [11].

4.4 Tracking Point Features

The MHT-IMM (MHT coupled with a multiple model Kalman filter, as discussed in [13]) algorithm can be applied for tracking key point features through an image sequence. The measurements for the tracking filter in this case will be the key features extracted (from and near the object of interest) from every frame of a given image sequence (key points are searched within a region of interest surrounding the estimated object centroid). The object centroid position is initially calculated in the first frame by taking the mean of the sum of object key point positions. In the subsequent frames the object centroid is estimated using the MHT-IMM tracker. The extracted measurements are then matched to predictions based on the *Mahalanobis* distance.

The advantage in using the key point tracking algorithm is that we can verify each of the temporally translated key points on the object (selected) against the likely contour of that object in every frame. By doing so, we examine to see whether the object as a whole is tracked correctly (the process for doing this is explained in the next section). In the next section we give the procedure involved in combining the point feature tracking algorithm and the contour grouping algorithm for object tracking.

5 Object Tracking

This section shows how the contour grouping and key point feature tracking procedures combined can be applied for object tracking in image sequences. One cycle of the algorithm recursion is displayed in Fig. (4).

For every frame in an image sequence we first apply the contour segmentation algorithm. This process will group segments of edges that are likely to be from the same object. The result of such a process applied to our test sequences are given in colour Figures 5(a) – 7(a). The procedure as seen from these figures, fail at high curvature contour regions, or is unable to bridge a gap in edgels extracted. As a result, contours from the same object are often broken or separated. To overcome this limitation we applied the contour merging algorithm, which resulted with recognisable object contours (colour Figs. 5b – 7b).

Once the object contours are categorised separately, we can now track a selected object (selection of object can be automated by using a snake type algorithm (eg: Gsnake [9, 10]) or any other suitable algorithm) from the initial frame through the sequence. To track the selected object, we first select some key features (points) from the object (these are selected using the edge map information and then applying the CSS algorithm) as discussed section (4).

The key features of the object are extracted in every frame and the object centroid calculated (this is the mean position of the sum of key points of the object contour). The key points (and the centroid) from the first frame are now tracked through the sequence using the MHT-IMM algorithm (as discussed in [13, 14]). The tracking process is achieved by predicting the object centroid position in the following frame, and then searching a region of interest surrounding the centroid to look for the key points, this process is followed by matching the key points to a grouped object contour within that region. This procedure will provide trajectories for every key point of the object. Each trajectory point is validated against a grouped contour in each frame. By imposing a distance threshold between the tracked key points and the key points on the segmented contour (in each frame), we can verify whether the points have been tracked to an acceptable level of precision. If an acceptable number of key points tracked are identified to lie on or near the object contour (that is passing the threshold test) in each of the frames, we conclude that the object has been tracked successfully. If a key point fails the threshold test, then that point will not be considered as part of that feature trajectory any further.

6 Results

Image Sequence (frames length)	**Number of key points selected**	**Attempted number of points tracked** (& percentage)		**Features tracked for more than 2/3's of the seq. length**	
UMASS (11)	8	8	100 %	8	100 %
PUMA (30)	4	4	100 %	4	100 %
Outcones(20)	12	10	83.3%	10	100 %

Table 1: Object tracking statistics for the 3 test image sequences considered.

The 3 sequences considered give a variety of scenarios to test our algorithm. In all 3 cases the tracking results are promising (see Figures 5 – 7). Table 1 provides quantitative performance values for the object tracker. For the UMASS lab sequence 100% of the key-points selected as forming the object (posters) in the first frame are successfully tracked for the entire sequence length.

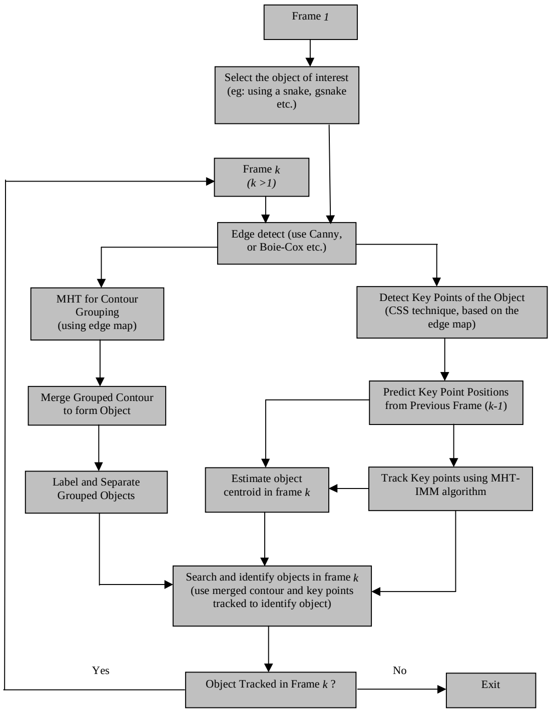

Frame *1*
Select the object of interest (eg: using a snake, gsnake etc.)
Frame *k* (*k* >*1*)
Edge detect (use Canny, or Boie-Cox etc.)
MHT for Contour Grouping (using edge map)
Detect Key Points of the Object (CSS technique, based on the edge map)
Merge Grouped Contour to form Object
Predict Key Point Positions from Previous Frame (*k-1*)
Label and Separate Grouped Objects
Estimate object centroid in frame *k*
Track Key points using MHT-IMM algorithm
Search and identify objects in frame *k* (use merged contour and key points tracked to identify object)
Yes
Object Tracked in Frame *k* ?
No
Exit

Figure 4: Overview (1 cycle) of the object tracker.

Similar observations can be made for the PUMA sequence. Finally, a multiple object example is demonstrated. For the outdoor cone sequence, 4 cones are considered as part of an object. As the result suggests, 10 out of the 12 key corner features are tracked successfully for more than 2/3's of the sequence length.

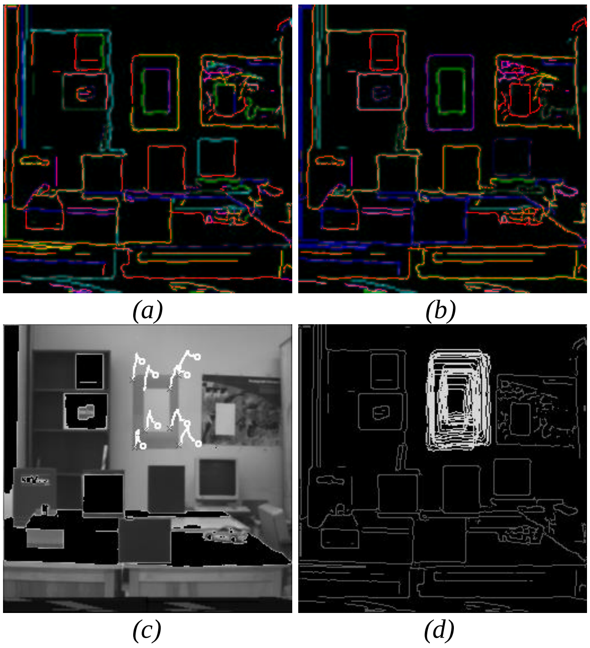

(a) (b)
(c) (d)

Figure 5: UMASS lab sequence result.

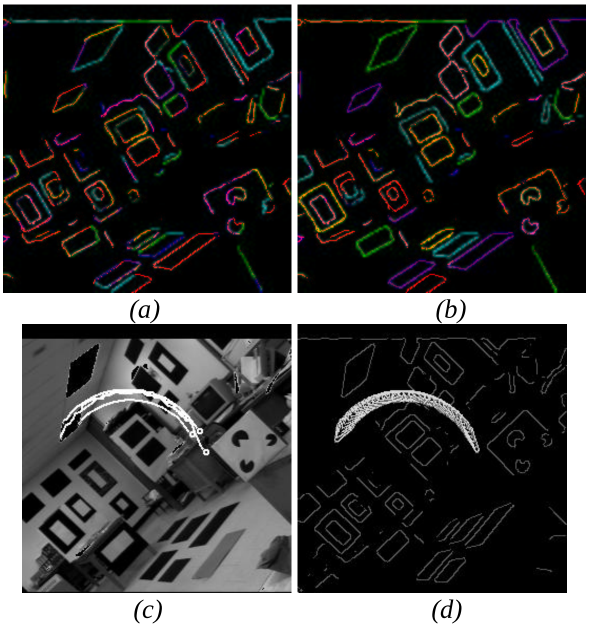

(a) (b)
(c) (d)

Figure 6: PUMA lab sequence result.

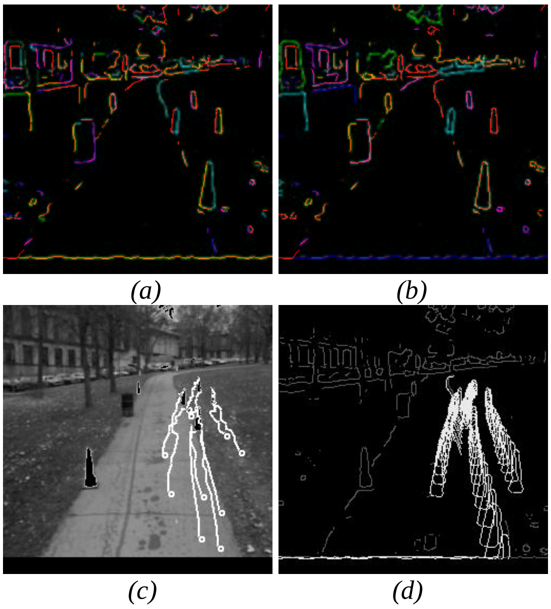

(a) (b)
(c) (d)

Figure 7: Outdoor cone sequence result.

Figures 5-7 (colour figures): (a) Contours grouped by applying the contour segmentation algorithm based on the edge map obtained (one frame of the sequence is displayed). Each segmented contour (grouped edges) is shown with a different colour. (b) Result after the application of segment merging algorithm (observe that the segments that are identified as forming the same object are merged together in most instances). (c) The trajectory of the key points by applying the MHT-IMM algorithm. The 'x' shows the start of the trajectory while the little white circle indicates the end of trajectory. (d) The identified object trajectory. The white contours (identified as belonging to the same object in each frame) are superimposed on the first frame of the sequence to show the motion of the object.

7 Discussion

Colour figure 5(a) – 7(a) shows the result of applying the contour segmentation algorithm. It can be seen that the segmentation algorithm fails to group segments of the same edge around sharp curves. Since the algorithm scans the edge image by "walking" along the contours, it may encounter a new contour at any point along its length. When tracking begins in the interior of a curve, it is usually partitioned, erroneously into two or more segments sharing common boundary points. As a result of this, contours belonging to the same object can be grouped as separate objects. To overcome this limitation we applied the contour-merging algorithm (as described in section 4) which provided better results (colour figures 5(b) - 7(b)). It can be clearly seen that most of the segments belonging to the same object have now been grouped together successfully (the quality of the segmentation also depends on the thresholds that are used for both algorithms [6-8]).

For the PUMA sequence, the window on the top left corner of frame 1 (see Figure 6) was tracked through the sequence. The result of the tracking is given in Figure 6(d) and the corresponding trajectories of the key points are given in Figure 6(c). From visual inspection the results are promising. Similar results are observed for the UMASS lab sequence (Fig. 5(c, d)), despite the short irregular translation of the posters (top right corner of frame 1). The qualitative results are supported by quantitative results presented in Table 1.

Figure (7) shows the result of the outdoor cone sequence. In this case, multiple objects are tracked (4 cones on the right). Each cone is treated as a separate entity, while all 4 cones combine to form a 'grouped-object'. Each of the key points from the 4 cones are tracked and matched to the segmented shape (the 4 cones). Apart from the last frame, where the cone in the front gets segmented with the road, 83% of the key points have been tracked correctly, thus successfully tracking the 4 cones.

8 Conclusion

In this paper we have shown how the multiple hypothesis technique can be used for rigid object tracking in image sequences. The contour of object tracked is achieved by first applying the MHT approach to group segments of the same object. This process is followed by applying the contour merging algorithm to identify recognisable object contours. Then by selecting key point features of this object, temporal tracking (matching) of key points is achieved by using the MHT again. The validity of the trajectory of the key points is verified by inspecting whether the key points were lying on or near the contour of the tracked object (searched within a region of interest). The results are promising for objects that are not occluded and can be recognised clearly in every frame.

One of the main drawbacks of the system is that the contour grouping process can break down due to occlusion of the object being tracked. The MHT can predict possible trajectory for the key points despite the occlusion [7] and thus retain the trajectory (as shown in [13, 14]). But the contour segmentation and grouping process will fail, as it considers only the edge map to group contours. As a result the object tracker fails in its primary purpose. The tracking process presented may also fail for deforming objects. This is because the key point tracking phase will not be robust enough to track unexpected deformation of object contours.

Recognising and tracking objects using point features as presented in this paper is possible for relatively simple objects (as demonstrated in the results). For complex objects the process is inefficient, and can lead to errors in object identification and tracking. A more versatile method of object tracking will require an object contour to be represented using a parameterised curve, such as using Snakes, Deformable templates, or using B-splines (see [3, 9, 10]).

References

[1] Y. Bar-Shalom and X. R. Li, Estimation and Tracking: principles, techniques, and software, *Artech House*, Boston, MA, 1993

[2] Y. Bar-Shalom and Y. Fortmann, Tracking and Data Association, volume 179 of *Mathematics in Science and Engineering*, AP, Boston, 1988

[3] A. Blake and M. Isard, Active contours, *Springer*, 1998

[4] R. A. Boie and I. J. Cox, "Two dimensional optimum edge detection using matched and wiener filters for machine vision", In *ICCV*, pp.450-456, June 1987

[5] J. Canny, "Computational Approach to Edge Detection", *PAMI*, Vol.8(6), pp.679-698, 1986

[6] I. J. Cox, J. M. Rehg, and S. Hingorani, "A Bayesian Multiple Hypothesis Approach to Contour Grouping and Segmentation", *IJCV*, vol. 11(1), pp.5-24, 1993

[7] I. J. Cox and S. L. Hingorani, "An Efficient Implementation of Reid's Multiple Hypothesis Tracking Algorithm and Its Evaluation for the Purpose of Visual Tracking", *PAMI*, vol. 18, no. 2, pp.138-150, Feb. 1996

[8] I. J. Cox, J. M. Rehg, and S. Hingorani, "A Bayesian Multiple Hypothesis Approach to Contour Grouping", *ECCV*, pp.72-77, 1992

[9] M. Kass, A. Witkin and D. Terzopoulos, "Snakes: Active contour models", *IJCV*, pp.321-331, 1988

[10] K. F. Lai and R. T. Chin, "Deformable contours – modeling and extraction", *IEEE Trans. PAMI*, Vol.17(11), pp.1084-1090, 1995

[11] F. Mokhtarian and R. Suomela, "Robust image corner detection through curvature scale space", *PAMI*, Vol.20(12), pp.1376-1381, 1998

[12] D. B. Reid, "An Algorithm for Tracking Multiple Targets", *IEEE Trans. on Automatic Control*, Vol.24, no. 6, pp.843-854, Dec. 1979

[13] P. Tissainayagam and D. Suter, "Visual Tracking with Automatic Motion Model Switching", *International Journal of Pattern Recognition*, Vol(34), pp.641-660, 2001.

[14] P. Tissainayagam and D. Suter, "Tracking multiple object contours with automatic motion model switching", *ICPR-2000*, pp.1146-1149.

COMPUTER VISION AND PATTERN RECOGNITION II

(This page left blank intentionally)

Polytopes, Feasible Regions and Occlusions in the n-view Reconstruction Problem

David McKinnon[1], Barry Jones[2] and Brian C. Lovell[1]
[1]School of Information Techology and Electrical Engineering
Intelligent Real-Time Imaging and Sensing (IRIS) Group
University of Queensland
[2]Department of Mathematics
University of Queensland

Abstract

This paper asseses the question, given a arbitrary point in $\mathbb{P}^3$, can it be reconstructed by a given camera orbit? *We show that a solution to this problem can be found by intersecting the view frustums of the cameras in the sequence creating a polyhedron that bounds the area in $\mathbb{P}^3$ observed by all cameras. For a projective set of cameras this can be considered as an expansion of the cheiral inequalities. We also show an exception to this basic principle is encountered when the point in $\mathbb{P}^3$ is occluded. Thus giving a* weak *condition for occlusion of an arbitrary point in $\mathbb{P}^3$.*

1 Introduction

This work addresses the motivation to find the regions in $\mathbb{P}^3$ observed by two or more cameras in a given sequence. We will refer to this common region as the *feasible region* for the given set of cameras, it represents the part of space where depth fusion may occur through triangulation. This is an expansion of the basic theory of *cheirality* [3], where in this case the inequalities are expanded to constrain a point to lie with the view frustum of each camera of interest. This work is also partly inspired be the treatment of polyhedral models in the computer graphics literature [5]. The following subsections introduce the basic concepts and notation maintained throughout the rest of the paper.

1.1 Set Notation

The feasible region for n views $(\alpha_1, \ldots, \alpha_n)$ will be denoted as the set $\mathcal{S}^{\alpha_1 \cdots \alpha_n}$ of points (where $\mathcal{S}^{\alpha_1 \cdots \alpha_n} \in \mathbb{P}^3$) that project (un/occluded) into each of the n-images. The problem of finding the feasible region for a sequence of n-images has strong analogues with the set of the relevant subproblems of finding the intersection of 2 or more views $\alpha_a, \ldots, \alpha_b \in \{\alpha_1, \ldots, \alpha_n\}$. Where $\alpha_a, \ldots, \alpha_b$ depicts the choice of $2 \leq k \leq n$ unique views from the set of n, there are $\binom{n}{k}$ such choices. The notation for the $\binom{n}{k}$ subsets of $\mathcal{S}^{\alpha_1 \cdots \alpha_n}$ is given as $\mathcal{S}^{\alpha_1 \cdots \alpha_n}_{\alpha_a, \ldots, \alpha_b}$. Feasible regions and their $\binom{n}{k}$ *subregions* have the inclusion relation $\mathcal{S}^{\alpha_1 \cdots \alpha_n} \subset \mathcal{S}^{\alpha_1 \cdots \alpha_n}_{\alpha_a, \ldots, \alpha_b}$.

We can now answer our initial questions with the set notation. Firstly, which set of points in $\mathbb{P}^3$ can be reconstructed by the given camera orbit?

$$S^n_{\cup} \equiv \bigcup_{\substack{perm_k(\alpha_a, \ldots, \alpha_b) \\ 2 \leq k \leq n}} \mathcal{S}^{\alpha_1 \cdots \alpha_n}_{\alpha_a, \ldots, \alpha_b} \tag{1}$$

Secondly, which set of points in $\mathbb{P}^3$ can be reconstructed using *all* the given views from the camera orbit?

$$S^n_{\cap} \equiv \bigcap_{\substack{perm_k(\alpha_a, \ldots, \alpha_b) \\ 2 \leq k \leq n}} \mathcal{S}^{\alpha_1 \cdots \alpha_n}_{\alpha_a, \ldots, \alpha_b} \tag{2}$$

Section 2 will explore the inequalities that bound these sets and section 3 will present a rudimentary algorithm to compute the bounds for an arbitrary camera orbit in projective space. Section 4 discusses some applications of the *feasible regions*.

1.2 Projective Geometry of Linear Features in $\mathbb{P}^2$ and $\mathbb{P}^3$

This section will outline the notation and the basic building blocks for the representation of linear features in $\mathbb{P}^2$ and $\mathbb{P}^3$. The development of the ideas in this section is heavily influenced by the notation and stucture of the linear features presented in [6], also drawing some similarities to the oriented matching constraints in [7].

The first consideration when dealing with the multi-view geometry of linear features is their notation. Consistantly we will refer to features as any type of geometric object observed in a scene, be this points, lines or planes.

Table 1 summarises the notation and degrees of freedom (DOF) for the group of linear features in the projective plane ($[A, B, C] \in \mathbb{P}^2$).

Hyperplane	$\mathbb{P}^2$	$\mathbb{P}^{2*}$	DOF
Points	x^A	$x^{[A]} = \epsilon_{ABC}x^A = x_{[BC]}$	2
Lines	$x^{[AB]}$	$x^{[AB]} = \epsilon_{ABC}x^{AB} = x_C$	1

Table 1. Linear features and there duals in $\mathbb{P}^2$

Similarly, Table 2 summarises the notation and the DOF for linear features in projective space ($[a, b, c, d] \in \mathbb{P}^3$).

Hyperplane	$\mathbb{P}^3$	$\mathbb{P}^{3*}$	DOF
Points	x^a	$x^{[a]} = \epsilon_{abcd}x^a = x_{[bcd]}$	3
Lines	$x^{[ab]}$	$x^{[ab]} = \epsilon_{abcd}x^{ab} = x_{[cd]}$	2
Planes	$x^{[abc]}$	$x^{[abc]} = \epsilon_{abcd}x^{abc} = x_d$	1

Table 2. Linear features and there duals in $\mathbb{P}^3$

These tables demonstrate the process of dualization for linear feature types via the antisymmetrization operator $[\ldots]$. The antisymmetrization operator should be considered as a determinantal method to generate the algebra for linear features, by performing an alternating tensor contraction over the space/s to which the operator is applied [**?**, **?**]. The tensor notation will also convert directly into a computational scheme to evaluate the geometric entity in question.

An important aspect of the tensor notation, is the ease at which linear features can be manipulated to form geometrically intuitive combinations of each other. The tables given above demonstrate this concept clearly, where lines are constructed from two points and planes (in $\mathbb{P}^3$) are constructed from 3 points or a line and a point. This facet of the tensor notation (Grassmann-Cayley algebra) greatly increases it efficacy in solving for complex geometric configurations.

2 Feasible Regions

In this section we will introduce the basic camera model (pinhole camera) and describe the process of projecting it's boundaries to form the view frustum. We will then discuss the general intersection problem for 2 or more view frustum and give bounds for the number of differentally oriented solutions (cheirality), of which we will only be concerned with the positively oriented solutions.

We will round-off this section by discussing the *weak* condition for occlusion of an arbitrary point in $\mathbb{P}^3$. We call this condition weak because it not geometrically based, rather it is based on a lack of the observation of the point in the image, when the feasible region suggests that the point should be observable within the view frustums.

2.1 Camera Model and View Frustum

The pinhole camera model is the standard for modelling perspective distortion in image sequence. Perspective distortion is most prolific when a small focal length thus a greater Field of View (FOV), is employeed to model a scene that is both close to the camera and with a high depth variation.

Firslty, we must consider the projection operator (P^{α}_{β}) or *camera matrix* that denotes the projection of linear features from the scene to the image plane ($P^{\alpha}_{\beta} : \mathbb{P}^3 \rightarrow \mathbb{P}^3$). Table 3 summarises the range of projection operators for linear features.

Hyperplane	$\mathbb{P}^3$	$\mathbb{P}^{3*}$
Point	$x^A \sim P^A_a x^a$	-
Line	$x^{[AB]} \sim P^{[A}_{[a} P^{B]}_{b]} x^{[ab]}$	$x_{[BC]} \sim P^{[c}_{[B} P^{d]}_{C]} x_{[cd]}$
Plane	-	$x_A \sim P^d_A x_d$

Table 3. Projection operators for linear features

Generally it may be stated that $\lambda x^{\alpha} = P^{\alpha}_{\beta} x^{\beta}$, where λ is an arbitrary scale factor. The form of P^{α}_{β} for a regular point-to-point projection is,

$$P^{\alpha}_{\beta} \equiv [R^c_w | - R^c_w t^w] \qquad (3)$$

where R^c_w is rotation from the center of the world reference frame to the orientation of the camera and t^w is the location of the camera in the world reference frame. This depiction of the camera (3), is defined in a Euclidean metric and also assumes calibration of the internal parameters. The more general projective uncalibrated model differs according to an affine ($K^A_{\hat{A}}$) and projective collineation ($H^{\hat{a}}_a$) [1]. Thus, the calibrated model for point projection is,

$$P^A_a \simeq K^A_{\hat{A}} P^{\hat{A}}_{\hat{a}} H^{\hat{a}}_a \qquad (4)$$

where $K^A_{\hat{A}}$ represents the internal parameters of the camera in an affine collineation,

$$K^A_{\hat{A}} = \begin{pmatrix} \alpha_A & s & A_0 \\ 0 & \alpha_B & B_0 \\ 0 & 0 & 1 \end{pmatrix} \qquad (5)$$

$\alpha_{A/B}$ encorporate the focal length f into scale factors for conversion between metric distances (mm) and pixels, s is the skew of the CCD imaging array and A_0 and B_0 are the coordinates of the cameras principal point [4].

From this point, no assumptions will be placed upon whether the cameras are un/calibrated. Instead, we will implicitly assume that problems dealing with calibrated cameras will present image coordinates as they would be read from the image plane ($A \in [0, \mathrm{cols}]$ and $B \in [0, \mathrm{rows}]$) and uncalibrated cameras present normalised image coordinates in a square centered fashion ($\hat{A} \in [-1, 1]$ and $\hat{B} \in [-1, 1]$) [2]. The ranges given for pixel coordinates typify the *bounds* for the camera. The projective (uncalibrated) assumption removes the ability to measure metric distances but can be made to preserve the orientation of an point-plane pair which we will see is the only requirement needed to build a proper view frustum.

The view frustum for the camera is found by an inverse projection of the bounds of the camera into the scene. That is, a projection of the lines that bound the image plane, into their corresponding planes in the scene. Figure 1 demonstrates the labelling of the vertices and edges for the standard camera.

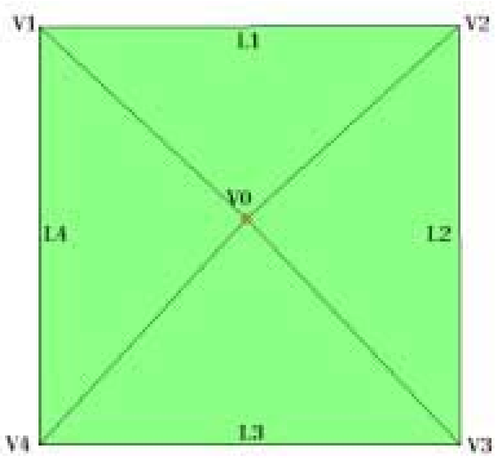


Figure 1. The labelling of the boundaries of projection for the standard view frustum.

Moving clockwise from the top left of the camera, the vertices V_1, V_2, V_3 and V_4 represent the corners of the image plane and V_0 is assigned to the cameras optical center (for finite cameras this is found as the nullvector of the camera matrix P^A_a or $\epsilon_{ABC} P^A_b P^B_c P^C_d \epsilon^{abcd}$). The polyhedron formed by the planes passing through V_0 and the joins of $V_1 \wedge V_2 = L_1, V_2 \wedge V_3 = L_2, V_3 \wedge V_4 = L_3$ and $V_4 \wedge V_1 = L_4$, gives a set of boundary conditions for a point in $\mathbb{P}^3$, where $L_\gamma = L^{[a_\alpha b_\beta]} = V^{[a_\alpha} V^{b_\beta]}$ are the lines (edges) joining the vertices of the image plane. A further set of inequalities should also impose the image plane (although this is a largely redundant constraint) as another boundary and the cheiral inequalities ($*$'s). These boundary conditions are expressed as a set of algebraic inequalities for a given point $x^d \in \mathbb{P}^3$, camera center V^{d_0} and plane-at-infinity $\pi^{[abc]}$.

$$\begin{array}{crcl} & \lambda_1 L^{[a_1 b_2} V^{c_0} x^{d]} & > & 0 \\ & \lambda_2 L^{[a_2 b_3} V^{c_0} x^{d]} & > & 0 \\ & \lambda_3 L^{[a_3 b_4} V^{c_0} x^{d]} & > & 0 \\ & \lambda_4 L^{[a_4 b_1} V^{c_0} x^{d]} & > & 0 \\ & \lambda_5 V^{[a_1} V^{b_2} V^{c_3} x^{d]} & > & 0 \\ * & \lambda_6 \pi^{[abc} x^{d]} & > & 0 \\ * & \lambda_7 \pi^{[abc} V^{d_0]} & > & 0 \end{array} \tag{6}$$

The scalars (λ_α) ensure the cheirality (orientation) of each plane with respect to the view frustum. This is achieved by making sure each point in the view frustum has a positive projective distance from the camera (only the sign is important). The last two inequalities ($*$'s) in (6) can be omitted from the boundary conditions if the scene points and cameras are know to be correctly oriented. Figure 2 shows some randomly generated positively oriented view frustums that have been clipped for display purposes. Note that not all of the cameras in this configuration are positioned to view the same part of space.

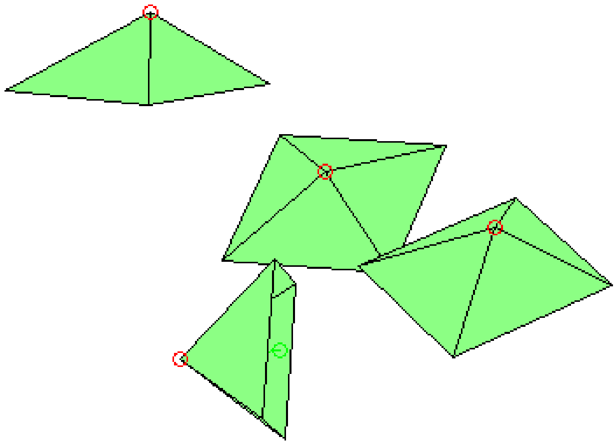


Figure 2. Four randomnly generated positively oriented view frustums for a square camera with a focal length of 1.

2.2 Feasible Regions and Cheiral Intersections

This section will move the discussion onto the case of multiple (n) cameras. Running out of meaningful indices, our first step will be to rewrite (6) in a symoblic vector notation, feeling comfortable that we have an effective means to compute the various vectors in the inequalities.

$$\begin{array}{rrcl} & \lambda_1 x^\top P_1^n & > & 0 \\ & \lambda_2 x^\top P_2^n & > & 0 \\ & \lambda_3 x^\top P_3^n & > & 0 \\ & \lambda_4 x^\top P_4^n & > & 0 \\ & \lambda_5 x^\top P_0^n & > & 0 \\ * & \lambda_6 x^\top \pi & > & 0 \\ * & \lambda_7 V_0^{n\top} \pi & > & 0 \end{array} \tag{7}$$

The superscript on the vectors denotes the image number and the subscript reflects the labelling (Figure 1) of the planes (P_β^α), points (V_β^α) and plane-at-infinity (π) where P_0^n is the n^{th} image plane. The inequalities given in (7) can now be considered to represent a valid set of points in the feasible region for the n-view reconstruction problem ($S_\cap^n$) and the entire set of points that are *reconstructable* ($S_\cup^n$) can be found as the union of the $\binom{n}{k}$ subproblems (where $2 \leq k \leq$ n).

This definition of the feasible regions for the n-view reconstruction problem is tractable, however we are compelled to find a more definitive statement for the regions that captures their geometric structure more precisely allowing visualisation, volume calculations and a more compact representation. This is where we introduce the language of polytopes.

Polytopes [8] allow us to represent the feasible regions as a convex hull of vertices (points), edges (lines) and facets (planes) in $\mathbb{P}^3$, thus providing a bounding polyhedra for the various linear programming problems implied by the application of the inequalities in (7). Since all planar intersections occur either before the plane-at-infinity or at the plane-at-infinity (for parallel planes) we will strip the last two inequalities ($*$'s) along with the 5^{th} (to cut down computation) from (7) leaving just the boundary conditions for the outer planes in the view frustums.

The bounding polyhedra will be constructed from the intersection of pairs of the four remaining planes from each camera (ie. lines $V_0^n \wedge V_\beta^n$ for camera n there is $\binom{4}{2}$ such lines for each camera) with the set of four remaining planes in each other camera (P_β^α where $\alpha \neq n$). In the general case there are two different ways that a line can intersect a plane. Firstly, if the points on the line and the plane are coplanar then their intersection will be the original line, or if the line and the plane are not coplanar then their intersection will result in a point. The points resulting from the line-plane intersections will form the vertices of the bounding convex hull representing the feasible region for a given set of cameras.

In summary we can consider the upper and lower bounds for the number of line-plane intersections for n-views (η_n) to be,

$$0 \leq \eta_n \leq 4 \times \binom{4}{2} \times \frac{n!}{(n-2)!} \tag{8}$$

the lower bound is the case of n-views with the same internal camera parameters all lieing in the same position with the same orientation, this is clearly a *critical configuration* where no depth can be recovered. The upper bound is formed by the general case given by n-views of a regular pinhole camera where the set of four lines from each camera must intersect each plane from each other camera. The intersection points will be positively and negatively oriented (ie. lie in front of the cameras and behind).

Of the collection of intersection points formed in the planes, only a certain number will be valid (feasible) points. Valid points are those that conform to the inequalities given in (7). This is to say that we are only interested in positively oriented intersection points that lie on the boundary of the n-view feasible region. An algorithm to find the intersection points that construct the bounding polyhedra will be considered in section 3.

2.3 A Weak Condition for Occlusions

The formulation for the feasible region given above will determine if a point lies within the intersection of the view frustums for the camera involved. However, this provides no indication as to whether the point of interest is occluded by some other object in the scene.

This is brings us to consider a statement about whether a point that lies in a valid feasible region for a set of cameras is *observable* or not. Firstly, a *strong* statement of observability requires a precise model of the scene in question and typical ray casting process (analogous to a image-to-world graphics engine) can resolve geometrically whether there exists a clear line of site from the point in the scene to the image.

Unfortunately, this strong statement of observability assumes that a complete 3D model has already been obtained. This is prohibitive for the purpose of building a 3D model from images, see we must seek another statement without this prerequisite.

The *weak* statement of observability requires that the assumed texture surrounding the point in the scene is observable in the expected location in image. This statement is clearly less precise, but practically quite an acceptable means to determined whether a point has been occluded. With the discussion of the feasible regions above and this statement of observability a robust method to locate the images where a feature is present ensues.

3 An Algorithm to Compute the n-view Intersection Polytope

We present a rudimentary algorithm to find the bounding polytope containing the set of points $S^n_\cap$ for n-views. The algorithm is geometrically intuitive, but far from optimal for task. For an optimal approach refer to [**?**].

The algorithm will follow along the lines of the discussion given in the previous section for isolating the line-plane intersection points. Then, application of the first 4 inequalities in (7) will reveal the superfulous intersection points which are not feasible. Of the remaining points which will now lie on the convex hull of the feasible region's polyhedra, edges must be selected to intersect points that lie only on a common plane of one of the view frustums. The proposed algorithm is given in Table (4).

	Algorithm for the polyhedra bounding $S^n_\cap$
1.	Find the η_n intersections, where $k = 2, \dots, n$
1.1	Find the vertices $\implies \epsilon_{abcd}x^{a_\gamma} = w_{[b_\alpha}y_{c_\beta}z_{d_\kappa]}$.
2.	Store the vertices that fulfill the first 4 inequalities (6).
3.	Find the edges from x^{a_γ} and the joins of x^{a_γ} in each plane.

Table 4. Proposed algorithm for calculating the bounding polyhedra for $S^n_\cap$

4 Applications

This section discusses some of the perceived applications of the concept of feasible regions in computer vision. It should be noted that this concept is very new and the authors are still considering the scope and application of the theory.

4.1 Object Querying

The motivation here is to be able to query a sequence of images of a static scene (and the corresponding 3D reconstruction of the scene) to see if a feature lies within the view frustum of a given camera or a given set of cameras.

Principally, we feel that this process may enhance the regular Structure From Motion (SFM) pipeline [**?**] with the ability to relocate lost or occluded tracks based on an expectancy for when they will return into the camera's field of view. Another related application would be to retrospectively update the tracking profile of a feature found in image α to all the previous images where the feature was observable.

This process could be used to great effect in the SFM pipeline for long sequences, where the camera completes a loop (ie. viewing the same area of the scene multiple times with some seperation in the exposure) which we will refer to as sub/sequence closure. In this case a very valuable alignment of the same set of features can be achieved to correct any drift in the egomotion estimation up until that point. This philosophy has been used with great success in closed turntable sequences.

4.2 Path Planning

Another perceived application of feasible regions is an autonomous robot fitted with either one or multiple cameras. If the robot is equipped with the mechanism to recognise certain objects in the scene (for our purpose we will imagine that these are polytope models), then the feasible region will allow the robots to plan a path around the object whereby the entire object of interest remains within the feasible regions of the camera at each time instant.

Complex polyhedral objects can be represented as the convex hull of vertices encompasing the poylhedra, if each of the vertices is visible un/occluded in each view of the camera/s views then the robot has maintained a suitable perspective on the object.

A robot fitted with cameras could then be instructed to model an entire area (at its own pace) making sure that it has gathered an adequate amount of information to fulfill the accuracy requirements of the mission.

4.3 Building Voxel Volumes

The final application we have considered is building an arbitrary indexed voxel volume given a sequence of camera motions. Currently, the authors are yet see an application of voxel volume modelling where the voxel space has not been predesignated to house the object of interest. We feel that if the voxel volume 3D reconstruction technique is to reach its full practical value, the voxel spaces must be defined purely by the path of the camera, not placed at a point in space as a priori to reconstruction.

Furhermore, information about the number of views that a feature is observable in and whether there are occlusions of the feature for a given path of the camera, is invaluable for the construction of the metric space and the subsequent optimisation problem for the building of the voxel volume.

5 Conclusions and Future Work

We have presented a novel idea that sets the boundary for points in space that are reconstructable by an arbitrary camera path. We have considered which points are reconstructable in the general case and which points are observable in every image. We have also shown that this formulation leads to a weak condition for occlusions.

As a matter of future work the authors would like to consider the quality of observation for a point and the subsequent quality of reconstruction for a point given multiple observations. The envisaged formulation is to consider the expected error in the observation of a feature in the image and the projection of this error into space forming a cone. The intersection of multiple error cones gives something akin to a set of geodesic probability surfaces, representing the likelihood of the corresponding 3D feature lieing in this part of space.

The introduction of such a concept would lead to a sophisticated method to determine the overall quality of reconstruction for the scene given a specific camera orbit. Then, points that are positioned poorly in the scene for reconstruction purposes can treated with a greater caution.

In this manor the quality of reconstruction can be set as an input parameter to the SFM pipeline and made to replicate the average depth fusion for human vision. This can be considered as the ideal quality of reconstruction for visualisation with VR glasses.

The authors are also looking into optimal algorithms to intersect the view frustums and create the resulting polyhedron.

6 Acknowledgements

The authors would like to thank Carlos Leung and Ben Appleton for many interesting discussions on the development of the ideas presented in this paper. Also we would like to thank Bob Andrews for his assistance in editing this paper.

References

[1] R. I. Hartley. Euclidean reconstruction from uncalibrated views. *Applications of Invariance in Computer Vision - LNCS*, 825:237–256, 1994.

[2] R. I. Hartley. In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):580–593, 1997.

[3] R. I. Hartley. Chirality. *Int. Journal of Computer Vision*, 26(1):41–61, 1998.

[4] R. I. Hartley and A. Zisserman. Multiple view geometry in computer vision. Cambridge University Press, 2000.

[5] S. Savchenko. *3D Graphics Programming : Games and Beyond*. SAMS Publishing, 2000.

[6] W. Triggs. The geometry of projective reconstruction i: Matching constraints and the joint image. *Int. Conf. on Computer Vision*, pages 338–343, 1995.

[7] T. Werner and T. Pajdla. Oriented matching constraints.

[8] G. M. Ziegler. *Lectures on Polytopes*. Graduate Texts in Mathematics, Springer-Verlag, 1994.

Improved Ensemble Training for Hidden Markov Models using Random Relative Node Permutations

Richard I. A. Davis and Brian C. Lovell
Intelligent Real-Time Imaging and Sensing Group,
School of Information Technology and Electrical Engineering,
The University of Queensland, Australia, 4072
{riadavis, lovell}@itee.uq.edu.au

Abstract

Hidden Markov Models have many applications in signal processing and pattern recognition, but their convergence-based training algorithms are known to suffer from over-sensitivity to the initial random model choice. This paper focuses upon the use of model averaging, ensemble thresholding, and random relative model permutations for improving average model performance. A method is described which trains by searching for the best relative permutation set for ensemble averaging. This uses the fit to the training set as an indicator. The work provides a simpler alternative to previous permutation-based ensemble averaging methods.

1 Introduction

The work of Davis and Lovell [1] focused upon Hidden Markov Model (HMM) ensemble learning using the well-known Baum-Welch procedure [2, 5, 6] for individual sequences and then averaging the resulting HMM ensemble parameters. It was demonstrated that this is a superior method to the standard method of converging a single model using the multiple sequences simultaneously [4, 2] in performing approximations to sequence distributions, and also for performing classification tasks. Thresholded Winsorization in which the best sub-ensemble is used was also shown to provide further improvements.

This paper investigates the potential for still further improvements on these methods based on searches through random relative permutations of the nodes of the ensemble of models, prior to ensemble parameter averaging. The methods investigated were based on varying the random permutation probability and the number of node transpositions applied to each member of the ensemble.

By relabelling states, it is easy to show that many different models can achieve the same probability despite large differences in configuration. The possibility of finding equivalently good models with large structural differences argues against the parameter averaging method to obtain improved models. In other words, it is quite possible that the average of two good models may be a very poor model in the same way that that the point midway between two mountain peaks is quite often a valley. This was the motivation for the method of searching through relative random node permutations.

Levinson, et al. (see [3], appendix B) first proposed that an approximation to bipartite graph matching be used for permutating an ensemble of trained models to achieve a good permutatation match, and good parameter estimates through averaging the permutation-aligned ensemble. This paper demonstrates that simpler randomised methods can be used with good results for node matching when the best method is selected according to the overall fit to the training data set.

2 Permutations of representations of Hidden Markov Models

The focus of this paper is on methods for permuting the alignment of models in an ensemble of models with the same structure.

A hidden Markov model ([2] chapter 6) consists of a set of n nodes or states, each of which is associated with a set of m possible observations (the structure of the model). The parameters of the model include an initial state π which describes the distribution over the initial node set, a transition matrix a_{ij} for the transition probability from node i to node j conditional on node i, and an observation matrix $b_i(O_h)$ for the probability of observing symbol O_h given that the system is in state i. Rabiner uses $\lambda = (A, B, \pi)$ to denote the model parameters.

This paper is based around the method of ensemble aver-

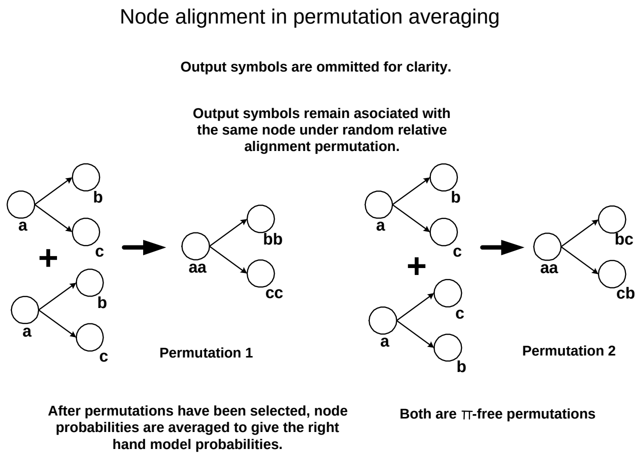

Node alignment in permutation averaging
Output symbols are ommitted for clarity.
Output symbols remain asociated with the same node under random relative alignment permutation.
a
b
c
+
a
b
c
aa
bb
cc
Permutation 1
a
b
c
+
a
c
b
aa
bc
cb
Permutation 2
After permutations have been selected, node probabilities are averaged to give the right hand model probabilities.
Both are π-free permutations

Figure 1. Illustration of the permutation averaging method for two different permutations.

aging suggested by Levinson et al. [3] and tested in [1]. The method is as follows. Consider an ensemble of N HMMs $\{\lambda_k = (A, B, \pi)_k; k = 1 \ldots N\}$ which are each trained using the Baum-Welch method to a specific observation sequence in a set of training sequences. Select an alignment of nodes across all N models, and then form a single model λ by averaging the matrix elements of the ensemble, using the selected node alignment (see figure 1).

In this paper we focus upon permutations which only affect those nodes with zero entries in the starting matrix π. These permutations will be termed *π-free permutations.*

Theorem. Random π-free permutations $S = \{S_k\}$ of HMM ensembles $\Lambda = \{\lambda_k\}$ trained using multiple Baum-Welch convergence have the same parameter-averaged model performance as unpermuted ensembles.

Proof. For every model λ_k, partition the set N of nodes into nodes Q with non-zero π values and a subset R with zero π-values. π-free permutations only act on R. Since initialisation of the BW procedure randomly selects π-free nodes, then the Baum-Welch process produces models for which the nodes in R are randomly permuted. Therefore, a permutation of these nodes after convergence will not change the result of the method. ■

This result means that the insertion of relative random permutations at any stage of the process does not affect the model performance. This paper concentrates on the potential for improvement in the model performance when a good permutation (as measured by its P_{all} score on the training set) rather than a random permutation is applied.

In cases where a permutation might involve a node with a very small starting probability, it is not possible to use the above proof. Permutations involving Q as well as R are less easy to analyze and will not be tackled here. However many models of interest have a well-defined, small starting node set.

3 Classes of permutations in HMM ensemble averaging

All methods investigated involve the use of the *joint emission probability P_{all}* to quantify the model quality, and also to select the permutation which maximises this function. P_{all} is defined as the product (over all sequences in an ensemble) of the probability that the model generated those sequences individually. This is used both in the training ensemble and in the test ensemble for final evaluation of the method.

The algorithms are all based upon the following set of main design components:

- Winsorization threshold level search based on P_{all}.
- Permutation fraction: the fraction of models in the ensemble being permuted which actually receive a permutation prior to averaging
- Number of transposition: the number of nodes being transposed in a given permutation in each model prior to averaging
- Type of permutation: excluding certain types of permutations from the set being considered, such as permutations involving the starting nodes in π or permutations which would change the transition structure of the resulting average model (from left-right to ergodic, for example)

There are many different ways in which these features may be combined, so we restrict ourselves to those which seem to represent the most important aspects of the ensemble permutation idea.

The following different permutation methods were evaluated:

- **VariableProbPerm** - this method scans through a probability scale from 0 to 1 representing the probability that a given model will be permuted in the ensemble.
- **NumTrans** - this method scans through the number of random transpositions applied to each model. This is limited by the size of the model, as applying too many relative interchange operations per model in the search has no extra benefit.

The ensemble used in both cases is a Windsorised ensemble - with the fraction of models used being determined before these trials were run. Code for training HMMs using these methods is available [7].

4 VariableProbPerm trial

Davis and Lovell [1] gave an empirical study of the following idea (initially suggested by Levinson et al. [3]): In HMM ensemble averaging, because the individual Baum-Welch convergence runs are all initialized to a set of random model parameters (random seeds) then applying another set of random permutations to the models, either before or after training to the observation sequence, will have no effect on the final model formed using parameter averaging.

The next step is to investigate the effect of other, more refined random permutation schemes. A similar strategy to the Winsorization approach will be studied in which a small set of relative permutation sets is compared in terms of the performance of the ensemble-permuted average model on the training data, as measured by P_{all}.

Methodology. In this trial of the VariableProbPerm method, a set of 20 test and 20 training sequences was generated from an initial generating model. The models were randomly generated left-right models with 5 nodes and 4 observation symbols. P_{all} values for the training data were used to compare permutation sets. The best permutation averaged model from those considered was selected. This selected model was then evaluated on the test data. The trial was repeated for 100 initial generating models. Permutations were ensured to be $\pi - free$ as they did not involve the left-right model.

Results. The performance of the variable probability of the small permutation method VariableProbPerm presented in the previous section was investigated and the results are displayed in figure 2. It shows the performance on 20 test sequences of the average of ensembles in which each member of the ensemble is permuted. The method shows steady improvement up to 90% permuted. Clearly there are significant improvements to be obtained by varying the probability of permutation of models in the ensemble and selecting the best fraction.

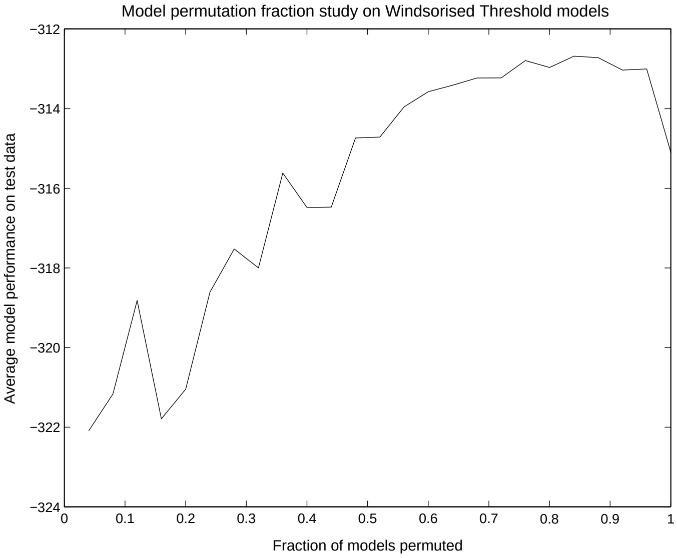


Figure 2. Small permutations are applied to a variable fraction of models in the ensemble.

5 Trial of NumTrans

This trial of the NumTrans method was designed to investigate the best scale of permutations. The major issues of interest were:

1. Is it worth probing the entire range of relative permutations, and
2. Are there any gains to be made by looking at large numbers of transpositions?

Methodology. In this trial of the NumTrans method, a set of 20 test and 20 training sequences was generated from an initial generating model. P_{all} values for training data were used to compare permutation sets. Permutations of the ensemble were constructed using a VariableProbPerm method with a 50% probability of permuting a given model. The parameter under investigation was the number of transpositions applied per model, when that model was selected for permutation. The number of transpositions ranged from 1 to 15. It was anticipated that large numbers of transpositions would not show any benefit over small numbers because of the existence of permutation inverses (so a large number of interchange operations is equivalent to its inverse operation, which can consist of a small number of interchange operations).

The best permutation averaged model from those considered was selected. This selected model was then evaluated on the test data. The trial was repeated for 100 initial generating models.

Results. The performance of the variable probability of the small permutation method NumTrans presented in the previous section was investigated and the results are displayed in figure 2. It shows the performance on 20 test sequences of the average of ensembles in which each member of the ensemble is permuted. As can be seen from the results of figure 3 transposition searches with more transpositions can be superior to smaller searches. Importantly, the best performance was found when the number of random transpositions was set to the number of states in the model. There was not a very clear trend in this instance. However we can deduce that applying large numbers of relative transpositions is not particularly helpful. This is due to the following:

1. the models already have highly random relative per-

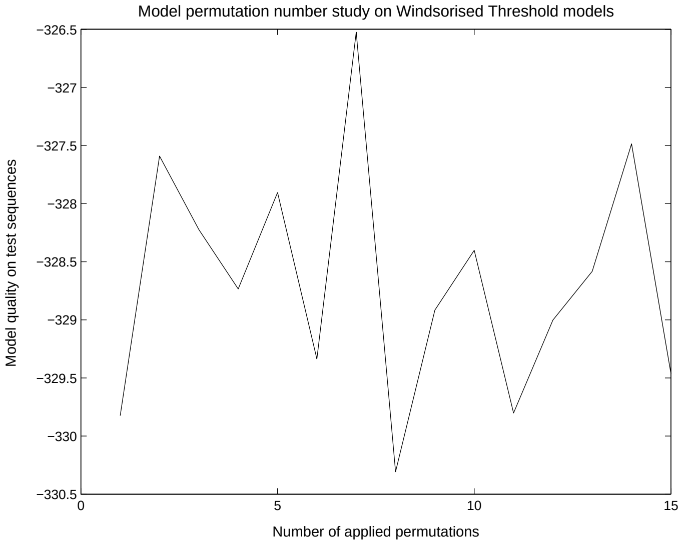


Figure 3. A variable number of permutations is applied to each member of the ensemble.

mutations due to the random initial seed choice

2. it is counterproductive to apply large numbers of transpositions to all models as it is the relative permutation which matters (so smaller numbers will give superior performance)

3. As the number of transpositions approaches the size of the model, performance will be similar to that in the case of small numbers of transpositions.

That said, searches using more permutations do have a slightly superior search ability. However, unless a high level of computing power is available, it seems best to only use permutations of one or two transpositions, in combination with the VariableProbPerm search method.

6 Conclusions

We have demonstrated that substantial gains in averaging can be made using either of two very simple random permutation alignment schemes, VariableProbPerm and NumTrans. These are simpler alternatives to the more complex scheme proposed by Levinson et al. in [3].

There are clear benefits to be found by applying random permutations of the ensemble, using both methods. A search is necessary to locate the best permutation size. In general, the gains obtainable by varying the probability of permutation are greater than the gains obtainable by varying the number of permutations.

The use of the joint emission probability P_{all} for the training set is a useful indicator in selecting the best permutations.

It may be possible to construct more advanced schemes based upon the findings of this paper. For example, a gradient-descent method in which good permutations are retained as the process continues may be a faster way of locating the best ensemble permutation set.

A modification of this technique to include permutations involving non-zero π vector elements may be possible, but is likely to be more complex. From the success of this method however, it seems a promising avenue for further investigation.

7 Acknowledgements

Thanks must go to Ben Appleton and Peter Kootsookos for helpful discussions.

References

[1] R. I. A. Davis and B. C. Lovell, “Improved Estimation of Hidden Markov Model Parameters from Multiple Observation Sequences”, *International Congress on Pattern Recognition*, Quebec City (2002)

[2] L. R. Rabiner, and B. H. Juang, *Fundamentals of Speech Recognition* New Jersey Prentice Hall, 1993.

[3] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, *An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition*, The Bell System Technical Journal 1035-1074, Vol. 62, No. 4, April 1983.

[4] Xiaolin Li, Marc Parizeau, Réjean Plamondon, “Training Hidden Markov Models with Multiple Observations - A Combinatorial Method”. *IEEE Transactions on PAMI*, vol. PAMI-22, no. 4, pp 371-377, April 2000.

[5] D. J. C. Mackay, “Ensemble Learning for Hidden Markov Models”, *Technical report*, Cavendish Laboratory, University of Cambridge, 1997.

[6] A. Stolcke and S. Omohundro. “Hidden Markov Model induction by Bayesian model merging.” In *NIPS 5*, pages 11-18. 1993.

[7] C Walder, R.I.A. Davis, ”IRIS source for estimating Hidden Markov Models”.

Available at:

http://www.itee.uq.edu.au/ iris/CVsource/source.html

3D Reconstruction through Segmentation of Multi-View Image Sequences

Carlos Leung and Brian C. Lovell
Intelligent Real-Time Imaging and Sensing Group
School of Information Technology and Electrical Engineering
The University of Queensland, Brisbane, Queensland, 4072, Australia

Abstract

We propose what we believe is a new approach to 3D reconstruction through the design of a 3D voxel volume, such that all the image information and camera geometry are embedded into one feature space. By customising the volume to be suitable for segmentation, the key idea that we propose is the recovery of a 3D scene through the use of globally optimal geodesic active contours. We also present an extension to this idea by proposing the novel design of a 4D voxel volume to analyse the stereo motion problem in multi-view image sequences.

1. Introduction

The reconstruction of a dynamic, complex 3D scene from multiple images is a fundamental problem in the field of computer vision. While numerous studies have been conducted on various aspects of this general problem, such as the recovery of the epipolar geometry between two stereo images [10], the calibration of multiple camera views [30], stereo reconstruction by solving the correspondence problem [24], the modelling of occlusions [9], and the fusion of stereo and motion [14], more work needs to be done to produce a unified framework to solve the general reconstruction problem.

Given a set of images of a 3D scene, in order to recover the lost third dimension, depth, it is necessary to compute the relationship between images through correspondence. By finding corresponding primitives such as points, edges or regions between the images, such that the matching image points all originate from the same 3D scene point, knowledge of the camera geometry can be combined in order to reconstruct the original 3D surface.

One approach to the correspondence problem involves the computation of a disparity map, where each pixel in the map represents the disparity of the matching pixels between two images. The optimisation of a cost function is a common approach in order to obtain the disparity map [8, 21, 22]. Taking advantage of the epipolar constraint, which enables the search area to collapse from a 2-dimensional image to 1-dimensional epipolar lines, along with the ordering [28], uniqueness and continuity constraint [18], algorithms have been proposed which compute the disparity map to sub-pixel accuracy. However, when factors such as noise, lighting variation, occlusion and perspective distortion are taken into account, stereo disparity algorithms are still challenged to model accurately discontinuities, epipolar line interactions and multi-view stereo [6, 11].

Roy and Cox [22] and more recently Kolmogrov [15] developed an algorithm for solving the multi-view stereo correspondence problem. By stacking the candidate matches of range disparity along each epipolar line into a cost function volume, maximum flow analysis and graph cuts are used in order to determine the disparity surface. While these approaches to stereo analysis provide a more accurate and coherent depth map than the traditional line-by-line stereo, these methods remain dependent on and sensitive to the uniqueness and the accuracy of the matching correspondence stage. Although the optimisation of the cost function is performed in a three-dimensional space, the computation of a disparity surface remains only a 2.5-D sketch of the scene [18].

While the aforementioned techniques operate in 1 or 2D space, there also exists a class of stereo algorithms that operate in 3D scene space. Introduced by Collins [5] and Seitz and Dyer [23], these algorithms, instead of using disparity to compute the depth of an image point, directly project each image into a 3D volume, such that the locations of 3D world points are inferred through analysis of each voxel's relationship in 3D space. Kutulakos and Seitz recently proposed the Space Carving Algorithm aimed at solving the N-view shape recovery problem [17]. The photo hull, the volume of intersection of all views, is determined by computing the photo-consistency of each voxel through projections onto each available image. While these approaches produce excellent outcomes, apart from the fact that they require a vast number of input images, improvements can be made by imposing spatial coherence, replacing the voxel-based analysis with a surface orientated technique.

Classical active contours such as snakes [12] and level sets [1] have mainly been applied to the segmentation problem in image processing. The recent introduction of fast implicit active contour models [16], which use the semi-implicit additive operator splitting (AOS) scheme introduced by Weickert et al. [26], is an improved version of the geodesic active contour framework [4]. Given such advancements in active contour analysis, multi-dimensional segmentation is becoming not only more robust and accurate, but computationally feasible. The application of surface evolution and level set methods to the stereo problem was pioneered by Faugeras and Keriven [7]. Although Faugeras' approach is limited to binocular stereo and the epipolar geometry, their novel geometric approach to the stereo problem laid the foundation for a new set of algorithms that can be used to solve the 3D reconstruction problem.

In this paper, we will present two new techniques for 3D scene reconstruction. Firstly, we propose a new approach to 3D reconstruction through the use of globally optimal geodesic active contours. In order to formulate the 3D reconstruction problem suitable for segmentation analysis, we explicitly describe the design of a 3D voxel feature space, which integrates all the information available from each camera view into one unified volume for processing. Rather than solving the correspondence problem between the images by computing disparity and matching feature primitives, and instead of using photo-consistency constraints to determine the colouring of each voxel, our approach projects and integrates all the feature information about each image into one voxel volume. By collapsing the voxel space into a metric space, segmentation algorithms can then be applied to directly reconstruct the complex 3D scene.

Secondly, we propose a new approach for the recovery of 3D models and its motion from multi-view image sequences. While there are many studies in the area of stereo and motion analysis from stereo rigs, we propose the use of a 4D voxel volume to recover not only 3D and motion information from stereoscopic image sequences, but an algorithm capable of processing multi-view image sequences. By augmenting the design of our 3D voxel feature volume to a 4-D feature space, we present a novel approach to the analysis of stereo motion for the reconstruction of dynamic 3D scenes.

Section 2 of this paper will explain in detail the design of the 3D voxel volume and how segmentation can be used to compute the 3D scene. Section 3 will further describe how the 3D voxel volume can be extended to solve the stereo motion problem. The application of 4D reconstruction to analysis multi-view image sequences will be presented. Finally, section 4 will provide a summary of the proposed techniques and directions for future research.

2. 3D Reconstruction

A common approach to stereo reconstruction is the optimisation of a cost function, computed by solving the correspondence problem between the set of input images. The matching problem involves establishing correspondences between the views available and is usually solved by setting up a matching functional for which one then tries to find the extrema. By identifying the matching pixels in the two images as being the projection of the same scene point, the 3D point can then be reconstructed by triangulation, intersecting the corresponding optical rays. Our proposed method differs from this approach by projecting all the images into a common space prior to analysing the correspondence between the images. The matching problem is then solved not as a correspondence problem between images, but as a matching functional, computed for each voxel in the volume. This functional is optimised through segmentation to recover the 3D structure of the scene.

Prior to the construction of the 3D voxel volume, the camera geometry of the images needs to be computed through camera calibration. Knowledge of the camera geometry not only enables the construction of the projection matrix, but also allows the computation of the polyhedral intersection of the camera views. Although solving the bounding region of interest of the images is similar in idea to solving the space carving problem, our proposed method differs greatly from space carving. Rather than deciding on the likelihood of a photo-consistent match for each voxel in 3D space, our method does not perform any computation at the projection stage. Instead, all the feature information is stored inside each voxel and is dependent on the solution to the segmentation problem in order to decide on the 3D surface of the scene.

Assuming a pinhole camera model, the 3D voxel volume is created by projecting all of the images into a 3D polyhedron, such that each voxel contains a feature vector of all the information contained in each camera view. For example, the feature vector can include the RGB values of the voxel's projection into each camera image, the gradient of the image, and even information relating to the projected pixel's neighborhood. A metric volume can then be derived from this voxel space to become the input to the segmentation stage. Cost functions such as the variance between the projections or even a probability density function can be used. Furthermore, by altering the resolution of the voxel volume, the segmentation can output either a dense or a sparse reconstruction of the 3D scene.

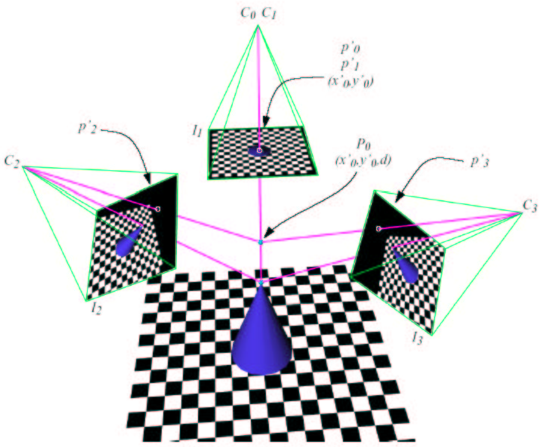


Figure 1. Multi-View Projection. 3D point P_0 is projected onto Images I_1, I_2, I_3. (Image courtesy of S. Roy and I. J. Cox, Figure 1 [22])

2.1. Voxel Volume

We briefly present the well-studied general framework of projective geometry required in the construction of the 3D voxel volume [10]. A set of n input images $I_1, \dots, I_n$ of a 3D scene are projected from n cameras $C_1, \dots, C_n$, as depicted in Figure 1 with $n = 3$. In our formulation, we will assume a pinhole camera model and that all surfaces are Lambertian (i.e. the intensity of a 3D point is independent of viewing direction). The projective coordinate of a 3D point P_w in world space is expressed with homogeneous coordinates as

$$P_w = \begin{bmatrix} x_w & y_w & z_w & 1 \end{bmatrix}^T$$

while the projective image coordinate of a pixel in image I_i is

$$p_i = \begin{bmatrix} x_i & y_i & z_i \end{bmatrix}^T$$

such that the corresponding pixel coordinate p'_i of the projected point p_i can be obtained by applying a homogenising function H where

$$H(\begin{bmatrix} x \\ y \\ z \end{bmatrix}) = \begin{bmatrix} x/z \\ y/z \end{bmatrix} \qquad (1)$$

Given the volume of interest of the 3D space for reconstruction, we can obtain each voxel's feature vector by projecting every P_w in the 3D volume onto each of the n images available. With f features for every pixel in the image, each voxel will contain an $f \times n$ matrix, such that the collection of all voxels will contain all the information all the images. In other words, given the 3D voxel volume, all the processing and analysis can be achieved without the need of the original images. We define 4 matrices to describe the projection from a voxel in the volume to a pixel in the image.

Given a volume of M voxels, each voxel will be indexed by its voxel coordinates, $v_m = [v_a, v_b, v_c, 1]^T$, where v_a, v_b and v_c will range from 1 to the dimensions of the volume. The extra parameter appended at the end of the voxel coordinate is for consistency with the augmented homogeneous coordinate in projective space. To transform from voxel coordinates to 3D world coordinates, we compute

$$P_w = \boldsymbol{V} v_m$$

where

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{I_3 k} & \boldsymbol{t_v} \\ \boldsymbol{0^T} & 1 \end{bmatrix}$$

and $\boldsymbol{I_3}$ is the 3×3 identity matrix, $\boldsymbol{t_v}$ the translation vector for specifying the world coordinate of the voxel volume's origin, and $\boldsymbol{k} = [k_x, k_y, k_z]^T$ is the stride in each voxel dimension, i.e. the number of units between each voxel in 3D world space. The choice of $\boldsymbol{k}$ and $\boldsymbol{t_v}$ is dependent on the resolution desired for the voxel volume and the origin of the volume of interest respectively.

From world coordinates, the classical 3×4 perspective projection matrix, $\boldsymbol{P}$, can be applied to obtain the projection of the 3D world point in image coordinates. In the case where we define the optical centre of the base camera, C_0, to coincide with the origin of the world coordinate system, the projection matrix will be simplified to be

$$\boldsymbol{P_0} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

subsequently, we can define a transformation matrix $\boldsymbol{W_i}$

$$\boldsymbol{W_i} = \begin{bmatrix} \boldsymbol{R_i} & \boldsymbol{t_i} \\ \boldsymbol{0^T} & 1 \end{bmatrix}$$

with rotation $\boldsymbol{R_i}$ and translation $\boldsymbol{t_i}$ to define the position and orientation of camera C_i relative to the base camera, C_0. The relative positions and orientations of each camera i is determined by a calibration procedure. Thus for C_i, the projective projection matrix will be

$$\boldsymbol{P_i} = \boldsymbol{P_0 W_i}$$

From the image coordinates, the pixel coordinates of a projective point can be recovered up to a scaling factor, given knowledge of the internal parameters of the camera. Neglecting radial distortion from calibration, a matrix of intrinsic parameters can be computed such that

$$\boldsymbol{A} = \begin{bmatrix} -f_x & 0 & o_x \\ 0 & -f_x/\alpha & o_y \\ 0 & 0 & 1 \end{bmatrix}$$

where f_x is the focal length in effective horizontal pixel size units, α the aspect ratio (i.e. the vertical to horizontal pixel size), and (o_x, o_y) the image centre coordinates.

Combining all the described matrices, each voxel can be projected into each image i by computing

$$p_i = \boldsymbol{A}\boldsymbol{P_i}\boldsymbol{V} v_m$$

From the obtained projective pixel coordinates, the actual pixel coordinates can be obtained by applying the homogenising function described in Eq. 1.

2.2. The Correspondence Problem

The construction of the voxel volume is purely based on image projections and thus does not require the solution to the correspondence problem in order to produce a dense reconstruction. Unlike algorithms that use disparity maps to guide the 3D reconstruction, dense feature correspondence or area-based matching of every pixel is no longer necessary. However, while the computation of the 3D volume does not require dense correspondences, feature correspondence is needed in establishing the camera geometry and for the purpose of camera calibration. The accuracy of the projections and the reliability of the volume are highly dependent upon robust and accurate camera calibrations. As noted by Medioni [19], in multi-view stereo, there is an imperative need for camera calibration and consistency of matches between multiple-views. Therefore full calibration information needs to be provided with the image set or techniques similar to [31] must be used to obtain the calibration parameters.

The construction of our voxel volume is similar to the concept proposed by Kimura, Saito and Kanade [13] in recovering the 3D geometry from the camera views. Their method is, however, restricted to three images since it is dependent on dense feature correspondences between the input images in order to model the 3D surface and cannot overcome the problem of occlusion since there are no matching points in those regions. Algorithms that depend on dense feature correspondence have much difficulty modelling occlusions. Our proposed method overcomes this problem by redefining the problem, using a new approach that does not depend on pixel to pixel feature correspondence. One of the advantages of our approach is that occlusion does not need to be explicitly modelled. Occluded regions visible in a limited number of images are still projected validly into 3D space for analysis, with the major difference being that less images project to that region. The occluded regions, however, can still be modelled, only that the 3D reconstruction for those region depends on less data. This scheme subsequently also allows for occluded region to be iteratively improved as more images of the occluded scene are available.

2.3. Segmentation

The development of algorithms that can provide globally optimal solutions to segmentation problems makes its application in image processing very attractive. By designing a volume appropriate for maximum-flow analysis, the minimum-cut associated with the maximum flow can be viewed as an optimal segmentation. While Roy and Cox have demonstrated a version of maximum-flow to analyse stereo images, a more computationally feasible method was recently proposed by Sun [24]. A two-stage dynamic programming (TSDP) technique was introduced to obtain efficiently a 3D maximum-surface, which enables the computation of a dense disparity map.

In our voxel volume formulation, since our projected volume enables us to work directly in true 3D coordinates, we aim to output a 3D surface representative of the complete 3D scene rather than using a disparity map to obtain a 2.5-D sketch of the scene [18]. Formulating the 3D reconstruction problem as a segmentation problem has many advantages over the use of the classical dynamic programming technique. In segmentation, optimisation is performed along a surface rather than along a line. This subsequently provides segmentation methods with the advantage of outputting contours that wrap back on themselves, while dynamic programming will have difficulty following these concave surfaces. Rather than reformulating dynamic programming or similar techniques in order to model occlusions and concavity, we propose the use of segmentation to approach 3D reconstruction from a new point of view.

Active contours have been demonstrated to be a useful tool in the segmentation problem. Geodesic active contours that use a variational framework have been shown to obtain locally minimal contours [4]. Fast implicit active contour models, that use the semi-implicit additive operator splitting (AOS) scheme introduced by Weickert et al. [16, 26], and shortest path algorithms [3], have been used to avoid the variational framework producing optimal active contours. By formulating a volume appropriate for 3D segmentation, we propose the use of a form of geodesic active contours recently introduced by Appleton and Talbot [2] which has been demonstrated to be globally optimal. By choosing a positive scalar metric, g, such that g can be assured to be always greater than zero, the minimisation of the energy functional E, can be formulated to describe the segmentation

$$E(C) = \int_C g(C(s))ds$$

where C is the segmentation contour.

3. 4D Reconstruction

The fusion of stereo and motion has been recognised by many researchers as a means of providing additional information that were not previously obtainable through their independent analysis. Waxman and Duncan pioneered the analysis of stereo motion by considering binocular image flow [25]. Many studies have subsequently tackled this problem through the use of Kalman filtering, optical flow and feature tracking [14, 20]. While these methods have demonstrated reasonable success, they are limited by problems inherent in the correspondence problem, as described in section 2.2. Similar to our proposed approach in the use of segmentation, rather than reformulating optical flow or Kalman filtering to model stereo motion, we propose the use of a 4D voxel volume in order to analyse the stereo dynamics in stereoscopic image sequences.

By embedding the design of our 3D voxel feature volume into a 4-D feature space, we present a novel approach to the analysis of stereo motion for the reconstruction of dynamic 3D scenes. Given a set of images captured over different time frames, we can compute the camera geometry and projection parameters for each image through the use of the many calibration techniques developed, such as Zhang's four point algorithm for stereo rig analysis [29]. From the set of projection matrices computed, we can construct a voxel volume for each time frame. Since geodesic active contours can be applied to segment multi-dimensional volumes, similar to the analysis of our 3D voxel volume, we can compute a segmentation in 4D in order to produce a 3D surface in time. The use of this 4D voxel volume also has the advantage of not only recovering the 3D and motion information from stereoscopic image sequences, but is capable of processing multi-view image sequences. The computational feasibility of multi-dimensional segmentation makes this 4D approach to stereo motion an attractive alternative to the analysis of dynamic, complex 3D scenes.

4. Summary and Future Directions

We have proposed two novel approaches to the 3D reconstruction problem through the design of a 3D and 4D feature voxel volume. While current techniques depend heavily upon the solution to the correspondence problem in order to guide the 3D analysis, by taking advantage of the recent developments in segmentation and the introduction of globally optimal algorithms, our method reformulates the computation of correspondence as a segmentation problem. Furthermore, we present a novel approach to the analysis of stereo motion by transforming the problem into a 4D segmentation analysis.

The success of this approach is dependent on the accuracy of the construction of the 3D and 4D voxel volume. Subsequently, the shortcomings of this method are related to its sensitivity to errors in camera calibration and projection. However, with the many developments and studies completed in this area, the error can be minimised. The accuracy of this method also increases as the number of input images increases, making this approach well suited for analysing multi-view image sequences.

The size, shape and location of the voxel volume is also currently manually estimated. Although a difficult and complex problem, a dramatic improvement to the algorithm will be the direct computation of a polyhedral volume of interest. Similar to solving the space carving problem, the polyhedral volume can be obtained by computing the intersections of all camera's field of view. Assuming a pinhole camera model, each camera projection will be a rectangular pyramid, thus the bounding polyhedron will be the solution to the problem of intersecting multiple rectangular pyramids of varying orientations.

The design of a multi-dimensional voxel volume also lays the foundation for 3D or even 4D recognition. A ball for example in 3D space would occupy a spherical volume, while a ball in trajectory can be recognised as a cylindrical tube with hemispherical ends in 4D space. Previous works by Xu [27] have attempted to unify stereo, motion and object recognition into one approach by observing their common use of feature correspondence. Using our new proposed approach, feature correspondence is replaced with a voxel volume that contains all feature information. Thus, through the analysis of a multi-dimensional feature volume, it is possible to design a unified framework for multi-view, motion and object recognition.

Acknowledgements

We would like to thank Ass. Prof. Peter Kootsookos, IRIS Group, School of ITEE, UQ, and Dr. Changming Sun, CSIRO Mathematical and Information Sciences, for the many inspirational discussions and expert advices throughout this research.

References

[1] D. Adalsteinsson and J. A. Sethian. A fast level set method for propagating interfaces. *Journal of Computational Physics*, 118(2):269–277, 1995.

[2] B. Appleton and H. Talbot. Globally optimal geodesic active contours. *Journal of Mathematical Imaging and Vision*, 2002. Submitted.

[3] M. Buckley and J. Yang. Regularised shortest-path extraction. *Pattern Recognition Letters*, 18(7):621–629, 1997.

[4] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61–79, 1997.

[5] R. T. Collins. A space-sweep approach to true multi-image matching. In *CVPR*, pages 358–363, 1996.

[6] I. Cox, S. Hingorani, S. Rao, and B. Maggs. A maximum likelihood stereo algorithm. *Computer Vision and Image Understanding*, 63(3):542–567, 1996.

[7] O. Faugeras and R. Keriven. Variational principles, surface evolution, pde's, level set methods and the stereo problem. *Special Issue on Partial Differential Equations and Geometry-Driven Diffusion in Image Processing and Analysis of the IEEE Transactions on Image Processing*, 7(3):336–344, March 1998.

[8] P. Fua. From multiple stereo views to multiple 3d surfaces. *International Journal of Computer Vision*, 24(1):19–35, 1997.

[9] D. Geiger, B. Ladendorf, and A.Yuille. Occlusions and binocular stereo. *International Journal of Computer Vision*, 14:211–226, 1995.

[10] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 1998.

[11] H. Ishikawa and D. Geiger. Occlusions, discontinuities, and epipolar lines in stereo. In *ECCV*, pages 232–248, Freiburg, Germany, June 1998.

[12] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1998.

[13] M. Kimura, H. Saito, and T. Kanade. 3d voxel construction based on epipolar geometry. In *ICIP*, volume 3, pages 135–139, October 1999.

[14] R. Koch. 3-d surface reconstruction from stereoscopic image sequences. In *ICCV*, pages 109–114, Cambridge, MA., USA, June 1995.

[15] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *ECCV*, volume 3, pages 82–96, 2002.

[16] G. Kühne, J. Weickert, M. Beier, and W. Effelsberg. Fast implicit active contour models. In L. V. Gool, editor, *Pattern Recognition*, Lecture Notes in Computer Science. Springer, Berlin, 2002. To appear.

[17] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. Technical Report TR692, Computer Science Dept., U. Rochester, 1998.

[18] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman and Co., 1982.

[19] G. Medioni. Binocular and multiple-view stereo using tensor voting. Technical report, USC IMSC, 2001.

[20] T. Moyung and P. Fieguth. Incremental shape reconstruction using stereo image sequences. In *ICIP*, 2000.

[21] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(4):353–363, 1993.

[22] S. Roy and I. J. Cox. A maximum-flow formulation of the n-camera correspondence problem. In *ICCV*, pages 492–499, 1998.

[23] S. M. Seitz and C. R. Dyer. Photorealistic scene reconstruction by voxel coloring. In *CVPR*, pages 1067–1073, 1997.

[24] C. Sun. Fast stereo matching using rectangular subregioning and 3d maximum-surface techniques. *International Journal of Computer Vision*, 47(1/2/3):99–117, May 2002.

[25] A. M. Waxman and J. H. Duncan. Binocular image flows: Steps toward stereo-motion fusion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(6):715–729, Nov. 1986.

[26] J. Weickert, B. ter Haar Romeny, and M. A. Viergever. Efficient and reliable schemes for nonlinear diffusion filtering. *IEEE Trans. Image Proc.*, 7(3):398–410, 1998.

[27] G. Xu. Unification of stereo, motion, object recognition via epipolar geometry. In *ACCV*, volume I287-291, 1995.

[28] A. L. Yuille and T. Poggio. A generalized ordering constraint for stereo correspondence. AI Memo 777, MIT, AI Lab, 1984.

[29] Z. Zhang. Motion and structure of four points from one motion of a stereo rig with unknown extrinsic parameters. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(12):1222–1227, December 1995.

[30] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.

[31] Z. Zhang, R. Deriche, L. T. Luong, and O. Faugeras. A robust approach to image matching: Recovery of the epipolar geometry. In *ECCV*, pages 179–186, 1994.

Player Tracking and Stroke Recognition in Tennis Video

Terence Bloom
School of Electrical Engineering and
Telecommunications,
The University of New South Wales,
Sydney, NSW 2052, Australia.

Andrew P. Bradley
Centre for Sensor Signal and Information
Processing (CSSIP),
The University of Queensland,
St Lucia, QLD 4072, Australia.

Abstract

In this paper we present an investigation into the computer vision problem of tracking humans in digital video. The investigation domain is digital tennis footage and the aim is to track the tennis player and recognise the strokes played. The motivation behind this investigation is to eventually automate the task of digital tennis footage annotation so that metadata, such as the time codes and a description of the strokes played, are automatically appended to the video. This then enables a number of compelling applications, from simple search facilities for the home viewer, to more complex analysis tools suitable for a tennis coach. The system developed solves the problem of tennis player tracking and stroke recognition using relatively simple, and well known, image processing operations constrained by an a priori *knowledge of the image capture conditions, the background scene, and the application domain.*

Keywords: Player Tracking; Video Annotation; Metadata; Digital Tennis Video**.**

Introduction

Automatic human tracking is a computer vision problem. It is the problem of getting a computer to analyse the digital video stream of a scene, detect when a person enters that scene, and then to subsequently track that person's movement through the scene. There have been many investigations into this research topic and for various applications: From systems designed to track a single person and find their body parts, e.g., Pfinder [1] or W^4 [2], to systems designed to track multiple people and monitor their interactions, e.g., Computers Watching Football [3].

A computer capable of tracking humans would be useful for a wide range of applications from automated tracking for security or TV cameras, to a vision-based human-computer interface. A different type of application for this technology is to automate the task of annotating digital video with metadata that denotes the presence, and a description of the actions, of people in the video. With the increasing use of digital multimedia there is a corresponding increase in the need for tools that enable the fast and efficient indexing, querying, and browsing of multimedia databases. Emerging standards, such as the MPEG-7 audiovisual format, will support this concept by providing a standard language for metadata description schemes for multimedia content.

Annotating digital tennis footage with metadata such as time codes and a description of the strokes played would provide easy access to digital tennis footage in a video archive. A tennis coach, for example, would benefit from this by being able to easily retrieve training footage of certain strokes of his/her protégé in order to track their improvement over time, or more importantly, to analyse match footage of an opponent to identify their weaknesses. Another possible application for tennis player tracking and stroke recognition is the automation of statistics tallying to provide match commentators or home viewers with direct access to current match statistics.

The System

The aim of this research was to develop a computer vision system for tracking tennis players and recognising their strokes. This aim is the tactical precursor to the automatic annotation of digital tennis footage with metadata. The approach taken was to build an entirely software-based system and to work 'off-line' with digital video in the standard uncompressed AVI format. In this way, the work was focused on algorithm development rather than an efficient real-time implementation.

Some specific assumptions were made in order to be able to realistically achieve the aim. The most significant assumption is that the camera is essentially fixed in position with no panning or zooming used (although camera jitter is accounted for). Further, the only people on the court are assumed to be the two players and their shadows are considered to be part of the player. Finally, the tracking system requires some initialisation in the form of capturing the background scene with no players present. Identification of the key frames, i.e., the frames when a stroke is actually played (when the racquet makes contact with the ball), is supplied by monitoring the audio stream for the distinct sound of the impact. In addition, the raw footage is manually edited into the individual points (rallies) and the system only attempts to track a single forecourt player.

As the tennis domain is very predictable, these assumptions are clear-cut, and can be relaxed one at a time by adding complexity to the system at a later stage.

The Algorithm

The algorithm is broken down into three units:

1. *Player Finding*: The player to be tracked is identified using a model of the background scene, standard image processing operations, and various *a priori* size and colour constraints.
2. *Player Tracking*: The player is tracked from one frame to the next by utilising the size and position of the player being tracked and their movement between consecutive frames.
3. *Stroke Recognition*: The key frames are identified as those frames where the racquet makes contact with the ball. These key frames are then further analysed to classify the tennis stroke as either: A forehand or backhand ground stroke, a volley, a smash, or a serve. Stroke recognition is performed using a three-stage algorithm based on the player's position in the court, finding the position of the player's racquet, and finally by finding the racquet arm of the player.

The structured and predictable nature of the tennis domain lends itself to exploitation similar to the notion of "closed-worlds" as applied to the football domain by Intille & Bobick [3]. That is, the visual processes that will be used to find the players are tailored using tennis domain knowledge and information that has already been learned about the player from previous processing. Furthermore, the complexity of the actual tracking problem is greatly reduced by some simple domain knowledge, e.g., that there are only two moving players, in two spatially distinct regions of the frames, i.e., we can expect them not to interact at all.

Player Finding

The traditional conceptual approach for tracking humans is to build and store a model of the scene, and then segment the humans in video by watching for variations from the scene model. In the tennis domain, the scene is the tennis court and the humans are the tennis players. A reference frame of the tennis court without any players present can be thought of as the simplest possible background model. Given a reference frame and a single frame from the tennis footage, the aim of the player finding unit is to derive, by visual means, the coordinates of the bounding box that encapsulates the player being tracked. Looking for large variations from the background, confirmed by the expected size and colour of the player, are used to achieve this.

The basic method for identifying large regions of variation is to first low-pass filter both the background frame and the current frame to remove random noise and reduce spatial detail. Then the pixel-by-pixel absolute difference is taken between them. The histogram of the difference image will exhibit a large dominant mode near zero (representing the static background pixels), and a smaller mode at a higher level (primarily representing the variation due to the player). Thus, by thresholding the difference image we can separate the player from the background by using a global threshold value found from the following equation:

threshold = mean + standard deviation / PSI.

Here PSI (Player Size Index**)** is a constant that is assigned based on the expected size of the player in the frame. In this way, the required percentage of pixels will be assigned to the player region after thresholding. The PSI has a limited number of values that relate to possible player sizes, e.g., when in the forecourt, on the baseline, or appearing or disappearing from the scene. If the player falls into a particular size range, the corresponding PSI is designated to that frame and used as the initial estimate for the next frame. The actual PSI values are calibrated manually for a given set-up.

Once the binary image has been obtained via thresholding morphological opening and closing operations, with square (4 x 4) structuring elements, are then used to remove small areas of noise. Then a connected component analysis is performed on the resulting image to identify the binary large objects (blobs) present. Finally, the blobs found are iterated through to find the one that is of the correct dimensions to be the player.

When the player is found in the first frame, a colour sample is taken as the average of a 5 by 5 grid of pixels from the centre of the player bounding box. Therefore, when more than one candidate player blob is identified (often the case when the player and their shadow are identified as two separate blobs) colour matching can be used to confirm the player blob. Colour matching is implemented as the sum of the absolute difference in each colour channel – sometimes referred to as a city-block distance [6]. The blob with the minimum distance to the initial colour sample of the player is then considered to represent the player. The operation of the three steps involved in the player finding unit are illustrated in Figure 1.

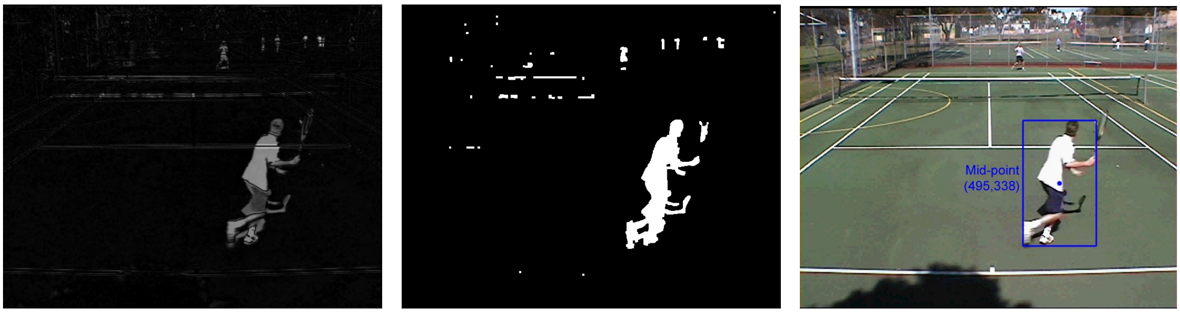


Figure 1: From left to right; the difference image produced from the pixel-by-pixel difference of a frame with the background reference frame. The difference image after thresholding and morphological filtering. The original frame with the player's position confirmed by identifying the bounding box of the blob that represents the variation in the difference image due to the player.

Player Tracking

Given the background scene image and a set of sequential frames of tennis footage, the aim is to be able to track the forecourt tennis player. That is, knowing at all times whether the player is present, and following their movement from frame to frame by maintaining the coordinates of their bounding box while they are present. The player finding unit is deployed to find the coordinates of the player's bounding box in each frame. While the player tracking unit, described here, applies robust inter-frame tracking techniques that further improves the reliability of the player finding unit.

Tracking requires a manual initialisation step for each point (rally) in the form of simply confirming the approximate position and size of the player found in the first frame. A flag indicating whether the player is present is then set accordingly, and maintained throughout. The player presence flag is only changed if the player disappears from, or re-appears into, view. The player is considered to have disappeared if the player finding unit reports the player as not being present in the current frame and the player bounding box had one side lying on the edge of the frame in the previous frame. Similarly, the player is said to have re-appeared if the coordinates of the player bounding box have one side lying on an edge of the frame and the player was not present in the previous frame.

The mid-point of the player bounding box is considered to be a good representation of the player's current position. Therefore, when the player is present, the bounding box found is confirmed by the player tracking unit if its mid-point has moved less than a pre-set distance between consecutive frames.

Stroke Recognition

Given the original set of blobs found by the player tracking unit for a key frame, i.e., a frame known to correspond to a tennis stroke, and the coordinates of the already confirmed player position, the aim of the stroke recognition unit is to make an elementary attempt at recognising the stroke played. Five different generic stroke types are defined in the current system, although there are many possible variations. Currently, the classified stroke types are: Forehand ground stroke; backhand ground stroke; volley; smash; and serve. On a conceptual level the player size, racquet orientation, and stroke timing can all be used to recognise these generic stroke types.

The stroke recognition unit utilises a three-stage algorithm. Initially, the player's relative size and position within the frame is used to distinguish the volleys from any other stroke (as a volley, by definition, is the only stroke played at the net). Hence, if the player's PSI is in the small size range and the player is at the net, the stroke is considered a volley. Next, if the stroke is not a volley, an attempt is made to find the position of the racquet head relative to the player's position. There are three distinct possibilities considered here: Above, to the right, or to the left, of the player. Assuming that we are tracking a right handed player and that we are looking at the forecourt player from behind, if the racquet head is found to the right of the player, the stroke must be a forehand ground stroke; if the racquet is found to the left of the player, the stroke must be a backhand ground stroke; and if the racquet is found above the player, the stroke must either be a smash or a serve. Finally, the timing of the stroke within a point is used to distinguish a serve from a smash; a serve being the only shot that is played at the beginning, or with the first few frames, of a point starting.

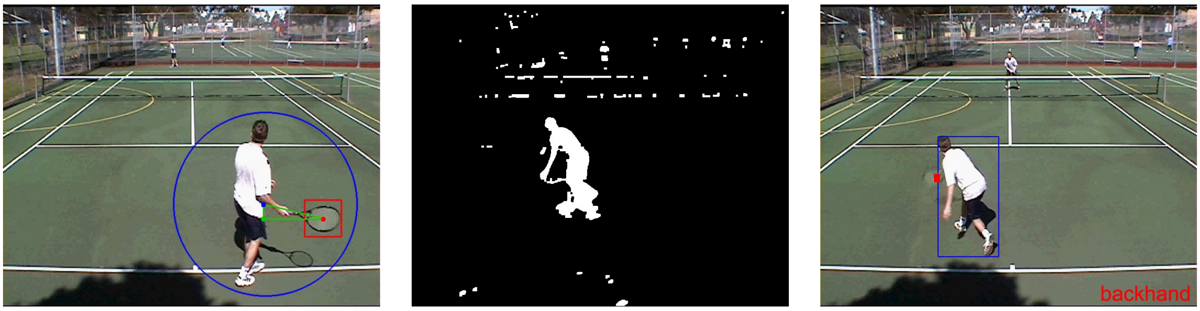


Figure 2: Left: The maximum possible racquet head size just fits into the square bounding box in red. A maximum allowable radius from the mid-point of the player is illustrated in blue. The green right-angled triangle formed from the mid-points of the player and the racquet is used to calculate the distance of the racquet from the player. Middle and right: A blob representing the racquet is found on the left of the player, thus the stroke is recognised as a backhand ground stroke.

The robust determination of the racquet position is the crux of reliable stroke recognition. This is achieved through further image processing on the already segmented binary difference image. It is expected that the disturbance created by the racquet head would often have been disconnected from the player's disturbance during previous visual processing. Thus, the first approach for finding the racquet is to look for a separate blob that could possibly represent it. As the time of a stroke is defined as the instant the racquet makes contact with the ball, it frequently happens that the disturbance of the racquet and the ball coincide, in fact, the ball's disturbance is often more solid and distinctive than that of the moving racquet. Hence, the racquet blob found is likely to be a somewhat blurred combination of the racquet, ball, and the court background.

The problem of finding the racquet blob is similar to that of finding the player blob. The full list of blobs present in the image is iterated, eliminating those that either contain less than a minimal number of pixels, or are too large (in either dimension) to be the racquet. These size criteria, which are easily established, discount most small noise blobs and other large disturbances (such as the net). In addition, to the size criteria, blobs beyond a pre-determined distance from the mid-point of the player can also be eliminated. These visual criteria are illustrated in Figure 2. The final stage for confirming a racquet blob is to take a 3 x 3 colour sample from the centre of each blob and to compare them to a known colour sample of a racquet. A blob is considered to be a match in colour only if all three channels are within a required range. If a blob is found to match, it is regarded as representing the racquet head. The mid-point of the racquet relative to the mid-point of the player is then used to classify the stroke played.

If none of the remaining blobs are found to be a good colour match, then it is assumed that either the racquet is contained within the player blob, or its disturbance is too distorted or discoloured. In this case, the stroke recognition unit undertakes further analysis of the player blob in order to find the arm holding the racquet. The player blob is first further morphologically closed through another iteration each of a dilation and erosion with a large, 7 x 7, structuring element. Then the player blob is skeletonised, using the Skeleton Zhou algorithm [4], to derive a 'stick figure' of the player, an example of which is shown in Figure 3. The major assumption used here is that the longest branch originating from a node in the top half of the player's skeleton (which discounts branches representing legs and shadows) represents the arm holding the racquet. Thus, the end-point is a good representation of racquet position relative to the player and so the stroke can be classified.

There were a few different approaches that could have been used for interpreting a player skeleton, such as the Hough Transform [5]. However, a more simplistic approach was used here and was found to be robust. The end-points of the skeleton are found by looking for pixels that have exactly one connected neighbour. Nodes are pixels with three or more connected neighbours. The length of the branch for each end-point is measured by counting pixels while iterating through to the closest node. If this node is in the upper half of the player, this branch is considered as possibly representing the player's arm holding the racquet. Once all of the branch lengths have been found, the longest branch is selected as the end-point representing racquet position.

The mid-point of the racquet blob, or end-point of the player skeleton, is therefore used to represent the position of the racquet. When compared to the mid-point of the player, using the heuristics described earlier, the stroke can be classified as a forehand, backhand, or overhead. The five generic strokes can thus, in the majority of cases, be recognised, especially when they are played clearly and in textbook style.

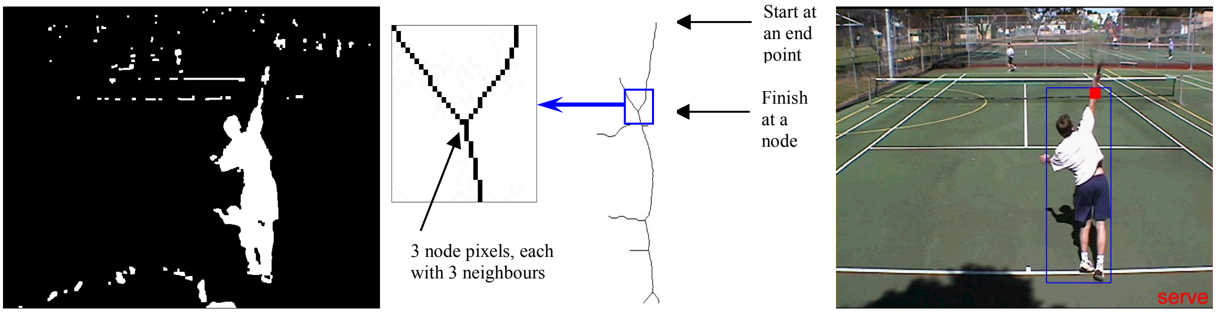


Figure 3: A frame where the racquet's disturbance was unable to be confirmed as all candidate blobs failed colour matching. The skeleton of the player is instead analysed in an attempt to find the arm holding the racquet. The top right branch was identified, and thus the position of the racquet relative to the player was found, helping identify the stroke as a serve.

Discussion

The stated aim of the work has been achieved, albeit under some restrictive assumptions. The forecourt player in the available footage was tracked quite robustly, even when exiting from, or re-entering into, view. The generic strokes were correctly recognised for the majority of examples tested. However, the algorithm indeed mis-recognised a stroke when either of the two assumptions broke down, i.e., that the colour sample used to match the racquet blob is accurate and representative; and that the skeleton's longest branch originating from the upper half represents the arm holding the racquet. However, as a first attempt the system served well in bringing to light the issues to be considered in future research.

It is believed that through further work to relax the necessary assumptions, the system would be capable of working well on real tennis training or even match footage. Thus, the algorithms described here could form the core technology of a system used for automatic tennis archive annotation. Significant further work is also required to improve the computational performance of the system so that it works in real-time. However, the methods used and described here are well studied and so there are a number of hardware and software solutions already available for this task.

The system's limitations are easily identified, and thus able to be addressed. To achieve more dynamic and reliable tracking, accounting for shadows, tracking both of the players, and handling non-players should be addressed. To be able to work with match footage the system should be able to work under camera movement, such as pan and zoom, and quickly and adaptively build up the required background scene image. To be more useful for annotation and statistics tallying, the system would be required to be capable of selecting the best key frame of when a stroke is played and preferably be able to distinguish between a greater variety of strokes than the five used in this study.

Conclusions

The point of departure for this research was the knowledge that current computer vision technology can be used to analyse a digital video stream to find a human moving in a scene. One possible application of this technology is to perform the useful, real-world, task of automatically annotating digital video streams with metadata that can then be used to search or summarise the video content. This research has demonstrated that for video footage of a single forecourt tennis player, captured and analysed under certain constraints, it is possible to solve this computer vision problem and hence to develop these types of applications. This was achieved by breaking the problem down into conceptual parts, solving these parts one by one by the application of relatively simple image processing techniques and some domain specific knowledge.

References

[1] C. R. Wren, A. Azarbayejani, T. Darrell & A. P. Pentland, Pfinder: real-time tracking of the human body, *IEEE Transactions on Pattern analysis and Machine Intelligence*, vol. 19, 7, pp.780-785, 1997.

[2] I. Haritaoglu, D. Harwood, & L. Davis, W^4: Who? When? Where? What? A real time system for detecting and tracking people, *International Conference on Face and Gesture Recognition*, Nara, Japan, 222-227, 1998.

[3] S. S. Intille & A. F. Bobick, Visual tracking using closed-worlds, Massachusetts Institute of Technology, Media Lab Perceptual Computing Group Technical Report No. 294, 1994.

[4] C. Quek & G. S. Zhou, A novel single-pass thinning algorithm and an effective set of performance criteria, *Pattern Recognition Letters*, 16:1267-1275, 1995.

[5] R. C. Gonzalez & R. E. Woods, *Digital Image Processing*, Addison-Wesley Publishing Company, Inc, 1993.

(This page left blank intentionally)

Image Searching Tool Using Category-Based Indexing

Aster Wardhani
School of Software Engineering and Data Communication,
Queensland University of Technology
Brisbane, QLD 4001
a.wardhani@qut.edu.au

Abstract

Searching for an object in a general image collection using current image retrieval systems, is still a problem. The retrieval results contain many unrelated images. In providing an effective and robust image database, objects in an image need to be extracted. Since the number of stored images can be very large, automation is an important aspect. Image indexing is a technique that extracts objects in an image automatically. The aim of this research is to propose a new object based indexing system based on extracting salient region representative from the image and categorising an image into different types.

Different image has different characteristics and often require different image processing techniques. Currently, most content based image retrieval (CBIR) systems operate on all images, without pre-sorting these images into different types. This resulted in limitations on retrieval performance and accuracy. Categories described here are of statistical and syntactical descriptions rather than semantical. By analysing which features are dominant in an image, two outcomes will be obtained: category for that image and salient object. Identifying salient object further reduce the retrieval results into relevant images.

1. Introduction

Tools available for searching for an image within a collections are still far from satisfactory. General images, such as ones found in the internet can be very complex. Currently, standard internet image searching tools such as *Google*, use image filenames as indexing attributes. This result shows that a search keyword “mango” can results in a large number of retrieved images, mostly do not contain the fruit mango. To index an image using its content, there are currently three general approaches: *object recognition*, *statistical analysis* and *image segmentation*. Object recognition techniques are limited to specific domains, e.g., images containing simple geometric objects. This approach has been used to retrieve images of tools and CAD geometric objects [13] and also medical images [10]. For most other types of images, such as images containing people, sceneries, etc, object recognition techniques are infeasible.

To pursuit the complexity of these images, researches then employed statistical indexing based on colour and texture [7], [3]. Images can be retrieved by specifying a combination of RGB colour values, textural measures, and more recently using other features such as shapes and spatial relations between regions in the image [4], [11], [12]. Colour is a low level feature, and by itself cannot adequately describe objects in images. To enable objects to be extracted and indexed within the image, image segmentation technique are used [9]. However, segmentation results of general images are noisy and contain too many regions. Thus, this approach are still limited to simple objects, and thus for general images currently do not provide a meaningful object based representation.

Techniques used by general CBIR systems are generic and aimed to handle all types of images. This is not optimal, since different images have different level of complexity and may require different features and analysis techniques. For example, shape retrieval is not suitable for images containing mostly textures or irregular shapes, such as landscape images. Currently, most content based image retrieval (CBIR) systems operate uniformly on all images, without pre-sorting these images into different types. This has resulted in limitations on retrieval performance and accuracy.

2. Proposed System

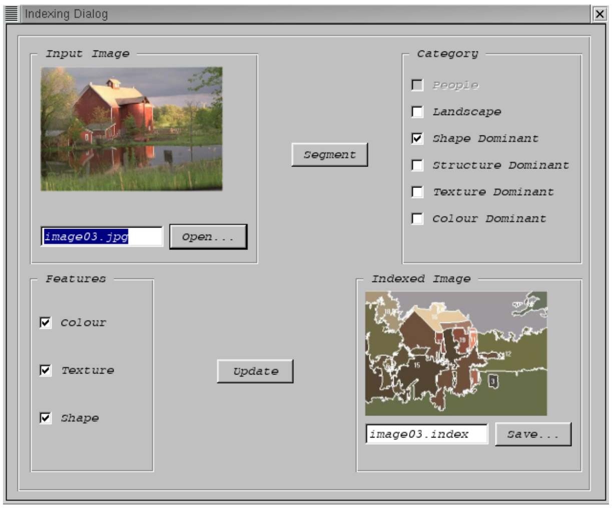
Rather than matching the whole image, it is more sensible to firstly *categorise* the image into different types. This is performed by finding the dominant characteristics of the image, such as how much texture, how complex the shapes are, and the presence of a dominant region. This strategy is supported by psychophysical evidence showing that hu-

Category Name	Feature Characteristics
Landscape	Colour = green and blue Spatial relation = vertical layer

People	Colour = human skin
	Shape = oval

Shape dominant	Number of regions = small Shapes = non complex Figure/background image = yes
Colour dominant	Number of regions = large Colour distribution = smooth
Texture dominant	Number of regions = large Colour distribution = non smooth
Structure dominant	Number of regions = large Shapes = complex

Table 1. Image Categories



□ People
□ Landscape
☑ Shape Dominant
□ Structure Dominant
□ Texture Dominant
□ Colour Dominant

image03.jpg

☑ Colour
☑ Texture
☑ Shape

image03.index

Figure 1. Category Indexing Tool

mans holistically classify visual stimuli before recognising the individual parts [6].

Based on the above intuition, the following approach to index images using categories is proposed. To provide image retrieval in the internet, this system can be implemented in two steps. Firstly, images are retrieved using filename such as using Google's image searching tool. Google search engine is used as an example as it index a large collections of images in the internet. Results from this search will then be classified into four different general and two semantic categories, shown in Table 1.

Typical indexing system using the proposed category is illustrated in Figure 1. The dialog box shows the features extracted from an image and the measured category for that image. The category is obtained automatically by analysing the composition of colour, texture and structure from the main regions. The regions are produced using the perception-based image segmentation system [14]. By implementing perceptual grouping, the results achieved are clean and only containing significant regions. Some segmentation results using this technique are illustrated in Figure 2 (different colour indicates different region).

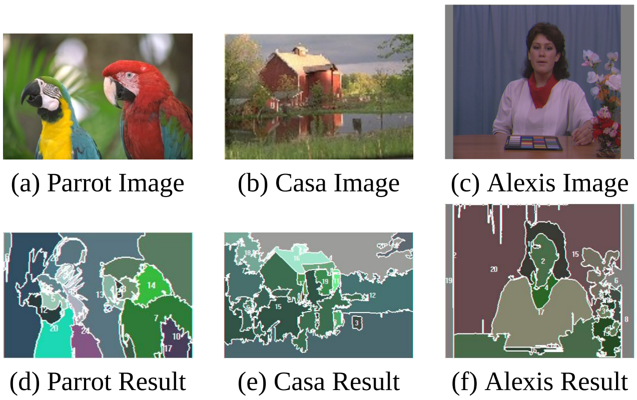
Using this image categorisation, a large retrieval results can be organised into groups that are based on the features of the image content. This would make navigation of results easier. The sheer number of uncorrelated retrieval results makes the searching task difficult and tedious. Using the category, a user could exploit the organisation of images into categories to locate images of interest. The meaning of each category will be explained in Section 3.4. For example, searching for the image "mango" would result in the following categories.

1. *shape dominant* (images likely containing a mango fruit)
2. *colour dominant* (images containing smooth areas, such as mango slices, some mango fruits)
3. *texture dominant* (images containing textural areas, such as mango trees)
4. *structure dominant* (other more complex images, containing structural and geometric regions)

If by "mango" the user interest means a picture of a single mango fruit, she/he would select the shape dominant category. However, if mango refers to a mango tree, the texture dominant category is more relevant. In many cases the categories may coincide with different semantic meanings of the search term. In searching for *people*, colour dominant (of skin colour) may be the most relevant category, whereas in searching for a *house*, structure dominant may be the most relevant.

The features described in Table 1 involves colour histogram, size and location of regions, number of regions and textural descriptions. Using this classification, the image shown in Figure 2(a), can be classified as a *Colour dominant* image, because of the appearance of large areas of smooth colour. Figure 2(b) will be classified as a *Landscape* image with the large tree areas. Figure 2(c) will be classified as a *People* image with the appearance of skin colour region.

Since images are segmented into a set of meaningful regions, retrieval results can be further pruned by performing object based query. In the example above, if the search for a "mango" aimed to retrieve all images that contain (a) mango fruit(s), the user initially is given six different groupings of retrieval results. To retrieve all images that contain a single or a collection of mango fruit(s), further object based query can be performed by firstly select the shape dominant group to choose images that have a single mango object. The user can then select the individual mango region from the image as the query object. The matching will then be performed to


(a) Parrot Image (b) Casa Image (c) Alexis Image

(d) Parrot Result (e) Casa Result (f) Alexis Result

Figure 2. Indexing Results using Perception-Based Segmentation Technique

all images that have similar shape and colour combination of the selected region. This will result in retrieval images containing not only a single mango fruit but also images containing (a) mango(s) within other objects. Additionally, within each groupings, further classification can be made, creating a so called "parse-tree" query.

3. Methodology

Based on the above considerations, the proposed object based image indexing system is described in Figure 3. This system consists of the following stages: segmentation and grouping stage, dominant region extraction and category generation. An input image is firstly segmented. The segmentation results are then grouped together, using the technique proposed in [14]. After this process, both low level and high level features are extracted from each region and by analysing these features, a category and dominant regions will be obtained for that image. To determine which region is dominant, some heuristics will be generated. A relevance feedback such as used in [5], can be used interactively by users to change the chosen object and category within the selected retrieved images.

3.1 Image Segmentation

Although image segmentation techniques have been studied extensively, there are still some drawbacks and issues that need to be solved. One of the drawback is the quality of the segmentation results for natural images. There are various complexities in natural images that need consideration, such as variation of texture and lighting, and complex object shapes. Results from existing techniques have the following properties:

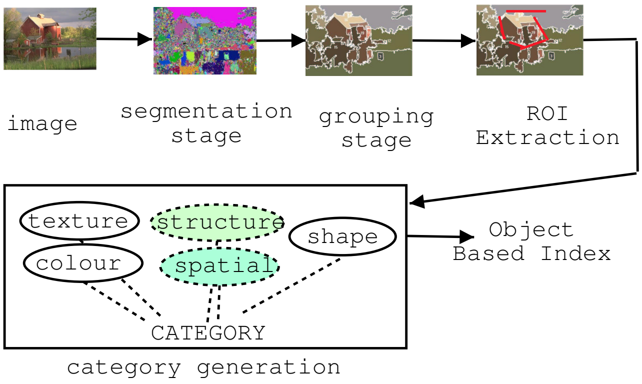
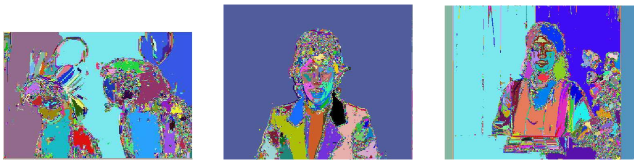


Figure 3. Proposed Image Indexing System

- Over segmented results, containing noisy regions at objects' boundaries and textural areas.
- The demarcation of regions do not always follow perceptual intuitions
- Results are very sensitive to threshold and requiring manual tuning

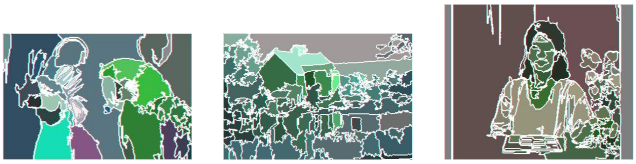
Selecting the best segmentation technique is an important issue. The difficulty in providing a meaningful segmentation is due to the non-correlation of existing colour distances with human perception of objects. Properties of natural images are too complex, requiring new perceptually intuitive metrics, which can adapt to intensity variations. Additionally, since each object in an image consists of a hierarchical composition, the segmentation process needs to be performed hierarchically. Some perceptual measures, therefore, should be added to improve the segmentation results.

To provide some perceptual foundations to the image segmentation implementation, the use of HVC-based region growing segmentation has been proposed, as reported in [15]. This segmentation approach uses adaptive threshold thus eliminates the need for manual tuning. This is performed by dividing an image into blocks, at each block, the presence of a strong edge indicates pixels requiring high threshold value. This value is then used as a threshold for all pixels in that block. To avoid missing any regions, the threshold is increased by a small amount to produce a slightly over-segmented results. It was perceived that it is better to have an over-segmented image than under-segmented results. Most of the noisy regions will be grouped later in the region grouping stage. Example of results from this segmentation technique is shown in Figure 4. This approach will be used in the proposed image indexing system.



(a) Parrot Image (b) Claire Image (c) Alexis Image

Figure 4. Segmentation Results



(a) Parrot Image (b) Casa Image (c) Alexis Image

Figure 5. Size Grouping Results

3.2 Region Grouping

Gestalt laws state that visual elements that belong to the same object, have the properties of *similarity, proximity, good continuation, closure, common fate, surrounded-ness and relative size, and symmetry* [1]. In the area of psychology, these principles have been accepted as the perceptual grouping laws. For this reason, Gestalt principles are used as the basis for the region grouping stage. The principles of proximity, similarity and good continuation can be used to group segments into regions. The principles of closure, common fate, symmetry and surrounded-ness, however, can be used for higher level grouping. This high level grouping is applied to find relationships between regions or objects' components. This was then translated into a grouping hierarchical formation. The issue would then how should the grouping operation should be performed? What are the grouping algorithm and rules required? To solve this issue, three different features were considered: *texture, colour* and *line continuation* [18]. The first grouping performed is the size grouping. The aim of this grouping is to merge noisy areas (region whose size is less than 100) using the similarity of region HVC means. Example of results from this grouping stage is shown in Figure 5.

The next grouping stage is the colour histogram grouping. This is performed by comparing the similarity of two region colour histograms. This is then followed by line continuation grouping. Regions are grouped based on comparing line continuation surrounding regions. Examples of final grouping results is shown in Figure 2. At each grouping step, the number of regions are reduced at each stage of grouping, finally leaving only significant components. These results are more meaningful and provide more data abstraction than the standard image segmentation techniques.

3.3 Dominant Region Extraction

It is difficult to know which object is of interest. An image can contain background objects which may look the same from the computer's point of view with the regions that belong to main objects. There might also be multiple objects. However, generally, there is a process of selection of important from less important objects in an image in human perception. Often the selection is not always based on semantic reasonings. Syntactical relations between regions, such as difference in size and texture can create a point of interests. A small house in the forest as depicted in Figure 3 is an example of region of interest. In Gestalt principles, this is described as figure-background principle (smaller figure against bigger surrounding objects).

The aim of dominant region extraction is to eliminate background, non-important regions, producing the most essential region. The reduction of non useful regions are required to reduced the matching and expensive structural analysis. The background will be eliminated by applying the figure/background principle of Gestalt laws. By analysing that the largest region surrounding other objects entirely can conclude that this region is the background thus eliminated. Similar idea was also used in the form of Region of Interests (ROI), used in a CBIR system proposed by Moghaddam et. al. in [8]. In this system, however, the regions are extracted manually by users.

In the proposed system, dominant region will be extracted automatically by analysing the size (largest), location (center) of the regions and appearance of interesting shapes and structures. Interesting shapes will be judged on the region shape regularity and its geometric properties.

3.4 Category Generation

Category generation is then responsible to assign an image type. Image type is aimed to provide sufficient groupings of images with similar characteristics. The issue would then be: "What are the succinct category that can capture different image characteristics?"

To categorise images without performing object recognition, the classification shown in Table 1 is used. It is based on the strength of different features that can be exploited from different image type. For example, images with texture dominant can be handled more effectively with a robust texture matching, whereas images under shape dominant can concentrate on good shape matching.

Shape dominant category is for images containing a small number of regions. These regions also have simple and regular shapes. Images categorised under colour dominant contains large number of smooth (non-textured) regions with less regular shapes. Texture dominant images are images that contain a large number of textural regions. There are many ways in detecting and describing texture. Using results from image segmentation, textural areas can be defined as neighboring regions whose size are small and spurious. These regions were marked as texture map and have been used successfully in segmenting texture for general images as reported in [16]. By measuring how many regions from segmentation are spurious regions, we can estimate whether the image is texture dominant. Using this texture map, texture dominant images can be classified as images containing mostly textural regions.

Images used here is assumed to be general whose semantic are unknown a priori. However, out of these images, many of them contain images which have distinct features, whereby from these features, semantic meaning can be derived. Such example of images are landscape images and images containing people. Landscape images usually contain large regions of sky (certain blue colour) and grass or trees (certain green colour). Using the colour combination, it has been proven successful in retrieving landscape images. Images containing people can be indexed and retrieved by analysing the presence of skin colour in the image. Since the ultimate goal is to perform image classification, using both facts another categories listed in Table 1 are added under classification of people and landscape. It would be possible to add further classification to domain specific images, such as botanical data, museum artefact, etc, whose retrieval has been demonstrated in systems such as [10].

Each of the different category is derived based on the analysis of the features extracted, described above. To allow such classification, each dominant region extracted is analysed for its various features. Both statistical and structural features will be used. This will include colour, texture and shape. For each shape some measures will be generated such as number of corners, degree of curvedness, etc, to distinguish between regular and complex shape.

Another new contribution in this project is the use of object structure [17]. Researches in psychology stated that classification of a scene may remain valid as long as the relative relationships between the image regions remain the same [2]. In the category of *structure dominant*, the existence of certain "interesting" or "prototype" structure will be used to represent an image and used for matching. Rather than matching the whole image or even the whole object (since same object can appear differently in different images), regions and their relations can be used instead. An example of a relational tree extracted is shown in Figure 6.

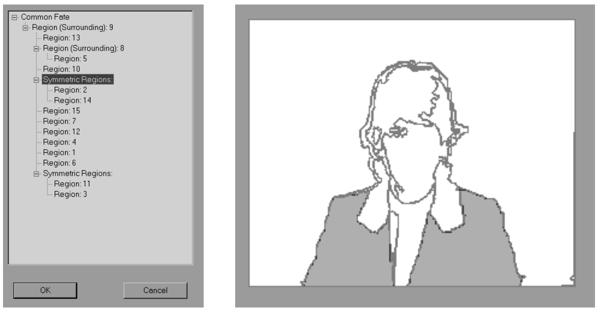


Figure 6. Image structure describes using relational tree

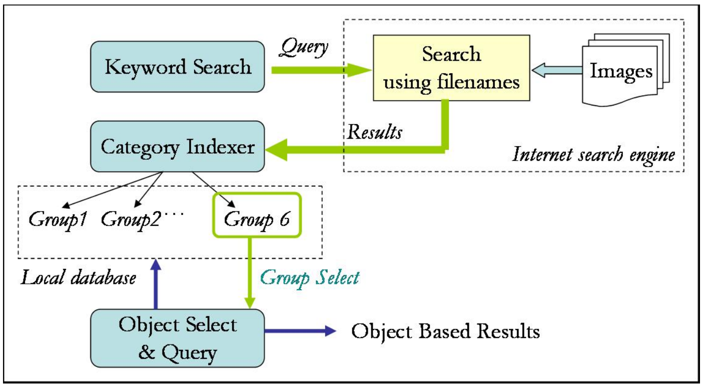


Figure 7. Retrieval System Implementation

3.5 Query Interface

There are many types of queries in CBIR, such as using keyword (semantic), or graphical descriptions (query by example, sketch, etc.). In this system, both queries are combined. Figure 7 shows the design prototype. The system will be implemented to retrieve images in internet. The search starts with query by keyword. The system will then sends this query to standard internet search engine such as *Google* image search tool to retrieve all images based on *image filenames*. The system will download all the links as well as all the thumbnails to be analysed. The analysis will group the retrieval results into 6 different categories described previously. Users can then prune the searching, by navigating the classification and selecting an object from an image. Since each region is segmented in the image, *object based query* can then be performed, reducing further to images that relates to the both *semantic keyword* and *graphical descriptions*.

Currently, segmentation and feature extraction stages have been performed and the overall retrieval system is currently being developed.

4. Conclusions

In order to retrieve images from large collections, a robust object based CBIR is crucial. This research aims to develop an image retrieval system that is based on extracting the dominant figure / region in the image, which subsequently placing an image into one of the proposed generic categories. The indexing information consists of not only low level but high level features. Images will be classified into a set of types. An indexing template will be generated automatically for each type, based on visual observation of which combination of features occurs for that type of image. Such template will be used to match images against the query information. We need to investigate the suitable and succinct set of features for each type of template.

This research will provide a new image retrieval system that provides users with the ability to further classify the content of an image. The impact from this method is more accurate retrieval results. An image will be represented by rich descriptions that relate directly to the content of the image.

References

[1] V. Bruce and P. Green. *Visual Perception: Physiology, Psychology and Ecology*. Lawrence Erlbaum Assoc. Hove and London, 2nd ed edition, 1990.

[2] C. Cave and S. Kosslyn. The role of parts and spatial relations in object identification. *Perception*, 22:229–248, 1993.

[3] T. S. Chua, K.-L. Tan, and B. C. Ooi. Fast signature-based color-spatial image retrieval. In *Proceedings of IEEE conference on multimedia computing and systems*, 1997.

[4] M. Flickner, H. Sawhney, H. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The QBIC system. *IEEE Computer*, 28, No.9:23–31, 1995.

[5] F. Jing, B. Zhang, F. Lin, W.-Y. Ma, and H. jiang Zhang. A novel region-based image retrieval method using relevance feedback. In *ACM Workshop on Multimedia*, pages 28–31, 2001.

[6] P. Lipson, E. Grimson, and P. Sinha. Configuration based scene classification and image indexing. In *Proceedings of IEEE conference on computer vision and pattern recognition*, 1997.

[7] H. Lu, B. Ooi, and K. Tan. Efficient image retrieval by color contents. In *Proceedings of 1994 international conference on applications of databases*, 1994.

[8] B. Moghaddam, H. Biermann, and D. Margaritis. Region-of-interest and spatial layout for content-based image retrieval. Technical Report TR-2000-35, MERL - mitsubishi electric research laboratory, November 2000.

[9] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE transactions on pattern analysis and machine intelligence*, 23(4):349–361, April 2001.

[10] A. Mojsilovic and J. Gomes. Semantic based categorization, browsing and retrieval in medical image databases. In *Proceedings International Conference Image Processing*, 2002.

[11] S. Sclaroff, L. Taycher, and M. La Cascia. Imagerover: A content based browser for the world wide web. In *IEEE Workshop on Content-based Access of Image and Video Libraries*, pages 2–9, 1997.

[12] J. Smith and S. Chang. Visualseek: a fully automated content-based image query system. *ACM Multimedia '96*, 1996.

[13] G. Srinivas, E. Fasse, and M. Marefat. Retrieval of similarly shaped parts from a cad database. *Systems, Man, and Cybernetics*, 3, 1998.

[14] A. Wardhani. *Application of psychological principal to automatic Object identification for CBIR*. PhD thesis, Information technology, Griffith University, 2001.

[15] A. Wardhani and R. Gonzalez. Automatic object extraction for content based retrieval. In *Proc. of International Conference on Telecommunication*, volume 2, pages 1005–10, 1997.

[16] A. Wardhani and R. Gonzalez. Using high level information for region grouping. In *Proc. of the IEEE Region 10 Conference (Tencon'97): Speech and Image Technologies for Computing and Telecommunications*, pages 339–342, 1997.

[17] A. Wardhani and R. Gonzalez. Automatic image structure analysis. In *Proc. of the IEEE International Conference of Multimedia Systems (ICMCS'98)*, pages 180–188, 1998.

[18] A. Wardhani and R. Gonzalez. Automatic object identification for image indexing. In *Proc. of the International Wireless Telecommunications Symposium (IWTS'98)*, pages 330–334, 1998.

Statistical significance of A_z scores: Classification of masses in screening mammograms as benign or malignant based on high dimensional texture feature space

Gobert N. Lee and Murk J. Bottema
School of Informatics and Engineering
Flinders University
PO Box 2100, Adelaide SA 5001, Australia
and
Cooperative Research Centre for Sensor Signal and Information Processing
SPRI Building, Mawson Lakes Blvd, Mawson Lakes, SA 4095, Australia
glee@infoeng.flinders.edu.au
murkb@infoeng.flinders.edu.au

Abstract

In order to develop a method for classifying masses in digitised screening mammograms as benign or malignant, 260 image texture features were measured on 43 images of known malignant masses and 28 images of known benign masses. A genetic algorithm was used to select the optimal subset of k features based on A_z scores where k is a natural number. The leave-one-out A_z score for the optimal k features ranges from 0.80 to 0.95 for $k = 2, 3, ...12$. Since feature space reduction can result in optimistic estimates of classifier performance, the statistical significance of these scores were estimated by computing the empirical distribution of A_z scores in the context of the experimental parameters. For $k = 6, 7, 8$, the A_z scores were found to be significant at the $p = 0.05$ level.

1. Introduction

In the field of computer-assisted diagnosis, many authors have demonstrated that the texture of image intensity surfaces of screening mammograms provides information regarding the disease state of tissue [3, 4, 5]. Generally, these texture features are ones that are not seen by radiologists during visual inspection of the mammogram and are not based on models of the appearance of cancer in mammograms. Accordingly, the nature of texture features that are likely to provide positive predictive power is not well constrained. The result is that researchers are obliged to search far afield in order to discover optimal combinations of features. In addition, obtaining large numbers of training images on which to base the development of algorithms is not trivial and so a natural consequence is that studies comprise relatively small numbers of training images compared to the dimension of the feature space [2, 4].

Classification based on large number of features and a small training set can be optimistically biased. It can be shown that if the dimension of the feature space is greater than, or equal to, one less than the number of training images, then for any assignment of the training images into two groups, there exists a hyperplane which separates the two groups perfectly. Moreover, the hyperplane can be chosen so that the distance between an image in the feature space and the hyperplane (magnitude of the discriminant score) is the same for each training image.

One way to overcome this problem is to extract from the original feature space, a low-dimensional subspace that is realistic with respect to the size of the training data set. Selecting an arbitrary low-dimensional subspace defeats the purpose of considering many features, so the natural choice is a subspace that is optimal in some sense with respect to distinguishing between benign and malignant cases. However, the performance of the selected subspace in terms of classification is bound to be high. This is because in order to find the optimal subset, many different combinations of features are tested. It is expected that some of them will have a performance higher than average while others below average. From this collection, the feature combination that has the highest performance is selected, hence the performance will be high. The question is: is the performance of the optimal feature subspace selected greater than could be expected by chance?

The above question can be answered by performing a significance test. This will require knowledge of the distribution of the maximal performance scores obtained by repeating the selection process described above many times for data where there is no difference between the two groups. The distribution of these maximal A_z scores is not known and so was estimated using simulations.

Here we report on an experiment in which 260 texture features were measured on 71 training images. A genetic algorithm was used to select the k-dimensional feature subspace that is optimal with respect to A_z score for $k = 2, \ldots, 12$. The significance of the A_z score was measured by constructing empirical distributions of A_z scores for each k based on the full feature selection process.

2. Methods and materials

2.1 Data set

The data set comprises a total of 71 screening mammograms of which 43 contain malignant masses and 28 contain benign masses. The mammograms were obtained from the archives of *BreastScreenSA*, the South Australia branch of the National Screening Program in Australia. All malignant masses were biospy proven and the benign cases had a three years elapse time showing no sign of malignance. As the primary objective of the project is to assist diagnosis of clinically difficult cases, only the recall cases were included in the data set.

Electronic copies of the selected mammograms were acquired with a *Lumisys Lumiscan 150* laser digitiser. The resulting images have a spatial resolution of 50 μm and a depth resolution of 12 bits (4096 gray-level resolution). The images were reviewed and annotated by a radiologist experienced in mammography. Corresponding to the radiologist's annotation, regions of interest (ROIs) with a centering or near-centering mass were located. The size of each ROI is 1024×1024 pixels at full spatial resolution.

2.2 Texture measures

A total of 260 texture features were measured. These included 12 features based on image energy (see below), 8 based on gradients, and 240 based on co-occurrence matrices.

2.2.1 Textures Based on Co-occurrence Matrices

The co-occurrence matrix at distance d and direction θ is the array, P, where $P(i, j)$ is the joint probability that a pixel has image intensity value i and that the pixel at distance d in direction θ has value j. In addition to a choice of direction and distance, a co-occurrence matrix also requires a choice of quantisation of image intensity values. If the range of image intensity values is quantised to q bins, the co-occurrence matrix will be of size $q \times q$.

Co-occurrence matrices were constructed for distances $d = 11, 15, 21, 25, 31$, directions $\theta = 0, \pi/2$ and quantisation resolutions $q = 400, 100, 50$ for 40×40 half overlapping blocks in the straightened border region. The straightened border region, also called the rubber band straightened image, is an 80 pixel wide ring about the mass [4]. The directions $\theta = 0, \pi/2$ were chosen because they represent the directions perpendicular and parallel to the boundary of the mass. Radial structures near the mass boundary are known signatures of malignant masses. These 30 co-occurrence matrices were computed on two versions of the straightened border region. The first, called the polygon method, is found by connecting user defined points by line segments. The second, called the threshold method is found by finding a threshold for the ROI semi-automatically [2]. Hence a total of 60 co-occurrence matrices were constructed for every 40×40 block. The number of blocks varied from image to image depending on the size of the straightened border region, which, in turn varied according to the size of the mass.

For every co-occurrence matrix, the inverse distant moment (IDM) was computed according to the following formula.

$$IDM = \sum_{i=0}^{q-1} \sum_{j=0}^{q-1} \frac{1}{1 + (i-j)^2} P(i, j) \qquad (1)$$

For fixed values of d, θ, q, and choice of boundary method, the distribution of IDM values for all the 40×40 blocks in the straightened border region was recored. The first four moments of this distribution were recorded as features on which to base classification. Hence there were a total of 240 features based on co-occurrence matrices.

2.2.2 Intensity Gradient Features

The mass border was determined in each ROI using a polygon method. Background subtraction was performed. The geometric center of the polygon was used to define an 80 pixel wide annulus containing the border of the mass. The ROI was subsampled by a factor of 5 reducing the size from 1024×1024 to 204×204. The directional derivatives of the image intensity surface were computed in the directions both normal and tangential to the mass boundary. The first four moments of the distributions of the magnitudes of these directional derivates result in eight gradient features.

2.2.3 Local Image Energy Features

The local energy image y was computed from the image x by

$$y_{i,j} = \frac{1}{(2m+1)(2n+1)} \sum_{h=-n}^{n} \sum_{k=-m}^{m} x^2_{i-h,j-k}. \qquad (2)$$

For the region within the mass, the values $m = n = 12$ were used to produce an energy image restricted to the mass region. Two energy images were derived from the straightened border region. One with values $m = 3$ and $n = 10$, and the other with values $m = 10$ and $n = 3$. These choices were made to enhance features normal to the mass boundary in the first case, and tangential to the boundary in the second case.

For each of the three resulting energy images, the first four moments of the distribution of energy values were recorded resulting in a total of 12 image energy features on which to base classification and an over all total of 260 features from all three classes of features combined.

2.3 Genetic algorithm

A genetic algorithm [1] was used for feature subsets selection. The fitness criterion is based on the area under the receiver operating characteristic (ROC) curve, A_z, computed using the trapesoidal rule. (Technically, A_z referred to the area under a binormal ROC curve.) The genetic algorithm was initialised with a population of 1000 and is allowed to evolve over 500 generations. The mutation rate was set to 0.1. For each generation, the chromosomes with A_z score higher than the average A_z score of the current generation were retained in the parent pool. The remaining chromosomes were deleted from the population.

3. Classification results

The classification performance was evaluated using ROC methodology and the area under the ROC curve A_z was measured. As the optimal number of features k is not known a priori, classification using a range of feature number was performed. Figure 1 shows the training and the leave-one-out cross-validated A_z scores corresponding to $k = 2, 3, ...12$. The feature subsets correspond to the leave-one-out A_z scores are shown in Table 1. None of the optimal feature subsets include intensity gradient features, and therefore, these are not shown in Table 1.

4. Statistical significance estimation

In estimating the statistical significance of the classification results, the null hypothesis was that there is no difference between the two groups with respect to the k selected features. The null hypothesis implies that the observed classification results are no better than would be expected by chance. An empirical distribution of the maximal A_z scores based on the null hypothesis was simulated for each k. This was done by using a bootstrap method to generate 500 different 260 dimensional feature spaces with 71 data points and randomly assigning the data points to one group of 28 and the other of 43. For each of the 500 feature spaces, the genetic algorithm was used to search for the optimal k-dimensional subspace where $k = 3, 4, ...10$, and the associated optimal A_z score was recorded. Figure 2 shows the training maximal A_z distributions for $k = 3$ and $k = 10$. For $k = 4, 5, ...9$, the distributions were intermediate to the ones shown in the figure.

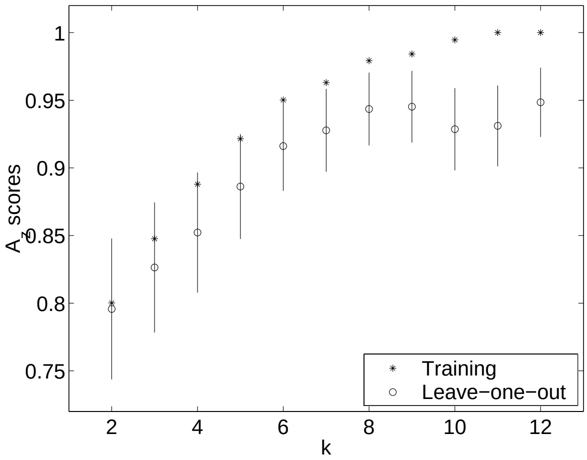


Figure 1. The classification results, both training and cross-validated A_z scores, are plotted against the number of features k. The cross-validated A_z scores are shown with error bars of one standard deviation.

The statistical significance of the cross-validated A_z scores are the ones of interest since the cross-validated A_z scores provide a better (less biased) estimate of the classification performance. In order to estimate such statistical significance, the cross-validated A_z scores should be compared to the cross-validated A_z score distributions. Unfortunately, the cpu time needed to compute the cross-validated A_z score distributions was prohibitively large. Instead, the training A_z distributions were used. This gives a conservative estimate of the significance because A_z scores based on training data are positively biased when compared to the leave-one-out A_z scores. The statistical significance estimates of the leave-one-out A_z scores $k = 3, 4, ...10$ are shown in Figure 3. For $k = 6, 7, 8$ the A_z scores were found to be significantly larger than predicted by the null hypothesis at the $p = 0.05$ level.

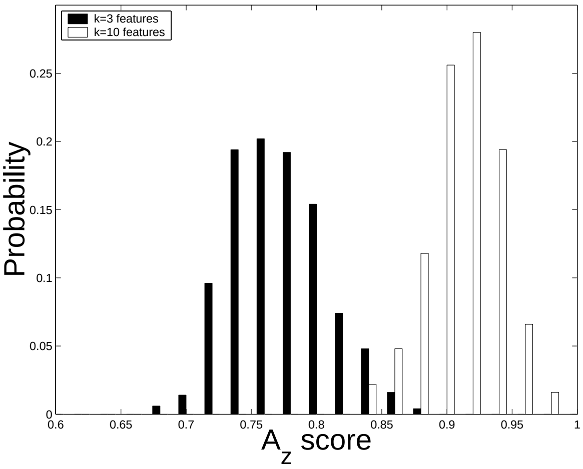


Figure 2. The empirical distributions of the training A_z score for $k = 3$ and 10. Each distribution consists of 500 data points.

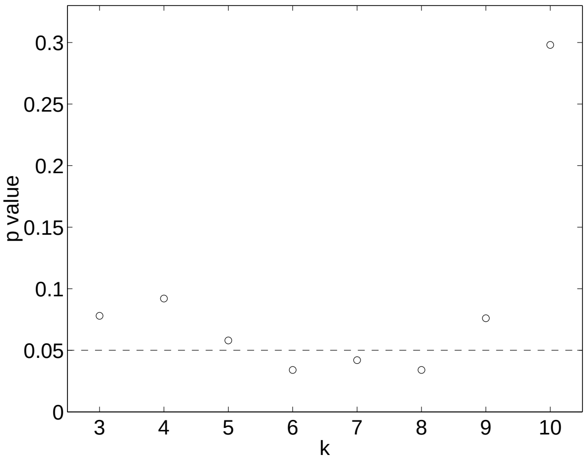


Figure 3. For each k, the p-value of the A_z score was computed by generating an empirical distribution of A_z scores (see text). For $k = 6, 7, 8$ the A_z values are significant at the $p = 0.05$ level.

5. Discussion and conclusion

Drawing conclusion regarding classification experiments comprising data sets that are small in size relative to the dimension of the feature space is always tenuous. Merely reporting the classification scores without incorporating the bias originating from the optimisation steps used to arrive at these scores does not provide sufficient information to judge the classification. Reporting confidence intervals for the classification performance score still does not acknowledge the bias of the method used to arrive at the score. This is borne out by the fact that the A_z score for some values of k were high but not significant. For example, for $k = 10$, $A_z = .9286$ but the p-value was near 0.3. By generating an empirical distribution of classifier performance values, it is possible to address the question of the significance of the classification performance measured experimentally.

References

[1] J. Holland. *Adaptation in natural and artificial systems.* Ann Arbor: The University of Michigan Press, 1975. Reprinted: Cambridge, Massachusetts: MIT Press, 1992/1994.

[2] G. Lee and M. Bottema. Classification of masses in screening mammograms as benign or malignant. In M. J. Yaffe, editor, *Proceedings of the 5th International Workshop on Digital Mammography, June 11-14, 2000, Toronto, Canada*, pages 259–263. Madison, Wisconsin: Medical Physics Publishing, 2001.

[3] I. Magnin, A. Bremond, F. Cluzeau, and O. C. Mammographic texture analysis - an evaluation of risk for developing breast cancer. *Optical Engineering*, 25(780–784), 1986.

[4] S. Sahiner, H.-P. Chan, N. Petrick, M. Helvie, and M. Goodsitt. Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis. *Medical Physics*, 25(4):516–526, 1998.

[5] D. Thiele, T. Johnson, M. McCombs, and L. Bassett. Using tissue texture surrounding calcification clusters to predict benign vs malignant outcomes. *Medical Physics*, 23(4):549–555, 1996.

Table 1. Optimal feature subsets corresponding to the leave-one-out A_z scores in Figure 1. Table entries are the values of k for which that feature appeared in the optimal k feature subset. Q, d and θ are parameters of the co-occurrence matrices where Q is the gray-level scale quantisation, d is distance in pixels and θ is the direction measured anti-clockwise with 0 pointing vertically downward. M1, M2, M3 and M4 are the first 4 moments of the distribution of the inverse difference moment measured on co-occurrence matrices. Local images A,B and C are mass center region 25 $\times$ 25 pixels, straightened border regions 7 $\times$ 21 pixels and 21 $\times$ 7 pixels, respectively. m1, m2, m3 and m4 are the first four moments of the distribution of the energy values.

Co-occurrence matrix based features											Local image energy features				
			Border region by threshold method				Border region by polygon method								
Q	d	θ	M1	M2	M3	M4	M1	M2	M3	M4		m1	m2	m3	m4
400	31	0							3	4	A		7-12	9-11	
		$\pi/2$													
	25	0					8,10,12								
		$\pi/2$													
	21	0			2,4-12	3			2						
		$\pi/2$													
	15	0		12	11	8-10									
		$\pi/2$	6-7,12												
	11	0													
		$\pi/2$		4											
100	31	0					5-10,12				B				
		$\pi/2$													
	25	0					6-7,9								
		$\pi/2$	11		11										
	21	0				12									
		$\pi/2$													
	15	0				12									
		$\pi/2$	8-11	5											
	11	0													
		$\pi/2$	8-10	11	5-7,12										
50	31	0					11			11	C			12	
		$\pi/2$		3	4										
	25	0						11							
		$\pi/2$	8-10	5-7,12											
	21	0				12									
		$\pi/2$	10												
	15	0													
		$\pi/2$													
	11	0													
		$\pi/2$													

(This page left blank intentionally)

Whole of Word Recognition Methods for Cursive Script

C. Weliwitage
S2114330@student.rmit.edu.au

A. Harvey
harvey@rmit.edu.au

A. Jennings
ajennings@rmit.edu.au

School of Electrical & Computer Systems Engineering, RMIT
GPO Box 2476V, Melbourne, Victoria 3001, Australia

Abstract

Most cursive script recognisers, segment the words into characters, either prior to recognition or during recognition. Whole of word recognition removes the needs for segmentation of the word into characters, eliminating problems associated with poor placement or missing segmentation points. The clear problem with this is that instead of a finite alphanumeric A-Z, 0-9 vocabulary, an unbounded word vocabulary is needed for the unrestrained case. However, in many cases, the application context means that there will be a strictly finite number of words in the application vocabulary. Therefore word recognition becomes feasible. Some examples are signature recognition and the words indicating the dollar amount on a cheque.

This paper examines the current methods of whole word or holistic word recognition methods giving special attention to useful features in recognition algorithms. Some useful features mentioned in the literature are ascenders, descenders, holes, loops, near loops, number and direction of strokes, the direction and orientation of the outer contour of the word, endpoints, cross points, and word length.

An attempt is made to introduce a new feature, 1D word profile, based on horizontal average of the word image. Method is tested with a sample of 72 machine printed cursive words and the results are compared with existing holistic features. Applying the technique for handwriting is under evaluation.

Current Word Recognition Methods

According to a survey on off-line cursive word recognition by **Steinherz, Rivlin and Intrator [1]**, features useful in recognition of off-line segment-free recognition of cursive word recognition can be classified into three categories based on representation level, ie:

1. low level
2. medium level
3. high level

Low level features include, smoothed traces of the word contour, pieces of strokes between anchor points, edges of the polygonal approximation etc.

Medium level category is an aggregation of low level features to serve as primitives. Medium level features are continuous in nature in contrast to low level features.

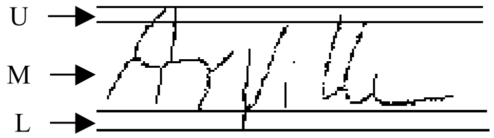
High level features are holistic or global features such as ascenders, descenders, loops, i dots, t strokes etc.

The paper summarises the different algorithms proposed for off-line cursive word recognition. They include minimum edit distance calculations based on dynamic programming, Hidden Markov Models or other specialised methods.

Govindaraju and Krishnamurthy [2] presents an algorithm which uses temporal information derived from off-line word images in the form of uptrends and downtrends of each stroke for the holistic recognition of off-line cursive words. This method basically applicable, only to cursive words and small lexicons. Method represents the off-line cursive word as a set of strokes. Each stroke is traversed to extract global contour features relating to the upstroke and downstroke movements of the pen. First the binarised word image is skeletonised using a standard thinning algorithm. Then the image is subdivided into three zones: Upper, Middle and Lower [Fig. 1]. For the recognition process, an array of inception and terminal points of each stroke is generated. A feature vector [Fig 2] is created using the peak and valley points identified along the word contour. The attributes of the feature vector subset of each stroke piece between two feature points are as follows:

1. Orientation (Up or Down)
2. Slope of Stroke
3. Length of Stroke
4. Stroke piece start zone (Upper, Middle and Lower)
5. Stroke piece end zone (Upper, Middle and Lower)

To minimise the Influence of ligatures, components of the vectors, which have the same start and end zones and the slope that is almost zero are deleted.



U – Upper zone
M – Middle zone
L – Lower zone

Fig. 1 Reference lining on word image (Ref 2)

$$\bar{F} = \left\{\begin{array}{ccccc} 1 & 47 & 5525 & 2 & 1 \\ 0 & 71 & 1145 & 1 & 2 \\ 0 & 3 & 257 & 2 & 2 \\ 1 & -87 & 485 & 2 & 2 \\ 0 & 26 & 5 & 2 & 2 \\ 0 & 78 & 2197 & 2 & 3 \\ 1 & -85 & 2417 & 3 & 2 \\ 1 & -55 & 3826 & 2 & 1 \\ 0 & -74 & 2696 & 1 & 2 \\ 1 & -74 & 2696 & 2 & 1 \\ 0 & 10 & 457 & 1 & 1 \\ 0 & 82 & 1885 & 1 & 2 \\ 1 & -80 & 148 & 2 & 2 \\ 0 & -50 & 485 & 2 & 2 \\ 1 & 37 & 269 & 2 & 2 \end{array}\right\}$$

where:
Column 1: Orientation (Up: 1; Down: 0)
Column 2: Slope of stroke piece
Column 3: Length of stroke piece
Column 4 & 5: Stroke_piece_Start and End_zones
(Upper: 1; Middle:2; Lower: 3)

Fig.2 The feature vector matrix of the image in Fig. 1 (Ref 2)

Parisse [3] presents a method of use of simplified profiles of word shapes for the global recognition of off-line handwriting.

This method, first extracts the complete contour of a digitised word. By eliminating internal contours, the upper and lower parts of the contour of the image are obtained [Fig 3], and transformed into a series of vectors. The upper and the lower profiles extracted correspond to two series of vectors representing the top and the bottom of the word. Vectorization, ie. series of points obtained through contour extraction is transformed into a series of vectors, is the next process before attempting comparison. Dynamic time warping technique is used for comparison of two vector series, ie. upper profile of the word to be recognised with the upper profile of the known word.

This global method is limited to a smaller lexicon as training of each individual word is required. To generalise it to a larger lexicon, the use of sub profiles [Fig 4], that are the profiles of strings of two or three letters or n-grams are extracted and the extraction procedure is totally automatic. In contrast to global comparison of word profiles, recognition will be based in seeking of all the profiles of known n-grams in the shape of the unknown word using dynamic time warping algorithm.

Guillevic and Suen [4] propose a method for recognizing unconstrained, writer independent, handwritten cursive words belonging to a small static lexicon, ie amounts written in bank cheques. After pre-processing, slant correction mainly, amount segmentation into words and extraction of global features for the recognition module are performed.

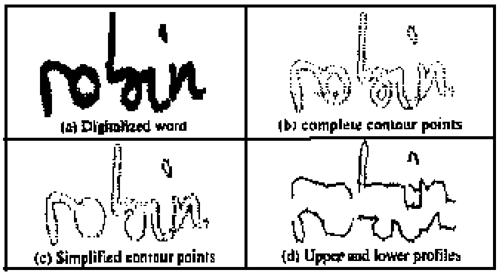

(a) Digitalized word
(b) complete contour points
(c) Simplified contour points
(d) Upper and lower profiles

Fig 3 Processing of the word 'robin' to extract upper and lower profile vectors (Ref 3)

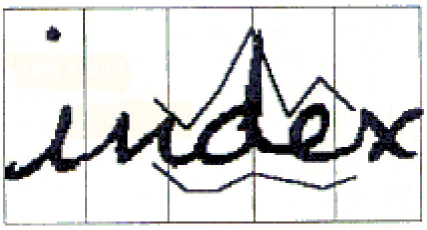


Fig 4 Extracted sub-profile for the 'de' in the word 'index' (Ref 3)

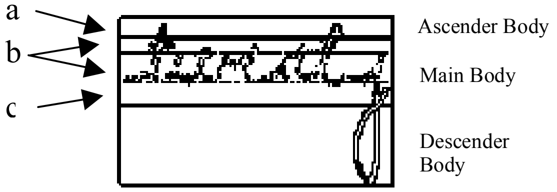
Seven types of global features are extracted from the word image, [Fig 5] and [Fig 6].
They are:

1. Ascenders
2. Descenders
3. Loops
4. Estimate of the word length
5. Vertical Strokes
6. Horizontal Strokes
7. Diagonal Strokes

Thresholds for ascenders and descenders are determined empirically and are expressed as a percentage of the main body height. Ascenders and descenders are detected by following the upper and the lower contour of the word respectively. Word length is estimated as the number of central threshold crossings. Strokes are extracted using mathematical morphology operations.

Input feature vector is different to class feature vector. Input or word feature vector consists of eleven features, relative position of ascenders, descenders, loops, strokes, number of ascenders, descenders, loops and the word length. When classifying, this input feature vector is converted to the class feature vector, which is build up of eleven sub vectors. Class feature vector is compared to the vectors obtained from the training sample.

Nearest neighbour classifier is the classifying technique used. Minimum shift distance is defined to get the distance between two feature vectors



a. Ascender threshold
b. Reference lines
c. Descender threshold

Fig 5 Ascender, Descender and Loop features (Ref 4)

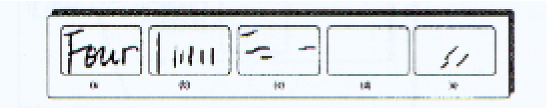


Fig 6 Stroke features: (a) Original Image (b) Vertical (c) Horizontal (d-e) Diagonal Features (Ref 4)

Madhvanath and Krpasundar [5], present a technique for pruning of large lexicons for recognition of cursive script words. This technique involves extraction and representation of down-ward strokes from the cursive word to obtain a generalised descriptor, which is matched with ideal descriptors.

A new approach to the method presented in [11] is used in obtaining shape descriptors (M-medium, A-Ascender, D-Descender, F-f-stroke and U-unknown), from the down ward strokes of off/on line strokes. In this approach, each downward stroke has been represented by an ordered pair (u, l) where u and l are in the range [-1, +1], and a word is represented as a sequence of such (u, l) pairs. Limiting contour extrema has been used to approximate the end points as shown in Fig 7.

To compute the distance between the descriptor extracted from the image and the 'ideal' descriptor corresponding to a given word (ASCII string derived), elastic matching technique is used and implemented using a trie-representation, ie. organising the lexicon entries and their ideal descriptors as a trie of stroke classes.

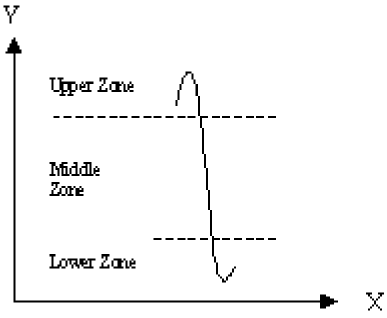


Fig 7 Upper and lower extensions of a stroke (Ref 5)

Guillevic and Suen [6], describe a fast reader system in which the recognition is done at the sentence level. Information from the graphical input is supplemented with the knowledge of context, orthography, syntax and semantics as the skilled human reader does in text reading. Method is tested on bank check processing. Theoretically this method is more robust against spelling mistakes, missing letters, unreadable letters etc.

The primary features extracting should be invariant enough and at the same time discriminative enough. So the features selected here are

1. Loops
2. Near loops
3. Ascenders
4. Descenders
5. Horizontal and vertical strokes

When these primary features are not sufficient, secondary features such as the characters that are adjacent to blank spaces and the estimation of the number of characters in an input word are useful.

Cai and Liu [7] present a new method to automatically determine the parameters of Gabor filters to extract structural features from word images. Features used in this system are the parameters of word image line segments, ie orientation, length and line centroid. Extraction of these line segments would be based on the output of the Gabor filter. Since it is difficult to order 2D word image features in 1D domain, all line segments in a word are divided into eight groups according to their orientation and each group is further divided into three sub-groups based on four base lines. Dynamic programming is used to calculate the distance between two words.

Madhvanath and Govindaraju [8], discuss the use of holistic features in their address classifier implemented at CEDAR. Features used by the system are the word length, number and positions of ascenders, descenders, loops and points of return. Macro features, or composite features such as 'ff' and 'ty' are also extracted and used in the classifier to enhance the scores. Feature equivalence rules provide means of normalization among different writing styles.

Madhvanath, Kim and Govindaraju [9], present a method of chain code based representation and manipulation of hand written images. Techniques that are applicable to word level recognition as well as image level and character level , are described.

For chain code representation, binary image is first scanned and the contour is traced and expressed as an array of contour elements which contain x,y coordinates of the pixel, slope/direction of the contour into the pixel and auxiliary information such as curvature. Slope convention is as shown in Fig 8.

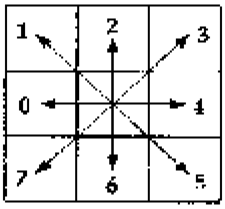


Fig 8 Slope convention ***(Ref 9)***

Determination of following word recognition techniques are described:

- Upper and lower contour
- Local contour extremas
- Reference lines
- Word length

Useful features such as word length, number and the location of ascenders and descenders are extracted from the extremas of the upper and lower contours of the word image.

Madhvanath and Govindaraju [10], discuss the method of applying holistic recognition techniques to a large, dynamic lexicon of handwritten words. In the past, these methods are mainly applied to small, static lexicons. For large, dynamic lexicons, this approach can be used for lexicon reduction which will eventually improve computational efficiency as well as combining this technique serially with analytical classifiers will improve the performances of word recognizers.

In this paper, two different lexicon reduction methods are described.

1. Use of constrained, bipartite graph matching scheme to match perceptual features such as ascenders, descenders, for unconstrained handwritten words.
2. A system that operates on pure cursive script, which captures relative heights of downstrokes in the word and form a string descriptor and matched with lexicon entries using a syntactic matching scheme [12].

In first scheme, perceptual features, either scalar (eg. word length) or positional (eg. ascenders) are collected. A confidence and a weight is associated with every feature type. Features used in the system are:

- Natural word length
- Ascenders
- Descenders
- Facts, ie. Description of the existence of certain features in specific regions of the word

The images and the lexicons are represented by wordgraphs, or sets of feature nodes, and a match between nodes of the image and of a given lexicon entry is obtained using constrained bipartite matching.

In second method, the word contour is extracted and represented as chain code. Downstrokes are then extracted from the contour and a shape descriptor is created using the relative heights of downstrokes. This descriptor is matched against predicted descriptors.

New Feature for holistic HWR based on centroid running average of word profile

A novel feature is introduced in this paper, which will be useful in holistic word recognition. Preliminary testing has been done using printed script characters, and applying this method to handwriting is under investigation.

As the first stage of the method, the average vertical distance of the pixels in each column of the word image is calculated and converted to a 1D profile of the word image [Fig 10]. Running average of this profile is determined and used as the new feature. For the calculations of this average, centroid or the base line is selected as the peak line of row-wise histogram of each word image.

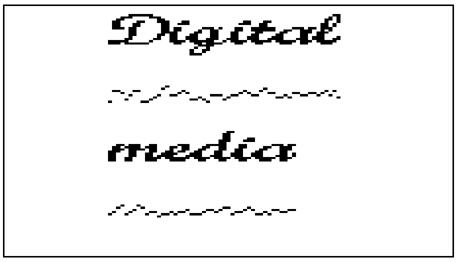


Fig. 10 Abstraction of 1D word profile

Results are analysed and compared with the word length feature, which is a widely used **basic** feature in holistic word recognition. Table 2 summarises the preliminary statistics of the results obtained for a sample of 72 words. Results are analysed for a lexicon of 10 words selected randomly. Table 3 shows few examples of word classes having the same word length and combination of word length and the running average of 1D centroid profile improve the discrimination factor of these word classes.

Conclusion

Existing methods of offline whole-word recognition have been summarised in this paper. These are listed in references [2] to [10]. According to the survey, most useful features for the holistic word recognition are, word length, number and positions of ascenders/descenders, holes, loops, near loops, number and direction of strokes, information on upper and lower contour of the word profile, endpoints and cross points.

A new feature method based on the word horizontal average is presented here. Preliminary studies with machine generated cursive text show promising results. According to the results shown in Table 2 and Graph 1, it can be proved that this new feature is as powerful as word length feature in HWR. Most importantly, as shown in Table 3, combination of these two features will enhance the discrimination factor among word classes.

An evaluation of this method using handwritten cursive script is now in progress.

Table 1: A summary of methods presented in the literature under survey

Ref.	Pre-Processing Method(s)	Recognition Method(s)	Test Data	Result
2	Thinning, zoning	Not Given	552 test words, 10 word lexicon	80% -92%
3	Upper/lower contours, n-grams	Dynamic time warping, dynamic programming	16,200 words	50% - 96%
4	Slant correction	Nearest neighbour classifier, genetic algorithms	5,322 training words, 2, 515 test words	72% - 98.5%
5	Down-ward stroke descriptor	Elastic matching, Trie implementation	21,000 words lexicon, reduced to 1000 words	> 95%
6	Line removal, slant correction	Mathematical Morphology	Not Given	Not Given
7	Slant/tilt correction, baseline finding, image normalisation, line width calculations	Dynamic programming, fuzzy inference	105 training words, 113 test words	64.6% – 94.7%
8	Baseline skew correction, character slant correction, finding reference lines	Euclidean distances between the feature vector and the description vector	103 images of street names	88% - top choice, 97% within top three
9	Noise removal, Slant correction, smoothing contours	Chain code processing	768 images of city names, lexicon of 1000 random city names	> 98% for the reduction of half of the lexicon
10	Chain coding, skew and slant correction	Constrained bipartite matching, syntactic matching	1. 768 images of city names, lexicon of 1000 words 2. 825 cursive words	1. > 98% for the reduction of 50% of lexicon 2. 75% for the reduction of 99%

Table 2: Preliminary results showing a comparison of word length feature with the suggested new feature for a page of 72 machine generated cursive script with a lexicon of 10 words (15 words belong to the lexicon)

			Feature 1 (Word Length)			Feature 2 (Running average)		
	Word Class	No. of Occurrences	Word Length in pixels	No. of detected words with similar features	% Discrimination[1]	Running average of 1D profile	No. of detected words with similar features	% Discrimination[1]
1	image	2	48	3	67%	-6	3	67%
2	media	1	48	3	33%	11	7	14%
3	extract	1	56	1	100%	14	1	100%
4	Digital	1	59	1	100%	8	3	33%
5	specific	1	60	3	33%	-5	1	100%
6	features	3	63	1	100%	-4	7	43%
7	difficult	1	67	2	50%	13	1	100%
8	different	1	70	1	100%	7	1	100%
9	database	3	72	2	50%	15	2	50%
10	corresponding	1	112	1	100%	-21	1	100%

Graph 1: Graphical representation of the results shown in Table 2

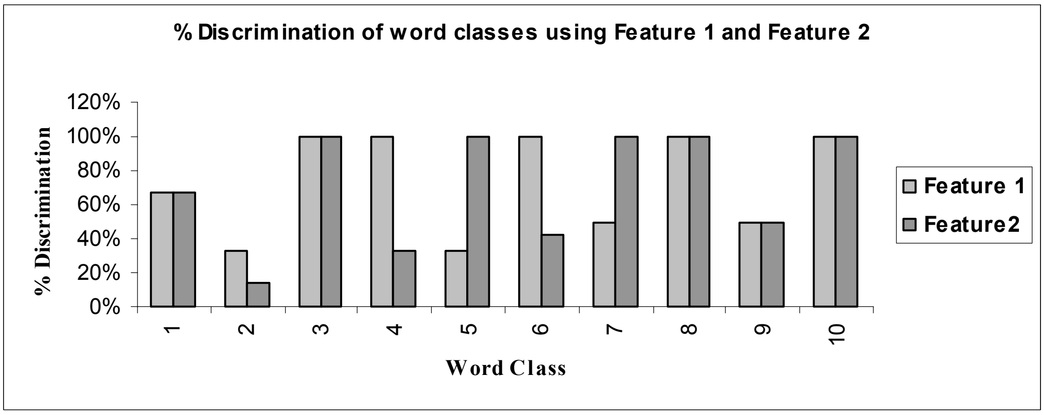


Table 3: Examples of few word classes having same word length

Word Class	Feature 1	Feature 2	% Discrimination using both features
are	26	0	100%
for	26	-2	100%
from	40	3	100%
these	40	6	100%
image	48	-6	100%
media	48	11	100%
specific	60	-5	100%
domain	60	8	100%
format.	60	1	100%

References:

[1]. T Steinherz, E Rivlin, N Intrator, Off-line cursive word recognition – A survey, International Journal on Document Analysis and Recognition, Volume 2, Issue 2-3, 1999, Pages 90-110

[2]. V. Govindaraju, R. K. Krishnamurthy, "Holistic handwritten word recognition using temporal features derived from off-line images", Pattern Recognition Letters, 1996, Vol 17, No. 5, pg 537 – 540.

[3]. C. Parisse ,"Global word shape processing in off-line recognition of handwriting", IEEE transactions on pattern analysis and machine intelligence, 1996, Vol 18 , No 4, pg 460 – 464.

[4]. D Guillevic , C Y Suen, "Cursive script recognition applied to the processing of bank cheques", Proceedings, International Conference on Document Analysis and Recognition, 1995, pg 11-14.

[5]. S Madhvanath, V Krpasundar, "Pruning large lexicons using generalised word shape descriptors", Proceedings, International Conference on Document Analysis and Recognition, 1997, pg 552-555.

[6]. D Guillevic , C Y Suen, "Cursive script recognition : A fast reader scheme ", Proceedings, International Conference on Document Analysis and Recognition, 1993, pg 311-314 .

[7]. J Cai, Z Q Liu , "Off-Line Unconstrained Handwritten Word Recognition", Proc. 1996 Australian New Zealand conf. On Intelligent Information Systems, 1996, pg 199-202.

[8]. S Madhvanath, V Govindaraju, "Using Holistic Features in Handwritten Word Recognition", Proceedings of U.S. Postal Service 5th advanced Technology Conference, 1992, pg 183-198.

[9]. S Madhvanath, G Kim, V Govindaraju, "Chaincode Contour Processing for Handwritten Word Recognition", IEEE transactions on pattern analysis and machine intelligence, 1996, Vol 21 , No 9, pg 928 – 932.

[10]. S Madhvanath, V Govindaraju, "Holistic Lexicon Reduction for Handwritten Word Recognition", SPIE, Vol 2660, pg 224- 234.

[11]. S Madhvanath and S N Srihari, "Effective reduction of large lexicons for recognition of off-line cursive script, Proceedings of the Fifth International Workshop of Frontiers in Handwritten Recognition, 1996.

[12]. G. Seni, N. Nasrabadi and R. Srihari, "An online cursive word recognition system", Proceedings of the IEEE CVPR-94, Seattle, Wa, Jun 17-23, 1994.

[1] *% Discrimination =(no. of Occurrences of word class/ no. of detected words with similar features)*100*

Intrinsic correspondence using statistical signature-based matching for 3D surfaces

Birgit M. Planitz and Anthony J. Maeder
CRC for Satellite Systems
Queensland University of Technology
GPO Box 2434 Brisbane, QLD 4001
{b.planitz, a.maeder}@qut.edu.au

John A. Williams
School of ITEE
University of Queensland
St Lucia, QLD 4072
jwilliams@itee.uq.edu.au

Abstract

A wide variety of applications including object recognition and terrain mapping, rely upon automatic three dimensional surface modelling. The automatic correspondence stage of the modelling process has proven challenging. Intrinsic correspondence methods determine matching segments of partially overlapping 3D surfaces, by using properties intrinsic to the surfaces. These methods do not require initial relative orientations to begin the matching procedures. Hence, intrinsic methods are well-suited for automatic matching.

This paper introduces a novel intrinsic automatic correspondence algorithm. Local feature support regions are described using distance and angular metrics, which are used to construct cumulative distribution function signatures. Local correspondences are hypothesised by comparing the signatures of two surfaces. A geometric consistency test is then applied to select the best local correspondences. Finally, registrations are computed from the remaining correspondences and the best alignment is selected. Results demonstrating the algorithm's accuracy in selecting correspondences for mutual partially overlapping surfaces, are presented. The algorithm's parameters prove robust, with only the local region size being surface dependent.

1 Introduction

Automatic correspondence is an important step in three dimensional (3D) modelling. Automatic correspondence is essential in applications where the position of the sensor, with respect to the scene, is unknown. A typical example is terrain mapping [8]. It is desired that the images of all views (i.e. 3D surfaces) are input into the modelling system, where they are automatically manipulated to form a 3D model of the object/scene. This section outlines the 3D modelling process, discusses intrinsic correspondence, and briefly highlights existing intrinsic correspondence algorithms.

The 3D modelling process consists of four main stages. First, a sensing device is used to obtain 3D surfaces of different views of an object/scene in the **data acquisition** stage. Secondly, the matching segments of different surfaces are found using a **correspondence** algorithm. Thirdly, the corresponding surfaces are aligned by applying a **registration** scheme. Finally, the aligned surfaces are merged to form a complete 3D model of the object/scene, in the **integration and reconstruction** phase. The automatic data acquisition, registration, integration and reconstruction stages have been more or less solved. Automatic correspondence however, has proven challenging.

Correspondence methods can be categorised as either *intrinsic* or *extrinsic* [12]. Intrinsic methods form correspondences by comparing the intrinsic properties of surfaces, whereas extrinsic methods form correspondences using the relative orientations between the surfaces being matched. Extrinsic methods require a rough initial alignment between the surfaces to converge to the correct solution [3, 5]. Intrinsic correspondence-registration methods automatically form these initial alignments within the algorithm. Therefore, intrinsic techniques are the key to developing automatic correspondence algorithms, because theoretically no user interaction is required. However, no *fully* automatic technique exists.

Some key intrinsic methods are highlighted as follows. The Random Sample Consensus based Data Aligned Rigidity Constrained Exhaustive Search method defines regions on one surface X, and searches for regions of similar size on the other surface Y [4]. Triangles comprised of selected control points make up the regions. A similar method is graph matching, whereby a graph using distances between points is constructed on X [6]. The algorithm then attempts to build the same (or part of the same) graph on Y. Methods such as spin-image and geometric histogram matching treat the correspondence problem slightly differently. The former creates signatures on each surface, that are based on the horizontal and vertical distances from selected points to every other point on the surface [9]. The latter examines the angles and vertical distances from a given mesh facet, to other facets within a predefined distance [2].

Aspects of the outlined correspondence methods are referred to in other sections of this paper. Intrinsic correspondence is detailed in the following section. Section 3 then introduces a novel intrinsic correspondence algorithm. Section 4 provides results of matches between mutual partially overlapping surfaces. Finally, Section 5 summarises the work outlined in this paper.

2 Background

As discussed developing an intrinsic method is the best solution to constructing an automatic correspondence method. In this section a typical intrinsic algorithm is outlined, and its main components are reviewed in detail.

2.1 Intrinsic Correspondence Dissected

Figure 1 illustrates a typical approach to intrinsic correspondence. This method assumes pairwise correspondence and registration. When registering a set of surfaces, it is sufficient to do the initial alignment in a pairwise manner, matching each surface with every other surface individually. The final accurate alignment between all surfaces can be obtained using a multiview registration scheme [14].

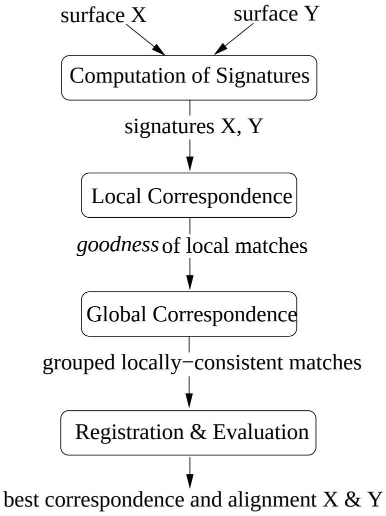

surface X
surface Y
Computation of Signatures
signatures X, Y
Local Correspondence
goodness of local matches
Global Correspondence
grouped locally−consistent matches
Registration & Evaluation
best correspondence and alignment X & Y

Figure 1. The steps in a typical intrinsic correspondence algorithm.

Figure 1 illustrates a procedure, where the number of potential correspondences between two surfaces are continuously reduced, such that only a few remaining hypothesised matches are passed to the registration-evaluation process. Essentially, this algorithm is an exercise in *correspondence pruning*. The first step of the pruning process is to select small local feature support regions on both surfaces X and Y. Then, these regions are given signatures based on their local features. The signatures of X are matched with the signatures of Y in the local correspondence phase, and the *goodness* of each match is passed to global evaluation procedure. Here, bad matches are discarded, reducing the number of possible local correspondences. The remaining matches are accumulated to provide evidence for consistent local matches. The sets of consistent matches are then used to compute registrations to bring X and Y into common coordinate systems, where their alignments are evaluated. The best alignment gives rise to the optimal correspondence between the two surfaces.

A more detailed review of the steps shown in Figure 1 is discussed in the following sections.

2.2 Signatures

Signatures of surface regions are constructed from intrinsic surface properties. The signature construction process encompasses two very important steps: choosing a viable descriptor of a region, and storing this descriptor in a signature that can easily be compared to other signatures. Some considerations when selecting descriptors, region sizes and signatures are highlighted below.

It is desired that a surface descriptor is robust, and unique to its local neighbouring region. Regions can be uniquely described using angular [2], distance [2, 4, 6], and differential [10] features. The following examples concern robustness. Distances between points are robust, as they use the original surface properties [11]. However, features such as normals, are less robust because smoothing is usually required so that the descriptors are accurate [5].

Selecting the size and shape of the local support regions, is a process that concurs with choosing a region descriptor. In some cases the regions' size and shape may be constant [2], and in others variable [4]. It is essential that each local support region is large enough to store a unique description of the area, but not too large, as the entire region may not be in the overlapping portion of the corresponding surfaces.

After selecting one or more surface descriptor(s) and defining the region size and shape, the descriptors of each support region must be stored as signatures. Typical signatures include graphs [6] and histograms [2]. The signatures must be of reasonable size to ensure that local matching is efficient. Signatures such as spin-images are less favourable because they require large data storage space [9].

Once the signature computation process is complete, the signatures of surfaces X and Y are passed to the local match algorithm.

2.3 Local Correspondence

The function of the local correspondence algorithm is to determine how well the signatures of surface X match with the signatures of surface Y. A suitable match metric is required for this operation. Examples include the Minkowski norm for shape distribution matching [11], and Bhattacharyya distance for histogram matching [2].

Local correspondences are generally computed in batches, as shown in Figure 1, where every signature of one surface is compared with every signature on the other. The *goodness* of all local matches is then passed to a global evaluation algorithm.

2.4 Global Correspondence

Global correspondence algorithms prune all local matches, so that only good locally-consistent matches are selected. An example is pruning a probability matrix, where the entries of the matrix are *probabilities* of matches between signatures on X and Y. The global correspondence algorithm is used to *thin* the p-matrix by only selecting matches with a probability greater or equal to an acceptance level $p0$.

There are a number of ways of selecting global correspondences. This process is often integrated into the local matching procedure, such that bad matches are discarded immediately [4], and not all matches are passed in a batch for global evaluation. However, analysing a batch of matches can be extremely valuable when using pruning methods, because the correspondences can then be evaluated at different acceptance levels.

The correspondences that pass global evaluation are a much smaller set of possible matches between two surfaces, and these are passed to a registration-evaluation algorithm.

2.5 Registration and Evaluation

The final analysis of potential matches between two surfaces is completed by evaluating how well two surfaces align when using selections of the remaining correspondences. Three corresponding pairs of regions must be selected to uniquely align two surfaces in a common coordinate system.

A number of registration-evaluation methods are used to select the best possible transformations to align two surfaces. In most cases all possible combinations of three corresponding region pairs are used to compute registrations between the surfaces. This is generally followed by another evaluation where extrinsic-type metrics [3, 5] are computed to further prune the correspondence space. The final registrations are examined by using methods such as an evidence accumulation scheme to test for consistent transformations [2], or a model based scheme to test for good alignments [4].

The best alignment between the two surfaces is selected as the final outcome of the intrinsic correspondence-registration algorithm.

3 Statistical Signature-based Matching

In this section, a novel intrinsic correspondence algorithm is introduced. The algorithm conforms to the structure outlined in Section 2, and the following subsections discuss signature selection, local and global correspondence, registration and alignment evaluation.

3.1 Signatures

In this correspondence algorithm, distance and angular descriptors, and statistical signatures are used to characterise regions on the surfaces being matched. It is assumed that the surfaces are stored as polyhedral (generally triangular) meshes. The following paragraphs discuss the signature derivation process, with respect to the mesh surfaces.

Two descriptors are utilised to describe a local region on the mesh. The $D1$ distance is a robust descriptor, and was derived for surface model matching [11]. The metric is the Euclidean distance between a centre vertex and points on the surface, in the local feature support region. The second metric is the $A1$ descriptor, which is the angle between the normal of the centre vertex and the normals of points, in the local feature support region. Angles between facet normals are used as feature metrics in geometric histogram matching [2]. Both metrics are used in a combined fashion to more uniquely portray each local support region on the surfaces being matched.

Local support regions are selected around each centre vertex as follows. First, the *border layers* of each mesh are determined. This is done by selecting the vertices on the border of the mesh (layer 1), then the vertices that connect to border vertices (layer 2), then the vertices that connect to layer 2 vertices (layer 3), and so on as shown in Figure 2. The mesh can now be evaluated at a certain level. For example, if there are many points on the mesh, it would be very inefficient to evaluate the potential match between every point on mesh X and every one on Y. Therefore, only the vertices of the innermost layers (say layer 3 and above are selected).

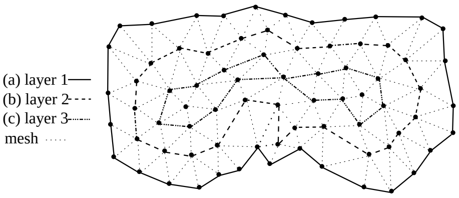


Figure 2. The border layers of a mesh, with (a) being the outermost layer, (b) the second, (c) the third and so forth.

Once the vertices to be evaluated have been selected, a neighbourhood radius is chosen (this is currently a user-defined value). This distance determines the surrounding neighbourhood of each vertex, by including all the points in the neighbourhood that fall into the sphere generated by the radius. The $D1$ and $A1$ metrics are then calculated for each local support region. The combination of distance and angular descriptors has proven powerful in correspondence algorithms [2, 9].

A separate signature is then built for both the $D1$ and the $A1$ metrics. The signatures are cumulative distribution functions (cdfs), and were selected in concurrence with the local match metric discussed in the following section.

3.2 Local Correspondence

A brute force local matching algorithm is employed, where the signature of every vertex under evaluation on mesh X is tested against the signature of every selected vertex on Y. The match metric employed is the Kolmogorov-Smirnov two sample test (KS-test). The KS-test tests whether two samples, that are drawn independently, belong to the same population [7]. The test statistic used in the two-sided KS-test is T, which is the greatest absolute distance between the two cdfs supplied for each match. The acceptance level, or probability of a match p, is calculated using T [7].

All $D1$ signatures of the two surfaces X and Y are compared, and all $A1$ signatures of X and Y are compared. The comparison results are stored in 2D probability matrices, P_D and P_A respectively. The respective elements of the matrices are then multiplied to form the p-matrix P, which is passed to the global correspondence algorithm for further evaluation.

3.3 Global Correspondence

The global correspondence method prunes local matches by examining geometric consistency. The first step of the global correspondence method however, is discarding local matches in the p-matrix that have probabilities below $p0$. The parameter $p0$ is the acceptance value that supports the hypothesis that two signatures belong to the same distribution.

Every possible combination of three local correspondences is then chosen from the remaining matches. The local correspondence pairs form triangles as shown in Figure 3. Geometric consistency must exist between the two surfaces if the formation of local correspondences are to be accepted. That is $d1_X \approx d1_Y$, $d2_X \approx d2_Y$, and $d3_X \approx d3_Y$. The measurement used to test similarity in the three distances is $|d_X - d_Y| \leq \tau$, where τ is a percentage (also user defined) multiplied by $max(d_X, d_Y)$. The varying nature of τ ensures that good matches based on larger triangles are more acceptable then those based on smaller ones. This increases the robustness of the algorithm, minimising the error introduced when considering small triangles as the best correspondences.

Every possible combination of three local correspondences is evaluated, and only the Q matches that pass the geometric consistency test are passed to the registration-evaluation algorithm, for further evaluation.

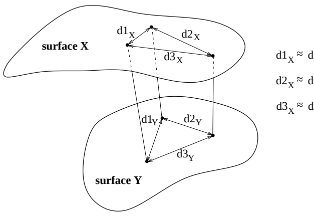


Figure 3. Three local correspondences between surface X and Y. The distance between the selected vertices on the surfaces must be similar for good geometric consistency.

3.4 Registration and Evaluation

Each possible combination of three centroids, of the Q remaining matches, is supplied to the alignment algorithm [1]. These registrations are then analysed visually. This process will soon be replaced by an extrinsic-type correspondence-alignment-evaluation scheme. A typical evaluation method is to compute the closest points from mesh X to mesh Y, and then determine the distances between them [13]. Finally, the number of point-pairs whose distances fall below a threshold τ are summed, and the match with the highest sum is selected as the best correspondence between the two meshes.

3.5 The Algorithm Summarised

The novel intrinsic correspondence method discussed in the previous sections is summarised in Figure 4.

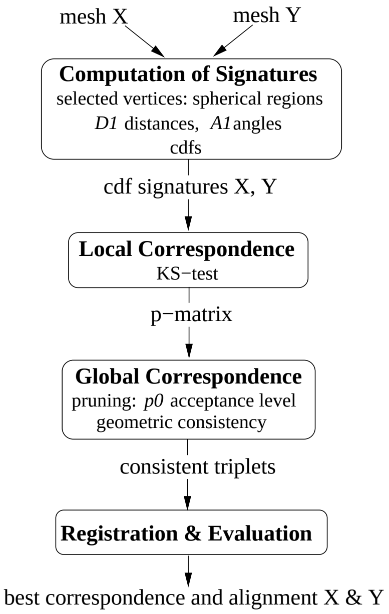


Figure 4. The steps of the statistical signature-based matching intrinsic correspondence algorithm.

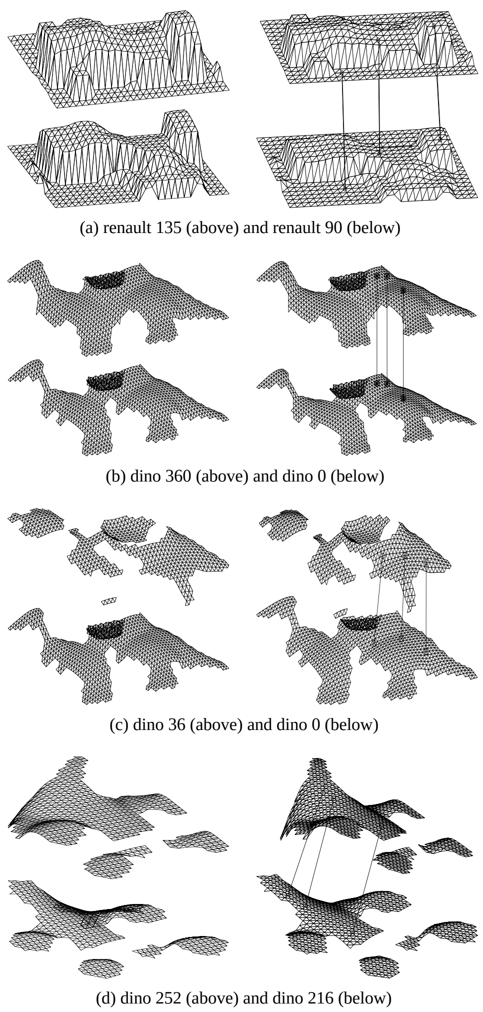
4 Results

In this section, the results of the statistical-based signature matching algorithm are illustrated for surfaces with mutual partially overlapping segments. Both visual results and an analysis of the effects of the parameters on the algorithm are highlighted.

The surfaces matched are triangulated meshes and are displayed on the left hand side in Figure 5. The best triplets of corresponding points are shown in the right hand column. Note that at least three matches are required to register two surfaces.

Figure 5(a) shows two views of a Renault figurine, seen at 90 and 135 degree viewing angles. The correspondence triplets demonstrate accurate matches between the two surfaces. Figure 5(b) shows two very similar views of a toy dinosaur (viewing angles 0 and 360 degrees). The parameter $p0 = 0.99999999$ was chosen to reduce the possible number of remaining matches to four. When two surfaces are almost identical, the algorithm results in many accurately matched local support regions. Layers four and above were chosen for the match, although far fewer could have been selected because of the similarities of the two surfaces.

Figure 5(c) illustrates the best match when two surfaces are matched that contain a smaller percentage of mutual partially overlapping segments (views 0 and 36 degrees of the toy dinosaur) . The triplet of matches are not as accurate as that in part (b), but correspond to regions of very similar surface variation. Figure 5(d) too shows the resulting correspondences when matching two surfaces contain a smaller percentage of mutual partially overlapping segments (views 216 and 252 degrees of a toy dinosaur). The triplet of best matches re-affirms the accuracy of the algorithm.



(a) renault 135 (above) and renault 90 (below)

(b) dino 360 (above) and dino 0 (below)

(c) dino 36 (above) and dino 0 (below)

(d) dino 252 (above) and dino 216 (below)

Figure 5. The surfaces tested (left), and the best resulting correspondence triplets (right).

Table 1 summarises the parameter selection for the surfaces used to test the algorithm. The table also includes the radius size selected for each match procedure. The radius size was constant for each of the dinosaur match procedures, highlighting the robustness of the algorithm. The radius size selection will be the most important factor when fully automating the intrinsic algorithm. Other parameters such as τ are robust for a variety of surfaces and do not require adjustment. The neighbourhood layer region is mainly selected to increase the computational efficiency of the algorithm, and will only affect the correspondences if too few points are

Parameters	renault (90,135)	dino (0,360)	dino (0,36)	dino (216,252)
$p0$	0.90	0.9999 9999	0.95	0.945
radius	50	25	25	25
layers	≥ 3	≥ 4	≥ 3	≥ 3
tau	0.05	0.05	0.05	0.05

Table 1. The values selected for the parameters of the novel correspondence algorithm.

selected for matching.

The algorithm is currently implemented in MATLAB and takes no more than 30 minutes to run for larger meshes (e.g. *dino 0* and *dino 360*, with 964 and 1038 vertices respectively), depending on the parameter selection. Note that MATLAB is very slow in comparison with C or C++. The final algorithm will be implemented in C or C++ to dramatically reduce the computation time.

The results in this section highlighted that the novel intrinsic correspondence technique accurately selects correspondences between partially overlapping surfaces. The algorithm provides results that can be input into an extrinsic algorithm to achieve accurate final correspondences. Some further analysis required before the algorithm is complete includes examining the effects of the algorithm's parameters, and the algorithm's efficiency.

5 Conclusion

This paper highlighted the importance of intrinsic correspondence techniques, and introduced a novel intrinsic algorithm. The new method is a signature matching technique that uses the $D1$ distances and $A1$ angular measurements as descriptors of local support regions. The signatures are cdfs, and are compared by the local correspondence algorithm. The comparisons are made using a KS-test and are stored in two probability matrices, whose respective elements are multiplied. The resulting p-matrix is pruned by accepting only those local matches which are greater or equal to the acceptance value $p0$. Global correspondences are then tested for geometric consistency to further reduce the match search space. Finally registrations are applied to the best matches, and visual evaluations of the best alignments are made.

The initial results of the statistical-based signature matching algorithm are promising. The algorithm provided accurate alignments that can be input as initial relative orientations in extrinsic methods. The algorithm's only surface dependent parameter is the local support region radius size, making the technique robust. Future work includes fully automating the parameter selection process and coding a registration-evaluation algorithm to finalise the automatic correspondence process.

ACKNOWLEDGMENT: The Cooperative Research Centre for Satellite Systems is established and supported under the Australian Government's Cooperative Research Centres Program.

References

[1] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-D point sets. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 9(5):698–700, Sept. 1987.

[2] A. Ashbrook, R. Fisher, C. Robertson, and N. Werghi. Finding surface correspondence for object recognition and registration using pairwise geometric histograms. In *Proc. European Conference on Computer Vision (ECCV98)*, pages 674–686, 1998.

[3] P. J. Besl and N. D. McKay. A method of registration of 3-D shapes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 14(2):239–256, Feb. 1992.

[4] C. Chen and Y. Hung. RANSAC-Based DARCES: A new approach to fast automatic registration of partially overlapping range images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(11):1229–1234, November 1999.

[5] Y. Chen and G. Medioni. Object modeling by registration of multiple range images. *Image and Vision Computing*, 10(3):145–155, Apr. 1992.

[6] J. Cheng and H. Don. A graph matching approach to 3D point correspondences. *Int. Journ. of Pattern Recognition and Artificial Intelligence*, 5(3):399–412, 1991.

[7] W. J. Conover. *Practical Nonparametric Statistics*. John Wiley & Sons, Inc., 1999.

[8] D. F. Huber and M. Hebert. A new approach to 3-d terrain mapping. In *Proceedings of the 1999 IEEE/RSJ Int. Conf. on Intelligent Robotics and Systems (IROS '99)*, pages 1121–1127. IEEE, October 1999.

[9] A. Johnson and M. Hebert. Surface registration by matching oriented points. *IEEE Proc. Int. Conf. on Recent Advances in 3-D Digital Imaging and Modeling*, pages 121–128, May 1997.

[10] P. Krsek, T. Pajdla, V. Hlavac, and R. Martin. Range image registration driven by hierarchy of surface differential features. In *Proceedings of the 22nd Workshop of the Austrian Association for Pattern Recognition*, pages 175–183, May 1998.

[11] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin. Matching 3d models with shape distributions. In *Int. Conf. on Shape Modeling and Applications 2001*, pages 154–166, Genova, Italy, May 2001.

[12] B. M. Planitz, J. A. Williams, and M. Bennamoun. Automatic correspondence of range images. In *The Fifth Asian Conference on Computer Vision (ACCV2002)*, pages 562–567, January 2002.

[13] G. Roth. Registering two overlapping range images. In *Proc. IEEE Conf. on 3-D Digital Imaging and Modeling*, pages 191–200, 1999.

[14] J. A. Williams and M. Bennamoun. Simultaneous registration of multiple corresponding point sets. *Computer Vision and Image Understanding*, 81(1):117–142, Jan. 2001.

Background Modeling For Tracking Object Movement

Yue Feng
School of Electrical and Computer Engineering
RMIT University, Australia.
yue.feng@innovonics.com.au

Alan L. Harvey
School of Electrical and Computer Engineering
RMIT University, Australia.

Andrew Jennings
School of Electrical and Computer Engineering
RMIT University, Australia.

Abstract

A background modeling method for tracking object movement in a cluttered scene is proposed in this paper. The emphasis of the system is on the analysis of background information gathered solely from an image sequence. Two main processes are implemented in this system. The first process uses a novel technique to extract an initial background model from an image sequence. The second process updates the background model, when the background situation changes at a later time. Extensive knowledge of segmentation and edge detection is used in the modeling. It is planned to implement the algorithm on a personal computer to produce a real time working system. The image capture rate was 10 images/second. The image sequence may contain the moving objects during the background extract in stage.

Keywords Image processing, background modeling, edge detection, difference image

INTRODUCTION

This paper presents a novel technique for a low-cost PC based read-time visual modeling system; called a background modeling system, for simultaneously tracking movement object, and monitoring their activities in monochromatic video. This system also constructs dynamic background models to isolate object movement. Background modeling system has been designed to work with monochromatic stationary video sources, either visible or infrared.

The remainder of the paper is organized as follows. Section 2 describes the segmentation that employs an adaptive double window modified trimmed mean (DW-MTM) filter for filtering the input image in our real time system and a local neighborhood edge operator in a low level image processing. In this section, we modified a novel cluster analysis technique to extract some interesting points to labeling object movement and issued background model. Section 3 describes the background model updating. Section 4, experimental results are reported and summarized a complete algorithm. Discussions and conclusions are provided in section 5.

A simple and commonly used background modeling method involves subtracting each new image from a model of the background scene and thresholding the resulting difference image to determine foreground pixels. The pixel intensity of a completely stationary background can be reasonably modeled with a normal distribution, and it can adapt to slow changes in the scene by recursively updating the model. However, those approaches have difficulty in modeling backgrounds in a complex and very varied environment such as some lighting changes. In this case, more than one process may be observed over time at a single pixel. In [1], a mixture of three normal distributions was used to model pixel values for traffic surveillance applications model road, shadow, and vehicle. In [2], pixel intensity is modeled by a mixture of K Gaussian distributions (typically, K is three to five). [3] uses a nonparametric background model by estimating the probability of observing pixel intensity values based on a sample intensity values for each pixel. [4] uses a model of background variation that is a bimodal distribution constructed from order statistics of background values during a training period. They apply to obtain the background model even if there are moving foreground objects in the field of view, such as walking people, moving cars, etc. Our background modeling method developed in our laboratory creates a background scene model which obtains a non-movement object binary image when there are moving foreground objects or if there are not moving foreground objects in the field of view.

The major features of background modeling are as follows:

- Leaning initial background model when the system starts up.
- Updating background model.
- Issue a background model for the future tracking.

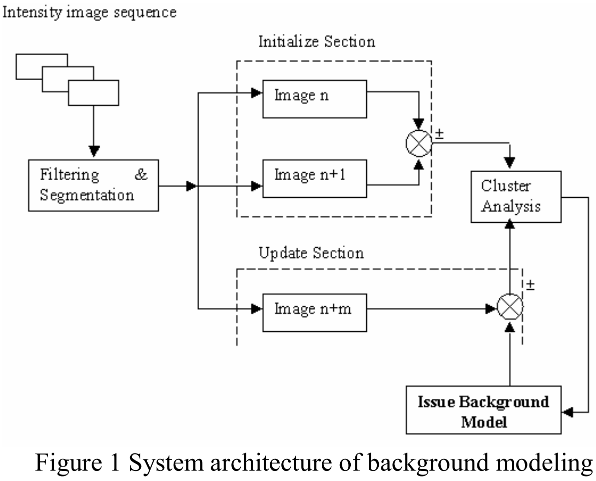
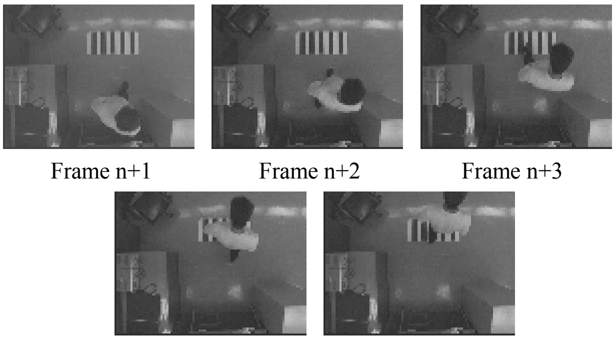


Figure 1 System architecture of background modeling

The block diagram in Figure 1 shows the system architecture for background modeling. In the first stage, a sequence of raw intensity images is first fed to a segmentation process to binaries and subtracts two input image to a binary image. In second stage, we perform a cluster analysis to obtain some interesting points from the binary image. Then, we issue a background model from the result of the cluster analysis.

The Image used in this research was taken from a vertical direction movement in the cluttered monitoring area. The image was digitized from a video camera with an initial pixel resolution of 768 × 576 pixels. The resolution was reduced to 384 × 288 pixels for processing. At this resolution, 384 × 288 pixels were equivalent to about 3 × 2.3 square meter on the monitoring area. The test image sequence (figure 2) was captured and digitized at the rate of 10 images / per second – the fastest rate possible with the computing equipment available for this work.


Frame n+1 Frame n+2 Frame n+3

Frame n+4 Frame n+5

Figure 2 a single movement object sequence

Learning Initial Background Model

In a real time system, its noise comes from anywhere. The noise overcomes the difficulties of using the median estimator to estimate the local mean in our image processing. A new local mean is then computed using only pixels within a small gray-level range about the median. This effectively removes outliners in the calculation of the mean estimate, hence improving the overall performance of the mean filter in the presence of outliners such as salt and pepper noise. In this section, we describe the sequence image filtering as first process. An adaptive double window modified trimmed mean (DW-MTM) filter is employed for the first step that filters the input image in our real time system. The adaptive DW-MTM filter algorithm is described as follows.

Given a pixel located at x, y within the image, a median filter (MED [g (x, y)]) is computed within an n × n local region surrounding the location x, y. The g (x, y) is the noise-corrupted image. The median value computed from this filter is used to estimate the mean value of the n × n local area. Next, a larger-size window surrounding the pixel at location x, y of size q × q is used to calculate the mean value. In computing the mean value in the q × q window, only pixels within the gray-level range of

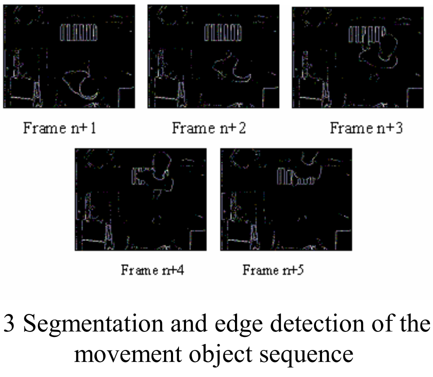
$$\text{MED}\,[g\,(x, y)] - c \text{ to } \text{MED}\,[g\,(x, y)] + c$$

are used, eliminating any outliners from the mean calculation. The output of the DW-MTM filter is the q × q mean filter. The value of c is chosen as a function of the noise standard deviation as

$$c = K \bullet \sigma_n$$

Where σ_n is the variance of the noise. Typical values of K range from 1.5 to 2.5. This range for K is based on the assumption that for Gaussian noise statistics the peak-to-peak gray-level variations will be in the range of $\pm 2\sigma_n$ 95% of the time and any values outside this range are more than likely outliers. For K=0, the DW-MTM filter to a n × n median filter, and for K very large, the DW-MTM reduces to a q × q mean filter. Hence, as K decreases, the filter does a better job of filtering impulsive noise, but a poor job of filtering uniform and Gaussian-type noise. In our system, we assume n=3 and q=5.

After the input image is filtered, we state that edge detection is part of a process called segmentation—the identification of regions within an image for second step processing. There are many varieties of edges; they may be classified into three major classes: a line-edge has a zero order discontinuity, a step-edge has a first order discontinuity, and a roof-edge has a second order discontinuity. It is compared with the Marr-Hildreth edge detection involve large windows and often needs floating-point calculations to maintain accuracy [5]. Canny edge detector that extracts not only step edges but also ridge and roof edges [6] and Shen-Castan edge detector. [7.] We are not interested in the direction of the edge but only in its presence, we should use direction-invariant edge detectors. On other way, we need to minimize the process time of edge detection. Here, we use a local neighborhood edge operator which is direction-invariant is the Laplacian 3 × 3 edge detector for this application. For the result refer to Figure 3.


Frame n+1 Frame n+2 Frame n+3

Frame n+4 Frame n+5

Figure 3 Segmentation and edge detection of the single movement object sequence

We use two consecutive frames to retrain the background model. After the sequence starts, we subtract two consecutive frames to obtain a result of a binary image (difference image) as figure 4 that shows to subtract two consecutive frames from the figure 3.

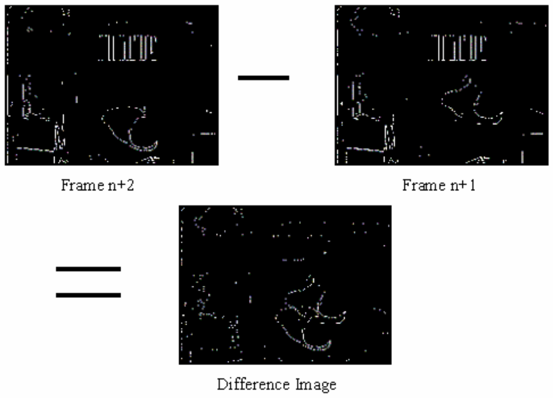


Figure 4 Subtract two consecutive frames from the single object movement sequence

The cluster analysis technique is used to extract some interesting points for labeling the object movement and some noise from the result image. In the result image, there is the object movement and some noise. For this noise, due to a great many factors such as light intensity, type of camera and lens, motion, temperature, atmospheric effects, dust, and others, it is very unlikely that two pixels that correspond to precisely the same gray level in the scene will have the same gray level in the image. Noise is a random effect, and is characterizable statistically only. The result of noise on the image is to produce a random variation in level from pixel to pixel, and so the smooth lines and ramps of the ideal edges are never encountered in real images. Subtracting the current image from the sequence image also produces the noise. As the result image is a binary image, the white pixels show the movement object or noise. There are some white pixels that merge together from pixel to pixel. We can state that the result image can response the noise only has a small number pixel merge as it comes from two consecutive frames. We can assume a threshold value to reject these noise from pixel to pixel in one gray-level at clustering process. This threshold value is called Noise_Threshold in number of pixels. The detail processing is as following.

Almost always, when information about an object or region class is available, a pattern recognition method is used to find out some interest points for the moving object. For this application, we investigate some related techniques: statistical pattern recognition, neural nets, syntactic pattern recognition, recognition as graph matching, optimization techniques in recognition, and fuzzy systems. They not only require considerable time but are also beyond our requirements. A cluster analysis technique is modified for this application as follows.

In the image analysis, clustering can be used to find groups of pixels with similar gray levels, colors, or local textures, in order to discover the various regions in the image. Clustering is the process of counting and labeling of objects within an image. Clusters consist of pixel groupings that are related to one another by a predetermined criterion. This criterion on measure can be defined as a distance between clusters or a similarity measure such as a pixel value, or it may be a complex set of identifiers and constraints that define membership in a cluster. The cluster may ultimately be mapped into a binary image as a unique object.

Formally, if p is a value of pixel, the cluster is defined as

$$C_i(p) = p \qquad (1)$$

Where i is an integer (0, 1, 2, ….) and cluster center point is defined as an interest point to response this cluster.

Here, our approach is to specify a connected in pixel that the clusters should be. In the binary image, the clustering algorithm processes each pixel and when one is found that is nonzero, it becomes part of the first cluster and is marked. The next nonzero pixel $f(x,y)$ found is tested to see if it is connected to one previous pixel of outer corners of its half_8_adjacent [$f(x-1,y-1)$, $f(x,y-1)$, $f(x+1,y-1)$, $f(x-1,y)$, $f(x-1,y+1)$]as shown in Figure 5. If it is, it is marked as a member of the first cluster and the search continues. If it is not, it becomes the first member of the second cluster and is marked accordingly. This process continues until all pixels have been evaluated.

$f(x-1,y-1)$	$f(x,y-1)$	$f(x+1,y-1)$
$f(x-1,y)$	$f(x,y)$	
$f(x-1,y+1)$		

Figure 5 Outer corners of a pixel's half_8_adjacent neighbors

If we set the Noise_Threshold = 4 (pixels) and maximum value of pixel number =11, we can obtain that $C_3(p)$ is a valid cluster and $C_1(p)$ &$C_2(p)$ are noise value with the equation (2),.

In order to minimize the processing time, we need to define a maximum value of pixel number for the cluster. Then we get one point to response these valid clusters as an interest point in the object as following relate equation.

Noise_Threshold < valid cluster $C(p)$ < maximum value of pixel number (2)

The example shows a graphic of a set of a set of clusters that have been labeled by grayscale value. After clustering, we found out there are three clusters $C_1(p)$, $C_2(p)$, and $C_3(p)$ as shown in Figure 6. From the equation (1), we known that $C_1(p) = 4$, $C_2(p) = 2$, and $C_3(p) = 10$.

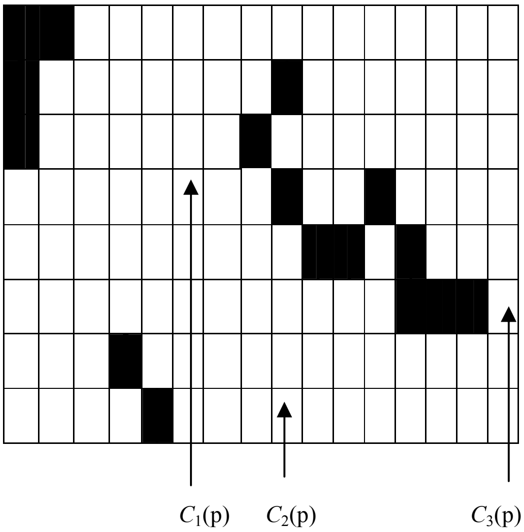


Figure 6 A graphic of a set of clusters

At the completion of the function, a 2D data array contains pixel values that reflect their membership in the clusters found that meet the criteria of being pixel connected apart. Cluster of pixels that are not connected with other cluster in outer corners of its 8_adjacent will be partitioned into pieces as they are found in the image. The difference binary image (after segmentation) is scanned from the upper left corner origin in column order. If a zero value gray-level is defined for object in the difference binary image (image [row][col]), we have that a chart of the function is shown in Figure 7.

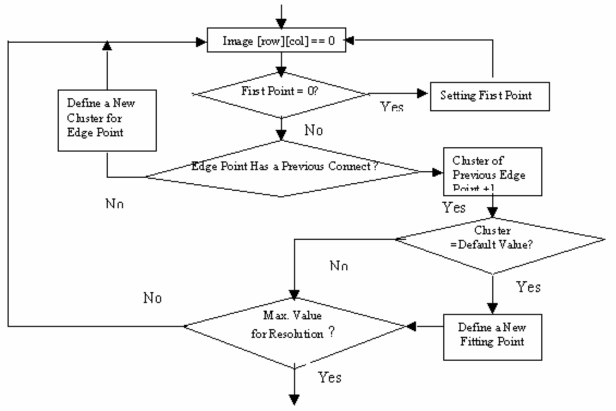


Image [row][col] == 0
First Point = 0?
Yes
Setting First Point
Define a New Cluster for Edge Point
No
Edge Point Has a Previous Connect ?
Cluster of Previous Edge Point +1
No
Yes
Cluster =Default Value?
No
No
Yes
Max. Value for Resolution ?
Define a New Fitting Point
Yes

Figure 7 A chart of cluster analysis function

During the cluster analysis, a valid cluster result is generated with some points to interrupt the movement object. These points are called movement object points. The valid cluster result is generated with some points in response to background noise. These points are called non-movement object points such as a value (3, 263) at figure 8 that shows a real time result of the cluster analysis. It was produced by clustering the difference image (figure 4).

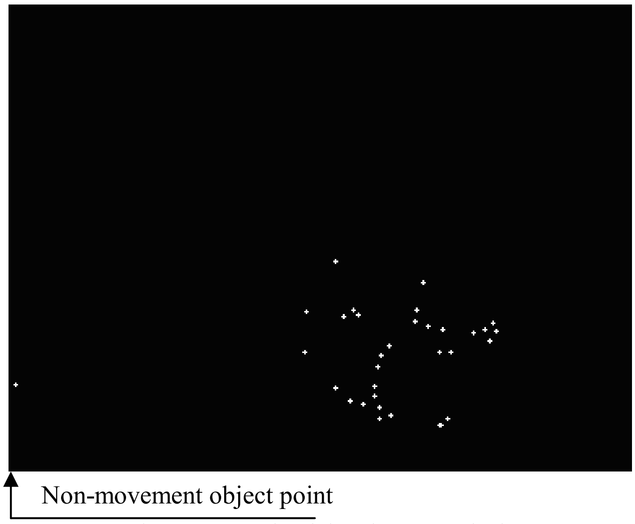

Non-movement object point

Figure 8 Result of the cluster analysis

Issue Background Model

The cluster result is generated with some points to interrupt the object movement and other points to response the background noise or without any point. If this result is zero there is not any point, we state one of both images is its background model for currently sequence. If this result is not zero such as figure 8, we state there is some object movement at the scene. The background model is issued from one of two consecutive frames binary images remove the movement object pixels. For the background modeling, we are interested in the outline points of the moving object as the figure 8. Removing all the moving object pixels that are enclosed from its outline points creates the binary background model. Figure 9 show a result of background model that remove the movement object of frame n+1 on figure 4.

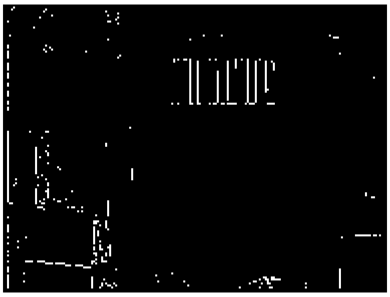


Figure 9 A background model for the single moving object sequence

Updating background Model

The background model cannot be expected to stay the same for long periods of time. There could be illumination changes, such as the sun being blocked by clouds causing changes in brightness, or physical changes, such as a deposited object. As the cluster analysis is applied to each frame for tracking object movement at our object tracking system. From the figure 8, we know the non-movement point that is out of the movement object. When we keep the background model to apply to it sequence, a cluster analysis result show as figure 11 at the frame n+7 of single movement object sequence as following figure 10. From the figure 11, the points of non-movement object are 6.

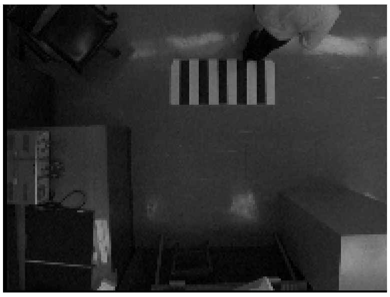


Figure 10 Frame n+7 of single movement object sequence

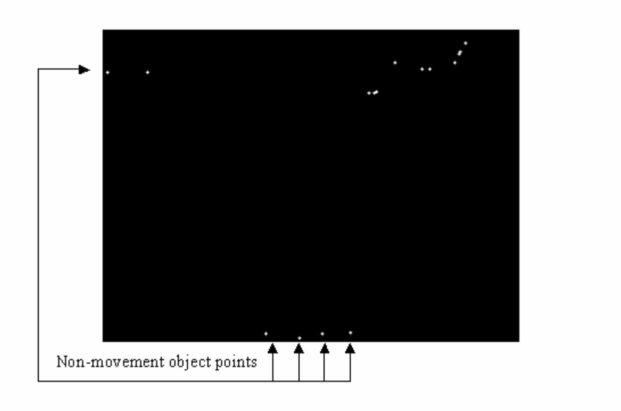


Figure 11 Cluster analysis result for frame n+7 of single movement object sequence

When there are more non-movement object points than a limit value, we need to update the background model. If this limit value is assumed set to 6 points, an update background processing will occur at frame n+7 of single movement object sequence. This update background-processing algorithm is summarized as follows:

- Get a cluster analysis result image by subtracting next frame.
- If this result is zero there is not any non-movement object points and movement object points, we state one of both frames is the background model for current sequence.
- If this result is non-zero, its background model for currently sequence is that remove all of movement object pixels that are enclosed from its outline points.

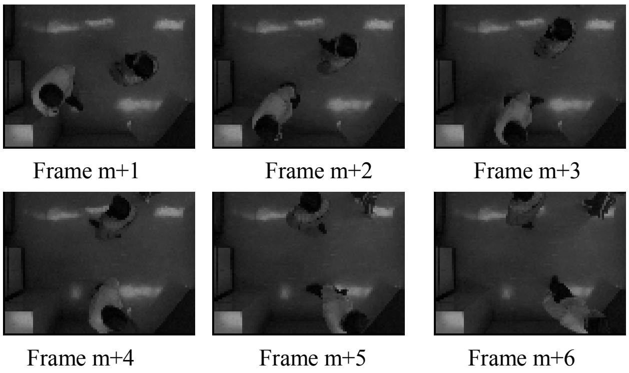
Experiments

The complete algorithm for our background modeling method is summarized in section 4.1. To evaluate our algorithm, two experiments are presented. The first experiments (section 4.2) demonstrate the performance of our background modeling method in the case of a multi-movement object environment. The second experiments (section 4.3) show the performances of our background modeling method to apply to our object tracking system.

The complete algorithm using the background modeling system is as follows:

- Capture image on and filter the sequence image,
- Edges detect two current consecutive frames and subtract them,
- Issue an initial background model from a result of the cluster analysis,
- If non-movement object points are over a limit value, update processing occurs for the background model.

When we apply this background model method to a multi-movement object sequence as figure 12, a result of background model is shown on figure 15. We subtract two consecutive frames as figure 13 at our background modeling. A result of the cluster analysis on multi-object movement sense is shown on figure 14. We are also success using the current background model to tracking the multi-movement object at this multi-movement object sequence. The figure 16 shows a cluster analysis result of the frame m+10 with same background model. From this result, we can state that the background model can be expected to stay the same for next tracking; there is not any non-movement point at this frame.



Frame m+1 Frame m+2 Frame m+3

Frame m+4 Frame m+5 Frame m+6

Figure 12 A sequence for multi-object movement

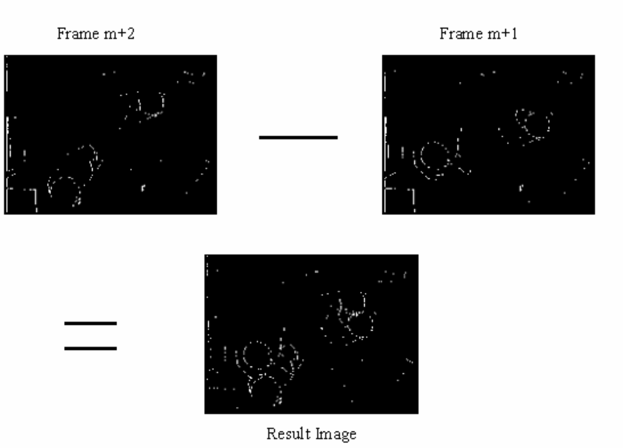


Figure 13 Subtract two consecutive images

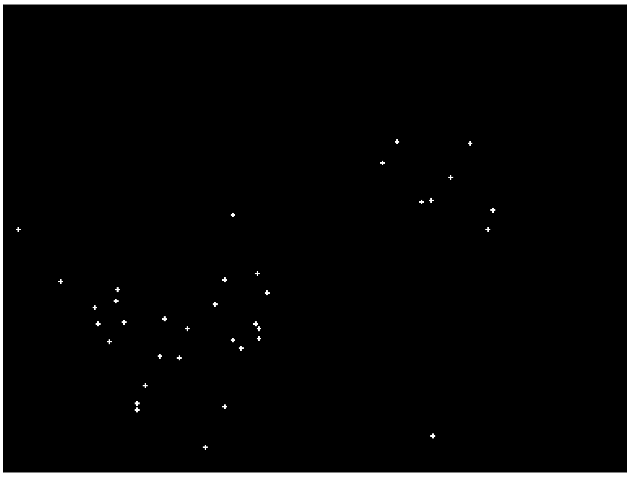


Figure 14 A result of the cluster analysis on multi-object movement sense

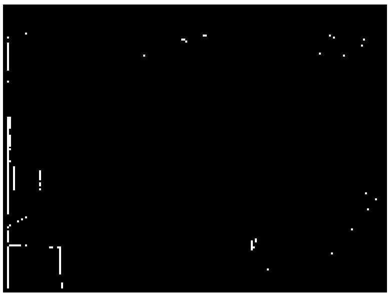


Figure 15 A background model for multi-movement object sequence

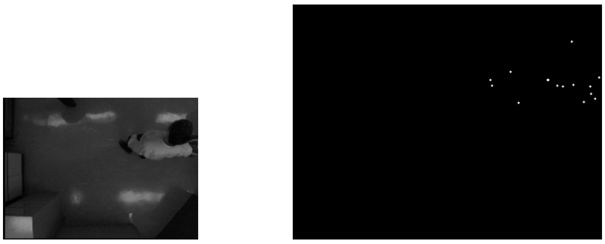
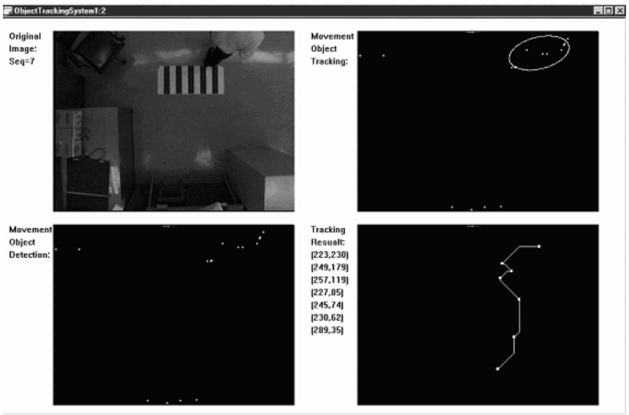


Figure 16 (a) Frame m+10

Figure 16 (b) A result of the cluster analysis for frame m+10

Figure 16 A result of the cluster analysis for frame m+10 with the background model on multi-object movement sense

The background model method was success fully implemented on our tracking system as in figure 16. The object tracking result shows the single movement object from frame n+1 to frame n+7 of the single movement object sequence (Figure 2). There are four windows to display an original image, movement object detection, movement objects tracking and tracking result.



```
[223,230]
[249,179]
[257,119]
[227,85]
[245,74]
[230,62]
[289,35]
```

Figure 16 a movement object tracking system

Conclusion

A real-time computer vision system able to model a stationary object background or a movement object background in cluttered environments has been presented. The proposed system is based on the modeling of the structure of the scene. The quality of the detection is improved when the background is highly textured.

Therefore in our future works we will use this modeling method in our object tracking system for tracking rigid and non-rigid movement object. It will be used for to tracking multiple non-rigid movement objects in a cluttered scene.

Reference

[1] N. Friedman and S. Russell, “Image Segmentation in Video Sequences: A Probabilistic Approach,” Uncertainty in Artificial Intelligence, 1997.

[2] E. Grimson, C. Stauffer, R. Romano, and L. Lee, “Using Adaptive Tracking to Classify and Monitoring Activities in a Site,” Proc. Computer Vision and Pattern Recognition Conf., pp. 22-29, 1998.

[3]A. Elgammal, D. Harwood, and L. Davis, “Non-Parametric Model for Background Subtraction,” Proc. IEEE Frame Rate Workshop, 1999.

[4]I. Haritaoglu, D. Harwood, and L. S. Davis, “W4: Real-Time Surveillance of People and Their Activities,” PAMI, August 2000.

[5] Marr, D. and E. Hildreth, “ Theory of Edge Detection” Proceedings of the Royal Society of London, Series B. Vol. 207.187-217, 1980.

[6]Canny, J., “A Computational Approach to Edge Detection” IEEE, PAMI, VOL8, 679-698, JUNE 1986.

[7] A. Elgammal, D. Harwood, and L. Davis, “Non-Parametric Model for Background Subtraction,” Proc. IEEE Frame Rate Workshop, 1999.

OFCat: An extensible GUI-driven optical flow comparison tool

Robert J. Andrews and Brian C. Lovell
School of Information Technology and Electrical Engineering
Intelligent Real-Time Imaging and Sensing
University of Queensland

E-mail: `s341199@student.uq.edu.au, lovell@itee.uq.edu.au`

Abstract

OFCat is an extensible GUI-driven optical flow computation and analysis tool. The user can add their own image sequences, filtering, differentiating and optical flow techniques. OFCat can process both grayscale and color images. It implements a number of low-pass filtering and differentiating techniques. Many traditional optical flow extraction techniques are available, in addition to some novel color-based methods.

Analysis of the recovered flow is performed using ground-truth analysis, image reconstruction and sparsity vs error analysis. Synthetic ground-truth can be created and used to create a test image sequence from a single image.

1 Introduction

Optical flow has many applications in computer vision and image-sequence processing. Among other areas, it has been used for robot navigation, face tracking and recognition, three-dimensional reconstruction and general analysis of motion.

Barron [1] and others wrote a review paper on optical flow in 1994 and provided code for other researchers to compare their results to. The tool for the analysis and comparison of optical flow described in this paper (OFCat) is intended to be a useful addition to the computer vision researcher's toolbox. It allows use of custom image sequences, lowpass filters, differentiating kernels and optical flow methods. Many techniques and algorithms are implemented in each of these categories. Some novel color-based optical flow techniques are available to the researcher using OFCat, as well as some previously theorised color techniques [5]. In addition, an error analysis section is provided which allows visualisation of the flow, analysis versus ground truth, image reconstruction error and density vs error threshold metrics. Both color and grayscale images are supported, with similar and different methods available for each.

Due to the script processing nature of Matlab® the speed of algorithms can often be increased by compilation. A script is provided for compilation of the algorithms that benefit most in the application. This script assumes a Matlab®-compatible compiler is installed and that Matlab® is appropriately configured.

In addition to automatic C-translation and compilation, it is possible to use MEX (Matlab® shared library file) wrappers to incorporate external C (or C++) code. This possibility is exploited in the application by incorporating functions which use optimized functions from the Intel® IPL® [3] (Image Processing Library) and Intel® OpenCV® [4]. These libraries take advantage of the enhanced multimedia extensions available on Intel Pentium-based chips. Hence, some functions are available for Intel processers only.

When assessing the accuracy of a system, it is necessary to investigate each constituent independently. The recovery of optical flow using gradient-based methods generally involves three steps:

- Low-pass filtering;
- Estimation of first and often second-order derivatives; and
- Recovering optical flow from the derivatives and other metrics

OFCat allows the user to quickly assess the benefits and drawbacks of various algorithms in each of these three categories, with the possibility of adding their own algorithms.

2 Filtering

2.1 Lowpass filtering

Several non-linear de-noising filters have been implemented, listed in table 1. All of these filters are available in color.

Filter
Median Filtering (3x3)
Median Filtering (5x5)
Wiener (Adaptive) filtering (3x3)
Wiener (Adaptive) filtering (5x5)

Table 1. Implemented non-linear de-noising filters

Filter	Color	IPL®
Box (Linear Avg) Filter (3x3)	Yes	Yes
Box (Linear Avg) Filter (5x5)	Yes	Yes
Gauss Filter (3x3) $\sigma = 0.85$	Yes	Yes
Gauss Filter (5x5) $\sigma = 1$	Yes	Yes
Gauss Filter (3x3) $\sigma \neq 0.85$	No	No
Gauss Filter (5x5) $\sigma \neq 1$	No	No

Table 2. Implemented linear de-noising filters

Table 2 lists the linear filters available to the user. In addition to these, the Matlab®'s FDATool (Filter Design and Analysis Tool) can be used to design IIR or FIR filters which can be easily imported into OFCat.

As Gaussian filters are quite popular among vision systems, an additional GUI is provided for the choice of sigma, the shape of the Gaussian.

2.2 Differentiating filters

Spatio-temporal derivatives are necessary for all gradient-based optical flow methods. OFCat computes temporal derivatives separately from spatial derivatives. Three temporal derivatives are implemented; one-sided finite difference, central finite difference and four-point central difference. The corresponding temporal supports are 2, 3 and 5.

Default spatial derivatives available to the user are listed in table 3. Five filters are denoted as Optimal. These are functions utilising filters from Intel®'s IPL®, which more closely approximate the ideal filter as the length of the filter increases [3].

Manduci provides an in-depth theoretical treatment of spatio-temporal filtering and its effect on the accuracy of optical flow methods in his paper [7].

3 Optical Flow methods

OFCat's default optical flow methods can be categorised as working with gray-scale data (one set of partial derivatives) or color data (three sets of partial derivatives). Each category is described below.

Filter	Color	IPL®
Sobel (3x3)	No	Yes
Sobel (5x5)	No	Yes
Prewitt (3x3)	No	Yes
Prewitt (5x5)	No	Yes
Matlab®'s gradient function	Yes	No
Matlab®'s difference function	Yes	No
Optimal 5x5	Yes	Yes
Optimal 7x7	Yes	Yes
⋮	⋮	⋮
Optimal 13x13	Yes	Yes

Table 3. Implemented spatial-differential filters

3.1 Gray-scale optical-flow

Some common methods of recovering optical flow have been implemented, these include many block-matching algorithms and many of the traditional gradient-based methods reviewed in Barron et. al.'s paper [1]. These are: Horn and Shunck, Lucas and Kanade, Nagel, Uras et. al., Uras et. al. with Barron's gaussian curvature and Anandan. Block matching metrics available to the user are: Normalized, Standard, Zero-mean and Zero-mean normalized versions of Correlation, Sum of absolute differences and Sum of squared differences. Most of the gradient-based methods have default parameter values which are modifiable via dialogs.

This large selection of optical flow techniques enables the researcher to quickly evaluate the performance of various techniques on their image sequence. Alternatively, if they develop a new algorithm, they can quickly and easily compare it's performance with existing techniques.

3.2 Color optical flow

The area of color, gradient-based optical flow has been largely overlooked by researchers. The three planes of data available to an optical flow algorithm are usually combined to form one intensity image. Golland [5] presented two methods which utilise this extra information; these and other color-based methods are presented for the user's appraisal.

The optical flow constraint equation is

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0 \qquad (1)$$

We can extend this to a system of equations, in the form $Ax = b$, with each row consisting of data calculated from different planes of the image, e.g. Red, Green and Blue.

This becomes

$$\begin{aligned}\frac{\partial I_R}{\partial x}u+\frac{\partial I_R}{\partial y}v+\frac{\partial I_R}{\partial t}&=0\\ \frac{\partial I_G}{\partial x}u+\frac{\partial I_G}{\partial y}v+\frac{\partial I_G}{\partial t}&=0\\ \frac{\partial I_B}{\partial x}u+\frac{\partial I_B}{\partial y}v+\frac{\partial I_B}{\partial t}&=0\end{aligned}\tag{2}$$

There are a few options available for the solution to this equation (2), the two that have been the major focus of color optical flow methods implemented are

- Least squares
- Simple gaussian elimination

If we disregard one plane of the image, we have two equations and two unknowns. This is the basis of Golland's "color constancy" equation, disregarding brightness and retaining the hue and saturation values of the image. Alternatively, we can retain the brightness plane (when we use the HSV or YUV color models) and incorporate one of the color planes.

Three different color models are available to the user. These are HSV, RGB and normalized RGB.

In addition to those mentioned above, two novel methods are included.

- Neighborhood least squares, taking a 3×3 neighborhood in each plane and using all of these measurements to create a system of linear equations;
- Application of a standard grayscale optical flow method (e.g. Horn and Shunck [6]) to each plane, then use the intrinsic error measure of each recovered flow to combine them into one flow field.

Figure 1 displays the GUI used for launching color optical flow algorithms. The corresponding grayscale GUI is very similar, though offers different optical flow methods.

4 Error Analysis

Four methods of analysing the estimated optical flow are provided. These are detailed in the next four sections.

4.1 Flow visualisation

Visualisation of the flow is performed via the dialog shown in Figure 2. The recovered flow, ground-truth flow and the difference of both can be viewed here. Additionally, the curl, divergence and magnitude of any of these vector fields can be displayed.

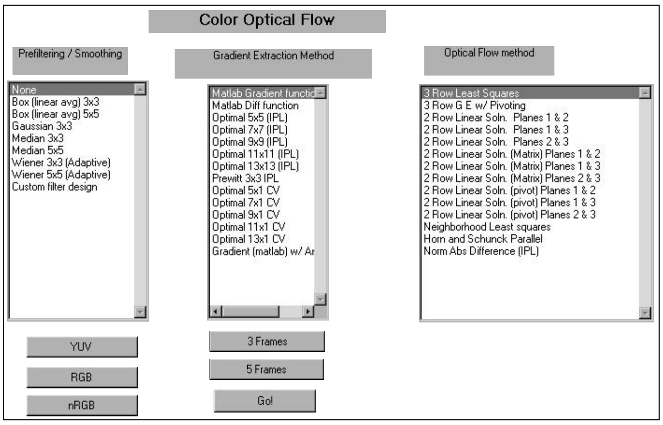


Figure 1. GUI used to select methods for filtering, deriving and estimating optical flow

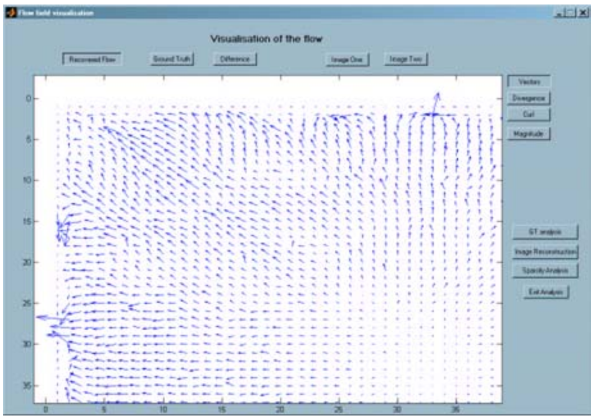
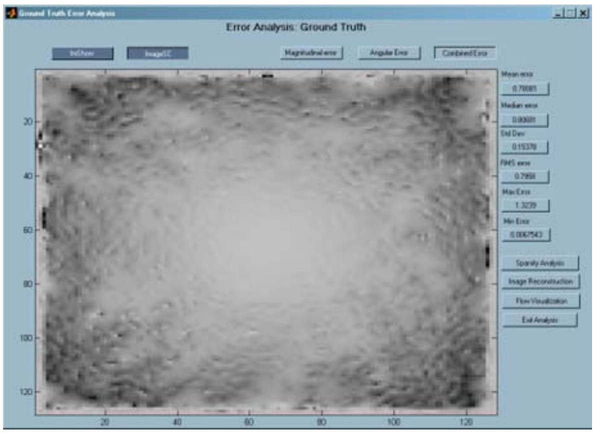


Figure 2. Flow visualisation



Mean error
0.78881
Median error
0.80681
Std Dev
0.15378
RMS error
0.7958
Max Error
1.3239
Min Error
0.0067543

Figure 3. Ground Truth Analysis

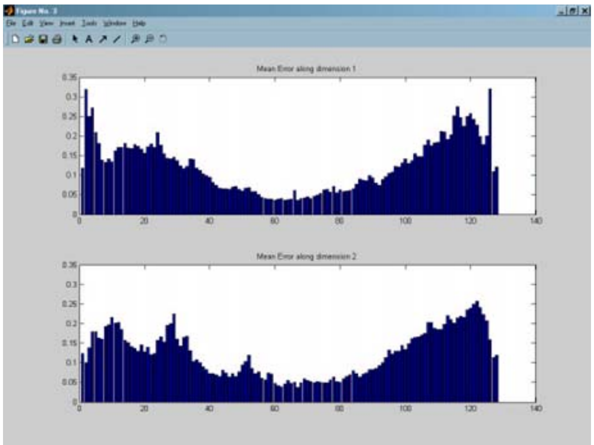
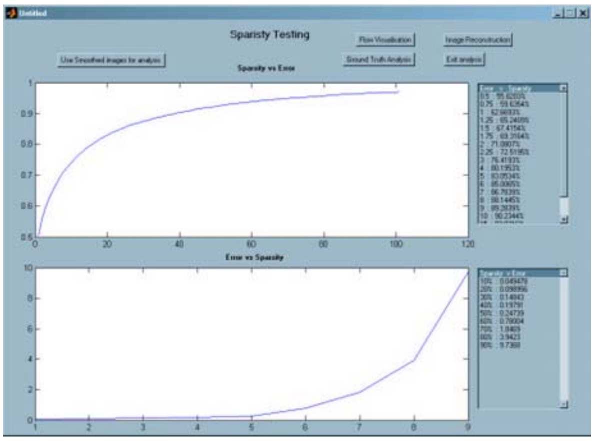


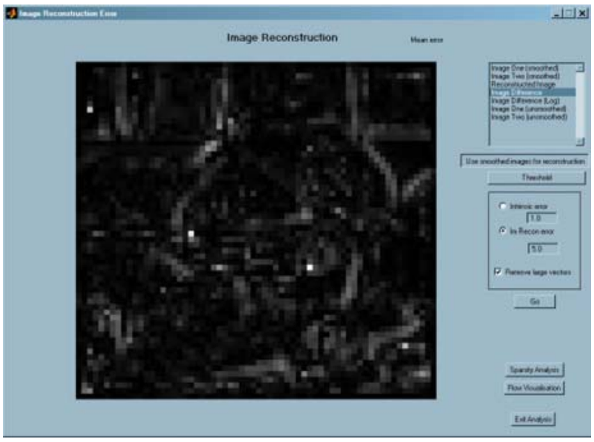
Figure 4. Mean error along each dimension



Error vs Sparsity
0.5 : 55.6203%
0.75 : 59.6354%
1 : 62.6693%
1.25 : 65.2409%
1.5 : 67.4154%
1.75 : 69.3164%
2 : 71.0807%
2.25 : 72.5195%
3 : 76.4193%
4 : 80.1953%
5 : 83.0534%
6 : 85.0065%
7 : 86.7839%
8 : 88.1445%
9 : 89.2839%
10 : 90.2344%

Sparsity vs Error
10% : 0.049479
20% : 0.098956
30% : 0.14843
40% : 0.19791
50% : 0.24739
60% : 0.79004
70% : 1.8469
80% : 3.9423
90% : 9.7368

Figure 5. Sparsity Analysis



○ Intrinsic error
1.0
● Im Recon error
5.0
☑ Remove large vectors

Figure 6. Image Reconstruction Interface

4.2 Ground Truth Analysis

When ground truth is available, the interface shown in Figure 3 enables the user to analyse the magnitudinal, angular and combined error both visually and statistically. The mean, median, standard deviation, RMS, maximum and minimum errors are calculated and displayed in their respective buttons. Each of the buttons opens a new figure, displaying two plots. The plots are each statistic, respectively, taken across each dimension (horizontally and vertically). An example of these plots is shown in Figure 4.

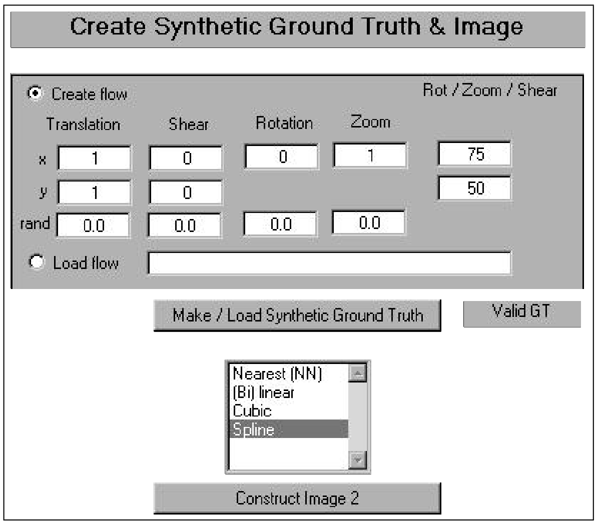
4.3 Sparsity Analysis

The dialog shown in Figure 5 illustrates the image reconstruction error of optical flow methods by calculating the density of flow fields thresholded at chosen errors. The resulting function is inspected with linear interpolation to estimate the image reconstruction error at certain levels of density. The user can also choose to view the error function when image reconstruction takes the smoothed images for input. The contrast between using smoothed and unsmoothed images illustrates the effect of smoothing on the process of calculating optical flow.

4.4 Image Reconstruction

This interface (Figure 6) allows reconstruction of the second image in the sequence from the first image and the optical flow as per Barron and Lin [2]. The user is able to view the first and second real images, the first and second smoothed images, the reconstructed image, the difference between the reconstructed and second image and the natural logarithm of the difference between reconstructed and second original image.

The natural logarithm of the difference is useful for visualisation of the error as most of the error is of a smaller order of magnitude than the largest errors. Using the natural



Create Synthetic Ground Truth & Image
Make / Load Synthetic Ground Truth
Valid GT
Construct Image 2
Nearest (NN)
(Bi) linear
Cubic
Spline

● Create flow

	Translation	Shear	Rotation	Zoom	Rot / Zoom / Shear
x	1	0	0	1	75
y	1	0			50
rand	0.0	0.0	0.0	0.0	

○ Load flow

Spline

Figure 7. Synthetic Ground Truth Creation Interface

logarithm enables the user to observe the sites perturbed by both large and moderate errors.

The mean of the difference between the reconstructed and correct image is also calculated and displayed.

Thresholding of large vectors (a magnitude greater than > 5 pixels) and thresholding on a chosen image reconstruction or intrinsic error level is supported. The mean returned after thresholding is the mean of only the accepted reconstructed sites. The density of the reconstructed image is also shown. This indicates how many sites were deemed unacceptable and consequently disregarded.

4.5 Creating synthetic error

When ground truth is not available for a sequence, it is still possible to perform ground truth analysis if we create a synthetic sequence based on the first frame of the original. Synthetic planar ground truth can be described by a global translation, shearing in the horizontal and vertical planes, rotation and zoom. If the user has their own synthetic ground-truth flow field saved in a .MAT file, OFCat can load and use this instead.

Figure 7 shows the dialog enabling creation of a synthetic image sequence and ground truth from parameter values. After the synthetic ground truth has been calculated, it is used to warp the first image of the image sequence. This process is repeated until we have a sequence of five frames. Both color and grayscale images are supported.

This sequence then becomes the input to the filtering / optical flow method. As ground-truth is now available, methods of ground-truth analysis can be applied.

5 Conclusion and Future Work

OFCat is a visual, extensible, optical flow analysis tool, intended for students learning about optical flow or for researchers implementing their own methods. Adding new algorithms for filtering, differentiating or computing optical flow to OFCat is simple. OFCat provides quantitative and qualitative error analysis, making it easy for researchers to compare their new method(s) with traditional ones.

Two extra color models will be implemented in the near future, these are the YUV and CIE (UCS) color systems.

OFCat is available to for download at
http://itee.uq.edu.au/ ˜iris/ComputerVision/OFCat/
The user must have Matlab® and Intel®'s IPL® installed. Scripts are included for recompilation of the optimized functions.

6 Acknowledgements

The authors would like to thank David McKinnon and Nic Geard for their assistance in creating this document.

References

[1] J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Systems and experiment performance of optical flow techniques. *International Journal of Computer Vision*, 12:43–77, 1994.

[2] J. L. Barron and T. Lin. Image reconstruction error for optical flow. 1995.

[3] I. Corporation. Intel ipl reference manual. Technical report, Intel Corporation, Feb 2000.

[4] I. Corporation. Intel opencv reference manual. Technical report, Intel Corporation, Feb 2000.

[5] P. Golland and A. M. Bruckstein. Motion from color. Technical report, Computer Science Dept, Technion, I.I.T., Haifa, Israel, 1997.

[6] B. Horn and B. Shunck. Determining optical flow. *Artificial Intelligence*, (17):185–203, 1981.

[7] R. Manduci. Improving the accuracy of differential-based optical flow algorithms. UCB//CSD- 93-776, University of California at Berkeley, 1994.

(This page left blank intentionally)

Color Optical Flow

Robert J. Andrews and Brian C. Lovell
School of Information Technology and Electrical Engineering
Intelligent Real-Time Imaging and Sensing
University of Queensland

E-mail: `s341199@student.uq.edu.au, lovell@itee.uq.edu.au`

Abstract

Grayscale optical-flow methods have long been the focus of methods for recovering optical flow. Optical flow recovery from color-images can be implemented using direct methods, i.e. without using computationally costly iterations or search strategies. The quality of recovered optical flow can be assessed and tailored after processing, providing an effective, efficient tool for motion estimation.
In this paper, a brief introduction to optical flow is presented, the optical flow constraint equation and its extension to color images is presented. New methods for solving this extended equation are given. Results of applying these methods to two synthetic image sequences are presented.

1 Introduction

Optical flow is a useful tool for many tasks in computer vision. It has been applied to problems of motion-segmentation, time-to-contact and three-dimensional reconstruction (structure from motion) among others. Traditionally, most researchers in this field have focussed their efforts on extending Horn and Shunck [8] or Lucas and Kanade's [9] methods, all working with grayscale intensity images.

Color image sequences have been largely ignored, despite three planes of information being available instead of one. Golland proposed and discussed two simple methods which incorporate color information [7]. She investigated the RGB, normalized RGB and HSV color models. Her results indicated that color methods provide a good estimate of the flow in image regions of non-constant color.

This paper compares traditional grayscale with Golland's methods and two new color methods. It also describes the proposed the extension of grayscale methods into color.

2 Optical flow

The optical flow of an image sequence is a set of vector fields, relating each image to the next. Each vector field represents the apparent displacement of each pixel from image to image. If we assume the pixels conserve their intensity, we arrive at the "brightness conservation equation",

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \qquad (2.1)$$

where I is an image sequence, $[dx, dy]$ is the displacement vector for the pixel at coordinate $[x, y]$ and t and dt are the frame and temporal displacement of the image sequence. The idea of brightness conservation and optical flow were first proposed by Fennema [6].

The obvious solution to 2.1 is to use template-based search strategies. A template of a certain size around each pixel is created and the best match is searched for in the next image. The best match is usually found using correlation, sum of absolute difference or sum of squared difference metrics. This process is often referred to as block-matching. Such a search strategy is computationally costly and generally doesn't represent sub-pixel displacements.

Most methods presented in the last twenty years have been gradient-based. They solve the differential form of 2.1, derived by Taylor expansion. After discarding higher order terms, this is

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0 \qquad (2.2)$$

Here we have two unknowns in only one equation, the problem is ill-posed and extra constraints must be imposed in order to arrive at a solution.

The two most commonly used and earliest optical flow recovery methods in this category are briefly outlined below, Horn and Shunck's [8] and Lucas and Kanade's [9] optical flow methods. These and other traditional methods are outlined and quantitatively compared in Barron et.al. [4][3].

2.1 Horn and Schunck

Horn and Shunck [8] were the first to impose a global smoothness constraint, assuming the flow to be smooth across the image. Their minimization function,

$$\int\int (I_x u + I_y v + I_t)^2 + \lambda^2(\parallel \nabla u \parallel_2^2 + \parallel \nabla v \parallel_2^2) dxdy \tag{2.3}$$

can be expressed as a pair of Gauss-Siedel iterative equations,

$$u_{n+1} = u_n - \frac{Ix\,[I_x u_n + I_y v_n + I_t]}{\alpha^2 + I_x^2 + I_y^2} \tag{2.4}$$

and

$$v_{n+1} = v_n - \frac{I_y\,[I_x u_n + I_y v_n + I_t]}{\alpha^2 + I_x^2 + I_y^2} \tag{2.5}$$

2.2 Lucas and Kanade

Lucas and Kanade [9] put forth the assumption of constant flow in a local neighborhood. Their method is generally implemented with neighborhoods of size 5×5 pixels centered around the pixel whose displacement is being estimated. Measurements nearer the centre of the neighborhood are given greater weight in the weighted-least-squares formulation.

2.3 Other methods

Later methods generally extended of these two traditional methods. More recently, researches have been focusing on using concepts of robustness to modify Lucas and Kanade's method [2][1]. These methods choose a function other than the squared difference of the measurement to the line of fit (implicit in least squares calculation) to provide an estimate of the measurement's contribution to the best line. Functions are chosen so that outliers are ascribed less weight than those points which lie close to the line of best fit. This formulation results in a method which utilises iterative numerical methods, e.g. gradient descent or successive over-relaxation.

3. Using color images

Recovering optical flow from color images seems to have been long overlooked by researchers in the field of image processing and computer vision. Ohta [11] mentioned the idea, but presented no algorithms or methods. Golland proposed some methods in a thesis and a related paper [7]. She proposed using the three color planes to infer three equations, then solving these using standard least squares techniques.

$$\begin{aligned} \frac{\partial I_R}{\partial x}u + \frac{\partial I_R}{\partial y}v + \frac{\partial I_R}{\partial t} &= 0 \\ \frac{\partial I_G}{\partial x}u + \frac{\partial I_G}{\partial y}v + \frac{\partial I_G}{\partial t} &= 0 \\ \frac{\partial I_B}{\partial x}u + \frac{\partial I_B}{\partial y}v + \frac{\partial I_B}{\partial t} &= 0 \end{aligned} \tag{3.1}$$

The other idea proposed by Golland was the concept of ”color conservation”. By constructing a linear system to solve from only color components, e.g. Hue and Saturation from the HSV color model, the illumination is allowed to change, the assumption is now that the color, rather than brightness is conserved.

3.1. Color Models

Three color models have been implemented and tested in this paper. These are RGB, HSV and normalized RGB.

The RGB (Red, Green, Blue) color model decomposes colors into their respective red, green and blue components.

Normalized RGB is calculated as

$$N = R + G + B, \quad R_n = \frac{R}{N}, \quad G_n = \frac{G}{N}, \quad B_n = \frac{B}{N} \tag{3.2}$$

each color being normalized by the sum of all colors at that point. If the color value at that point is zero the normalized color at that point is taken as zero.

The HSV (Hue, Saturation, Value) model expresses the intensity of the image (V) independently of the color (H, S). Optical flow based purely on V is relying on brightness conservation. Conversely, methods which are based on H and S rely purely on color conservation. Methods which combine the two incorporate both assumptions.

Similar to HSV, the YUV model decomposes the color as a brightness (Y) and a color coordinate system (U,V). The difference between the two is the description of the color plane. H and S describe a vector in polar form, representing the angular and magnitudinal components respectively. Y, U and V, however, form an orthogonal euclidean space.

An alternative to these spaces is *CIE perceptually linear color space* also known as UCS (Uniform Chromaticity Scale). This color system has the advantage of euclidean distances in color space corresponding linearly to perception of color or intensity change.

Neither YUV, nor UCS have been implemented, though this is the next step in analysing color optical flow.

3.2. Methods

Two obvious methods for arriving at a solution to the extended brightness conservation equation 3.2 are apparent:

- Disregarding one plane so as to solve quickly and directly, using Gaussian Elimination.

- Solving the overdetermined system as is, using either least squares or pseudo-inverse methods.

Disregarding one of the planes arbitrarily may throw away data that is more useful to the computation of optical flow than those kept. However, if speed of the algorithm is of the essence, disregarding one plane reduces memory requirements and computational cost. Another possibility is merging two planes and using this as the second equation in the system. Numerical stability of the solution should be considered when constructing each system. By using the simple method of pivoting it is possible to ensure the best possible conditioning of the solution.

The methods of least squares and pseudo-inverse calculation are discussed in nearly all linear algebra texts.

A simple neighborhood least-squares algorithm, akin to Lucas and Kanade's [9], though not utilising weighting, has also been implemented. Values in a $3 \times 3 \times 3$ neighborhood around the center pixel were incorporated into a large, overdetermined system.

Another option for the computation of optical flow from color images is to estimate the optical flow of each plane using traditional grayscale techniques and then fuse these results to recover one vector field. This fusion has been implemented here by simply selecting the estimated vector with the smallest intrinsic error at each point.

All of the the methods mentioned above have been implemented and compared in this study.

4 Error Analysis

Image reconstruction is a standard technique for assessing the accuracy of optical flow methods, especially for sequences with unknown ground truth (see Barron and Lin [5]). The flow field recovered from an optical flow method is used to warp the first image into a reconstructed image, an approximation to the second image. If the optical flow is accurate then the reconstructed image should be the same as the second image in the image sequence. Generally, the RMS error of the entire reconstructed image is taken as the image reconstruction error. However, it is advantageous to calculate the image reconstruction error at each point in the image. This enables a level of thresholding in addition to, or instead of culling estimates with high intrinsic error.

The density of the flow field after thresholding at chosen image reconstruction errors can also be used to compare different methods. This is the method applied for comparison herein.

Method	64×64	128×128	240×320
Color (Least Sq)	0.66	2.11	9.24
Color (G E, 3 rows)	0.05	0.24	1.58
Color (GE, pivot, 2 rows)	0.03	0.14	0.54
H & S (20 its)	0.23	0.45	2.18
Lucas & Kanade	0.73	3.52	15.19
Nagel	1.63	6.70	36.68
Uras, et. al.	2.06	8.19	37.07
NCC	3.00	9.08	40.20
Black & Anandan	0.60	2.94	19.08

Table 1. Time taken for computation

5. Results and Discussion

Table 1 compares the time taken for recovery of optical flow using Matlab®, excluding low-pass filtering and derivative calculation times. The times recorded from computation on a 700Mhz Pentium III® processor. This highlights the drastic decrease in computational cost of direct color methods. The two row partial pivoting Gaussian Elimination method is seen to perform at approximately 20Hz. Compared to Horn and Shunck's method [8], the best performer in the field of grayscale methods, this represents an approximately fourfold increase in speed.

Figure 5.1 compares three common grayscale optical flow methods; Horn and Shunck [8], Lucas and Kanade [9] and Nagel [10]. This figure illustrates the density of the computed flow field when thresholded at chosen image reconstruction errors. It is seen that Lucas and Kanade's method [9] slightly outperforms Horn and Shunck's [8] method, which itself performs better than Nagel's [10] method at image reconstruction errors $>\approx 1.35$.

Figure 5.2 compares the performance of Lucas and Kanade's [9] with three color methods. The first frame of this image sequence is shown in figure 5.3. This sequence was translating with velocity [-1,-1] pixels per frame. The three color methods shown here are gaussian elimination (with pivoting) of the saturation and value planes of HSV, Gaussian elimination of RGB color planes and neighborhood least squares. Neighborhood least squares is seen to perform the best out of the color methods, closely approximating Lucas and Kanade at higher densities. Both gaussian elimination versions performed worse than the others.

An image sequence displaying a one degree anticlockwise rotation around the center of the image was used to assess three other color optical flow methods. Pixel displacement ranges between zero and 1.5 pixels per frame. The methods compared were "Color Constancy" [7], least squares solution to 3.2 [7] and Combined-Horn and Shunck. Horn and Shunck's [8] (grayscale) algorithm was used as a yardstick for this comparison. The results are displayed in

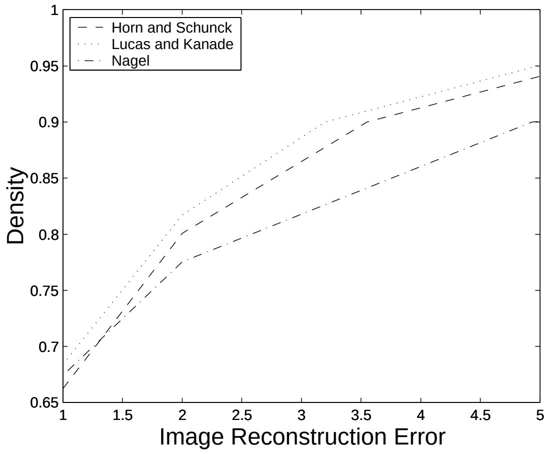


Figure 5.1. Comparison of Grayscale methods applied to translating colored clouds

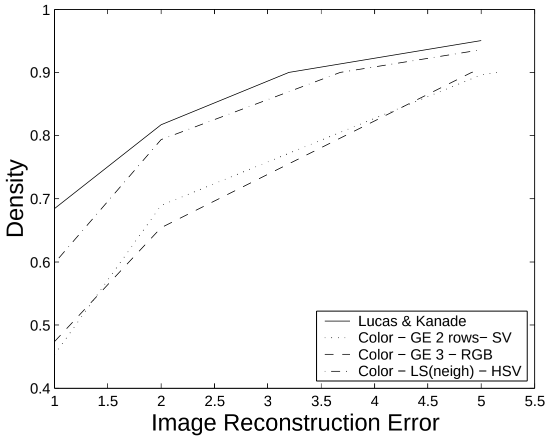


Figure 5.2. Comparison of gray and color methods applied to translating colored clouds

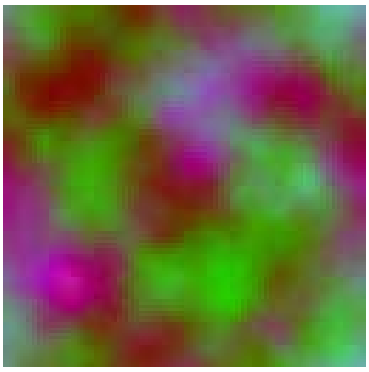


Figure 5.3. First frame of the translating RGB clouds sequence

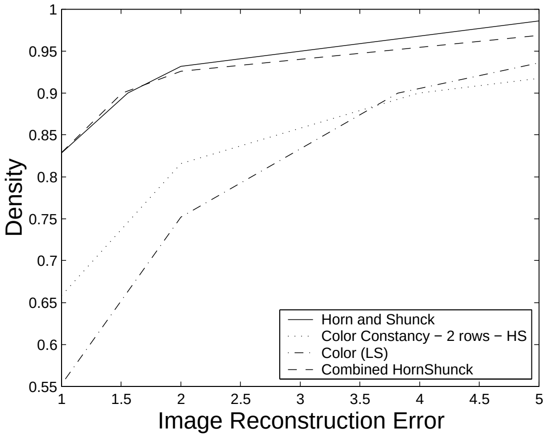


Figure 5.4. Comparison of techniques applied to a rotating image sequence

figure 5.4. Combined-Horn and Shunck applied Horn and Shunck optical flow recovery to each plane of the RGB image and fused them into one flow field utilising a winner-takes-all strategy based on their associated error. It can be seen that the Combined-Horn and Shunck method performed similarly to Horn and Shunck [8]. The methods of least squares [7] and direct solution of the color constancy equation [7] did not perform as well.

Figure 5.5 gives an example of the optical flow recovered by the neighborhood least squares algorithm. This corresponds to the rotating image sequence. Larger vectors (magnitude greater than 5) have been removedand replaced with zero vectors. This field has a density of 95%.

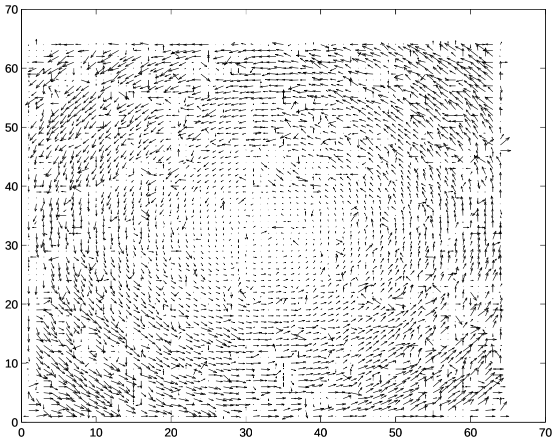


Figure 5.5. Optical flow recovered by direct two-row optical flow and thresholding

6 Conclusion and future work

Color optical flow has been shown to be quite simple to compute and to have a level of accuracy similar to traditional grayscale methods. The speed of these algorithms is a significant benefit; the linear optical flow methods presented run substantially faster than grayscale, non-linear methods.

YUV and UCS color models will be implemented and compared.

Accuracy of the neighborhood least squares approach can be improved in a number of ways. Using robust methods, e.g. least-median of squares [2], could provide a much better estimate of the correct flow. Applying the weighted least squares approach of Lucas and Kanade [9] could likewise improve the results.

A better data-fusion algorithm could be used to improve the Combined-Horn and Shunck method. The three flows being combined could be calculated using any grayscale method.

Methods that iterate towards a solution usually perform better with a good initial starting estimate. Color-optical flow could be used to provide this estimate, speeding the computation of some of the slower, well-known grayscale methods.

These issues will be investigated in future work.

7 Acknowledgements

The authors would like to thank David McKinnon and Nic Geard for their support in the creation of this document

References

[1] P. Anandan and Michael J. Black, *The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields*, Computer Vision and Image Understanding **63** (1996), no. 1, 75–104.

[2] Alireza Bab-Hadiashar and David Suter, *Robust optic flow estimation using least median of squares*, International Conference on Image Processing (1996).

[3] J. Barron, D. Fleet, S. S. Beauchemin, and T. Burkitt, *Performance of optical flow techniques*, RPL-TR-9107, Queen's University, July 1993.

[4] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, *Systems and experiment performance of optical flow techniques*, International Journal of Computer Vision **12** (1994), 43–77.

[5] John L. Barron and T. Lin, *Image reconstruction error for optical flow*, (1995).

[6] C. Fennema and W. Thompson, *Velocity determination in scenes containing several moving objects*, Computer Graphics and Image Processing **9** (1979), 301–315.

[7] P. Golland and A. M. Bruckstein, *Motion from color*, Tech. report, Computer Science Dept, Technion, I.I.T., Haifa, Israel, 1997.

[8] B. Horn and B. Shunck, *Determining optical flow*, Artificial Intelligence (1981), no. 17, 185–203.

[9] B. Lucas and T. Kanade, *An iterative image restoration technique with an application to stereo vision*, Proceedings of the DARPA IU Workshop (1981), 121–130.

[10] H. H. Nagel, *Displacement vectors derived from second-order intensity variations in image sequences*, CGIP (1983), no. 21, 85–117.

[11] N. Ohta, *Optical flow detection by color images*, Proceedings of IEEE International Conference on Image Processing (1989), 801–805.

(This page left blank intentionally)

Detection of Unknown Forms from Document Images

Andrew Busch a.busch@qut.edu.au
Wageeh W. Boles w.boles@qut.edu.au
Sridha Sridharan s.sridharan@qut.edu.au
Vinod Chandran v.chandran@qut.edu.au
Research Concentration in Speech, Audio and Video Technology,
Queensland University of Technology, Brisbane, Qld, Australia

Abstract

This paper presents a novel technique for distinguishing images of forms from other document images. The proposed algorithm detects regions which are likely to be used for text entry, such as lines, boxes, and character entry fields, and calculates a probability of the document being a form based on the presence of such structures. Experimental results from testing on both filled and unfilled forms, as well as a selection of non-form documents are presented. All document images are assumed to have been scanned at a known resolution.

INTRODUCTION

The extraction and processing of information contained in printed forms is a task of great importance in many areas of business and government alike. To date, the vast majority of form processing has been done manually, with human operators performing all of the associated tasks up to and including data entry. In recent times, a large amount of research has been undertaken in the fields of form identification, field location and data extraction. All of the research to date, however, makes the assumption that the image to be analysed is indeed a form, which may not always be the case in many applications. For this reason, this paper presents a technique for classifying a document image as either a form or non-form, and identifying likely field areas within any forms detected.

Previous work in the field of form field detection has provided an excellent starting point for this research. The technique proposed by Wang and Srihari [1] removes isolated characters, then searches for intersections of line segments. Yuan et al [2] present a method of detecting fields in forms that relies on segmentation algorithms to find text and straight lines, and uses adjacency graphs to detect possible entry fields in form images with no text entered. Xingyuan et al [3] propose a more robust technique which detects rectangular fields and lines regardless of text or other markings, but does not explicitly detect other form structures.

A number of techniques have also been proposed to remove the effects of noise and poor image acquisition, which can often cause unwanted line breaks, false intersections and broken junctions [4-6].

The work presented in this paper is in two parts. The first section describes a technique for detecting the primitive data entry structures that distinguish forms from other documents, namely lines, bounded rectangular areas, checkboxes, and character cell fields, or 'tooth' structures. In the second section we attempt to determine if an unknown document is likely to be a form. Using the presence of the previously detected structures, combined with the amount of text found in the document, a form probability score is proposed as an indication of the likelihood of the candidate document being a form.

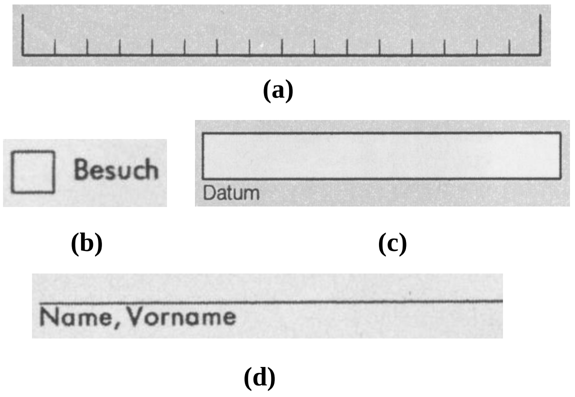
Results from experiments over 100 form and 200 non-form document images from a variety of sources are presented.

DETECTION OF FORM STRUCTURES

An initial investigation of documents contained in [7] has identified four major structural elements which can be used to identify forms. These are: horizontal lines (either solid or dotted), bounded rectangles, small checkboxes, and character cells or 'tooth' structures (Fig 1). Examination of all training data has shown that every form document contains one or more such structures. Detecting such structures in complex document images, however, is not a trivial problem. Attempting to segment a document image and classify regions is problematic due to frequent overlapping of neighboring regions, especially when dealing with completed forms. More traditional shape recognition techniques such as the generalized Hough Transform [8] are also inaccurate in the presence of noise, and also quite slow computationally. As all of the desired regions consist entirely of vertical and horizontal lines, our approach to the detection problem begins with finding all such lines in the candidate image. Once these lines are found, each is further processed to determine if it is a likely form structure.

Line Detection

We define a 'line' in a document image to be a contiguous or near-contiguous sequence of n 'on' pixels in the horizontal (vertical) direction, where n is directly proportional to the resolution of the image. As the smallest lines of interest are approximately the same width as a character, n is chosen as to correspond with a distance of 2mm in the original document. To detect such a sequence, we employ a one-dimensional summing filter in the horizontal (vertical) direction defined by the equation



(a)

(b) (c)

(d)

Figure 1. Four common form structures, (a) tooth structure, (b) checkbox, (c) rectangle, (d) line

$$S(x,y) = \frac{1}{n}\sum_{i=x}^{x+n} I(i,y) \qquad (1)$$

where *I(x,y)* is the original binary image. By applying a threshold to the resulting image, the starting points of all possible lines can be found.

$$L_H = S(x,y) \geq \tau \qquad (2)$$

Binary morphological operations can then be used to extend these starting points across all *n* pixels in the line segment. Figure 2 shows the result of line detection on a typical form image.

The detection process thus outlined is successful at detecting regions likely to contain lines, however also gives rise to a number of false positives. In particular, large regions black regions in the document such as images, thick vertical lines and large sections of text are often falsely detected as horizontal lines. To remove such regions we first segment the line image L_H into connected components, and calculate the height and width of each component. Components which do not satisfy a minimum width and width:height ratio are removed. This process also has the effect of removing valid horizontal lines which are connected to thick vertical lines, however as such lines are almost always borders or part of images, this is not undesirable.

Vertical lines by themselves do not constitute a possible text entry field. For this reason, all vertical lines which do not at some point cross a valid horizontal line are also removed. To allow for noise, small breaks in lines, and scanning errors, we relax this constraint somewhat, allowing vertical lines which are close (within *n* pixels) to either a horizontal line or another valid vertical line to be kept as well.

Line Grouping

Once all possible lines have been detected, we then attempt to combine these lines to form one of the four form structures.

In order to detect character cells or 'tooth' structures, each horizontal line is analysed for vertical line crossings, or near crossings. Such crossings must extend significantly in the vertical direction, since we assume that the horizontal line represents the bottom of the tooth structure. We then look for a periodic structure within these crossings, constrained by likely cell size. Due to noise, handwriting or other markings within the structure, it is possible that extra vertical crossings unrelated to the structure are present. In order to allow for this, an algorithm has been developed as follows:

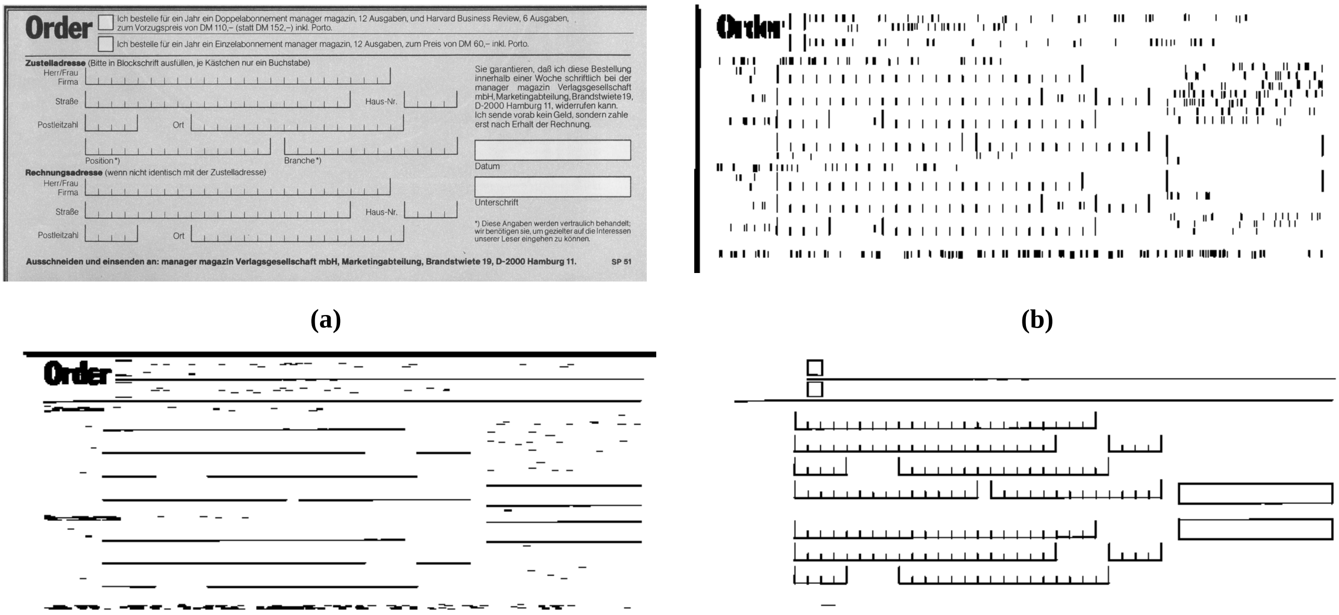
```
For every vertical line crossing not already part of structure:
    search for more crossings within search dist. x
    for each such crossing found:
        search line at same dist. ±5%
        if another crossing found,
            recalculate mean distance, search again
            if #crossings > 4, structure found.
```

The search distance *x* is proportional to the resolution of the document, and we have used a range of *n-5n* in our experiments with good results. In order to reduce the false detection rate, we have also enforced a criterion whereby a structure is *not* considered valid if more than half of its crossings are not fully joined. Finally, we search for a top bounding line, which is defined as a horizontal line within *n-5n* of the original lines, which crosses (or nearly crosses) each vertical segment of the structure. If two or more such lines are found, only the closest to the baseline is taken. Any such line found will still be considered for the baseline of further tooth structures.

For the detection of rectangles and boxes, we use a similar algorithm to that proposed in [3], whereby each set of candidate lines are checked to determine whether they form an enclosed area. In order to prevent rectangles being found in locations already covered by previously detected tooth structures, baselines of such structures are only considered as the *top* of a rectangle. Small breaks in the perimeter of rectangles are permitted, so long as they do not exceed 5% of the total distance. Rectangles that are completely covered by other rectangles are then removed. Regions whose area exceeds a certain size threshold are also removed, as these are unlikely to be text entry fields, and are more likely borders or frames.

A checkbox is defined as a special case of rectangle, where the following three criteria are met:

- The sides are of equal length (square)



Order □ Ich bestelle für ein Jahr ein Doppelabonnement manager magazin, 12 Ausgaben, und Harvard Business Review, 6 Ausgaben, zum Vorzugspreis von DM 110,– (statt DM 152,–) inkl. Porto.

□ Ich bestelle für ein Jahr ein Einzelabonnement manager magazin, 12 Ausgaben, zum Preis von DM 60,– inkl. Porto.

Zustelladresse (Bitte in Blockschrift ausfüllen, je Kästchen nur ein Buchstabe)

Herr/Frau Firma

Straße Haus-Nr.

Postleitzahl Ort

Position*) Branche*)

Rechnungsadresse (wenn nicht identisch mit der Zustelladresse)

Herr/Frau Firma

Straße Haus-Nr.

Postleitzahl Ort

Sie garantieren, daß ich diese Bestellung innerhalb einer Woche schriftlich bei der manager magazin Verlagsgesellschaft mbH, Marketingabteilung, Brandstwiete 19, D-2000 Hamburg 11, widerrufen kann. Ich sende vorab kein Geld, sondern zahle erst nach Erhalt der Rechnung.

Datum

Unterschrift

*) Diese Angaben werden vertraulich behandelt; wir benötigen sie, um gezielter auf die Interessen unserer Leser eingehen zu können.

Ausschneiden und einsenden an: manager magazin Verlagsgesellschaft mbH, Marketingabteilung, Brandstwiete 19, D-2000 Hamburg 11. SP 51

(a)

(b)

(c)

(d)

Figure 2. Results of form structure detection. (a) Original document image, (b) detected vertical line segments, (c) detected horizontal line segemnts, (d) final form structures

- Side length is within a given range (we use *n*- *5n*)
- Sides do not significantly extend beyond the corners of the rectangle

All horizontal lines that do not form part of any of the above structures are considered lines.

FORM CLASSIFICATION

The classification of documents into form and non-form classes is achieved using a score based on the presence of previously detected form structures combined with the amount of text contained in the document. Examination of a large number of forms has revealed that most do not contain as much text as other documents of a similar size. Thus, the presence of text in a document image has a negative impact on the probability of that document being a form. Numerous algorithms exist for the segmentation and extraction of printed text from documents , but for accuracy we have manually measured the amount of text present in each test document. As we are only interested in the body text of the document, any large segments such as headlines or titles are not included. We thus define the form probability score as:

$$p=\mathbf{w_1}d_{tooth}+\mathbf{w_2}d_{box}+\mathbf{w_3}d_{rect}+\mathbf{w_4}d_{line}-\mathbf{w_5}d_{text} \quad (3)$$

where d_{type} represents the total horizontal lineal distance covered by the given structure type, and **w** is a weighting vector. A positive *fps* value indicates that the document is likely to be a form. In order to obtain a true likelihood estimate, this value can be normalised, such that:

$$\hat{p}=\frac{p}{\left(d_{tooth}+d_{box}+d_{rect}+d_{line}+d_{text}\right)} \quad (4)$$

In order to calculate the weighting vectors we have processed a large number of both form and non-form documents, and examined the relationship between the amounts of each structure present. By constructing plots of dtext vs dtype for each structure type, it can be seen that there exists an almost linear separation between form and

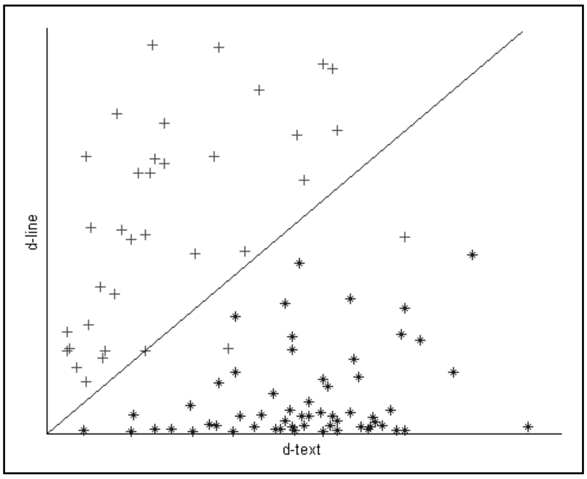


Figure 3. Plot of d-line vs d-text for a selection of form(+) and non-form(*) document images

non-form documents. We then find the gradient of this line and use it to calculate the corresponding weighting coefficient in **w**, assuming $\mathbf{w}_5 = 1$. Figure 3 shows an example of such a plot. It should be noted that we could find no non-form documents containing the tooth structure, meaning that the value of $\mathbf{w}_1$ would approach infinity. For this reason we have made this coefficient very large, approximately ten times the value of the next highest coefficient.

RESULTS

Experiments were conducted in two stages, using a set of 100 form and 200 non-form images acquired from a variety of sources, including the University of Washington database [7]. Firstly, the form structure detection algorithm was applied to all form images, and results compared to those calculated manually. Overall, 2443 of 2567 (95%) form structures were successfully detected as the correct type. Of those structures that were not detected successfully, approximately two thirds were due to misclassification of one structure as another, with the remaining missed entirely. An additional 181 form structures were falsely detected, with almost all of these being small lines. A typical form image with all detected structures is shown in Figure 2. Table 1 shows the confusion matrix for this experiment.

Table 1. Confusion matrix for detection of form structures

Actual Type	Detected Type				
	tooth	rect.	box	line	missed
tooth	229	0	0	3	0
rect	0	767	9	38	0
box	0	10	318	4	14
line	1	15	1	1253	19
none	0	9	4	168	x

The second stage of experiments involved calculating the normalised form probability score for each test document using the detected structures and known text amounts. Those documents obtaining a positive score were classified as forms, with the remaining classified as non-forms. From a total of 300 (100 form, 200 non-form) document images, 258 were correctly classified. Of those that were misclassified, 6 form images were missed, and 36 non-form images falsely detected. The overall error rate of the test was approximately 14%. Total processing time for both structure detection and form classification is approximately 5 seconds on a Pentium 3 600MHz computer.

CONCLUSIONS AND FUTURE RESEARCH

This paper has presented a technique to distinguish form documents from other types by identifying common structures usually present in such images. Experimental results have shown our algorithm for detecting such structures to be accurate and robust, with over 95% of structures detected correctly. Classification of form and non-form documents is accomplished by comparing the total number of such structures to the amount of text in the document, creating a form probability score. This statistic has shown to perform well, with almost all form images correctly identified and a false detection rate of under 15%.

Future research will aim to more accurately model the typical line and rectangle structure in forms by examining surrounding text. This should greatly reduce the number of false positive results.

REFERENCES

[1] D. Wang and S. N. Srihari, "Analysis of form images," Proc. of First International Conference on Document Analysis and Recognition, 1991.

[2] J. Yuan, Y. Tang, and C. Y. Suen, "Four directional adjacency graphs (FDAG) and their application in locating fields in forms," Proc. of Third International Conference on Document Analysis and Recognition, 1995.

[3] L. Xingyuan, D. Doermann, W.-G. Oh, and W. Gao, "A robust method for unknown forms analysis," Proc. of Fifth International Conference on Document Analysis and Recognition, 1999.

[4] H. Shinjo, K. Nakashima, M. Koga, K. Marukawa, Y. Shima, and E. Hadano, "A method for connecting desappeared junction patterns on frame lines in form documents," Proc. of 4th Int. Conf. on Document Analysis and Recognition, 1997.

[5] O. Hori and D. Doermann, "Robust table-form structure analysis based on box-driven reasoning," Proc. of Third International Conference on Document Analysis and Recognition, 1995.

[6] H. Fujisawa and Y. Nakano, "Segmentation methods for character recognition: from segmentation to document structure analysis," *Proceedings of the IEEE*, vol. 80, pp. 1079-1092, 1992.

[7] I. Phillips, S. Chen, and R. Haralick, "CD-ROM Document Database Standard," Proc. of Second International Conference on Document Analysis and Recognition, 1993.

[8] D. H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern Recognition*, vol. 13, pp. 111-122, 1981.

A Closed Form Solution to the Reconstruction and Multi-View Constraints of the Degree d Apparent Contour

David McKinnon[1], Barry Jones[2] and Brian C. Lovell[1]
[1]School of Information Techology and Electrical Engineering
Intelligent Real-Time Imaging and Sensing (IRIS) Group
University of Queensland
[2]Department of Mathematics
University of Queensland

Abstract

This paper presents a novel theoretical approach to calculating the apparent contour of a smooth surface. The problem is formulated as a dual space intersection of algebraic tangent cones, which we will consider to be the members of degree d hypersurfaces. The well known thereotical foundation for multi-view geometry is extended in light of this to solve the problems of triangulation and forming multi-view matching constraints for degree d apparent contours.

1 Introduction

The problem of reconstructing a static scene from mutliple images is rich field of research [3]. There are a wide range of algorithms that can be used to reconstruct linear features such as points and lines from there projections in mutliple images as well as the inverse problem of finding the egomotion of the camera observing the scene.

However there has been far less research on the reconstruction and multiple view geometry of arbitrary curves observed in the scene [5], although the simplest cases of the conic and quadric have been investigated to a greater extent [4, 6, 1].

Broadly speaking, there are two different catagories of curves commonly observed in a scene, these are the *static* and *apparent contours*. Static curves are rigid curves they may commonly occur as textures on a surface or a thin thread of wire or other such objects. Each point on a static curve obeys the regular epipolar transfer equations, however presently there is only one approach to finding a closed form algebraic solution for their geomerty in the general degree d case [5] and several for the special degree 2 case [4, 6, 1].

We say the degree 2 case of the static curve is a special case since it is the only type of 3D algebraic curve that can be described by one equation. The other class of geometric objects that can be described by one equation in 3D are the class of smooth degree d surfaces. The projection of a smooth degree d surface forms a degree d apparent contour in the image, the apparent contour in this sense is the intersection of the dual of the surface with the image plane [5, 1].

In this paper we consider the class of all degree d surfaces and their associated apparent contours to be representable as degree d hypersurfaces, thus creating a generic form of algebra for their manipulation. This paper will only present a brief theoretical overview of the concepts, although the computationally tractable cases of the apparent contour triangulation and multi-view geometry have been simulated in noise free conditions (up to degree 10). All the geometry and algebra presented in this paper is projective [8]

2 Linear Mutli-View Geometry

This section will outline the notation and the basic building block of linear mulit-view geometry. The development of the ideas underlining multi-view matching constraints and triangulation of linear features is heavily influenced by the notation and stucture of the linear matching constraints presented in [10]. Due to space considerations this paper assumes that the reader is familiar with majority of these concepts.

2.1 Features

The first consideration when dealing with the multi-view geometry of linear features is their notation. Consistantly we will refer to features as any type of geometric object observed in a scene, be this points, lines and planes in the linear case or hypersurfaces in the degree d case.

Table 1 summarises the notation and degrees of freedom (DOF) for the group of linear features in the projective plane ($[A, B, C] \in \mathbb{P}^2$).

Hyperplane	$\mathbb{P}^2$	$\mathbb{P}^{2*}$	DOF
Points	x^A	$x^{[A]} = \epsilon_{ABC} x^A = x_{BC}$	2
Lines	$x^{[AB]}$	$x^{[AB]} = \epsilon_{ABC} x^A = x_{AB}$	1

Table 1. Linear features and there duals in $\mathbb{P}^2$

Similarly, Table 2 summarises the notation and the DOF for linear features in projective space ($[a, b, c, d] \in \mathbb{P}^3$).

Hyperplane	$\mathbb{P}^3$	$\mathbb{P}^{3*}$	DOF
Points	x^a	$x^{[a]} = \epsilon_{abcd} x^a = x_{bcd}$	3
Lines	$x^{[ab]}$	$x^{[ab]} = \epsilon_{abcd} x^{ab} = x_{cd}$	2
Planes	$x^{[abc]}$	$x^{[abc]} = \epsilon_{abcd} x^{abc} = x_d$	1

Table 2. Linear features and there duals in $\mathbb{P}^3$

These tables demonstrate the process of dualization for linear feature types via the antisymmetrization operator $[\ldots]$. The antisymmetrization operator should be considered as a determinantal method to generate the algebra for linear features, by performing an alternating tensor contraction over the space to which the operator is applied [2].

2.2 Triangulation

Triangulation is the process of calculating a feature in $\mathbb{P}^3$ from two or more of its projections in $\mathbb{P}^2$. Firslty, we must consider the projection operator (P^α_β) or *camera matrix* that denotes the projection of linear features from the scene to the image plane ($P^\alpha_\beta : \mathbb{P}^3 \rightarrow \mathbb{P}^3$). Table 3 summarises the range of projection operators for linear features.

Hyperplane	$\mathbb{P}^3$	$\mathbb{P}^{3*}$
Point	$x^A \sim P^A_a x^a$	-
Line	$x^{[AB]} \sim P^{[A}_{[a} P^{B]}_{b]} x^{[ab]}$	$x_{[BC]} \sim P^{[c}_{[B} P^{d]}_{C]} x_{[cd]}$
Plane	-	$x_A \sim P^d_A x_d$

Table 3. Projection operators for linear features

Generally it may be stated that $\lambda x^\alpha = P^\alpha_\beta x^\beta$, where λ is an arbitrary scale factor.

Having observed $2 \ldots n$ image features, triangulation proceeds through the *reconstruction equations*,

$$\begin{pmatrix} P^{A_1}_a & x^{A_1} & 0 & \cdots & 0 \\ P^{A_2}_a & 0 & x^{A_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P^{A_n}_a & 0 & 0 & \cdots & x^{A_n} \end{pmatrix} \begin{pmatrix} x^a \\ \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix} = \mathbf{0} \quad (1)$$

where the resulting nullvector of these equations presents a solution for the scene feature and the scale factors λ_n. The stack of camera matrices on the left hand side of (1) is referred to as the *joint image projection matrix* ($P^\gamma_a \Rightarrow P^{A_1 A_2 \ldots A_n}_a$) and can be thought of as a vector of camera matrices that projects a common feature from the scene (x^a) to its *joint image* feature location ($x^\gamma \Rightarrow x^{A_1 A_2 \ldots A_n}$).

The reconstruction equations have,

$$(\sum_n DOF^n_i - DOF_s - 1)(DOF_s + 1) - n + 1 \quad (2)$$

DOF, where DOF^n_i and DOF_s denote the DOF of the nth image feature and scene features respectively. Furthermore the reconstruction equations are rank-$(DOF_w + n)$.

2.3 Multi-View Constraints

Mutli-View constraints provide a linear relationship between projections of scene features observed in two or more images. Multi-View constraints provide a means to calculate the structure of the scene and the egomotion of the camera. The approach to building multi-view constraints stems from the representation of a subspace in the *Grassmann* algebra. Here we wish to find a d_w-dimensional subspace for the scene (where the scene is embedded in $\mathbb{P}^{d_w}$), from the joint image projection matrix. This is achieved by antisymmetrizing over $d_w + 1$ of the joint images scene indeterminants, with corresponding unique choices of *any* $d_w + 1$ of the images indeterminants,

$$I^{\gamma_0 \cdots \gamma_d} \equiv \frac{1}{(d_w + 1)!} P^{\gamma_0}_a \cdots P^{\gamma_d}_d \epsilon^{a \ldots d} \equiv P^{\gamma_0}_{[a} \cdots P^{\gamma_d}_{d]} \quad (3)$$

(3) is known as the *Joint Image Grassmannian*. The selection of the image indeterminants $\gamma_0 \ldots \gamma_d$ from the rows of the joint image projection matrix determines which images the resulting matching constraint will represent. The choice of rows obeys the simple rules that for an image to be included in the matching constraint, it must be represented by at least one row, and less than $d_i + 1$ rows (where the image plane is embedded in $\mathbb{P}^{d_i}$). This leads to well known set of matching tensors (Table 4) and also explains why there is at most 4-view matching constraints for points and lines.

There are many variations of the atypical matching constraints given in Table 4, see [10, 3].

Views	Tensor	Constraint
2	$I^{B_1C_1B_2C_2}$	$I^{[B_1C_1B_2C_2}x^{A_1}x^{A_2]} = 0$
3	$I^{B_1C_1B_2B_3}$	$I^{[B_1C_1B_2B_3}x^{A_1}x^{A_2}x^{A_3]} = \mathbf{0}\ldots$
4	$I^{B_1B_2B_3B_4}$	$I^{[B_1B_2B_3B_4}x^{A_1}x^{A_2}x^{A_3}x^{A_4]} = \mathbf{0}\ldots$

Table 4. Atypical Linear Matching Constraint Tensors

3 Hypersurfaces

It is at this point that we enter into profitable new territory with the introduction of hypersurfaces into the tensor notation.

Definition A degree d hypersurface is denoted as the d-fold symmetric product (symmetrization) $(\ldots)$ of an indeterminant [2]. The resulting hypersurface is considered to be embedded in the space in which the symmetrization operator is applied. That is,

$$x_{\underbrace{(\gamma\ldots\gamma)}_{\mathrm{d-fold}}} \tag{4}$$

or in the Algebro-Geometric notation [9],

$$\underbrace{\mathbb{P}^n \times \ldots \times \mathbb{P}^n}_{\mathrm{d-fold}} \backslash \mathcal{S}_n^d$$

where $\mathcal{S}_n^d$ is the d-fold symmetric permutation group, this may also be considered as the degree d, n space *Veronese embedding* (ν_n^d).

We can state that hypersurfaces are generically points in a $\mathbb{P}^{\nu_n^d}$ dimensional space, where $\nu_n^d = \binom{n+d}{n} - 1$, thus they have $\nu_n^d - 1$ DOF. Some common examples of hypersurfaces are the conic $x_{(AA)}x^Ax^A = 0$, and the quadric $x_{(aa)}x^ax^a = 0$ hypersurfaces. Equivalent dual hypersurfaces are simply $x^{(AA)}x_Ax_A = 0$ where $x^{(AA)} \in \mathbb{P}^{\nu_2^2 *}$.

4 Degree d Multi-View Geometry of Hypersurfaces

Now we are ready to observe the degree d triangulation and Multi-View Geometry, of hypersurfaces. The development in this section will follow the exact path we took in Section 2, where in this case points and lines will be replaced by hypersurfaces and dual hypersurfaces.

4.1 Triangulation of Hypersurfaces

As in Section 2.1 our first step in solving the triangulation problem is addressing the nature of projection operators for degree d hypersurfaces. However, in this case we are concerned with the degree d embedding of hypersurfaces in $\mathbb{P}^2$ and $\mathbb{P}^3$ respectively. Firslty, we should note that this concept is not completely new, in [4, 1] an equivalent observation was made for the projection of degree 2 hypersurfaces. Table 5 summarises the range of projection operators for degree d hypersurfaces.

	$\mathbb{P}^{\nu_3^d}$	$\mathbb{P}^{\nu_3^d *}$
Hypersurface	$P^{(a\cdots}_{(A\cdots}P^{a)}_{A)}$	$P^{(A\cdots}_{(a\cdots}P^{A)}_{a)}$

Table 5. Projection operators for degree d hypersurfaces

Generally, it may be stated that projection of hypersurfaces is denoted as $\lambda x_{(A\ldots A)} = P^{(a\cdots}_{(A\cdots}P^{a)}_{A)}x_{(a\ldots a)}$ and dually $\lambda x^{(A\ldots A)} = P^{(A\cdots}_{(a\cdots}P^{A)}_{a)}x^{(a\ldots a)}$. We can also state that these projection matrices are the d-fold symmetric powers of the regular point projection matrix (and its dual), thus resulting in the dimension of these matrices being $((\nu_3^d+1)\times(\nu_2^d+1))$ and $((\nu_2^d + 1) \times (\nu_3^d + 1))$ respectively.

Since we are concerned with finding the the equation of the surface generating the apparent contour in the image. We must take the intersection of the dual hypersurfaces tangent cone with the image plane. For notational compactness we will assign $(A_n \cdots A_n) = \eta_n$ and $(a \cdots a) = \mu$. This leads us to the equivalent set of *dual* reconstruction equations for degree d hypersurfaces,

$$\begin{pmatrix} P_\mu^{\eta_1} & x^{\eta_1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ P_\mu^{\eta_n} & 0 & \cdots & x^{\eta_n} \end{pmatrix} \begin{pmatrix} x^\mu \\ \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} = \mathbf{0} \tag{6}$$

again the resulting nullvector of these equations presents a solution for the scene hypersurface (x^μ) and the scale factors λ_n.

The minimum number of image hypersurfaces required to reconstruct a degree d hypersurface is given as the lower bound of,

$$\nu_3^d \geq n \geq \frac{(d^2 + 6d + 11)}{3(d + 3)} \tag{7}$$

[5] the lower bound n must be rounded up to the closest integer value. The upper bound is the limit on the number of images for the resulting matching constraint. The DOF of these reconstruction equations and their rank are analogous to those stated for (1).

4.2 Multi-View Constraints for Degree 2 Hypersurfaces

Before we address the general formulation for the degree d matching constraints for hypersurfaces, we will tred gently by outlining the concepts for degree 2.

Firstly, it is not clear in general the make-up or practical relevance of features embedded in $\mathbb{P}^2$ or $\mathbb{P}^3$ that have DOF other than $\nu_n^d - 1$ (ie. hypersurfaces).

An application of (7) suggests the presence of degree 2 matching constraints for two through to ten image projections. Again, the object in building the matching constraints is select ν_3^2 unique rows from the joint image projection matrix to make up the matching constraints. The corresponding matching constraints are given in Table 6.

Views	Tensor	Constraint
2	$I^{B_1 \cdots F_1 B_2 C_2 D_2 E_2 F_2}$	$I^{[\cdots} x^{A_1} x^{A_2]} = 0$
3	$I^{B_1 \cdots F_1 B_2 C_2 D_2 E_2 B_3}$	$I^{[\cdots} x^{A_1} x^{A_2} x^{A_3]} = \mathbf{0}_{\cdots}$
4	$I^{B_1 \cdots F_1 B_2 C_2 D_2 B_3 B_4}$	$I^{[\cdots} x^{A_1} x^{A_2} x^{A_3} x^{A_4]} = \mathbf{0}_{\cdots}$
$\vdots$	$\vdots$	$\vdots$
10	$I^{B_1 B_2 \cdots B_{10}}$	$\ldots$

Table 6. Degree 2 Matching Constraint Tensors

It is not clear what the actual effect of selection of different rows from the joint image has on the resulting matching constraint (future work may included a thorough investigation of this uncertainty along the lines of [**?**]). But from the initial experimentation we have found that any combination of rows that meets the aforementioned requirements for defining a Grassmann subspace is adequate to construct the matching tensor. The most pertinent factor in selecting a number of rows to form the matching constraints, is minimising the size of the actual matching tensor.

The selection of k rows from a space of size n will result in the size of associated dimension of the matching tensor being $\binom{n}{k}$, so naturally values close to either n or 1 will yield smaller matching constraints.

4.3 Multi-View Constraints for Degree d Hypersurfaces

Finally, we can now see that an application of equation (7) will give the upper and lower bounds for the degree d multi-view constraints and an application of equation (6) will generate the reconstruction equations for the problem. Any selection of rows from the reconstruction equations meeting the aforementioned criteria of a valid subspace, will be adequate to reconstruct the degree d matching constraints.

5 Conclusions and Future Work

The authors have presented a general closed form linear method for the solution of degree 2 curves and surfaces which extends to the solution of degree d surfaces. The essential problems of triangulation and multi-view matching constraints for these features have been considered, unfortunately due space restrictions a full account of these geometries has been limited.

The authors have also considered a practical scheme to calculate the apparent contours through the fitting of cubic NURBS curves. NURBS are the only alternative since they have the essential property of projective closure. Once the NURBS curves have been fitted to the image data they must then be converted into their *implicit* representation via the process of implicitization [7]. This topic will be considered in future work.

References

[1] G. Cross. *Surface Reconstruction from Image Sequences : Texture and Apparent Contour Constraints*. PhD thesis, University of Oxford, Trinity College, 2000.

[2] W. Greub. *Multilinear Algebra*. Springer-Verlag, 1978.

[3] R. I. Hartley and A. Zisserman. Multiple view geometry in computer vision. Cambridge University Press, 2000.

[4] F. Kahl and A. Heyden. Using conic correspondences in two images to estimate the epipolar geometry. *Int. Conf. on Computer Vision*, 1998.

[5] J. Y. Kaminski, M. Fryers, A. Shashua, and M. Teicher. Multiple view geometry of non-planar algebraic curves. *Int. Conf. on Computer Vision*, 2001.

[6] L. Quan. Conic reconstruction and correspondence from two views. *Transactions on Pattern Analysis and Machine Intelligence*, 18(2), 1996.

[7] T. W. Sederberg. Applications to computer aided design. In *Proc. of Symposia in Applied Mathematics*, volume 53, pages 67–89. AMS, 1998.

[8] J. G. Semple and G. T. Kneebone. *Algebraic Projective Geometry*. Oxford University Press, 1952.

[9] I. R. Shafarevich. *Basic Algebraic Geometry*. Springer-Verlag, 1974.

[10] W. Triggs. The geometry of projective reconstruction i: Matching constraints and the joint image. *Int. Conf. on Computer Vision*, pages 338–343, 1995.