

Proceedings of the international workshop on computer vision applications (CVA), 23rd March, 2011, Eindhoven University of Technology

Citation for published version (APA):

With, de, P. H. N., & Shrestha, P. (Eds.) (2011). *Proceedings of the international workshop on computer vision applications (CVA), 23rd March, 2011, Eindhoven University of Technology*. Technische Universiteit Eindhoven.

Document status and date:

Published: 01/01/2011

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.



Proceedings of the International Workshop on Computer Vision Applications (CVA)

23rd March, 2011
Eindhoven University of Technology

Proceedings of the International Workshop on Computer Vision Applications (CVA)

Peter H.N. de With and Prarthana Shrestha (eds.)

Proceedings of a one-day workshop organized by the
*Werkgemeenschap voor Informatie en Communicatietheorie and
IEEE Benelux Chapters on Consumer Electronics and Information
Theory*

in conjunction with the
*Electrical Engineering Department of the
Technische Universiteit Eindhoven*

March 23rd, 2011

Sponsors



Copyright © 2011 by the authors. Considerable parts of this text have been or will be published by the IEEE or related institutes.

All rights reserved. No part of this publication may be stored in a retrieval system, transmitted or reproduced in any form or by any means, including but not limited to photography, magnetic, or other record, without prior agreement and written permission of the respective authors.

A catalogue record is available from the Eindhoven University of Technology Library

ISBN: 978-90-386-2474-7

Contents

Preface.....	v
Workshop Program.....	vii
Presenter Biographies.....	viii

Presentations

1. Medical Image Analysis: from Content to Care	1
<i>Marcel Breeuwer</i>	
2. Human Action Representation and Recognition	23
<i>Ling Shao</i>	
3. Visual Search: What's Next?.....	41
<i>Cees Snoek</i>	
4. High-tech Eyes for Industry and Society.....	61
<i>Jan Baan</i>	
5. Multi-camera Video Analysis for Activity Monitoring of People.....	65
<i>Peter Van Hese</i>	

Papers and Posters

1. Cost-efficient Nucleus Detection in Histopathology Images using AdaBoost	79
<i>Jelte Peter Vink and Marinus Bastiaan van Leeuwen (Philips Research)</i>	
2. Sparse Window Stereo Matching.....	83
<i>Sanja Damjanovic, Ferdinand van der Heijden and Luuk J. Spreeuwers (University of Twente)</i>	
3. Large Scale Detection, Classification and Localization of Traffic Signs	87
<i>Ivo Creusen and Lykele Hazelhoff (Cyclomedia Technology B.V.)</i>	
4. Face tracking camera system for surveillance.....	91
<i>Rick Peerlings and Rob Wijnhoven (ViNotion)</i>	
5. Automatic Assessment of Customers' Buying Behavior	95
<i>Mirela Popa, Leon Rothkrantz, Pascal Wigger (Delft University of Technology) Caifeng Shan and Tommaso Gritti (Philips Research)</i>	
6. Detection of Human Groups in Videos.....	99
<i>Selcuk Sandikci, Svitlana Zinger and Peter H.N. de With (Eindhoven University of Technology)</i>	
7. Towards Demographic Classification in Unconstrained Environments	103
<i>Caifeng Shan (Philips Research)</i>	

8. Vital Signs Camera	107
<i>Ingmar van Dijk, Adrienne Heinrich (Philips Research)</i>	
9. Instantaneously Responsive Subtitle Localization & Classification for TV Applications.....	111
<i>Bahman Zafarifar and Jingyue Cao (Trident Microsystems (Europe) B.V.)</i>	
<i>Peter H. N. de With (Eindhoven University of Technology)</i>	
10. Context analysis: sky, water and motion.....	115
<i>S. Javanbakhti, S. Zinger, J. Han, P. H. N. de With (Eindhoven University of Technology)</i>	
11. Seeing the user: CV in Support of Self Adaptive User Interfaces	117
<i>Hester Bruikman, Hao Wang, and Roy van de Korput (Philips Consumer Lifestyle)</i>	
12. Towards Multi-View Ship Detection for Maritime Surveillance.....	121
<i>Rob Wijnhoven and Kris van Rens (ViNotion)</i>	

Preface

These proceedings provide an overview of both invited lectures and poster presentations of International Workshop on Computer Vision and Applications, discussing the achievements of experts in the field of both disciplinary research and applications. The image and video analysis grows gradually into a mature field. This holds for both disciplinary researches where concepts come to the foreground that form the correct technical basis for multiple applications, but also in the application area, where system iterations give a driving force for improvements.

The Workshop program and its contents provides a sample moment of the state of the art, where the invited speakers come from different backgrounds, such as the medical, surveillance and multimedia fields, or similar. The program has been chosen such that representatives of those fields are among the presenters, in order to learn from each other's concepts and enhance cross fertilization.

With the third European project ViCoMo on analysis on its way and various national research projects just completed such as iCare in the area of analysis applications, the workshop is highly actual as ever and the potential of the area for the industry is still promising, besides some already established applications. We hope that you will enjoy the sampled results in one way or the other.

Finally, we would like to acknowledge the excellent cooperation with the IEEE Chapters of the Benelux on IT and CE, who always support us in organizing such an event and the University of Technology of Eindhoven, for hosting us at their premises.

Prof.dr.ir. Peter H.N. de With

Dr. ir. Prarthana Shrestha

Board member IEEE Benelux Chapter IT
Professor Video Coding and Architectures,
Electrical Engineering Faculty
University of Technology Eindhoven,
The Netherlands

Research scientist
Dept. SPS Video Coding and Architectures
Electrical Engineering Faculty
University of Technology Eindhoven,
The Netherlands

Earlier releases in this series:

- Proceedings Workshop on “Embedded Video Streaming Technology (MPEG-4) and the Internet”, IEEE Benelux Chapter on Consumer Electronics, ISBN 90-386-0991-4, P.H.N. de With (Ed.), Technische Universiteit Eindhoven, The Netherlands, December 2001 (155 pages).
- Proceedings Workshop on “The Design of Multimedia Architectures”, IEEE Benelux Chapter on Consumer Electronics, ISBN 90-386-0822-5, P.H.N. de With (Ed.), Technische Universiteit Eindhoven, The Netherlands, December 2003 (136 pages).
- Proceedings Workshop on “Resource Management for Media Processing in Networked Embedded Systems”, IEEE Benelux Chapter on Consumer Electronics, ISBN 90-386-0544-7, R.J. Bril and R.Verhoeven (Eds.), Technische Universiteit Eindhoven, The Netherlands, March 2005 (142 pages).
- Proceedings Workshop on “Content Generation and Coding for 3D-Television”, IEEE Benelux Chapter on Consumer Electronics, P.H.N. de With, C. Varekamp, D. Farin, Y. Morvan (Eds.), ISBN 90-386-2062-4, The Netherlands, June 2006 (CD-ROM).
- Proceedings Workshop on IP-television (IP-TV), IEEE Benelux Chapter on Consumer Electronics, Peter H.N. de With and Goran Petrovic (Eds.), ISBN 978-90-6144-988-1, Technische Universiteit Eindhoven, The Netherlands, January 2007 (99 pages).

Program of “International Workshop on Computer Vision Applications (CVA)”

One-day workshop at the Eindhoven University of Technology, on 23rd March, 2001. Organized by the *Werkgemeenschap voor Informatie en Communicatietheorie* (WIC), the *IEEE Benelux Chapter on Information Theory* and the *SPS-VCA group, TU Eindhoven*.

Workshop program

09.00-09.30 hrs. Registration and coffee
09.30-09.40 hrs. Opening by Prof.dr.ir. Peter H.N. de With (TU Eindhoven, SPS-VCA)
09.40-10.25 hrs. Prof.dr. Marcel Breeuwer (Philips Healthcare and TU/e), “Medical Image Analysis: from Content to Care”
Coffee break
10.55-11.40 hrs. Prof.dr. Ling Shao (Univ. of Sheffield, UK), “Human Action Representation and Recognition.”
11.40-12.25 hrs. Dr. Cees Snoek (Univ. of Amsterdam & UC Berkeley, USA), “Visual search: what's next?”
12.25-13.45 hrs. Lunch
13.45-14.30 hrs. Ir. Jan Baan (TNO Netherlands), “High-tech eyes for industry and society”
14.30-15.30 hrs. Extended break with poster session
15.30-16.15 hrs. Dr.ir. Peter Van Hese (Univ. of Ghent, Belgium), “Multi-camera video analysis for activity monitoring of people”
16.15-16.20 hrs. Closing t.b.d.

Organization committee

Prof.dr.ir. Peter H.N. de With (TU Eindhoven)
Dr. ir. Prarthana Shrestha (TU Eindhoven)
Dr. Jungong Han (CWI Amsterdam)
Dr.ir. Egbert Jaspers (ViNotion, Project leader ViCoMo)
Ir. Ralph Braspenning (Philips, Project leader iCARE)

Presenter Biographies

Marcel Breeuwer was born in Haarlem, The Netherlands, in 1957. In 1982 he received his degree in Electrical Engineering from the Technical University of Delft, The Netherlands. In 1985, he received his PhD from the Free University of Amsterdam, The Netherlands, for his research on supplementing lipreading with auditory information. From 1985 until 1997 he was Research Scientist at the Philips Research Laboratories, Eindhoven, The Netherlands, where he investigated data compression of audio, video and medical images and where he was heading the video coding cluster. In 1997 he started as Senior Scientist at Philips Healthcare, Best, The Netherlands, in the area of image-guided surgery and medical image processing. In 2006 he became Principal Scientist and head of the cardiovascular team in the Clinical Science & Advanced Development department of the Business Unit Clinical Informatics Solutions. Focus of this team was R&D on medical image analysis applications for supporting the care of patients with cardiovascular diseases. In January 2011 he moved to the MR Clinical Science department, where he is now responsible for the domain of MR Image Analysis & Visualization. He is (co)author of over 100 scientific publications and is (co)inventor of over 40 patent applications (27 in the domain of healthcare). He is part-time professor (1 day/week) in the Biomedical Image Analysis group of the Biomedical Engineering department of the Technical University Eindhoven, The Netherlands, with focus on cardiovascular image analysis and visualization. He is member of the Board of the Dutch Society of Pattern Recognition and Image Processing (NVPBV).

Ling Shao received the BEng degree in Electronic Engineering from the University of Science and Technology of China (USTC), the MSc degree in Medical Image Analysis and the PhD (DPhil.) degree in Computer Vision at the Robotics Research Group from the University of Oxford.

Dr. Ling Shao is currently a Senior Lecturer (Associate Professor) in the Department of Electronic and Electrical Engineering at the University of Sheffield, UK. Before joining Sheffield University, he worked for 4 years as a Senior Research Scientist in the Video Processing and Analysis Group, Philips Research Laboratories, Eindhoven, The Netherlands. Prior to that, he worked shortly as a Senior Research Engineer at the Institute of Electronics, Communications and Information Technology, Queen's University of Belfast. His research interests include Computer Vision, Pattern Recognition and Video Processing. He has published over 60 academic papers in refereed journals and conference proceedings and has filed over 10 patent applications. Ling Shao is an associate editor of the International Journal of Image and Graphics, the EURASIP Journal on Advances in Signal Processing, and Neurocomputing, and has edited several special issues for journals of IEEE, Elsevier and Springer. He has been serving as Program Committee member for many international conferences, including ICIP, ICASSP, ICME, ICMR, ACM MM, CIVR, BMVC, etc. He is a senior member of the IEEE.

Cees G.M. Snoek received the MSc degree in business information systems (2000) and the PhD degree in computer science (2005) both from the University of Amsterdam, The Netherlands, where he is currently a senior researcher at the Intelligent Systems Lab Amsterdam. He was a Visiting Scientist at Informedia, Carnegie Mellon University, USA (2003) and the Computer Vision Group at UC Berkeley, USA (2010-2011). His research interests focus on visual retrieval. He has published over 90 refereed book chapters, journal and conference papers in this field, and serves on the program committee of the major conferences in multimedia, computer vision, and information retrieval. Dr. Snoek is a lead researcher of the award-winning MediaMill Semantic Video Search Engine, which is a consistent top performer in the yearly NIST TRECVID evaluations. He is co-initiator and co-organizer of the annual VideOlympics, co-chair of: the Intelligent Multimedia

Mining Workshop, SPIE Multimedia Content Access conference 2010, 2011, Multimedia Grand Challenge at ACM Multimedia 2010, and area chair of the ACM International Conference on Multimedia 2011. He is a lecturer of post-doctoral courses given at international conferences and European summer schools. He is a member of ACM and IEEE. Dr. Snoek received a young talent (VENI) grant from the Netherlands Organization for Scientific Research in 2008, and a Fulbright visiting scholar grant in 2010. Both his PhD students have won best paper awards.

Jan Baan received the MSc degree in Technical Physics on the TU Delft in 1997. After that he continued his research on acoustical imaging in concert halls as research assistant in the research group Seismic and Acoustic of the TU Delft.

Since 2010 he works at TNO in the field of computer vision. TNO is a Dutch contract research organization. His work focuses on video processing and 3D reconstruction and understanding techniques. He developed computer vision applications in the domain of social security, mobility, sport and agriculture inspection and automation. One of his most important developments of the last years is video-based traffic monitoring (VBM), where individual vehicles are tracked with cameras alongside the road. It is successfully used for evaluation of traffic behavior. VBM is an important part of the new A270 test site between Helmond and Eindhoven, where fifty cameras follow vehicles over a distance of five kilometers.

Jan Baan is involved in various other projects, where his interest is to find innovative solutions, with a practical implementation approach.

Peter Van Hese received both the MSc degree in Electrical Engineering and the PhD degree in Engineering from Ghent University, Ghent, Belgium, in 2000 and 2008, respectively. His PhD research was based on a cooperation between the Medical Image and Signal Processing research group (MEDISIP) within the Department of Electronics and Information Systems (ELIS) at Ghent University, and the Department of Neurology (Epilepsy Monitoring Unit) at the Ghent University Hospital. Since 2010 he is a postdoctoral researcher at the Image Processing and Interpretation research group (IPI) at the Department of Telecommunications and Information Processing (TELIN) at Ghent University. His research interests include biomedical signal processing, and video processing and analysis in distributed smart camera networks.

Medical Image Analysis: from Content to Care

Prof. Dr. Marcel Breeuwer

Principal Scientist, Philips Healthcare



Medical Image Analysis

From Content to Care



Marcel Breeuwer

Part-time Professor
Eindhoven University of Technology
Biomedical Engineering
BioMedical Image Analysis

Principal Scientist
Philips Healthcare Best
Imaging Systems
MR Clinical Science



International Workshop on Computer Vision Applications - 23 March 2011



Overview

Medical image analysis

- Introduction:
 - Trends in healthcare
 - The patient care cycle
 - The need for clinical decision support
 - The role of medical image analysis & visualization
- Segmentation algorithms:
 - Active contouring
 - Vessel tracking
- Example applications:
 - Diagnosis of coronary-artery disease
 - Prediction of the risk of abdominal aortic aneurysm rupture

2

PHILIPS

TU/e Technische Universiteit
Eindhoven
University of Technology

Introduction

PHILIPS

TU/e Technische Universiteit
Eindhoven
University of Technology

Trends in healthcare


Cost control, quality improvement

- Aging population: (now: 15% 65+, 2040: > 25%)
 - Growing care demand (40% more CV patient in 2025)
 - Shortage of healthcare professionals (nr. professionals stable)
- Increasing amount of information per patient (medical imaging)
- Better informed patients (internet)
- Limited efficiency & effectiveness (errors do occur!)

⇒ Increasing cost of healthcare: 15% of GDP* by 2015


Potential solutions:


- Improve effectiveness (quality ↑)
- Improve efficiency (speed ↑, effort ↓)
- ...



*GDP = Gross Domestic Product (market value of all goods and services within the borders of a country per year)


4





The patient care cycle

A series of care steps





Between entering & leaving care, the patient goes through a series of **care steps**

In each step, **decisions** about the most appropriate care must be taken, based on available **patient-specific information & knowledge**

Decision support is needed to optimally benefit from the plurality of information involved

5





Clinical decision support

Using medical image analysis & visualization

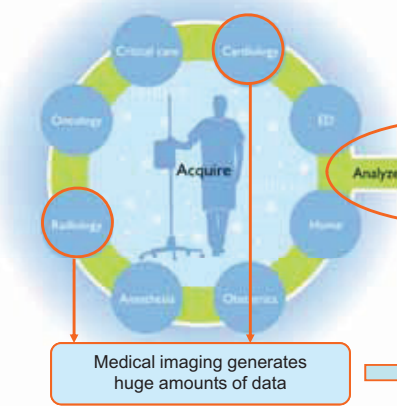



Image analysis & visualization to derive & present the essential information

Analyze Interpret Present



Medical imaging generates huge amounts of data

➡

From hundreds of Mbytes to a **few decisions**

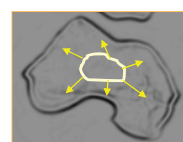
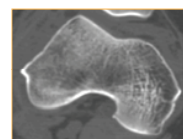
6

Algorithms

Segmentation with Active Contours – 1

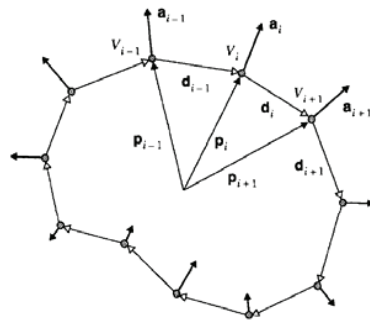
Active contouring

- An initial contour is available
- The contour is deformed by attracting it to specific image features (external force)
- The smoothness or curvature of the contour is constrained (internal force)
- Deformation is stopped when the contour does no longer change



Segmentation with Active Contours – 2

Discrete contour model



- The model consists of a set of vertices (nodes) V_i with locations p_i which are connected by edges (lines) d_i .
- Deformation is caused by acceleration forces a_i acting on the vertices.

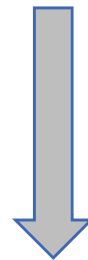
S. Lobregt and M. Viergever, "A discrete dynamic contour model", IEEE TMI Vol. 14, No. 1, 1995, pages 12-24.

Segmentation with Active Contours – 3

Newtonian displacement

- | | |
|----------------|-----------------|
| • Force | F |
| • Acceleration | $a = F / m$ |
| • Velocity | $v = a \cdot t$ |
| • Displacement | $s = v \cdot t$ |

$m = \text{mass}$



Segmentation with Active Contours – 4

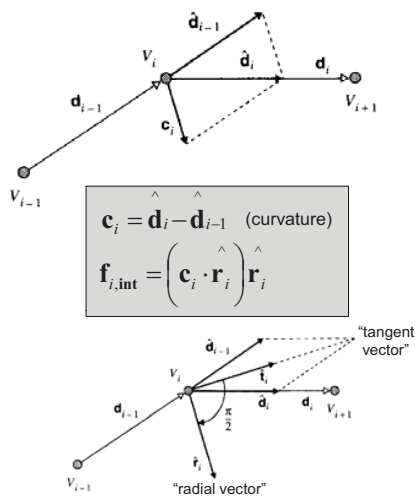
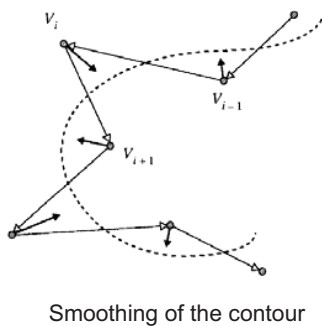
Iterative contour deformation

One time step Δt :

1. New location $\mathbf{p}_i(t + \Delta t) = \mathbf{p}_i(t) + \mathbf{v}_i(t) \cdot \Delta t$
2. New velocity $\mathbf{v}_i(t + \Delta t) = \mathbf{v}_i(t) + \mathbf{a}_i(t) \cdot \Delta t$
3. New forces $\mathbf{f}_i(t + \Delta t) = w_{\text{ext}} \mathbf{f}_{i,\text{ext}}(t + \Delta t) + w_{\text{int}} \mathbf{f}_{i,\text{int}}(t + \Delta t)$
4. New acceleration $\mathbf{a}_i(t + \Delta t) = \frac{1}{m_i} \mathbf{f}_i(t + \Delta t)$

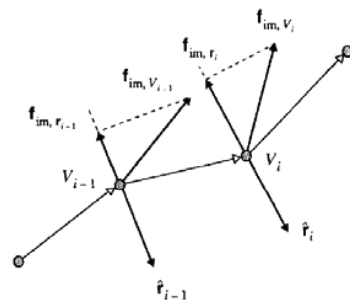
Segmentation with Active Contours – 5

Internal forces



Segmentation with Active Contours – 6

External forces



Only use component along radial

Energy image E_{im} derived from the original image, e.g. gradient magnitude

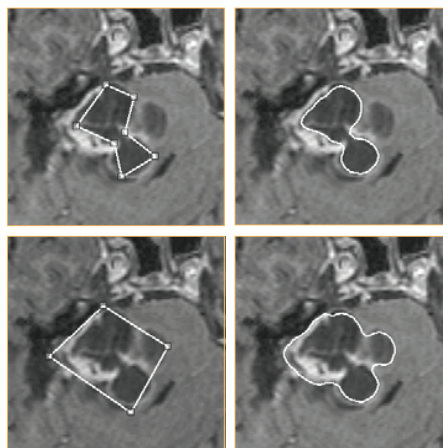
$$\mathbf{f}_{im} = -\nabla E_{im}$$

$$\mathbf{f}_{i,ext} = \left(\mathbf{f}_{im}(\mathbf{p}_i) \cdot \hat{\mathbf{r}}_i \right) \hat{\mathbf{r}}_i$$

Segmentation with Active Contours – 7

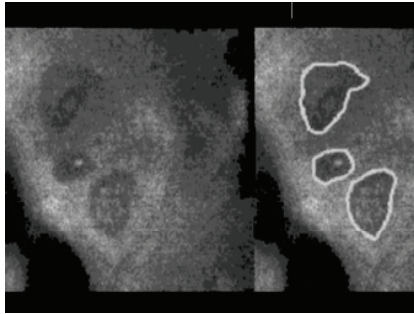
User is in control

- Initialization of contour determines final result
- Editing of contour, and new deformation to modify/correct result

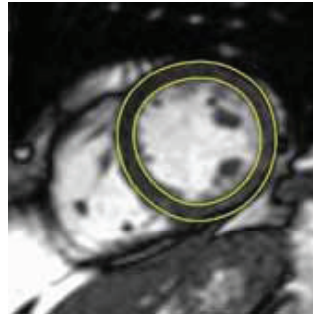


Segmentation with Active Contours – 8

Examples



Confocal microscope



Cardiac MR

Segmentation with Active Objects – 1

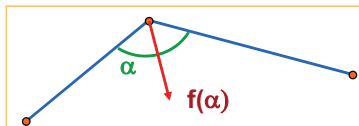
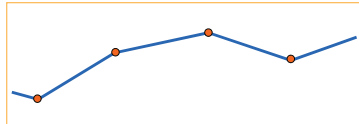
From 2D to 3D: Active objects

- An initial **surface** in 3D is available
- The **surface** is deformed by attracting it to specific image features (external force)
- The smoothness or shape of the **surface** is constrained (internal force)
- Deformation is stopped when the **surface** does no longer change

Segmentation with Active Objects – 2

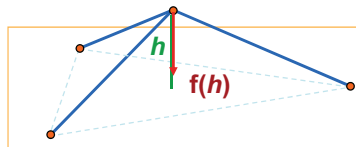
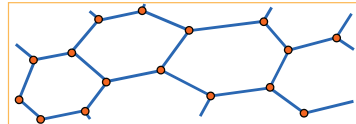
2D active contours versus 3D active objects

2D Discrete Contour:
vertex (node) has 2 neighbors



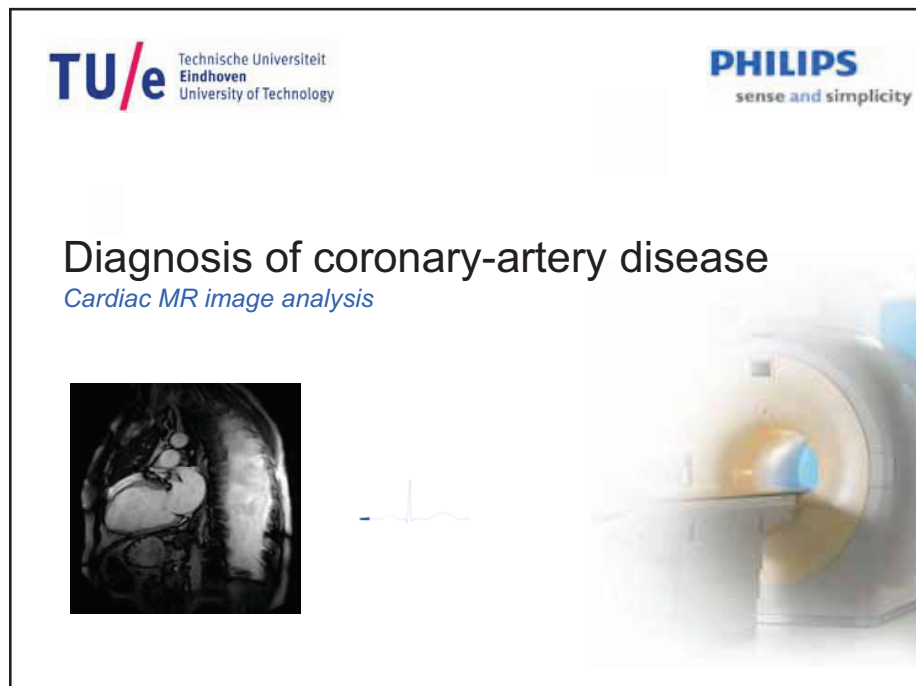
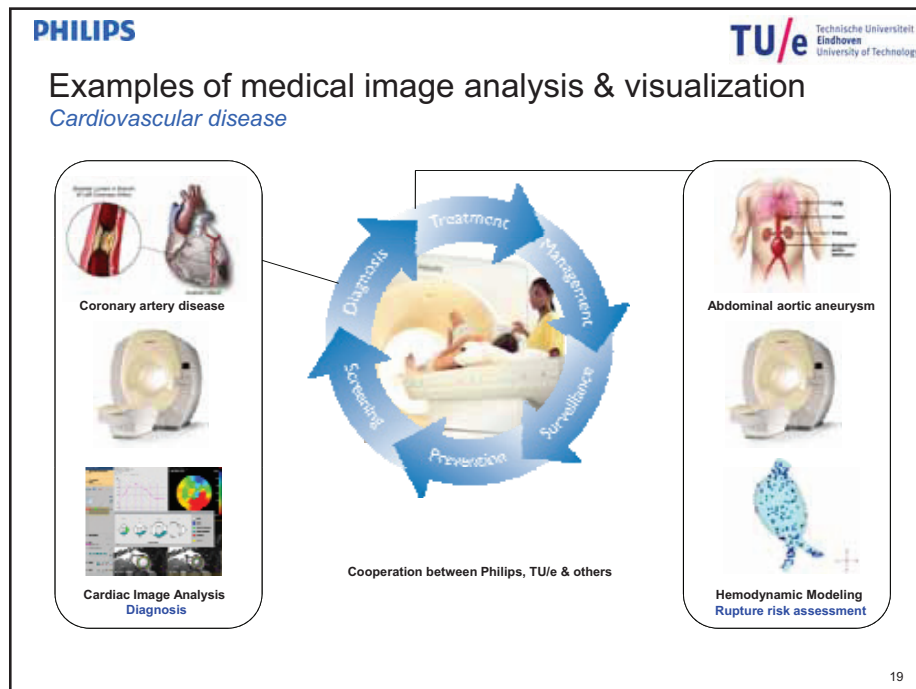
Internal force from vertex positions

3D Simplex Mesh:
vertex has 3 neighbors



Internal force from vertex positions

Example Applications




PHILIPS


Coronary-artery disease

The #1 cardiac disease in the western world


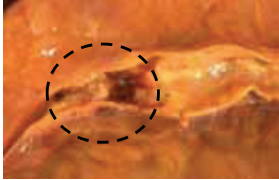
TU/e Technische Universiteit
Eindhoven
University of Technology



Blocked Lumen in Branch of Left Coronary Artery

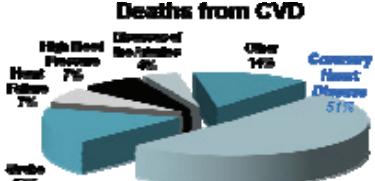


Anterior infarct

<http://www-medlib.med.utah.edu>

Deaths from CVD



AHA CVD Statistics (2003)

Prevalence	71.300.000 (2003)
Mortality	919.614 (2003) (1 out of every 2.7)
Cost	\$ 403.1 billion (2006)

21

PHILIPS

Coronary-artery disease

Two appearances


TU/e Technische Universiteit
Eindhoven
University of Technology

Partial obstruction (narrowing, stenosis):

- insufficient supply of blood to the myocardium (**ischemia**)
- reduced pump function of the heart → reduced blood supply to the body

Complete obstruction (occlusion):

- starvation of myocardial tissue (**infarction**)
- no muscle contraction in infarcted area → severely reduced supply to the body



infarcted (dead) tissue

<http://www-medlib.med.utah.edu>

22

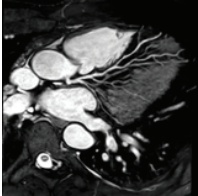
PHILIPS

TU/e Technische Universiteit
Eindhoven
University of Technology

Cardiac MR imaging

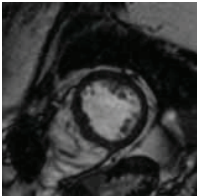
Comprehensive – Visualization of all disease aspects

Whole-heart / coronaries



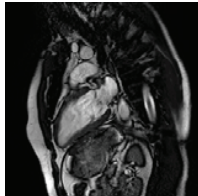
What is the patient's coronary anatomy?
Any stenosis?

Function – Short Axis

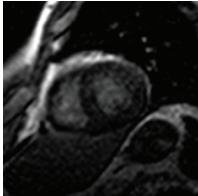


How well does myocardium contract (pump function ok)?

Function – Long Axis

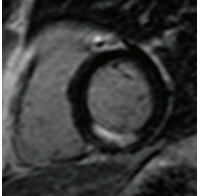


Perfusion



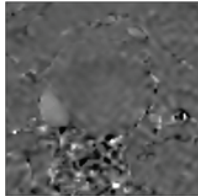
Is the myocardial blood supply ok?

Viability



Is there any dead myocardial tissue?

Flow



Blood flow in ventricles, aorta, ...ok?

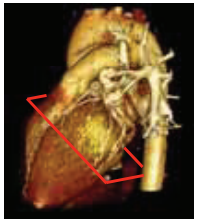
23

PHILIPS

TU/e Technische Universiteit
Eindhoven
University of Technology

Left-ventricular functional analysis

The most-frequently used cardiac MR analysis application



Key features:

- computer-assisted contouring
- volumetric analysis
- wall analysis
- 17-segment AHA scoring
- complete reporting
- diastolic functional analysis

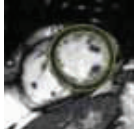
Volumetric analysis:

- ejection fraction
- stroke volume/index
- cardiac output/index
- ...

Wall analysis:

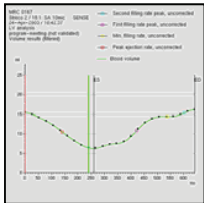
- mass
- thickness
- thickening
- time of max. thickness

Auto-contouring

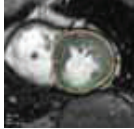
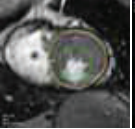
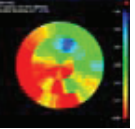


Gilion Hautvast et al.
(Philips Healthcare & TU/e)

Volumetric analysis



Wall contraction analysis

end diastole end systole bulls eye plot
of wall thickening

24

PHILIPS

TU/e Technische Universiteit
Eindhoven
University of Technology

Automatic contouring

Making the difference!

Early days of cardiac MRI (1990-2000)


- Fully manual contouring
- Cine CMR: 10 slices, 15 phases, 2 contours, 10 sec/contour → 83 min

Since ~2000

- Commercial semi-automatic methods → 15 min

Since 2009*

- From 15 to 30 phases per cardiac cycle
- Fully automatic methods → ~ 3 min
(automatic detection < 10 sec, rest manual corrections)



* Philips Healthcare's Cardiac Explorer software

Now used in clinical routine!


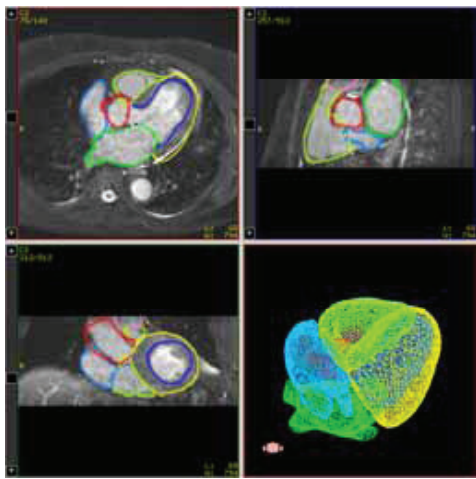
25

PHILIPS

TU/e Technische Universiteit
Eindhoven
University of Technology

Whole-heart cardiac MRI segmentation

Works in progress



Various:

- quantifications
- visualizations based on segmentation

Philips Research Hamburg/Aachen

After automatic refinement in CMR data

26

Coronary-artery tracking

Works in progress



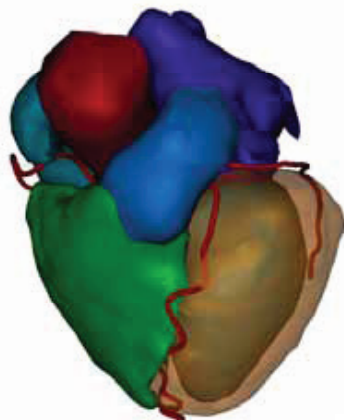
- Calculate “vesselness” feature
- Place two seed points
- Double wavefront propagation
- Rough backtracking of fronts
- Path recentering

Jeroen Sonnemans (Philips Healthcare)

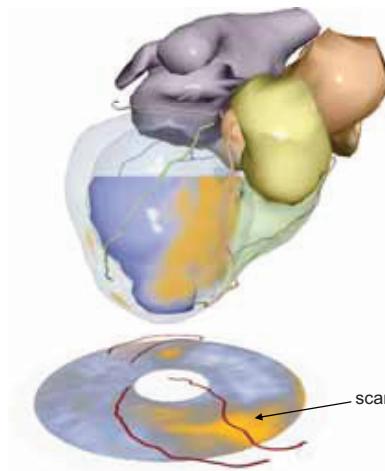
27

Comprehensive 3D visualization

Works in progress



Pierre Ermes (Philips Healthcare)
Maurice Termeer (TU Vienna, Austria – now Philips Healthcare)
Eduard Gröller (TU Vienna, Austria)
Anna Vilanova (TU/e, BMIA)



comprehensive visualization
(whole-heart, coronaries, scar)

28



PHILIPS

Erasmus MC
University Medical Center Rotterdam

TU/e Technische Universiteit
Eindhoven
University of Technology

Risk of abdominal aortic aneurysm rupture

Computer simulation of aortic flow & wall stress





Hemodyn

Funded by the Dutch Ministry of Economic Affairs (SenterNovem)

Research cooperation between the Technical University Eindhoven – BMT,
Erasmus Medical Center Rotterdam – Thoraxcenter / BME,
Philips Healthcare – Healthcare Informatics

Cooperating clinical centers:
Academic Hospital Maastricht and Catharina Hospital Eindhoven





29

PHILIPS


Hemodyn

TU/e Technische Universiteit
Eindhoven
University of Technology

Abdominal aortic aneurysm (AAA)




The disease

- Life-threatening dilatation of the abdominal aorta
- Most frequently occurring in elderly man
- USA statistics:
 - 1.5 million cases
 - 200.000 new diagnosis / year
 - 15.000 death / year



<http://www-medlib.med.utah.edu/WebPath>


30

AAA treatment




Current practice

- **Treatment:**
 - open surgery
 - endovascular stent placement
- **Decision based on geometry:**
 - diameter > 5.5 cm
 - diameter > 200% normal
- **Problem:**
 - rupture does occur for diameters < 5.5 cm
 - AAAs > 5.5 cm do not always rupture
 - diameter is not a good risk parameter!



Hypothesis: better rupture-risk predictors can be obtained by patient-specific **hemodynamic modeling**

31


The AAA modeling chain

A novel approach for risk assessment


Modeling of AAA flow / wall stress

- Rupture risk assessment
- Growth prediction

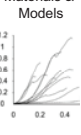
Blood inflow




Blood pressure



Materials & Models





Simulated Wall Stress


3D Imaging

Geometry Derivation


Volume Meshing

Hemodynamic Simulation

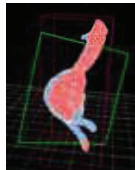
Results Visualization




3D CTA / MRI




Segmentation & Registration




Tetrahedral Mesh



Finite-element & volume modeling





Number Cruncher




Simulated Blood Flow

32

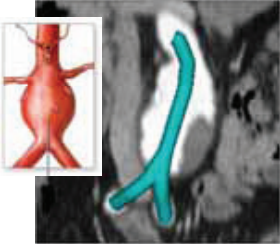




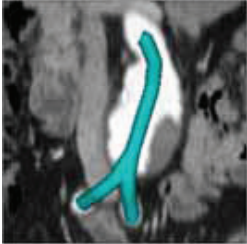


Segmentation of lumen & outer wall from CTA

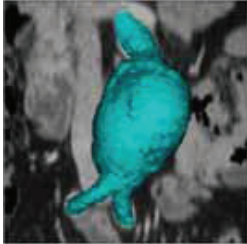
Getting the geometry




Initial tube around centerline




Automatic lumen segmentation




Final lumen segmentation



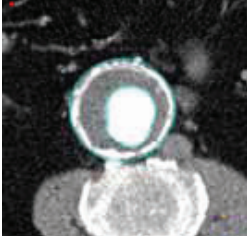
initial tube



lumen




outer wall





Lumen & outer wall segmentation

Ursula Kose (Philips Healthcare), Silvia Olabarriaga (UMC Utrecht)

33



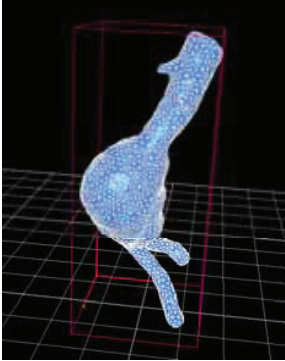




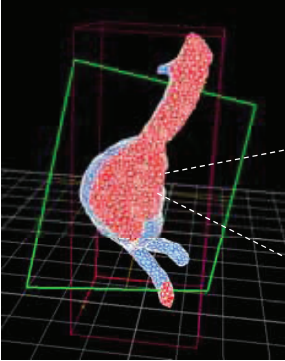
Volume meshing

Defining the 3D finite-element mesh

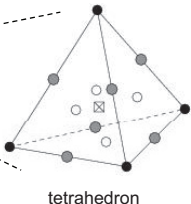
3D Delauney tetrahedralization:



Original segmented surface



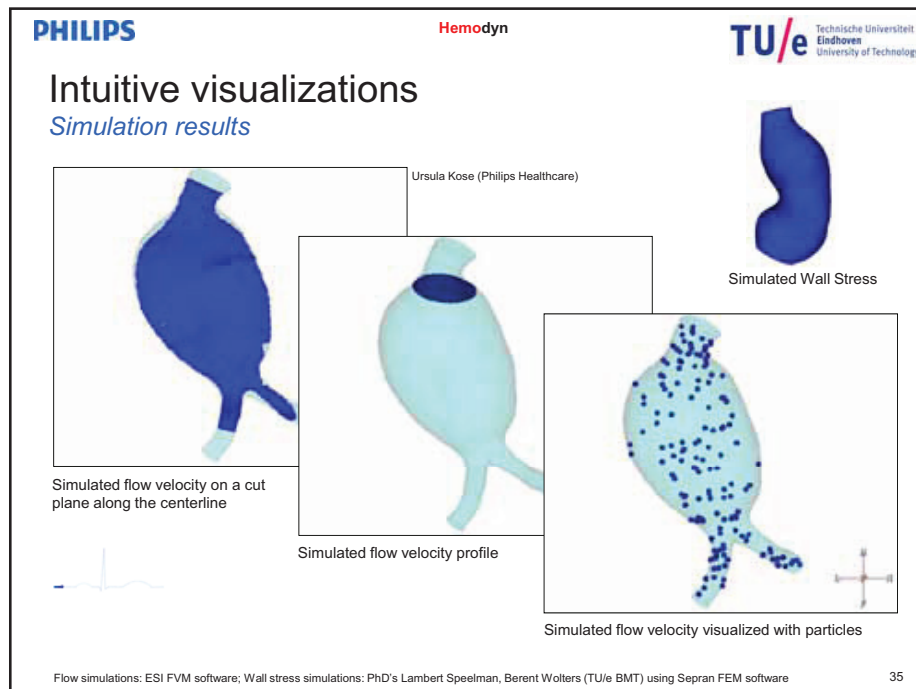
Surface with cutplane through volume mesh





tetrahedron


F. Laffargue
(Philips Healthcare Research France)

34









Clinical evaluation

Proof of clinical relevance

Clinical Participants

- Catharina Hospital Eindhoven
- Academic Hospital Maastricht

Methodology

- Each 20 patients / participant (aneurysm > 40 mm and < 55 mm)
- 4 times CTA and MRI at 4 months intervals
- Comparison of simulated wall-stress with (change in) AAA geometry

Funding

- Philips Healthcare Best (organization & imaging)

Results: significant correlation between 95-% percentile wall stress and aneurysm growth
(PhD Thesis Lambert Speelman, TU/e BMT)

Does increased wall stress lead to AAA growth?

36

Resumé

Resumé

- Aging population, increasing health care cost, shortage of clinicians
- Strong need for more effective & efficient healthcare
- Medical image analysis & visualization:
 - Strongly represented in NL
 - Assist diagnosis, therapy & follow/up
 - Key to higher quality, lower cost health care

From Content to Care !



PHILIPS

TU/e Technische Universiteit
Eindhoven
University of Technology



Thank you!

Questions?

39

Human Action Representation and Recognition

Prof. Dr. Ling Shao
Univ. of Sheffield, UK

Human Action Representation and Recognition

Ling Shao
The University of Sheffield

March 23, 2011

1

Introduction

- **Objective**

Recognizing human actions under viewpoint and illumination changes, intra-class variations, scaling, partial occlusion and background clutter, etc.

- **Possible applications**

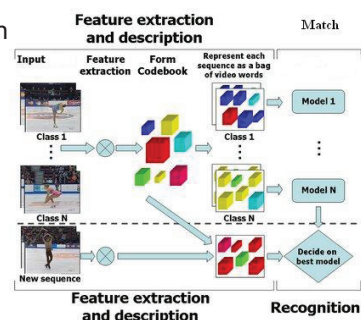
Human-computer interaction, Video search and mining, Video surveillance

- **Approaches**

Appearance-based

Optical flow based

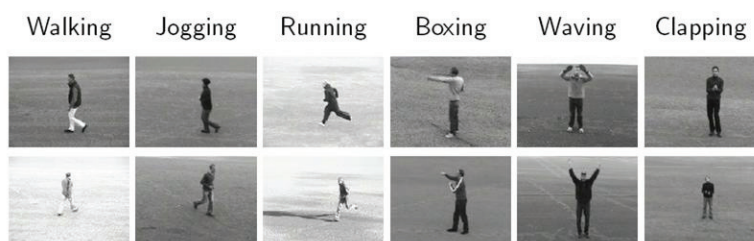
Spatio-temporal interest points based



2

Dataset: KTH-Actions

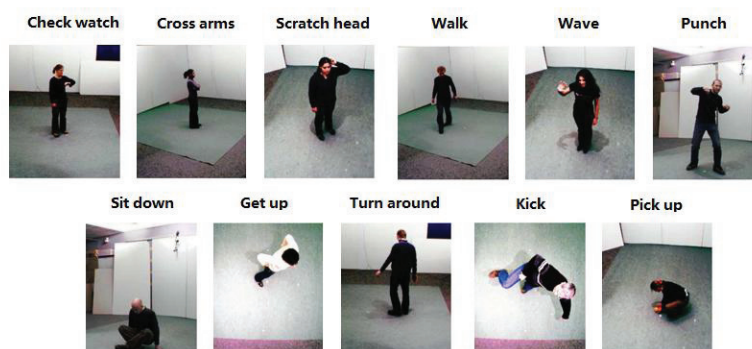
- 6 action classes by 25 persons in 4 different scenarios
- Total of 2391 video samples
 - Specified train, validation, test sets
- Performance measure: average accuracy over all classes



Schuldt, Laptev, Caputo ICPR 2004

Dataset: IXMAS

- 11 actions, 10 actors performing each 3 times
- Multiview: five cameras



Weinland et al. CVIU'2006

How to represent actions?

Holistic (global) representation

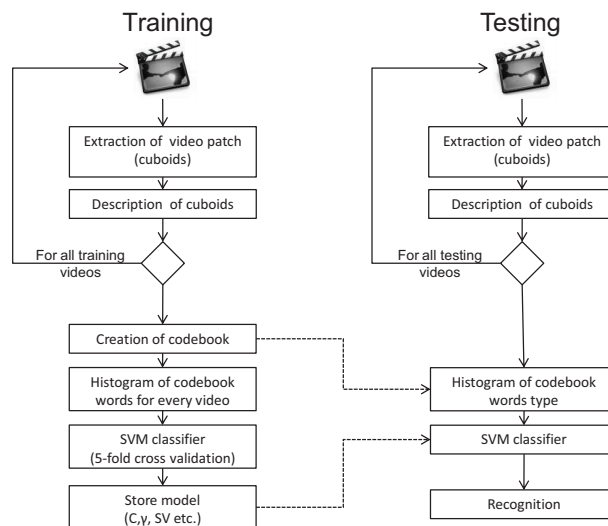
- Informative
- Better for constrained datasets
- Requires foreground segmentation

Sparse representation (local features)

- Less informative
- Robust to occlusion, clutter, camera motion
- Requires no segmentation

5

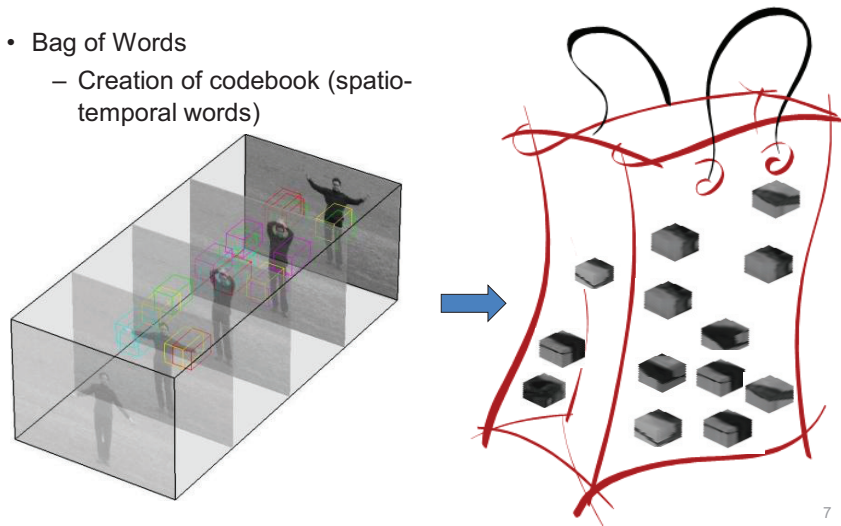
Methodology – block diagram



6

Sparse Representation – Bag of Visual Words

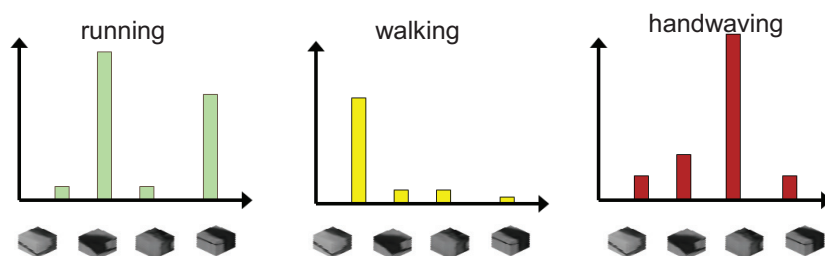
- Bag of Words
 - Creation of codebook (spatio-temporal words)



7

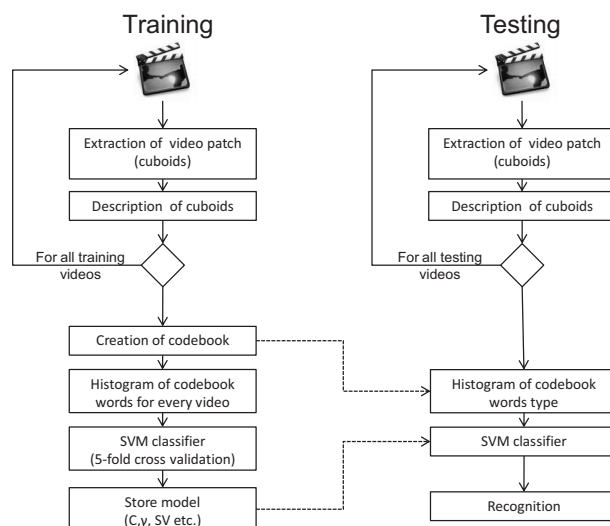
Sparse Representation – histogram of word occurrence

- Bag of Words
 - Creation of codebook
 - Histogram of codebook words for every video



8

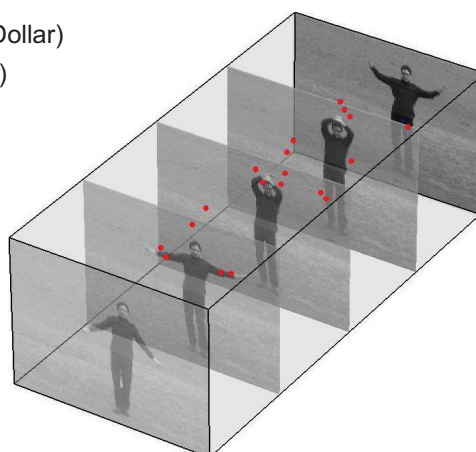
Recognition – block diagram



9

STIP extraction

- Feature extraction
 - Periodic feature detector (Dollar)
 - 3D corner detector (Laptev)
 - Bank of 3D Gabor filters
 - Space-Time DoG
 - 3D Hessian



10 10

Feature Extraction – Periodic Feature Detector

- Detector based on a set of separable filters

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2$$

- Spatial dimension: Gaussian filter

$$g(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right)$$

- Temporal dimension: Gabor filter

$$h_{od}(t; \tau, \omega) = -\sin(2\pi t\omega) e^{-t^2/\tau^2}$$

$$h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega) e^{-t^2/\tau^2}$$

- Any region with spatially distinguishing characteristics undergoing a complex motion will induce a strong response. Pure translation will not induce a response

Ref: P. Dollar et al, "Behavior recognition via sparse spatio-temporal features," *Proc. of ICCV Int. work-shop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VSPETS)*, pages 65-72, 2005.

11

Feature Extraction – 3D Corner Detector

- Extension of Harris's corner detection

$$\mu_{ST} = g(x, y, t; \sigma, \tau) * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix}$$

$$L_x = \partial_x(g(x, y, t; \sigma, \tau) * I(x, y, t))$$

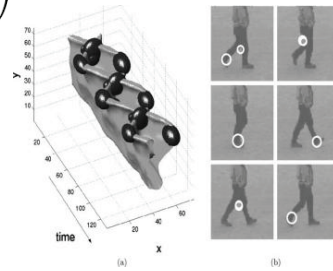
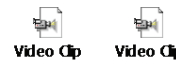
$$L_y = \partial_y(g(x, y, t; \sigma, \tau) * I(x, y, t))$$

$$L_t = \partial_t(g(x, y, t; \sigma, \tau) * I(x, y, t))$$

$$H = \det(\mu_{ST}) - k \cdot \text{trace}(\mu_{ST})$$



Search local positive maxima



12

Feature Extraction – Bank of 3D Gabor Filters

- The video is convolved with a bank of 3D Gabor filters with different orientations and different wavelength of the underlying cosine

$$G(x, y, t) = \cos\left(\frac{2\pi}{\lambda_x} X\right) \cdot \cos\left(\frac{2\pi}{\lambda_y} Y\right) \cdot \exp\left(-\left(\frac{X^2}{2\sigma_x^2} + \frac{Y^2}{2\sigma_y^2} + \frac{T^2}{2\sigma_t^2}\right)\right)$$

$$\begin{pmatrix} X \\ Y \\ T \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) \\ 0 & \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} \cos(\omega) & 0 & \sin(\omega) \\ 0 & 1 & 0 \\ -\sin(\omega) & 0 & \cos(\omega) \end{pmatrix} \begin{pmatrix} x \\ y \\ t \end{pmatrix}$$

13

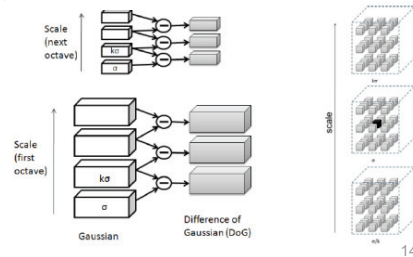
Feature Extraction – Space-Time DoG

- Video convolved with Gaussians

$$L(x, y, t, \sigma) = G(x, y, t, \sigma) * I(x, y, t)$$

$$\text{where } G(x, y, t, \sigma) = \frac{1}{(2\pi\sigma)^{\frac{3}{2}}} e^{-(x^2+y^2+t^2)/2\sigma^2}$$

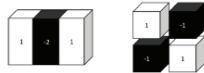
$$\begin{aligned} D(x, y, t, \sigma) &= (G(x, y, t, k\sigma) - G(x, y, t, \sigma)) * I(x, y, t) \\ &= L(x, y, t, k\sigma) - L(x, y, t, \sigma) \end{aligned}$$



14

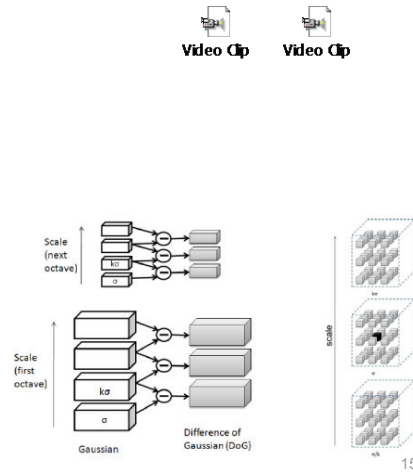
Feature Extraction – 3D Hessian

- Same concept as 3D SIFT,
- But: integral video + box filters



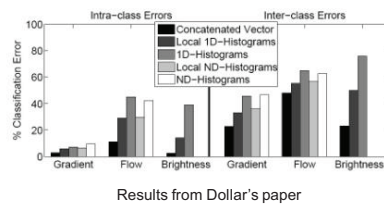
$$H(x, y, t; \sigma, \tau) = \begin{pmatrix} L_{xx} & L_{xy} & L_{xt} \\ L_{yx} & L_{yy} & L_{yt} \\ L_{tx} & L_{ty} & L_{tt} \end{pmatrix}$$

$$S = |\det(H)|$$



Description methods

- Gradient
 - Flattened Gradient vector
 - The gradient is computed for every slice of the cuboids and all the values are concatenated in a vector
 - The gradient was proved (by Dollar) to perform better than normalize pixel values or optical flow
 - The concatenated vector was proved (by Dollar) to perform better than the ND-Histograms and local ND-Histograms



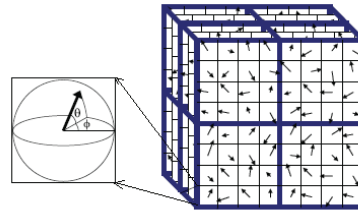
Ref: Dollár, P. et al. (2005). Behavior recognition via sparse spatio-temporal features.

Description methods

- Gradient

- 3D-SIFT

- It is an evolution of the common SIFT descriptor. Developed by Scovanner et al.
- The gradient magnitude and orientation in 3D are given



$$m_{3D}(x, y, t) = \sqrt{L_x^2 + L_y^2 + L_t^2},$$

$$\theta(x, y, t) = \tan^{-1}(L_y/L_x),$$

$$\phi(x, y, t) = \tan^{-1}\left(\frac{L_t}{\sqrt{L_x^2 + L_y^2}}\right).$$

Ref: Scovanner p. et al. (2007). A 3-dimensional sift descriptor and its application to action recognition.

17

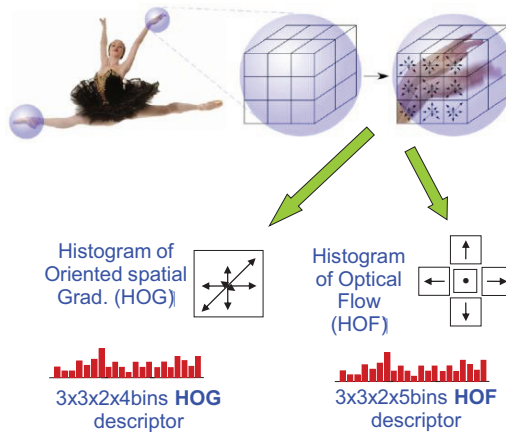
Description methods

- Gradient

- 3DSift

- Histogram of Oriented Gradient
- Histogram of Optical Flow
- Combination of them (HOGHOF*)

- Descriptors proposed by Laptev and used in his recent paper "Learning realistic human actions from movies"



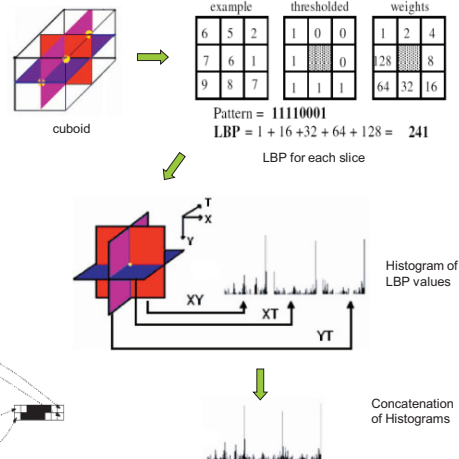
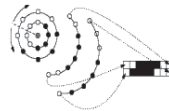
Ref: Laptev et al. (2008). Learning realistic human actions from movies. * Used Laptev binary file

18 18

Description methods

- Gradient
- 3D-SIFT
- HOG, HoF, HOGHoF
- LBP-TOP (proposed as descriptor of cuboids)

- I modified the LBP-TOP computing, for each cuboids, 3 slices in XY planes, 3 in XT plane and 3 in YT plane.
- I used a multiresolution LBP code ($R=2 \rightarrow \text{neighbors}=8 +$
 $R=3 \rightarrow \text{neighbors}=16 +$
 $R=4 \rightarrow \text{neighbors}=24$)

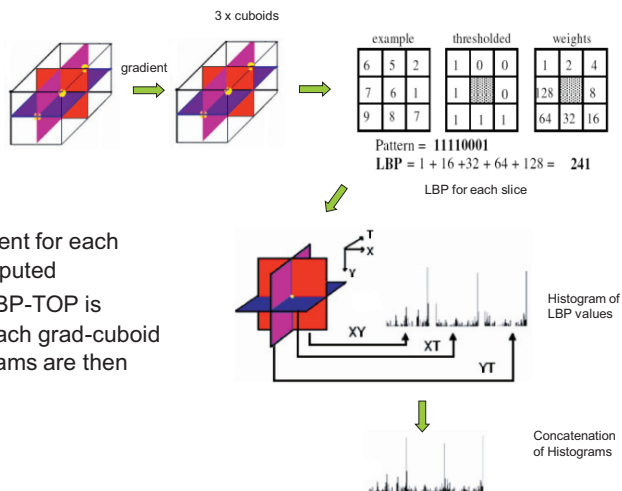


Ref: Zhao et al. (2008). Dynamic texture recognition using LBP with an application to facial expressions.

19

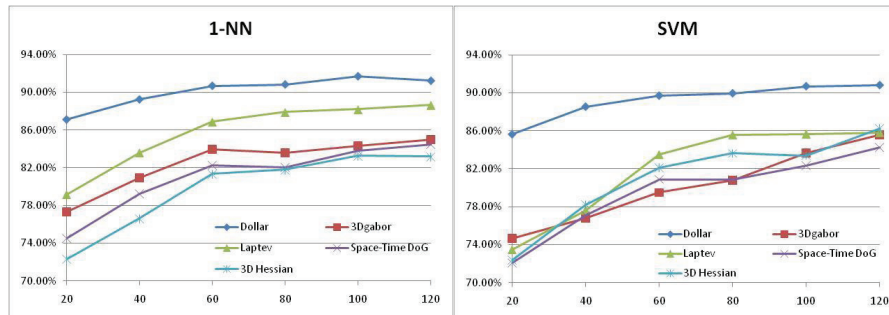
Description methods

- Gradient
- 3DSift
- HoG, HoF, HoGHoF
- LBP-TOP
- Grad LBP-TOP
 - The three gradient for each cuboid are computed
 - The modified LBP-TOP is computed for each grad-cuboid and the histograms are then concatenated



20 20

Results – feature extraction

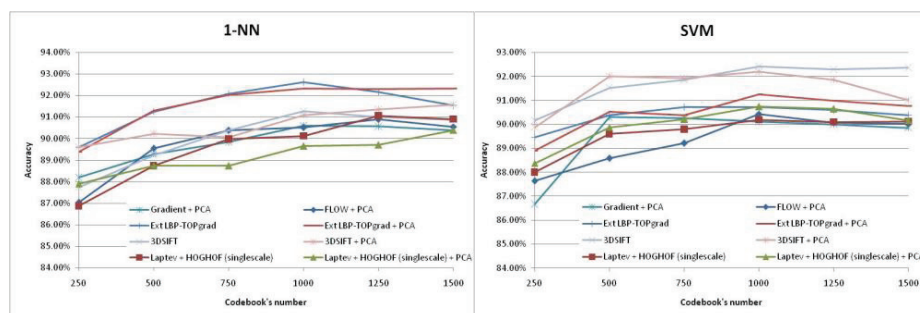


Extraction method	Environment	Computational time (s)
Periodic feature detector	Matlab	7.61
Spatio-Temporal Harris	C	33.53
Space-Time DOG	Matlab	353.9
3D Hessian	C	1.96*
Bank of 3D Gabor filters	Matlab	84.53

CONFIDENTIAL

21

Results – feature description



Description method	Environment	Computational time (s)
Gradient + PCA	Matlab	0.006
Optical Flow	Matlab	0.030
HOG-HOF	C	0.042
3D SIFT	Matlab	1.118
LBP-TOP	Matlab	0.0139
Ext Grad LBP-TOP	Matlab	0.1004

22

Results - comparison with state-of-the-art

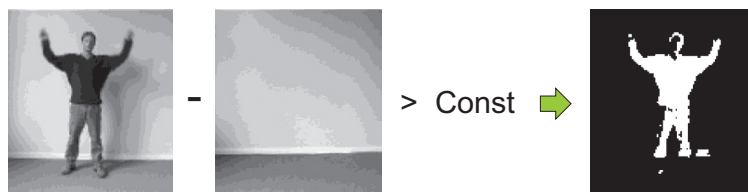
- Video file of 300 frames
- Extraction + description of 80 cuboids
- 1-NN (χ^2) and SVM (rbf kernel)

Extraction method	Description method	Feature length	Time (s)	Accuracy (SVM)	Accuracy (1-NN)
Periodic Feature Detector (Dollar)	Gradient Extended LBP- TOP _{8,8,2,2,2} + PCA	100	~24.8	91.25 %	92.32 %
Spatio-Temporal Harris (Laptev)	HOG-HOF (Laptev)	162	~68.7	89.88 %	89.05 %
3D Hessian	3D SURF	288	10.6	81.39 %	74.17 %

23

Global Representation: Foreground segmentation

Image differencing: a simple way to measure motion/change



Better Background / Foreground separation methods exist:

- Modeling of color variation at each pixel with Gaussian Mixture
- Dominant motion compensation for sequences with moving camera
- Motion layer separation for scenes with non-static backgrounds

24

Motion Templates

$$D(x, y, t) \quad t = 1, \dots, T$$



Idea: summarize motion in video in a
Motion History Image (MHI):

$$H_{\tau}(x, y, t) = \begin{cases} \tau & \text{if } D(x, y, t) = 1 \\ \max(0, H_{\tau}(x, y, t-1) - 1) & \text{otherwise} \end{cases}$$

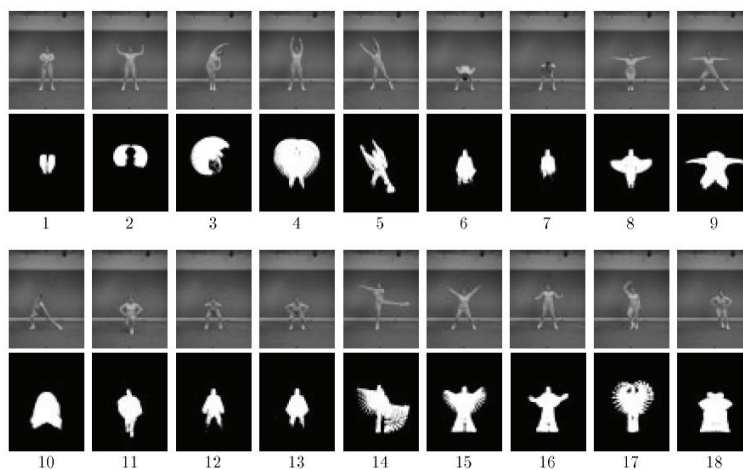
Descriptor: Hu moments of different orders

$$m_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q \rho(x, y) dx dy$$



[A.F. Bobick and J.W. Davis, PAMI 2001]

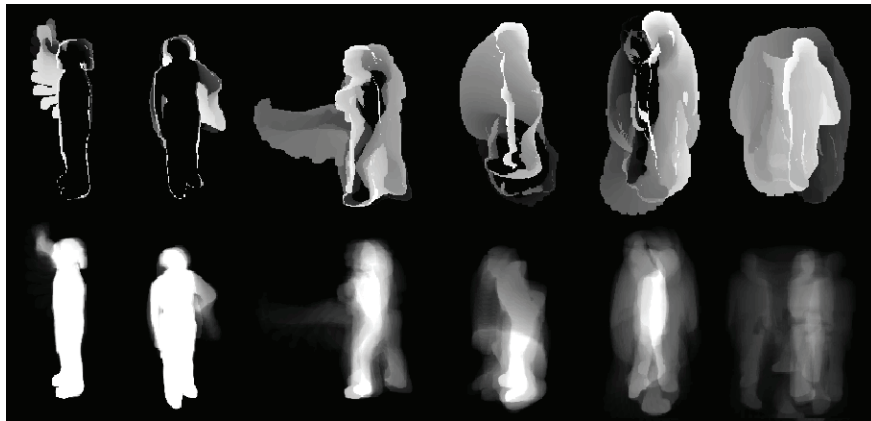
Aerobics dataset



Nearest Neighbor classifier: 66% accuracy

[A.F. Bobick and J.W. Davis, PAMI 2001]

Global Representation: MHI + GEI



Motion history images (top) and gait energy images (bottom)

27

Results on KTH

Comparison of two biologically-inspired methods

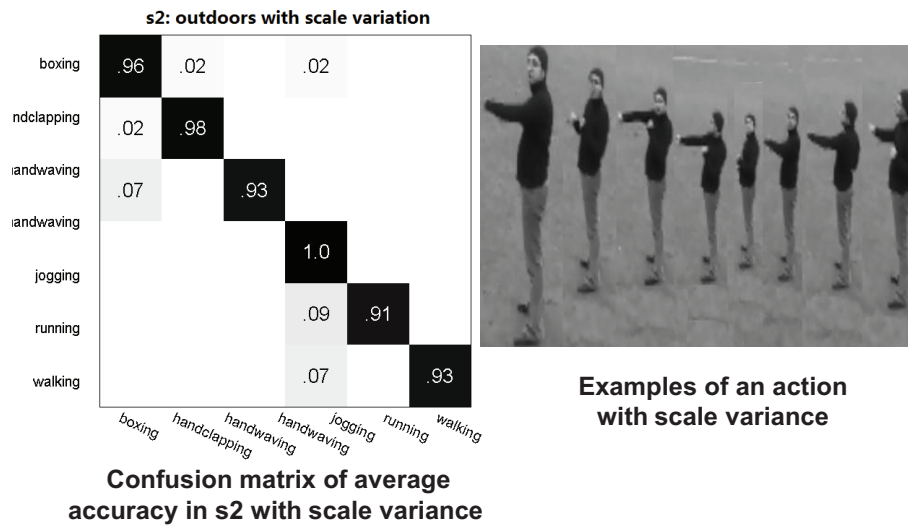
	KTH s1	KTH s1	KTH s1	KTH s1	Averg.
Our method	93.8	95.2	86.0	95.7	92.7
Jhuang et al.	96.0	86.1	89.8	94.8	91.7

Comparison of our method to others with the same evaluation scheme (split)

Method	Ours	Jhuang	Fathi	Ahmad	Nowozin	Schuldt
Evaluation	split	split	split	split	split	split
Accuracy (%)	92.7	91.7	90.5	88.3	87.0	71.7

28

Scale Invariance



29

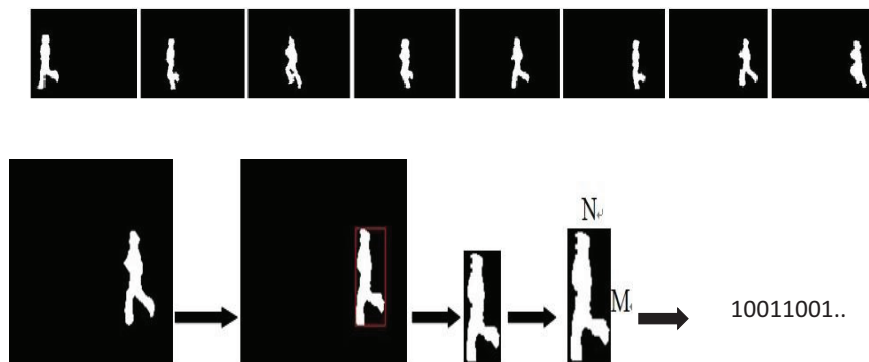
Results on IXMAS

Method	Cam1	Cam2	Cam3	Cam4	Cam5
Our method	83.3	83.2	90.5	85.2	75.2
Weinland'10	85.8	86.4	88.0	88.2	74.7
Weinland'10	84.7	85.8	87.9	88.5	76.2
Junejo	76.4	77.6	73.6	68.4	66.1
Yan	72.0	53.0	68.0	63.0	-
Weinland'07	55.2	63.5	-	60.0	-

Comparison with state-of-the-art methods (recognition rates in %)

30

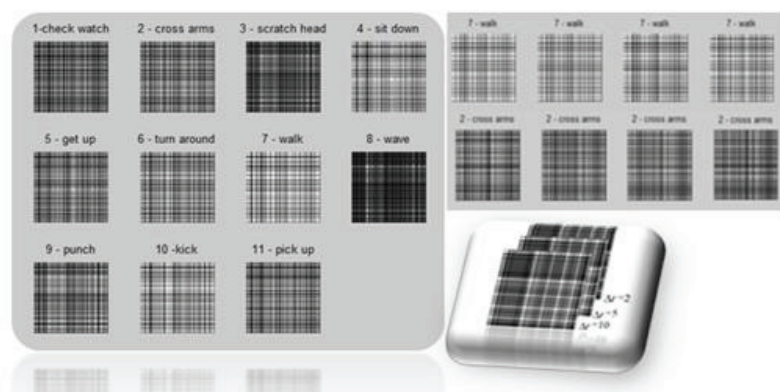
Representation on poses: Bag-of-Poses



Correlation between adjacent poses is lost.

Shao and Chen, BMVC 2010

Bag-of-Correlated-Poses



Left: Correlogram matrices of different actions performed by the same person

Top Right: Same actions performed by different persons;

Bottom Right: Correlogram matrices with different time offsets

32

Feature Fusion: BoCP + Extended MHI

Author	Recognition rate Single camera, Best result (%)	Method
<i>We</i>	93.3	<i>Our proposed method</i>
Lv, Fengjun et al.[15]	80.6	PMK-NUP
Junejo et al. [16]	77.6	Cross-View from Temporal
Weinland et al. [17]	81.3	3D Exemplars
Yan et al. [18]	68.0	4D Feature Models
Weinland et al. [17]	63.5	2D Exemplars

Comparison with other methods on the IXMAS dataset.

Visual search: what's next?

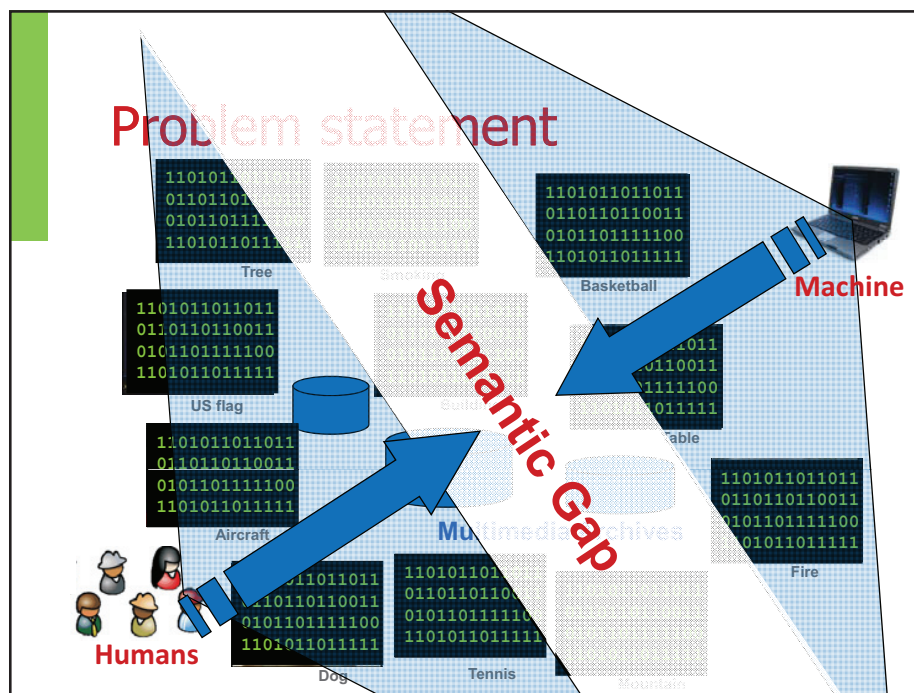
Dr. Cees Snoek

Univ. of Amsterdam & UC Berkeley, USA

Visual search: what's next?

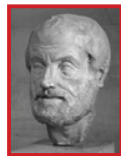
Cees Snoek

University of Amsterdam
The Netherlands



The science of labeling

- To understand anything in science, things need a name that is universally recognized



'categories'



living organisms



chemical elements

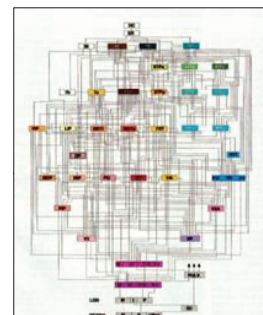
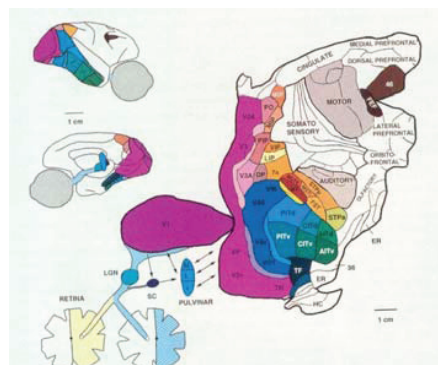


human genome

- Worldwide endeavor in naming visual information

How difficult is the problem?

- Human vision consumes 50% brain power...



Van Essen, Science 1992

Slide credit: Andrew Zisserman

Naming visual information

Focus of today's talk

- Concept detection
 - Does the image contain an airplane?



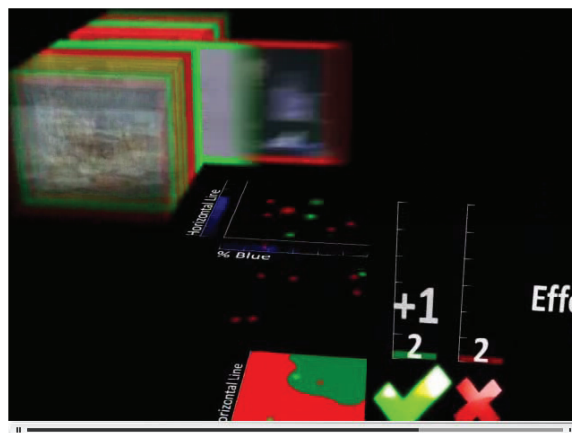
- Object localization
 - Where is the airplane, (if any)?



- Object segmentation
 - Which pixels are part of an airplane, (if any)?



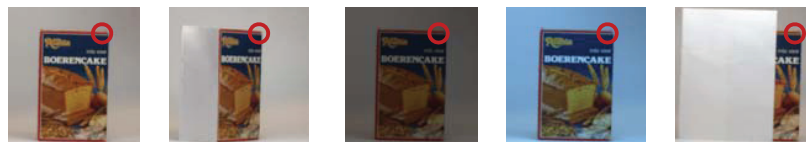
Concept detection in a nutshell



Visualization by
Jasper Schulte

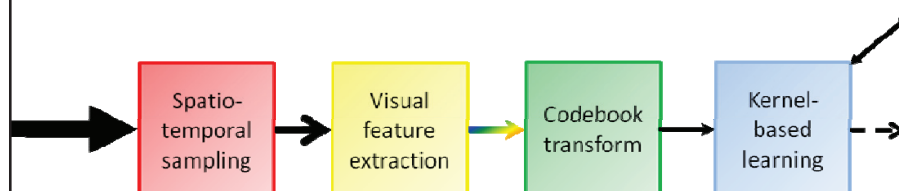
Variation in appearance

So many images of one thing, due to minor differences in:
illumination
background
occlusion
viewpoint, ...



- This is the sensory gap

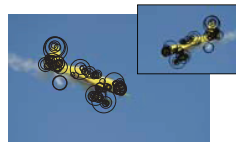
MediaMill concept detection



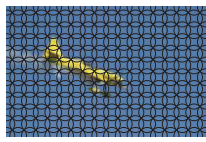
Mikolajczyk, IJCV 2005
Lazebnik, CVPR 2006
Zhang, IJCV 2007

Step 1: sampling points (standard)

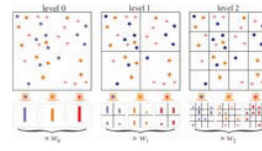
- Orientation and scale of concepts change
 - Salient point methods robustly detect regions
- Preferred for robust concept detection



Harris-Laplace



Dense sampling



Spatial Pyramid

Burghouts, CVIU 2009
van de Sande, PAMI 2010

Step 2: point description

- Lowe's SIFT descriptor measures intensity only
 - ...but illumination of concepts change
 - so, detection suffers from unstable region description
- Color descriptors
 - Increase illumination invariance
 - Increase discriminative power

Van de Sande, PAMI 2010

Illumination invariance

	Light intensity change $\begin{pmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & a \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$	Light intensity shift $\begin{pmatrix} R \\ G \\ B \end{pmatrix} + \begin{pmatrix} o_1 \\ o_1 \\ o_1 \end{pmatrix}$	Light intensity change and shift $\begin{pmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & a \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} + \begin{pmatrix} o_1 \\ o_1 \\ o_1 \end{pmatrix}$	Light color change $\begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$	Light color change and shift $\begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} + \begin{pmatrix} o_1 \\ o_2 \\ o_3 \end{pmatrix}$
RGB Histogram	-	-	-	-	-
O_1, O_2	-	+	-	-	-
O_3 , Intensity	-	-	-	-	-
Hue	+	+	+	-	-
Saturation	-	-	-	-	-
r, g	+	-	-	-	-
Transformed color	+	+	+	+	+
Color moments	-	+	-	-	-
Moment invariants	-	+	+	+	+
SIFT (∇I)	+	+	+	-	-
HSV-SIFT	-	-	-	-	-
HueSIFT	+	+	+	-	-
OpponentSIFT	+	+	+	-	-
C-SIFT	+	-	-	-	-
r, g SIFT	+	-	-	-	-
Transf. color SIFT	+	+	+	+	+
RGB-SIFT	+	+	+	+	+

Leung and Malik, IJCV, 2001
Sivic and Zisserman, ICCV, 2003
van Gemert, CVIU, 2010

Step 3: descriptor quantization

- Codebook model
 - Create a codeword vocabulary
 - Discretize image with codewords
 - Represent image as codebook histogram



Maji et al., CVPR 2008

Step 4: Efficient classification

$$K(a, b) = \sum_{i=1}^n \min(a_i, b_i)$$

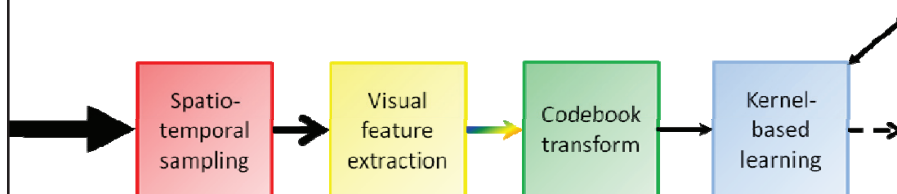
$$\begin{aligned} h(x) &= \sum_{i=1}^{\text{\#dim}} \left(\sum_{j=1}^{\text{\#sv}} \alpha^j \min(x_i, x_i^j) \right) + b \\ &= \sum_{i=1}^{\text{\#dim}} h_i(x_i) \end{aligned}$$

$$\begin{aligned} h_i(x_i) &= \sum_{j=1}^{\text{\#sv}} \alpha^j \min(x_i, x_i^j) + b \\ &= \sum_{x_i^j < x_i} \alpha^j x_i^j + \left(\sum_{x_i^j \geq x_i} \alpha^j \right) x_i \end{aligned}$$



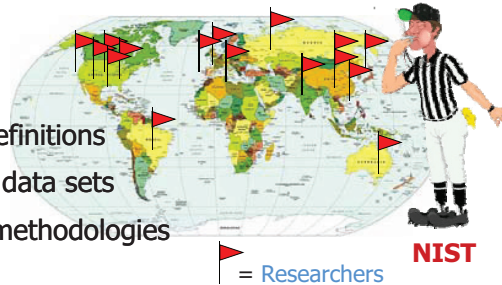
For the Intersection Kernel h_i is piecewise linear, and quite smooth, **blue plot**. We can *approximate* with fewer uniformly spaced segments, **red plot**. Saves time & space!

MediaMill concept detection



Evaluation best treated by TRECVID

- Situation in 2000
 - Various concept definitions
 - *Specific* and *small* data sets
 - Hard to compare methodologies
- Since 2001 worldwide evaluation by NIST
 - TRECVID benchmark



<http://trecvid.nist.gov/>

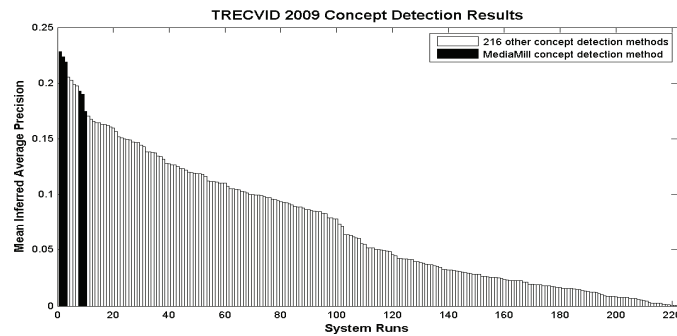
NIST TRECVID benchmark

- Promote progress in video retrieval research
 - Provide common dataset
 - Challenging tasks
 - Independent evaluation protocol
 - Forum for researchers to compare results



Snoek et al. TRECVID 09

TRECVID 2009 results



- Best performer for 10 out of 20 concepts
- Best overall performer

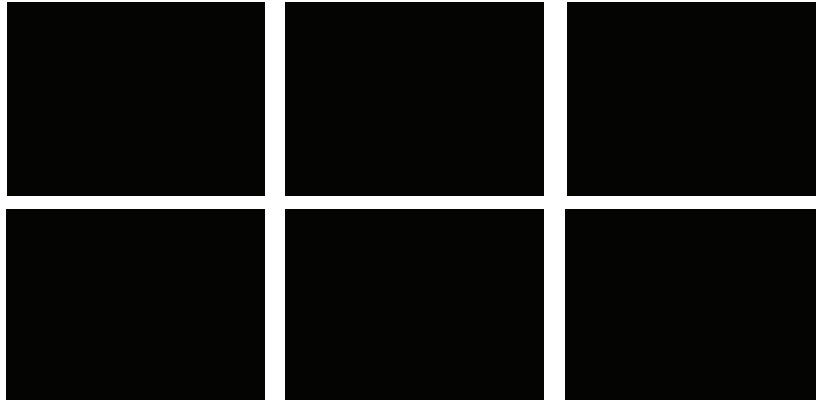
Snoek, TMM 2007

MediaMill video search engine

- CrossBrowser combines query results and time

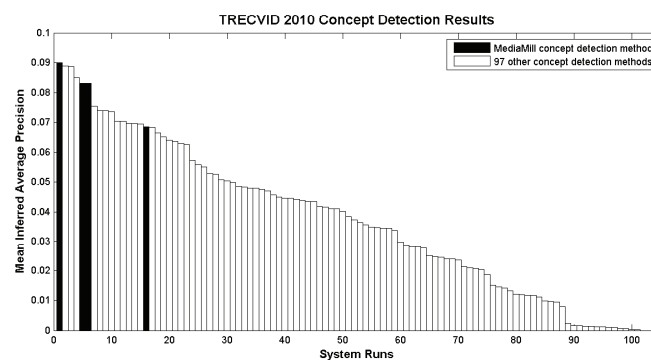


TRECVID 2010 Internet Archive web videos



Snoek et al, TRECVID 2010

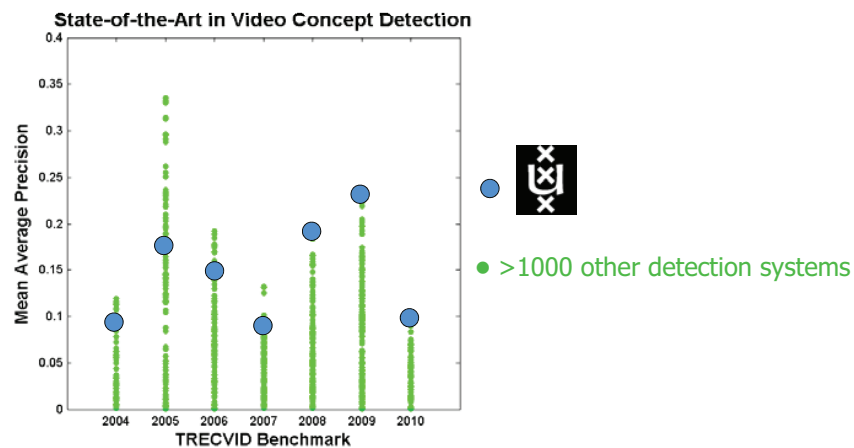
TRECVID 2010 results...



- Best performer for 6 out of 30 concepts
- Best overall performer

Snoek et al, TRECVID 04-10

Are we making progress?



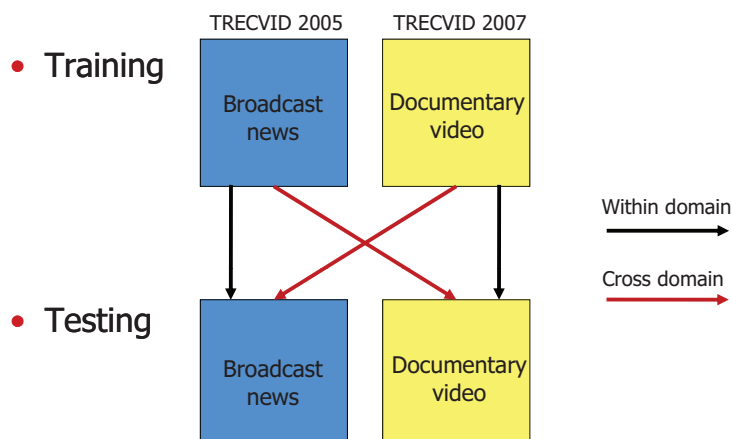
Community myths or facts?

- Chua et al., [ACM Multimedia 2007](#)
 - Video search is practically solved and progress has only been incremental
- Yang and Hauptmann, [ACM CIVR 2008](#)
 - Current solutions are weak and generalize poorly

We have done an experiment

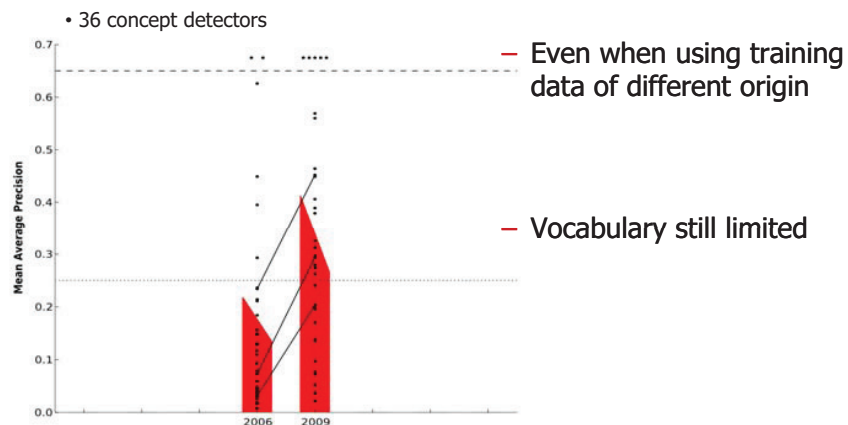
- Two video search engines from 2006 and 2009
 - MediaMill Challenge 2006 system
 - MediaMill TRECVID 2009 system
- How well do they detect 36 LSCOM concepts?

Four video data set mixtures



Snoek & Smeulders,
IEEE Computer 2010

Performance doubled in just 3 years



What's next?

**TOWARDS A HUMAN-SIZE
VOCABULARY**

massive amounts of

How to obtain labeled examples?

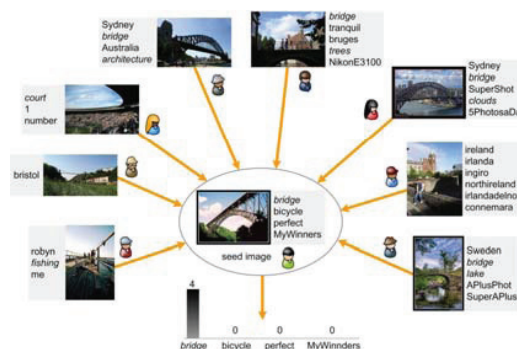


— ...but only human experts provide good quality examples

Xirong Li et al, TMM 2009

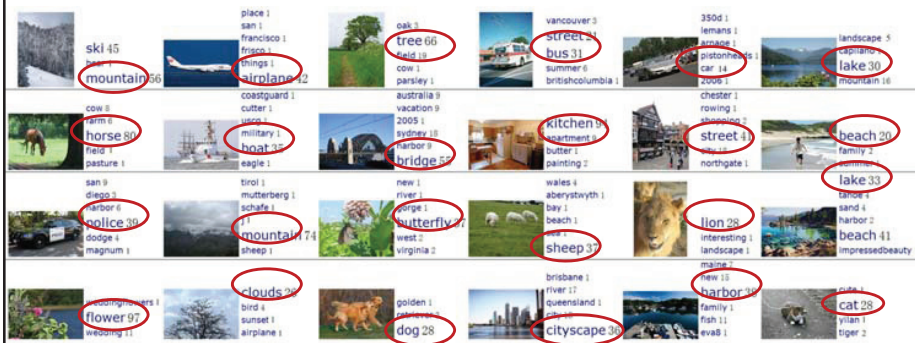
Learning social tag relevance by neighbor voting

- Exploit consistency in tagging behavior of different users for visually similar images



Updated tag relevance

- Objective tags are identified and reinforced



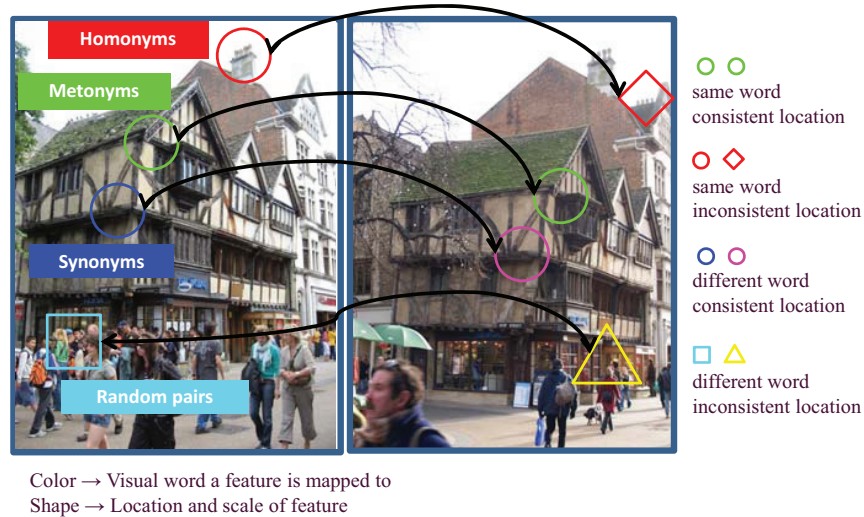
Based on 3.5 Million images downloaded from Flickr

What's next?

**TOWARDS MORE PRECISE
RECOGNITION**

Gavves et al, ACM MM 2010

Visual synonyms for landmark images



Gavves et al, ACM MM 2010

All Souls College - Oxford



What's next?

APPLICATIONS

Recent applications



SenseCam lifelogging
Byrne, MMTA 2010



Mobile video retrieval interfaces
Hürst, MM 2010



Crowdsourcing online concert video
Snoek, MM 2010



TV archive search
Huurnink, CIVR 2010

Conclusion

- Visual search is maturing quickly
- What's next?
 - a human-size vocabulary
 - more precise recognition
 - ...and great application potential

Thank you

- dr. Cees Snoek
<http://staff.science.uva.nl/~cgmsnoek>

• We are hiring!

- PhD students on visual event recognition

High-tech eyes for industry and society

Ir. Jan Baan
TNO Netherlands




TNO innovation
for life

High-tech eyes for industry and society

Jan Baan, 23 March 2011
International Workshop on Computer Vision Applications







TNO innovation
for life

Content

The presentation will show four applications from industry and society, where computer vision is an important part of the total system. The algorithms are not discussed in detail. The topic of this presentation is to show the computer vision in relation to the application.

4 computer vision applications

- DOS (Detection Surface Damage of Very Porous Asphalt)
- Video Based traffic Monitoring
- Industrial 3D röntgen inspection
- (Airborne) 3D reconstruction and classification






TNO innovation
for life

DOS (Detection Surface Damage of Very Porous Asphalt)

Zoab (very porous asphalt) is frequently used in the Netherlands. There is a maintenance planning for each road section. The maintenance planning is done on site inspection by experts. Stone loss is the most important factor.

A new system has developed, that is able to measure stone loss, with a good statistical relation with the expert assessment of the road.



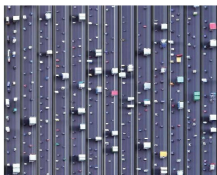




TNO innovation
for life

Video Based traffic Monitoring (VBM)

Cameras alongside the road follow each vehicle. The vehicle trajectories are used for traffic behavior analysis.

VBM is also an important part of the new A270 test site between Helmond and Eindhoven, where fifty cameras follow vehicles over a distance of five kilometers in real time. These vehicle trajectories are input for cooperative driving applications, which are proven on the test site A270.








TNO innovation
for life

Industrial 3D röntgen inspection

Agriculture product has wide variety in quality. Each product is different. A 3D röntgen inspection system is able to scan and sort these product in real time in a production line. For the 3D reconstruction Laminography techniques are used, computer vision techniques are used for the interpretation of the products.

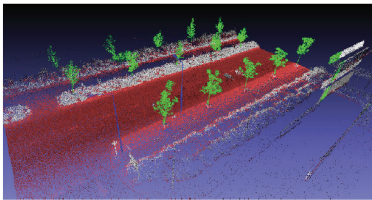




TNO innovation
for life

(Airborne) 3D reconstruction and classification

With high resolution photo images a 3D reconstruction of large areas is made. The application is change detection for the inspection of infrastructure. Classification techniques are developed for interpretation of 3D point clouds.



Multi-camera video analysis for activity monitoring of people

Dr.ir. Peter Van Hese
Univ. of Ghent, Belgium

workshop Computer Vision Applications, Eindhoven, 23/03/2011

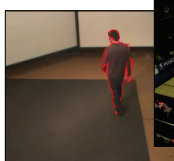
Multi-camera video analysis for activity monitoring of people

Peter Van Hese

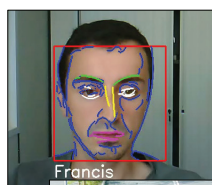


Overview of applications in this presentation

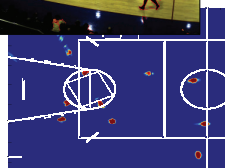
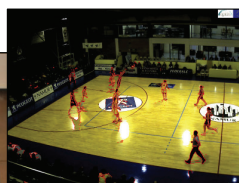
- Occupancy monitoring



- Face analysis



- Vehicle tracking



Features used:

moving edges

parabola edge maps

signatures

The Image Processing and Interpretation research group (IPI) at Ghent University, Ghent, Belgium

- 33 PhD-students, 5 postdocs, 1 technology developer, 1.5 professors, 3 associated professors
- Focus: image restoration and analysis, including video, 3D, medical, ...



Selection of IPI research topics

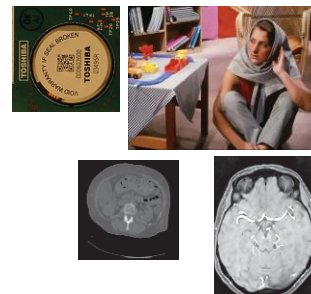
Image & (multi-camera) video analysis

- occupancy monitoring
- face analysis
- traffic analysis and security
- view selection
- mobile mapping and visual odometry
- segmentation (MR, CT, ...)
- ...



Image & video restoration

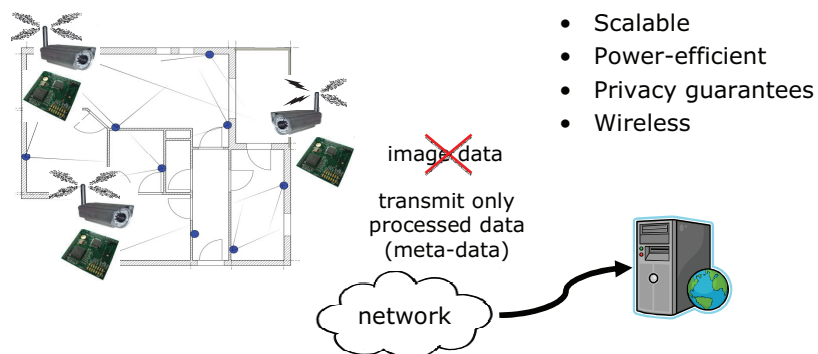
- noise removal (MRI, CT)
- super-resolution
- image deblurring
- 3D imaging
- ...



Overview

- Activity monitoring using smart camera networks
- Applications:
 - Occupancy monitoring and moving edges
 - Face analysis using parabola edge maps
 - Vehicle tracking using parabola edge maps
 - Vehicle tracking using signatures

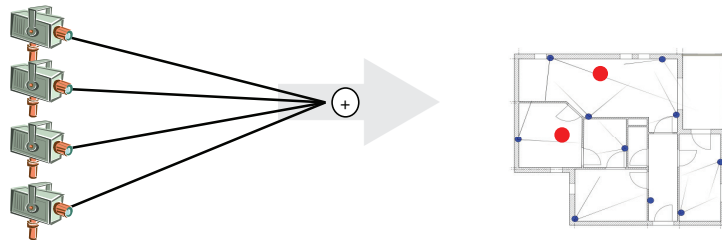
Smart camera networks provide a better solution



Requires

- (1) robust distributed video analysis and
- (2) power-efficient high-performance smart camera platforms

Video analysis for smart camera networks: key issues



computer vision algorithms

B/F segmentation

features: edges, parabolas, signatures

camera collaboration

data fusion (DS)
view selection

data for **applications**

occupancy maps
trajectories
recognition

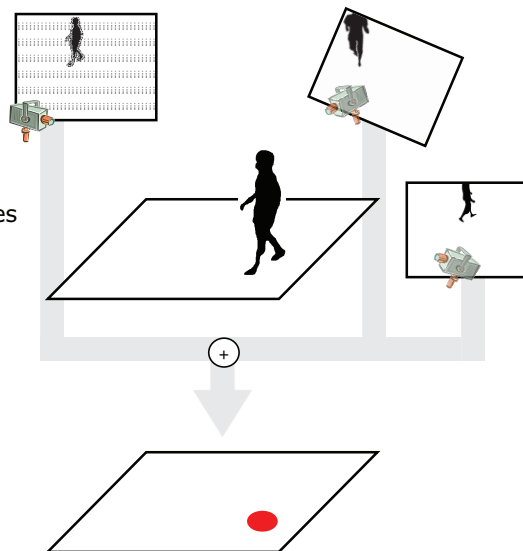
- smart cameras: local processing
- network of cameras: camera collaboration
- robustness: perform in realistic conditions, e.g., lighting changes

Occupancy monitoring and moving edges

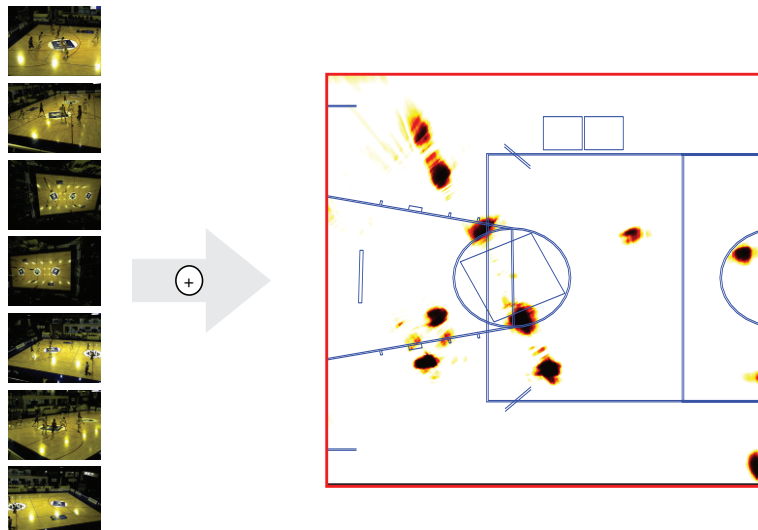
- Dempster-Shafer based multi-view occupancy maps
- Moving edges
 - as B/F segmentation and comparison to other B/F methods
 - robustness against lighting changes
 - as a compact representation
- Real-time occupancy monitoring
- Tracker

Dempster-Shafer based multi-view occupancy maps

1. B/F segmentation
2. Dempster-Shafer evidences per camera
3. data fusion

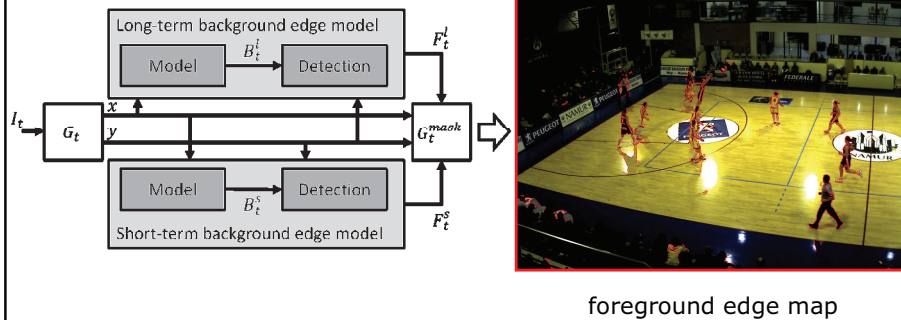


Fusion of all views



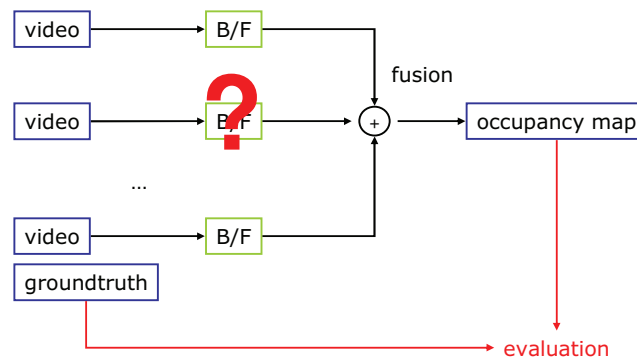
Moving edges

- Short-term and a long-term modeling of gradient image

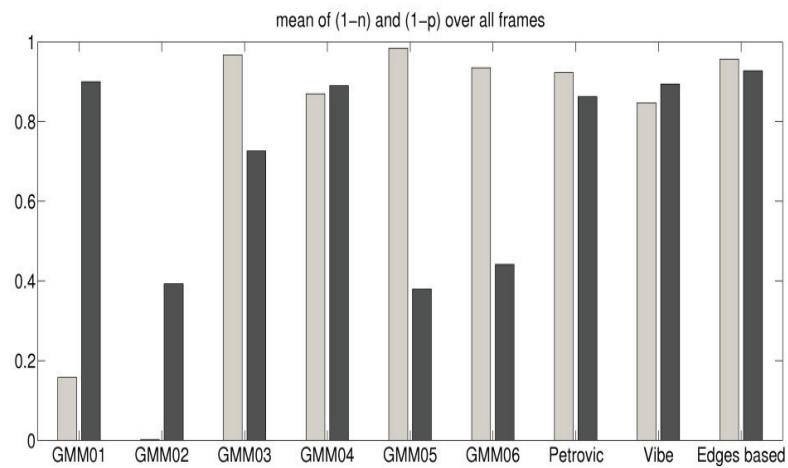


Evaluation of different B/F segmentation methods for Dempster-Shafer multi-view occupancy maps

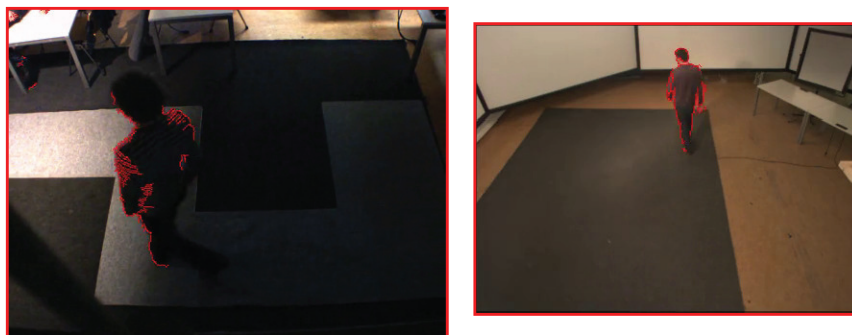
- ViBe
- gaussian mixture model
- Petrovic et al.
- moving edges based B/F method



All methods perform similar, except the new edges based segmentation method which outperforms all other methods

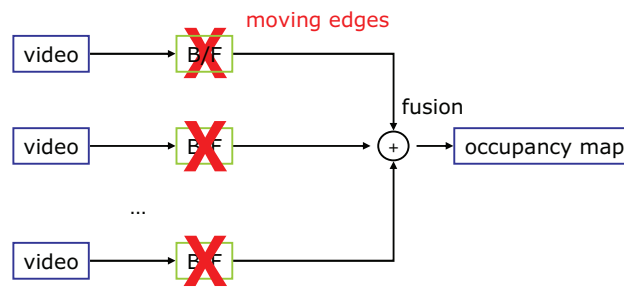


Moving edges: robustness against lighting changes



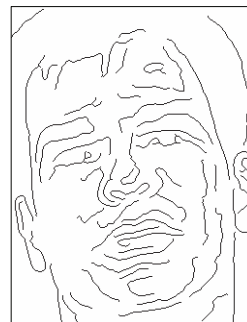
Moving edges as a compact representation

- Robustness against lighting changes
- Future: cameras send only meta-data, using edges as a compact representation



Face analysis using parabola edge maps

- Geometric features
 - constructive polynomial fitting
 - parabola matching
- Face recognition



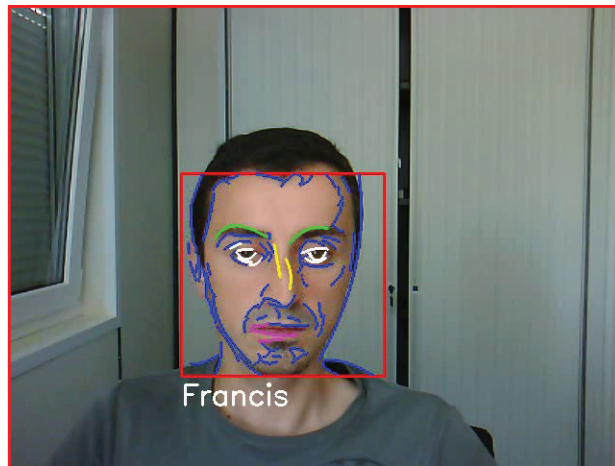
Face recognition



Face recognition

	LEM	PEM Distance	PEM Intensity	PEM Combination
<i>Controlled condition</i>				
GTFD	84,00%	72,00%	96,00%	98,00%
ATT	95,00%	90,00%	97,50%	100,00%
BERN	80,00%	93,33%	100,00%	100,00%
AR	96,40%	88,10 %	100,00%	98%
<i>Varying pose</i>				
BERN Right	55,00%	68,34%	90,00%	93,34%
BERN Left	48,33%	68,34%	91,67%	86,67%
BERN Up	46,67%	70,00%	88,33%	86,67%
BERN Down	45,00%	68,34%	78,34%	73,34%
<i>Size variation</i>				
AR with size variation	53,80%	70,56%	85,30%	90,21%
<i>Varying lighting condition</i>				
AR with left light on	92,86%	80,12%	93,27%	96,34%
AR with right light on	91,07%	82,10%	94,40%	94,84%
AR with both lights on	74,11%	78,30%	88,67%	92,10%
<i>Varying facial expression</i>				
AR with smiling expr.	78,57%	85,26%	98,34%	96,53%
AR with angry expr.	92,86%	82,30%	96,11%	97,30%
AR with screaming expr.	31,25%	71,62%	98,34%	96,21%

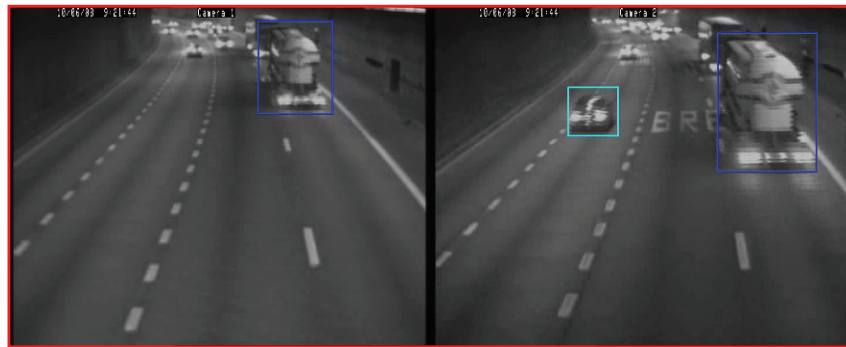
Real-time face analysis



Results: vehicle tracking using parabola edge maps



Results: vehicle tracking using signatures



Extended Abstracts and Posters

Cost-efficient Nucleus Detection in Histopathology Images using AdaBoost

Jelte Peter Vink and Marinus Bastiaan van Leeuwen
Philips Research Laboratories, Eindhoven, The Netherlands

Abstract—In March 2010, the Philips Digital Pathology Venture has introduced an ultra-fast scanner with which pathological imagery becomes available in digital format. This, in turn, enables automated assessments that enhance quality and throughput time of the diagnosis. Nucleus detection can be considered as one of the corner stones of a large number of applications for digital pathology.

We have addressed the problem of detecting nuclei. We have developed a framework to automatically train a nucleus detector that combines a high detection rate with low computational complexity. To this end, we have modified the supervised machine learning technique AdaBoost to include awareness of computational feature cost by bias towards previously selected features. By that, the computational complexity of the trained detector has dropped tremendously without decreasing the performance.

I. INTRODUCTION

Clinical pathology is a medical specialty that is concerned with the diagnosis of diseases, based on the laboratory analysis of tissue and bodily fluids such as blood and urine. In case of cancer, the diagnosis of the pathologist is the major indicator for the presence or absence of cancer and for the type of cancer. The pathologist studies the cell morphology, the staining pattern, the staining intensity, and the ordering of the cells in tissue (histopathology).

Automated nucleus detection has been widely studied [1], [2] resulting in many different methods, such as thresholding [3], watershed/water immersion, active contours, (generalized) Hough transform for circles/ellipsoids, h-maxima transform/h-dome, radial voting, graph-cuts, level set and mean shift. However, most methods are based on over-simplified segmentation concepts and are unable to achieve sufficient robustness to meet the application requirements.

Deterministic approaches to nucleus detection generally fail to deal with the highly heterogeneous character of pathological imagery, originating from both natural (e.g., life-cycle stadium) or procedural (e.g., tissue preparing, fixation and staining of tissue, and digitalizing of glass slide) differences. Machine learning strategies offer more flexibility and their ability for generalization render them more suitable for this particular application.

In our research we aimed at a pixel based detector to distinguish between nucleus pixels and background pixels based on AdaBoost, which we have modified to include awareness of computational feature cost by bias towards previously selected features.

In Section 2, we shall describe the design of this detector, in Section 3 we present an evaluation, while we draw our

conclusions in Section 4.

II. NUCLEUS DETECTOR DESIGN

To create the nucleus detector, we have used the supervised machine learning technique AdaBoost [4]. Viola and Jones [5] created a so-called attentional cascade which minimizes the computation requirements, while achieving high detection rates through advanced feature selection based on AdaBoost. We have recognized the potential of this approach to create a nucleus detector for digital pathology images.

From a training data set, AdaBoost creates a function H that maps feature values (\vec{x}_i) to desired outputs (y_i)

$$\begin{aligned} \vec{x}_i &\in X^M, X \in \mathbf{R}, \\ y_i &\in Y = \{-1, 1\}, 0 \leq i < N, \end{aligned} \quad (1)$$

where N and M are the number of samples and feature values, respectively. AdaBoost establishes a function H

$$H : X^M \rightarrow Y \quad (2)$$

that minimizes the error E [4]

$$E_H = \sum_{i=0}^{N-1} D(i) [y_i \neq H(\vec{x}_i)], \quad (3)$$

with respect to distribution $D(i)$. The classifier H is created using the following algorithm [6], [7] (see Algorithm 1).

Algorithm 1 AdaBoost [6], [7]

Initialize the distribution over the training set $D_1(i) = \frac{1}{N}$

For $t = 1 \dots T$

- 1) Train *Weak Learner* using distribution D_t
- 2) Calculate weight $\alpha_t \in \mathbf{R}$
- 3) Update the distribution over the training set:

$$D_{t+1}(i) = \frac{D_t(i)e^{-\alpha_t y_i h_t(\vec{x}_i)}}{Z_t}$$

where Z_t is a normalization factor for distribution D_{t+1}

Final classifier $H(\vec{x})$ is:

$$H(\vec{x}) = \begin{cases} 1 & \text{if } \left(\sum_{t=1}^T \alpha_t h_t(\vec{x}) \right) \geq \Theta \\ -1 & \text{otherwise} \end{cases}$$

where Θ represents a threshold

The *Weak Learner* selects the feature h which best separates the weighted positive and negative examples. For each feature, the *Weak Learner* determines the optimal threshold, such that the weighted error E_h is minimized [5].

Once a feature has been calculated, its residual computational cost in classifier H is negligible. Therefore, to reduce computational complexity of the trained detector, the *Weak Learner* should be biased towards previously selected features. Features that have not been selected yet are penalized to balance the performance increment they enable against the associated increase in computational cost. Therefore, *Weak Learner* has been adjusted to make use of penalty $\vec{\epsilon}_t \in \mathbb{R}^M$.

Adaboost requires a training set comprising a set of features and samples. We have semi-automatically labelled pixels representing nuclei using thresholding and manual adjustments. Different nucleus types are present in the images. In this work, we will focus on the nuclei of lymphocytes and small epithelial cells in immunohistochemistry (IHC) stained images. 30% of the pixels of 4 images has been used as training set. Next, potentially relevant features were calculated for each pixel, i.e.

- Standard deviation, StD , for a window W_n of $n \times n$
- Dynamic range, DR , for a window W_n of $n \times n$
- Local average, Avg , for a window W_n of $n \times n$

where

$$n = 3, 5, 7, \dots, 25 \quad (4)$$

Furthermore, we added 1D and 2D Haar-like features because these are inexpensive to calculate using an integral image [5].

III. RESULTS

The trained nucleus detector requires only 6 features, namely two 2D Haar-like features, two Avg features ($n = 5$ and $n = 19$), one StD feature ($n = 19$) and one DR feature ($n = 25$). Without our modification to *Weak Learner*, an additional 7 features were required.

Fig. 1 shows the performance of the nucleus detector on an IHC-stained image. As can be seen, almost all targeted nuclei are correctly detected as a solid cluster of pixels in the center of the nuclei. Based on 35 images, the pixel based detector has a detection rate of 99% and a false alarm rate of 5%. The high false alarm rate is related to the border region of the nuclei, which was not labeled but is detected occasionally.

Similar performance was achieved without our modification to *Weak Learner*.

The variation in appearance of the nuclei is large. As shown in Fig. 2, the current detector has difficulties in detecting the

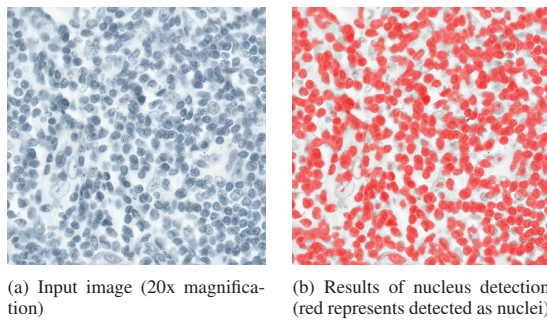


Fig. 1. Image of digitized glass slide of breast tissue using IHC-staining

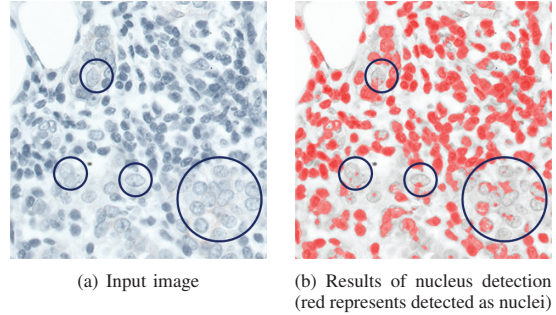


Fig. 2. Second image of digitized glass slide of breast tissue using IHC-staining (circles indicate nuclei that were not detected)

nuclei of large epithelial cells. Future work will focus on improving the detection of these nuclei.

IV. CONCLUSION

We have designed a generic framework, based on the supervised machine learning technique AdaBoost, to create a nucleus detector. AdaBoost has been used to select and combine useful features. By including awareness of computational feature cost by bias towards previously selected features, an improvement to AdaBoost has been made to reduce the computational complexity of the trained detector tremendously, without decreasing its performance.

The created nucleus detector has a detection rate of 99% and a false alarm rate of 5%. This detector requires only 6 local features.

The framework has a flexible design. In this work, we have focused on IHC-staining, but other staining can easily be included.

Future work will focus on further complexity reduction of the detector, as well as improving the detection of nuclei of especially large epithelial cells.

ACKNOWLEDGMENT

The authors would like to thank Erasmus MC and Philips Digital Pathology Venture for their support. This work was funded by the iCare project and will continue in Cyttron-II.

REFERENCES

- [1] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 146–168, 2004.
- [2] M. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE reviews in biomedical engineering*, vol. 2, pp. 147–171, 2009.
- [3] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [4] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *International Conference on Machine Learning*, 1996, pp. 148–156.
- [5] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.
- [6] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *European Conference on Computational Learning Theory*, 1995, pp. 23–37.
- [7] J. P. Vink and G. de Haan, "No-reference metric design with machine learning for local video compression artifact level," *IEEE Journal of Selected Topics in Signal Processing*, vol. PP, no. 99, p. 1, 2010.

Cost-efficient Nucleus Detection in Histopathology Images using AdaBoost

Jelte Peter Vink and Marinus Bastiaan van Leeuwen

Motivation

General

- Pathologists analyze tissue for diagnostic purposes
- In 2010, Philips Digital Pathology Venture launched an ultra-fast scanner, that
 - provides pathological imagery in digital format
 - enables automated nucleus detection

- Desire for high-throughput imposes constraints on computational complexity

- Nucleus detection is key component for many pathological analyses

- Nucleus detection is challenging due to large variation in appearance of nuclei

- Proposed methods do not meet robustness requirements

Solution

- Machine learning strategies offer flexibility and ability for generalization

Detector design

General

- Pixel based detector to distinguish between nucleus and background pixels

AdaBoost

- Included awareness of computational feature cost, bias towards previously selected features, to reduce computational complexity

Labeling

- Semi-automatically labeled IHC-images using thresholding and manual adjustments

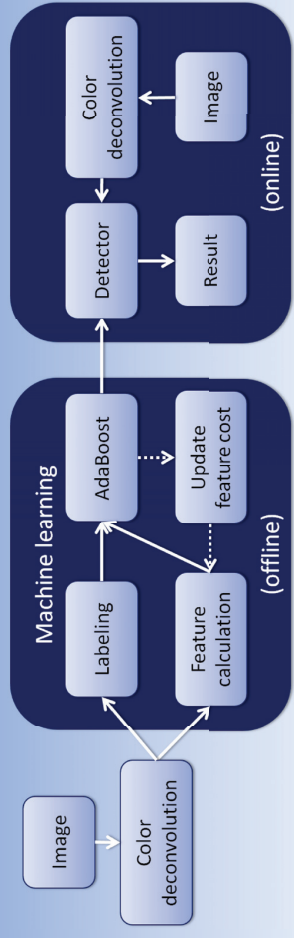
Feature set

- Local features like standard deviation, dynamic range and local average
- 1D and 2D Haar-like features using integral image

Sample set

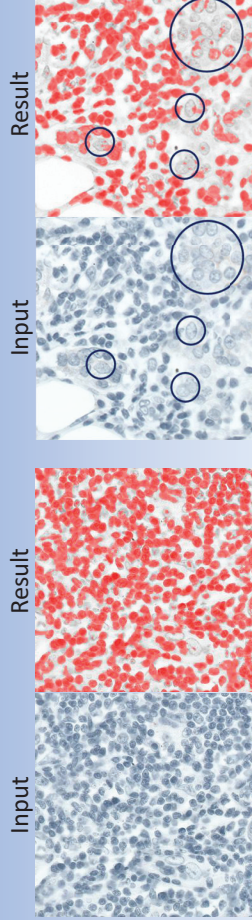
- Trained on 30% of 4 IHC-stained images

Overview framework



Results

- Trained nucleus detector requires only 6 features i.s.o. 13 features, with a detection rate of 99% and false alarm rate of 5% based on 35 IHC-images



Conclusion & future work

- Developed framework to train nucleus detector with high performance
- Tremendous complexity reduction by improvement to AdaBoost, without sacrificing performance
- Future work aimed at further improving performance and robustness

The authors would like to thank Erasmus MC and Philips Digital Pathology Venture for their support. This work was funded by the iCare project and will continue in Cytron-IL.

Sparse Window Stereo Matching

Sanja Damjanović, Ferdinand van der Heijden and Luuk J. Spreeuwers

Signals and Systems Group, Faculty of EEMCS, University of Twente, P.O.Box 217, 7500 AE Enschede, The Netherlands
 {s.damjanovic, f.vanderheijden, l.j.spreeuwers} @ewi.utwente.nl

Abstract—We proposed a new local stereo matching algorithm for dense matching of gray images. The algorithm is based on selection of a set of pixels from the matching windows which participate in the cost calculation and represents a hybrid approach in between the pixel based and the window based local stereo matching approach. The optimal choice of the window size and the threshold value in sparse window matching is important and depends on the stereo pair properties. We chose the optimal parameters for different stereo pairs and demonstrate the algorithm performance on the test images from the Middlebury stereo evaluation framework.

I. INTRODUCTION

Stereo matching algorithms can be classified into two categories: local and global [2]. In local stereo matching, the cost is aggregated over a support window which is most often rectangular. It is inherently assumed that all pixels within the matching window have the same disparity. This is not true for e.g. curved surfaces due to perspective distortion and occlusion. Also, the dimension of the objects whose disparity can be successfully recovered depends on the window size: the object's height and width in the image should be at least half the size of the window dimensions.

The ideal window for matching would be only one pixel. However, the one-pixel window does not provide sufficiently discriminatory cost for the local stereo matching. In order to combine the support of many pixels for cost aggregation as in the window-based matching but not to be limited by the window dimension like in the pixel-based matching, we introduce the hybrid support: a set of properly chosen pixels within the rectangular window i.e. "sparse window" [3].

We improve the matching results from [3] by choosing the optimal parameters in sparse window matching for different stereo pairs. Stereo pairs from the evaluation framework [1] have different properties. The stereo pairs differ in sizes, disparity range and level of details, see table I. We improve the postprocessing step from [3] by introducing the disparity consistency check and by filling in the missing disparities.

II. SPARSE WINDOW MATCHING

We consider a pair of gray valued, rectified stereo images I_L and I_R with disparity range D . We recover the disparity map which corresponds to the reference image I_L . In the matching process, we observe the rectangular $W \times W$, $W = 2 \cdot w + 1$, windows and select some pixels from the left and right matching windows as a suitable for matching if and only if the pixels lay within the fixed threshold T from the central pixels. The pixel from the left matching window declared as suitable is selected for the cost aggregation step only if the pixel at

the same position from the right window is also declared as suitable for matching. From the N_p selected pixels in each window, we form two $N_p \times 1$ vectors. The sum of squared differences normalized to N_p is used for the cost calculation. The adjusted Winner-Takes-All (WTA) method is applied to trustworthy disparity candidates [3].

A. Parameter selection

We vary the values of the parameters: the window size w and the threshold T . We calculated the disparity error rate for different w and T and chose those which give the smallest errors w.r.t. the ground truth disparity maps, Table I.

B. Postprocessing

We calculate disparity maps corresponding to the both images of the stereo pair using the optimal parameters. As the first postprocessing step, we apply 5×5 median filter to the both disparity maps. Next, we perform consistency check with the tolerance 1 for the disparity map corresponding to the left image. The inconsistent disparities are filled in by one of the four closest consistent neighbor disparities along vertical or horizontal direction. We chose the disparity of the neighbor pixel with the smallest intensity difference with the pixel with the inconsistent disparity. Finally, we apply 7×7 median filter to the final disparity map.

III. RESULTS AND CONCLUSION

Figure shows the resulting disparity maps obtained by our algorithm for the stereo pairs from the Middlebury database. The quantitative results within the Middlebury stereo evaluation framework are presented in Table II. For the stereo pairs *Teddy* and *Cones* we applied the central point subtraction step to compensate for the radiometric differences [4], [3].

The results show that with our hybrid technique edges of the objects are preserved. The disparities of some narrow structures are successfully detected and recovered, although their dimensions are much smaller than the size of the matching window. Such example of the narrow objects are most noticeable in *Tsukuba* disparity map (the lamp reconstruction) and in *Cones* disparity map (pens in a cup in the lower right corner). On the other hand, the disparities of the large low textured surfaces in stereo pairs *Venus* and *Teddy* are also successfully recovered with the same sparse window technique.

In comparison to our previous result in [3], the parameter optimization and the new postprocessing significantly reduced the error rates.

TABLE I
STEREO IMAGE PROPERTIES AND OPTIMAL PARAMETERS

Stereo pair	Size	Disparity range	w^{opt}	T^{opt}
Tsukuba	384x288	0 to 15	15	12
Venus	434x383	0 to 19	18	14
Teddy	450x375	0 to 59	12	16
Cones	450x375	0 to 59	15	12

TABLE II
ERROR PERCENTAGES WITH THE MIDDLEBURY RANKS [1] (MARCH 2011)

Stereo pair	Nonoccluded	Discontinuities	All
Tsukuba	1.88 (47)	3.10 (53)	8.96 (51)
Venus	0.21 (20)	0.71 (33)	2.84 (28)
Teddy	7.31 (44)	14.6 (61)	19.9 (64)
Cones	4.96 (62)	11.9 (62)	13.1 (69)

REFERENCES

[1] [Online]. Available: <http://vision.middlebury.edu/stereo/>

- [2] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 7–42, 2002.
- [3] S. Damjanović, F. van der Heijden, and L. J. Spreeuwens, "Sparse window local stereo matching," in *VISIGRAPP 2011, Vilamoura, Algarve, Portugal*. Vilamoura: INSTICC Press, March 2011, pp. 689–693.
- [4] —, "A new likelihood function for stereo matching: how to achieve invariance to unknown texture, gains and offsets?" in *VISIGRAPP 2009, Lisboa, Portugal*. Lisboa: INSTICC Press, February 2009, pp. 603–608.

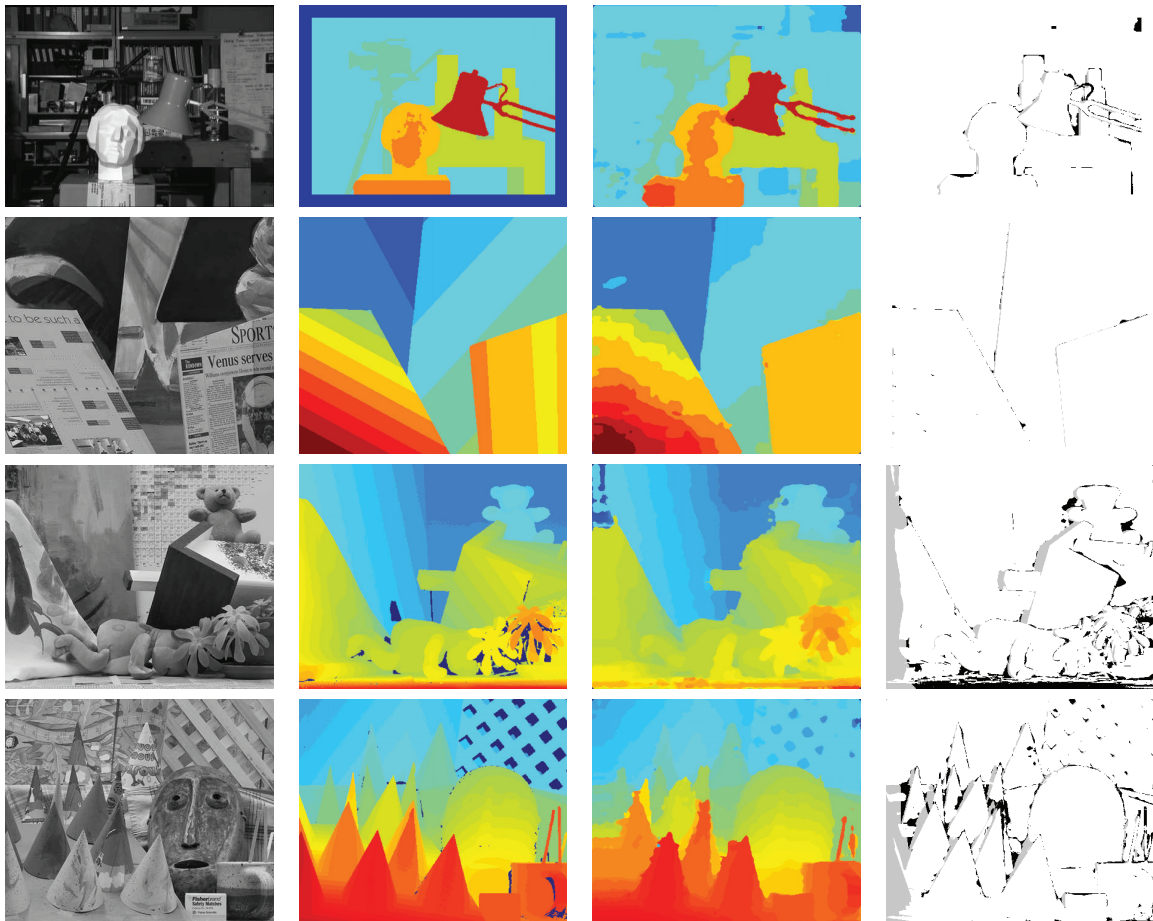


Fig. 1. Disparity results for the stereo pairs (1st row: Tsukuba, 2nd row: Venus, 3rd row: Teddy, 4th row: Cones) from the Middlebury database [1]. From left to the right columns show: The left image, Ground truth disparity maps, Result computed by the sparse window matching technique with postprocessing, Disparity errors larger than 1 pixel. The nonoccluded regions errors with ranking (March 2011) are respectively: *Tsukuba* 1.88% (47), *Venus* 0.21% (20), *Teddy* 7.31% (44), *Cones* 4.96% (62)

Sparse Window Stereo Matching

S. Damjanović, F. van der Heijden, L.J. Spreeuwers

Signals and Systems Group, Faculty of EEMCS, University of Twente, The Netherlands
s.damjanovic@ewi.utwente.nl

UNIVERSITY OF TWENTE.

Introduction

Local stereo matching

Rectangular window based:

- assumes the same disparity within window
- smooths disparity at discontinuities
- does not recover the disparity of thin objects

Area based:

- properly shaped support area
- requires costly segmentation

Pixel based:

- ideal window
- does not provide sufficient information

Sparse window matching:

- subset of pixels from the rectangular window
- assumes only pixels from the subset have the same disparity

Contribution

We introduced new hybrid matching approach sparse window matching and we further improve the results by introducing:

- parameter optimization
- new postprocessing

Experiments

- Sparse window matching technique evaluation on the stereo images from the Middlebury benchmark using windows of the optimal size for the stereo pairs $W = 2 \cdot w + 1$ and optimal threshold T
- Other parameters: $\sigma_n^2 = 0.5$, $N_D = 3$, $N_E = 5$, $K_p = 0.5$, $L = 5$

Algorithm

Pixel selection

Pixels $w_l^{i,j}$ and $w_r^{i,j}$, at position (i, j) in $W \times W$ matching windows, are selected, if $|w_l^{i,j} - c_l| < T_L \wedge |w_r^{i,j} - c_r| < T_R$ where c_l and c_r are central pixels.

Pixel selection is done for each pixel (r, c) in the reference image and for each possible disparity $d \in \{0, \dots, D\}$, resulting in $N_p = N_p^{r,c}(d)$ pixels selected for the cost aggregation.

Special cases:

1. $N_p \approx W^2$: erode low-textured window ($N_E \times N_E$)
2. $N_p \ll W^2$: dilate rich-textured window ($N_D \times N_D$)

Cost aggregation

z_l and z_r are $1 \times N_p$ matching vectors.

Sum of squared differences: $C_{nSSD} \propto \frac{1}{N_p} \cdot \frac{\|z_l - z_r\|^2}{4 \cdot \sigma_n^2}$

Adjusted Winner-Take-All

For trustworthy disparity candidates holds $N_p^{r,c}(d) > N_s^{r,c}$:

$$N_s^{r,c} = K_p \cdot \max_d \{N_p^{r,c}(d)\}, \quad 0 < K_p \leq 1$$

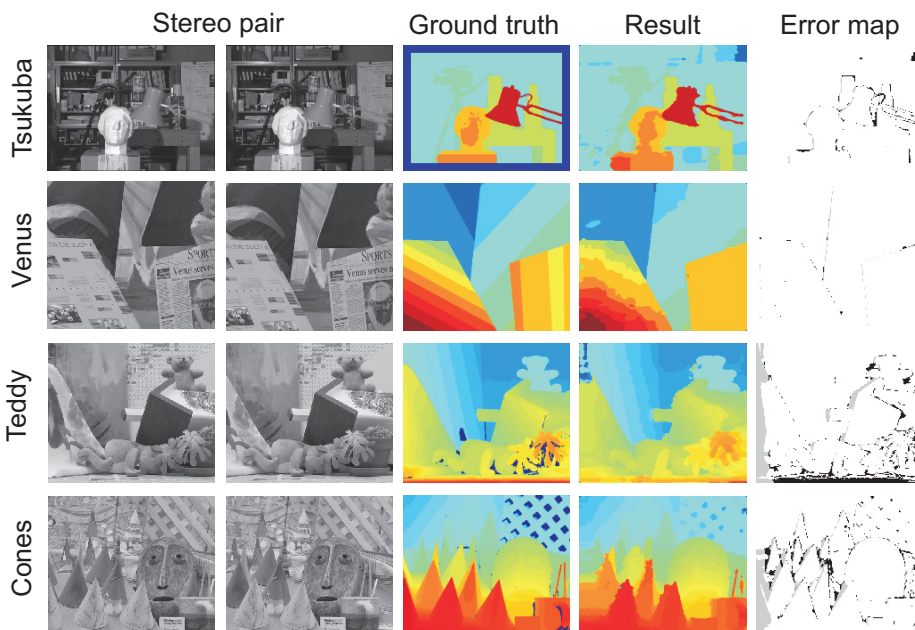
The optimal disparity at coordinates (r, c) :

$$d^{r,c} = \arg \min_{d \in \{0, \dots, D\}} \{C_{nSSD}^{r,c}(d) | N_p^{r,c}(d) > N_s^{r,c}\}$$

New postprocessing

- 5x5 median filtering of both disparity maps
- Consistency check with tolerance 1
- Filling in of the inconsistent disparities: we chose the disparity of the neighbor pixel with the smallest intensity difference with the pixel with the inconsistent disparity
- 7x7 median filtering

Results and Conclusions



Sparse window matching:

- simple implementation
- accurately recovers disparity of smooth surfaces
- preserves discontinuities
- scores well in the Middlebury ranking
- Future: occlusion treatment

stereo pair	size: RxC	disparity range 0 to D	optimal w	optimal T
Tsukuba	384x288	0 to 15	15	12
Venus	434x383	0 to 19	18	14
Teddy	450x375	0 to 59	12	16
Cones	450x375	0 to 59	15	12

stereo pair	NONOCC error [%] (rank)	DISC error [%] (rank)	ALL error [%] (rank)
Tsukuba	1.88 (47)	3.10 (53)	8.96 (51)
Venus	0.21 (20)	0.71 (33)	2.84 (28)
Teddy	7.31 (44)	14.6 (61)	19.9 (64)
Cones	4.96 (62)	11.9 (62)	13.1 (69)

Large Scale Detection, Classification and Localization of Traffic Signs

Ivo Creusen
Cyclomedia Technology B.V.
Achterweg 38, 4181 AE
Waardenburg
icreusen@cyclomedia.com

Lykele Hazelhoff
Cyclomedia Technology B.V.
Achterweg 38, 4181 AE
Waardenburg
lhazelhoff@cyclomedia.com

Abstract—Yearly inventories of traffic signs contribute to increased road safety and efficient sign maintenance. Whereas manual inventories are time consuming, automated localization of traffic signs is more efficient and less expensive. This paper describes the results obtained with an automated inventory system for traffic signs. Object-detection algorithms are applied to identify traffic signs in cycloramas, geo-referenced panoramic images that are captured from The Netherlands annually. We have employed multiple detectors, all capable of detecting one or more classes of traffic signs. Their output is fused, and afterwards, all detections are classified and positioned in the wgs84 coordinate system with an average positioning error of 20 cm. Within an area near Rotterdam, this process resulted in a list of 13.510 signs.

I. INTRODUCTION

Road safety and maintenance are important governmental responsibilities, aiming to prevent unnecessary collisions. One aspect is keeping up-to-date inventories of traffic signs and their maintenance status. Traffic signs can become damaged, dirty or go missing over time. Additionally, a good overview can reveal locations where traffic signs should be added or removed to improve road safety.

Performing an inventory of an extensive geographical region completely by hand is a labour intensive and error-prone task. Cyclomedia is in a uniquely suited position to propose a semi-automatic traffic-sign inventory-process. Cyclomedia has an annually updated database of *cycloramas*, geo-referenced panoramic photos taken every 5 meters on all public roads within the Netherlands. Based on these photos, the efficiency of the process is significantly improved, compared to searching for each sign to record its position. To further improve the efficiency, several algorithms can be applied to automate a significant part of the process. Object detection algorithms can be used to automatically locate traffic signs in the images. Once detected, an object classification algorithm determines the exact type of the sign. By combining detection results from multiple panoramic images, the 3D location of the sign is automatically estimated by a positioning algorithm.

II. SYSTEM OVERVIEW

The traffic-sign inventory-process consists of several stages, as shown in Figure 1. First, the cycloramas are processed by multiple detection algorithms. Two algorithms are currently

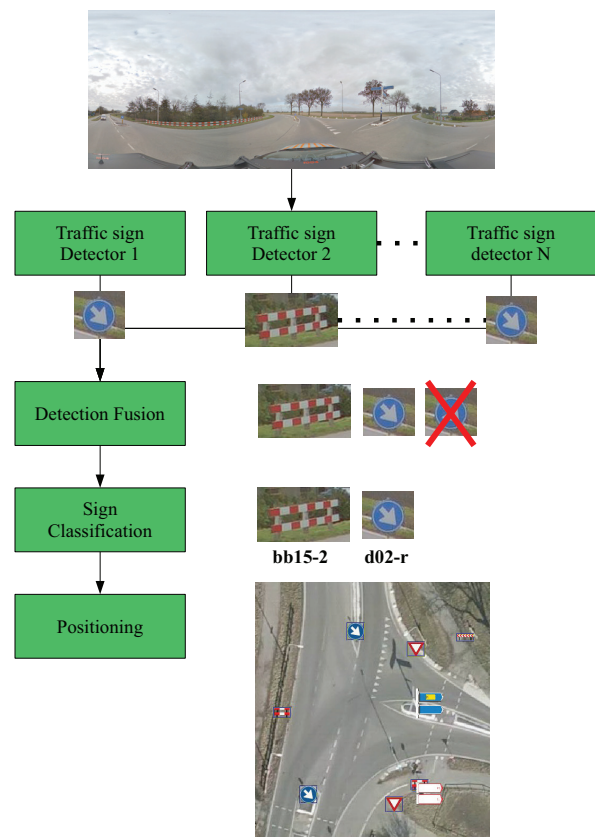


Fig. 1. System overview.

used in the system, one is based on the popular Histogram of Oriented Gradients (HOG) algorithm by Dalal and Triggs [1], and described in more detail in [2]. Multiple instantiations of HOG detectors are employed for detection. The other algorithm uses the Scale Invariant Feature Transform (SIFT) by David Lowe [3] in a preselected region, and matches these to a dictionary created from traffic-sign templates, for more details see [4]. Because some signs can be detected multiple times, the *detection fusion* step filters out these extraneous

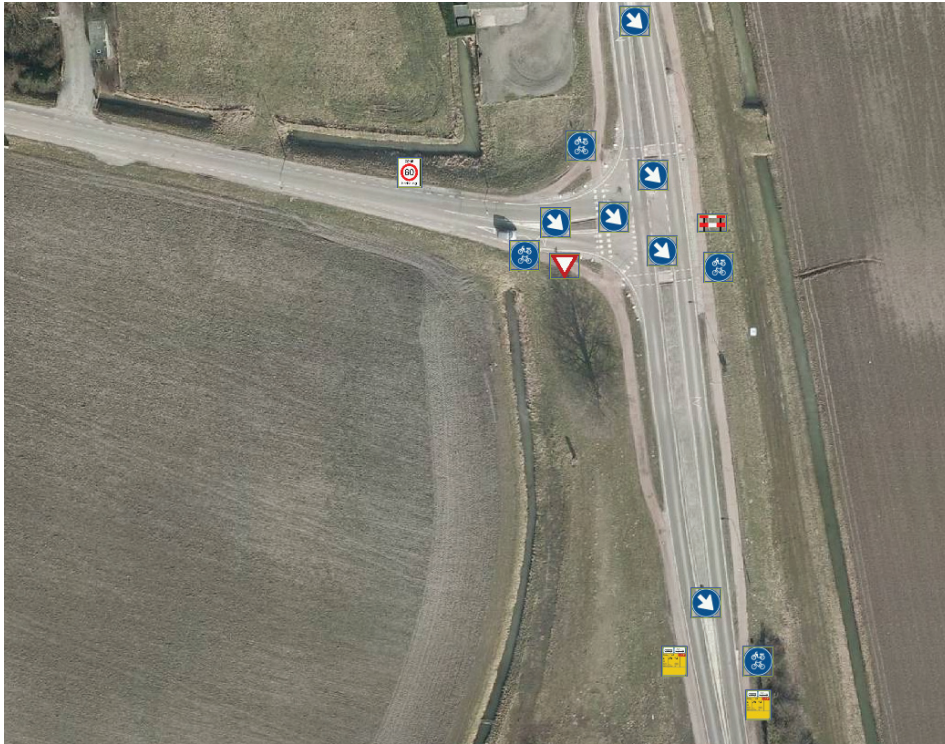


Fig. 2. Examples of localized traffic signs displayed on an aerial image of the respective road.

detections.

Afterwards, the *sign classification* algorithm identifies which exact traffic sign is in the detection bounding-box. To do this, the popular Bag of Words (BOW) approach is used [5], using SIFT features to match to a dictionary of features extracted from training images.

To determine the 3D coordinates of the signs, we utilize the recorded GPS position of the cycloramas. For each sign detected in one of the cycloramas, we know it's location must be on a virtual line extending from the GPS location, in the direction the sign was detected. By combining the results from multiple cycloramas, the position of the traffic sign can be estimated, for which meanshift, as described by Comaniciu and Meer [6], is employed.

III. RESULTS

The described system is applied to produce an inventory of all traffic signs located within the rural areas located near Rotterdam. This area included around 300.000 cycloramas, mostly captured at minor roads. All images are processed by our detection algorithms and the output is validated by means of manual inspection, including the insertion of sub-sign text. This procedure resulted in the localization of a total of 13.510 traffic signs, which are all positioned in 3 dimensions with an average accuracy of 20 cm. Fig. 2 displays examples of localized traffic-signs projected on an aerial image.

ACKNOWLEDGMENT

This work is partially sponsored by the ViCoMo project.

REFERENCES

- [1] N. Dalal and B. Triggs, "Histogram of oriented gradients for human detection", in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, June 2005, vol. 1, pp. 886-893
- [2] I. Creusen, R. Wijnhoven, E. Herbschleb and P.H.N. de With "Color Exploitation in HOG-based Traffic Sign Detection", in *Proc. IEEE International Conference on Image Processing (ICIP)*, Sept. 2010, pp. 2669-2672
- [3] D. Lowe "Distinctive image features from scale-invariant keypoints", in *Int. Journal of Computer Vision (IJCV)*, Jan. 2004, vol. 60, no. 2
- [4] E. Herbschleb and P.H.N. de With "Real-time traffic-sign detection and recognition" in *Proc. Visual Communications and Image Processing (VCIP), SPIE-IS&T Electronic Imaging*, Jan. 2009, vol. 7257, pp. 0A1-0A12
- [5] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray "Visual categorization with bags of keypoints" in *Workshop on Statistical Learning in Computer Vision (ECCV)*, May 2004, pp. 1-22
- [6] D. Comaniciu and P. Meer "Mean shift: A robust approach toward feature space analysis" in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, May 2002, vol. 24, pp. 603-619

Large Scale Detection, Classification and Localization of Traffic Signs

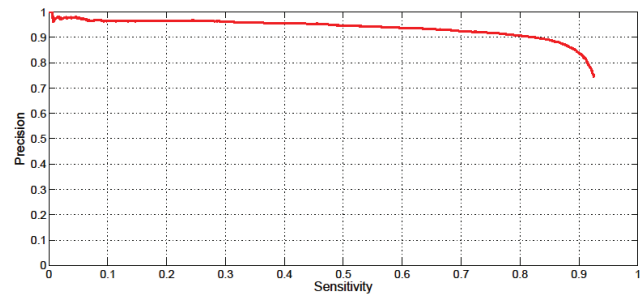
Ivo Creusen
Cyclomedia Technology B.V.
Achterweg 38, 4181 AE
Waardenburg
icreusen@cyclomedia.com

Lykele Hazelhoff
Cyclomedia Technology B.V.
Achterweg 38, 4181 AE
Waardenburg
lhazelhoff@cyclomedia.com

Abstract

Yearly inventories of traffic signs contribute to increased road safety and efficient sign maintenance. Whereas manual inventories are time consuming, automated localization of traffic signs is more efficient and less expensive. This paper describes the results obtained with an automated inventory system for traffic signs. Object-detection algorithms are applied to identify traffic signs in cycloramas, geo-referenced panoramic images that are captured from The Netherlands annually. We have employed multiple detectors, all capable of detecting one or more classes of traffic signs. Their output is fused, and afterwards, all detections are classified and positioned in the wgs84 coordinate system with an average positioning error of 20 cm. Within an area near Rotterdam, this process resulted in a list of 13.510 signs.

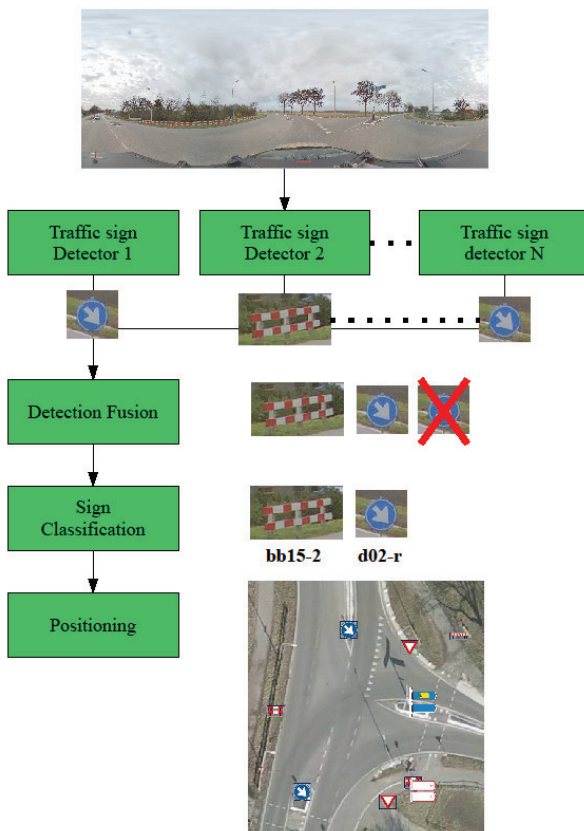
Example ROC curve for triangular signs



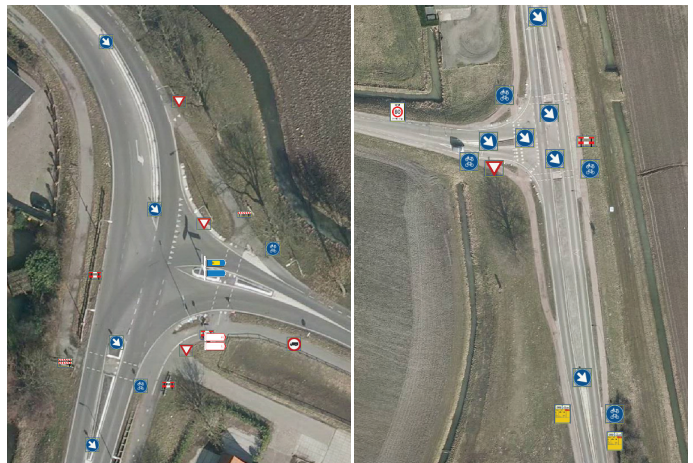
Examples of detected signs



System overview



Examples of localized signs



Results

- Processed 300.000 cycloramas.
- Localized 13510 traffic signs.
- Average position accuracy of 20 cm.



Face tracking camera system for surveillance

Rick Peerlings
ViNotion
P.O. Box 2346, 5600 CH Eindhoven
The Netherlands
rick.peerlings@vinotion.nl

Rob Wijnhoven
ViNotion
P.O. Box 2346, 5600 CH Eindhoven
The Netherlands
rob.wijnhoven@vinotion.nl

I. INTRODUCTION

Shoplifting is a recurring problem in retail stores. In order to identify the shoplifters, high quality imagery of their faces is required. Surveillance cameras typically monitor a large area, which results in low resolution images of faces. Pan Tilt Zoom (PTZ) cameras can be used by a security operator to manually zoom in on suspects. This has two limitations, it requires constant human attention, and it requires the operator to choose in real-time which person to follow. Moreover, although face detection algorithms exist [1], their application in low resolution surveillance images results in low quality face regions not suited for identification.

Within the ViCoMo project¹ we have designed a system that automatically zooms in on faces using face detection without any human interaction. In addition, it tracks people throughout their movement. Because the system records high resolution face images for every person, suspect identification can be done at a later time.

Also, automatic face recognition could be adopted to automatically detect repeat offenders, so security guards can take preventive measures.

II. PROPOSED SYSTEM

As a solution for the aforementioned problem, we present a face tracking camera system. This system scans each video frame to detect faces. If a face is detected, the PTZ camera zooms in on the face, and the face is kept centered in the image by moving the camera.

The system consists of five parts, see Figure 1. Each block is now discussed in detail. First, a video frame is captured by the camera block.

The face-detector block detects faces using a sliding window detector, that classifies each image window into face/non-face. To detect objects of different size, the detection process is repeated for scaled versions of the input image. We use scale steps of 1.05. Finally, a mean-shift mode-finding algorithm merges window-level detections. For this feature-based classification, we use our implementation of the Histogram of Oriented Gradients (HOG) algorithm, as proposed by Dalal and Triggs [2]. We use the following parameters: cells of 8×8

pixels, 4 block normalizations, 18 orientation bins using the sign, L2 feature normalization and a detector size of 48×48 pixels. The detector is trained to detect full-frontal faces using a fast learning algorithm [3].

Once a face is detected, tracking is started. Because objects have inertia, their movement between consecutive video frames is limited. Therefore, to find the face in the next image, only a limited region around the previous face position needs to be evaluated by the detector. To determine this search region, the current face position is predicted using the position in previous frames. For this prediction, we use the Kalman filter [4]. The filter temporally smoothes the face positions, and also extrapolates when the detector fails to detect a face. The accuracy (variance) of the prediction determines the search region. The scale of the previous detection limits the number of scales to search. Restricting the regions and scale results in faster detection and improves tracking robustness.

The P-control block uses the face positions from the Kalman filter to calculate control parameters for the PTZ camera movement. The controller keeps the face position centered in the video frame by moving the camera accordingly. Because face movement is limited, we found proportional control to be sufficient for our problem. The proportional controller has one configurable parameter, the gain, which needs to be set for the application. If the gain value is too small, the response is too slow, which can result in the face moving outside the camera view. On the other hand, if the gain value is too large, overshoots occur which results in a back and forth shaking motion around the face location.

The control block sends the parameters to the PTZ unit block, that actuates and moves the camera. Via this mechanical link between the PTZ unit and camera, the control loop is closed.

A. Operational modes

The face tracking camera system has three operational modes. At system start, no face position is known.

- 1) Scanning mode; the entire image is scanned for faces. When a face is found, the system switches to the tracking mode (Figure 2).
- 2) Tracking mode; the face is tracked by controlling the PTZ using the face detections (Figure 3).
- 3) Prediction mode; when in tracking mode and a face is not found, the Kalman filter predicts the face movement.

¹The Video Context Modeling (ViCoMo) project is an European ITEA project, focusing on the modeling of the context in which video-interpretation algorithms are used to improve the intelligence of these algorithms (<http://www.vicommo.org/>).

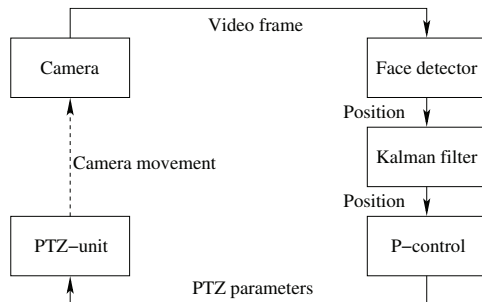


Fig. 1. Block diagram of the face tracking camera system.

If a face is not found for a certain amount of frames, the system goes back to scanning mode (Figure 4).

III. RESULTS

The proposed system has been implemented and performs detection and tracking in real-time. Our software implementation on a PC connects to a Pelco camera using the Pelco-D serial protocol.

The face detector in the tracking mode can be improved. Because it only detects full-frontal faces, track is lost once the face rotates. This will be solved as future work by continuously updating the detector for the person being tracked.

Because the face detector is based on shape descriptors, it is robust for different lighting conditions. The detector uses machine learning, so it can easily be trained to detect other objects, e.g. cars or humans.

The system has been successfully demonstrated at the Vision and Robotics fair in Veldhoven, the Netherlands in May 2010² and the ITEA co-summit in Ghent, Belgium in October 2010³. At both events, the majority of faces were correctly detected and tracked.

IV. CONCLUSIONS

In this paper we have presented a fully autonomous face tracking camera system for surveillance applications. The output of our robust face detection is filtered by a Kalman filter. Using the filtered detections, the control subsystem calculates the movement parameters for the PTZ camera. The tracking system is a control loop that keeps the face centered with a high zoom factor in the camera image. The obtained high resolution images enable the identification of faces.

REFERENCES

- [1] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 24, no. 1, pp. 34–58, January 2002.
- [2] N. Dalal and B. Triggs, "Histogram of oriented gradients for human detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 1. IEEE, June 2005, pp. 886–893.
- [3] R. G. J. Wijnhoven and P. H. N. de With, "Fast training of object detection using stochastic gradient descent," in *Proc. IEEE International Conference on Pattern Recognition (ICPR)*, August 2010, pp. 424–427.
- [4] G. Welch and G. Bishop, "An introduction to the kalman filter," 2001.

²<http://www.visionandrobotics.nl/>

³http://www.itea2.org/cosummit2010/_home



Fig. 2. Tracker in scanning mode.

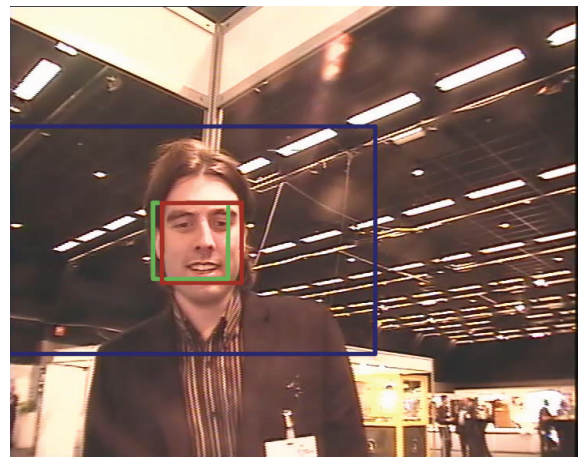


Fig. 3. Tracker in tracking mode. Red box: face detection; green box: Kalman prediction; blue box: region of interest.



Fig. 4. Tracker in prediction mode. Green box: Kalman prediction; blue box: region of interest.



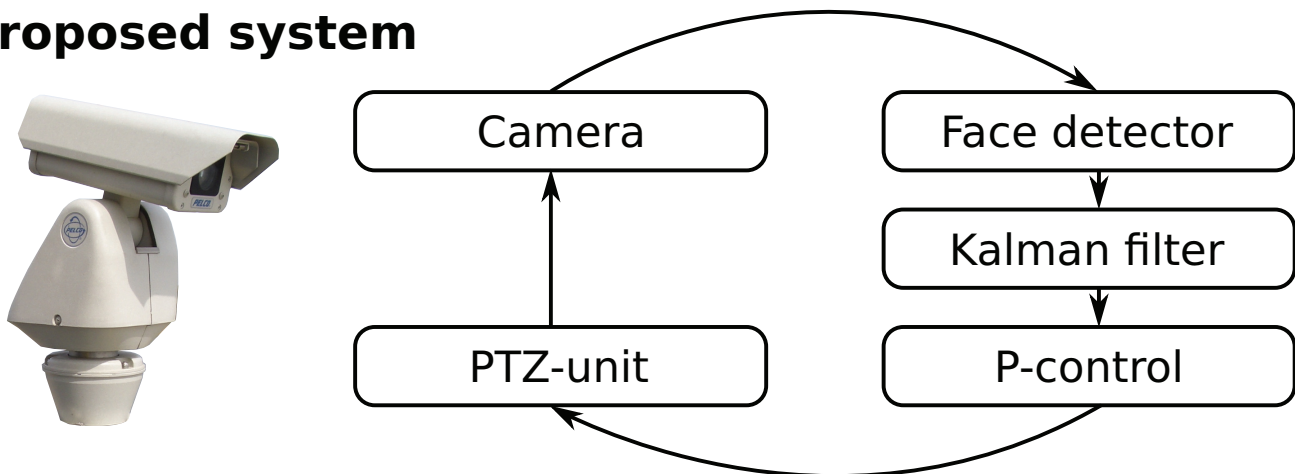
Face tracking camera system

Rick Peerlings, ViNotion

Introduction

- Identifying shoplifters requires high resolution face images
- But static cameras typically monitor large areas
- We propose a system that tracks faces using a PTZ camera

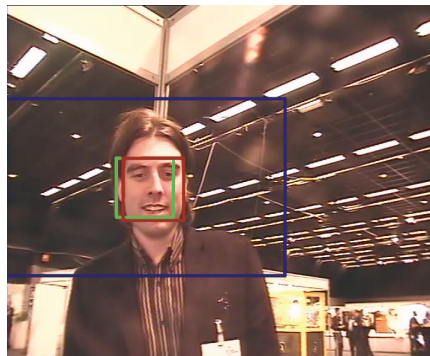
Proposed system



Operational modes



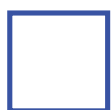
Scanning mode



Tracking mode



Prediction mode



*Region of interest
to search for faces*



*Kalman
prediction*



Detected face

Results

- Robust for different lighting conditions
- Can easily be retrained to detect and track other objects, e.g. cars
- Successfully demonstrated at two events

Automatic Assessment of Customers' Buying Behavior

Mirela Popa, Leon Rothkrantz, and Pascal Wiggers
MMI Department, TU Delft
Delft, the Netherlands
{m.c.popa;l.j.m.rothkrantz;p.wiggers}@tudelft.nl

Caifeng Shan and Tommaso Gritti
Video and Image Processing, Philips Research
Eindhoven, the Netherlands
{caifeng.shan;tommaso.gritti}@philips.com

Abstract—Video Analytics covers a large set of methodologies which aim at automatically extracting information from video material. In the context of retail, the possibility of effortlessly gather statistics on customer shopping behavior is very attractive. In this work, we focus on the task of automatic classification of customer behavior, with the objective to recognize buying events. The experiments are performed on several hours of video collected in a supermarket. Given the vast effort of the research community on the task of tracking, we assume the existence of a video tracking system capable of producing a trajectory for every individual, and currently manually annotate the input videos with trajectories. From the annotated video recordings, we extract features related to spatio-temporal behavior of the trajectory, and to the user movement and analyze the shopping sequences using a Hidden Markov Model (HMM). First results show that it is possible to discriminate between buying and non-buying behavior with an accuracy of 70%.

Key Words—Trajectory analysis, Optical flow, Hidden Markov Models, Shopping Behavior.

I. INTRODUCTION

THERE is an increasing amount of research in the area of video analysis and semantic interpretation as an application to automatic surveillance, traffic monitoring, video games, marketing, etc. In the field of marketing it is of primary concern to identify the most appealing products and services for customers and to maximize their impact. Computer vision provides multiple techniques which enable surveillance, action recognition, and behavior interpretation of customers.

People tracking, behavior analysis, and prediction were investigated by Kanda et al. in [3]. Accumulated people's trajectories over a long period of time provided a temporal use-of-space analysis facilitating the behavior prediction task performed by a robot. Hu et al. [2] used the Motion History Image (MHI) along with the foreground image obtained by background subtraction and the histogram of oriented gradients (HOG) [1] to obtain discriminative features for action recognition. Next a multiple-instance learning framework SMILE-SVM was build to improve the performance. This approach proved its effectiveness on a real world scenario from a surveillance system in a shopping mall aimed at recognizing customers' interest in products defined

by the intent of getting the merchandise from the shelf.

These approaches are suitable for action recognition under varying conditions in complex scenes such as background clutter or partial occluded crowds; still they require supervised learning based on a large reliable dataset.

Tracking people inside the shop can have many applications, such as global shopping behavior recognition, region of interest detection both individually and for a group of customers, measured at a specific moment or over time intervals. We plan to use the existing surveillance systems to observe the shopping behaviour of people [4], to get a better understanding of their needs.

Currently we are working on the action recognition module which can provide cues regarding customers' interest in products and can help interpreting different interaction patterns, such as grasping a product immediately, after a period time or even after more visits at the same place.

In this paper we propose an automatic surveillance system for detecting customers' buying behavior based on tracking and motion information and tested on real-life recordings in a shopping mall. Its applicability resides in identifying different buying patterns in terms of number of interactions and time spent but also in finding for which products categories the customers have trouble deciding. As a result, appropriate actions could be taken such as new products arrangements and more efficient usage of the store space.

Next we present the system overview, the data acquisition process and the experimental results. Finally we formulate our conclusions and give directions for future work.

II. OUR STUDY

A. System Overview

In this section, the design of our system for automatic assessment of customers' buying behavior is presented. We propose a modular approach and we describe next the functionality of each module. A diagram of the proposed system is shown in Fig. 1. First the video file is processed in order to extract the image sequences. Next the trajectory extraction module is employed. Currently the customers' trajectories are manually labeled, given that our goal consisted in the high-level analysis of behavior. In our future work, trajectories will be extracted by adopting person detection and

tracking. The next module is trajectory analysis, which detects segments of interest and includes potential buying segments. The analysis is done using relevant features such as speed, curvature, and duration of staying in a region. For each segment, the corresponding human image area is computed for further analysis. Next the motion analysis module is applied to each segment, by estimating optical flow between every two consecutive images in the corresponding human area.

Histograms of the motion-related vectors are extracted and fed to a Hidden Markov Model classifier. We chose this type of classification method due to its characteristics such as incorporating the dynamics of motion features during time.

Finally a conclusion is drawn regarding the buying behavior of a customer for the given trajectory segment, based on the maximum likelihood.

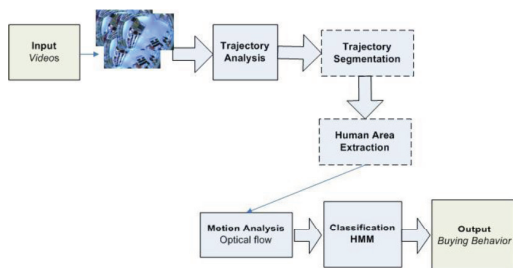


Fig. 1 System overview

B. Data acquisition

In order to test our system we used recordings in a shopping mall taken at different time intervals using a fish-eye camera attached to the ceiling.

An example of the acquired type of images is depicted in Fig. 2.



Fig. 2 Fish-eye camera image acquisition

We collected and manually annotated approximately 5 hours of recordings resulting in 270 customers' trajectories. Furthermore 100 trajectories contained buying segments while the rest were labeled as non-buying ones. We present next the experimental results obtained using the recorded data.

C. Experimental results

We performed a number of tests in order to find the best feature descriptor and HMM topology for our buying behavior analysis system as described in Section A.

A first experiment was based only on trajectory features using HMMs. We classified the trajectories set into buying and non-buying ones, using a 10-fold cross validation and we obtained an accuracy of 70%. Still this approach is not

discriminative enough, as it lacks information regarding customer's interaction with products.

Next we refined our analysis by employing motion analysis on the interesting trajectories segments. Normalized histograms of optical flow in 8 bins proved to be a better descriptor compared to average length and angle. In order to find the best HMM topology for the given problem we used an extensive set of states (1-10) and number of Gaussian Mixtures (1-10) as it is depicted in Fig. 3.

We found out that the best accuracy of 70% was obtained for 6 states and 2 mixtures topology.

HMM states	Accuracy
2	67.72%
3	67.73%
4	69.15%
5	69.18%
6	69.96%
7	68.41%
8	69.13%
9	68.07%
10	69.54%

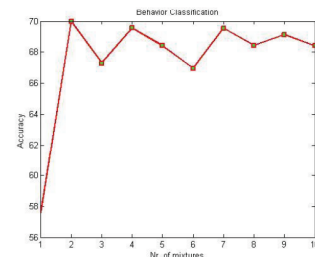


Fig. 3 Buying behavior classification results

III. CONCLUSIONS

We presented an approach towards understanding customers' shopping behavior applied to real-life recordings in a supermarket. We designed and implemented a first running prototype for detecting customers' buying behavior, achieving an accuracy of 70%.

As future work we plan to improve and refine the action recognition module by using different types of features such as interest-points models and an extended set of shopping related actions. We also aim at extending the system by fusing the data from cameras at different location and view angles.

ACKNOWLEDGMENT

This work was supported by the Netherlands Organization for Scientific Research (NWO) under Grant 018.003.017.

REFERENCES

- [1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893, San Diego, California, June 2005.
- [2] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. S. Huang, "Action detection in complex scenes with spatial and temporal ambiguities," in *Proceedings of International Conference on Computer Vision (ICCV '09)*, October 2009.
- [3] T. Kanda, D. F. Glas, M. Shiomi, H. Ishiguro, and N. Hagita, "Who will be the customer?: A social robot that anticipates people's behavior from their trajectories," *Int. Conference on Ubiquitous Computing*, 2008.
- [4] M. C. Popa, L. J. M. Rothkrantz, Z. Yang, P. Wiggers, R. Braspenning, and C. Shan, "Analysis of Shopping Behavior based on Surveillance System", *2010 IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC'10)*, Istanbul, Turkey, October 2010.
- [5] A. Valera and S.A. Velastin, "Intelligent distributed surveillance systems: A Review", in *IEEE Proceedings – Vision, Image, and Signal Processing*, 152(2), pp 192–204, April 2005.

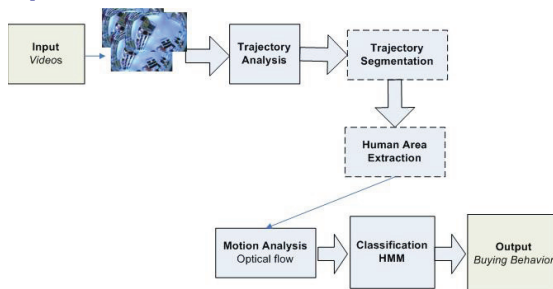
**M.C. Popa, L.J.M. Rothkrantz,
and P. Wiggers**
MMI, TU Delft
Delft, Netherlands

C. Shan and T. Gritti
Video and Image Processing,
Philips Research
Eindhoven, Netherlands

Project Goals

- o Investigate shopping behavior
- o Assess customers' interaction with products
- o Detect buying actions

System Overview



Recordings

In order to test our system we used real-life recordings in a supermarket, taken at different time intervals using a fish-eye camera attached to the ceiling.



We collected and manually annotated approximately 5 hours of recordings resulting in 270 customers' trajectories. Furthermore 100 trajectories contained buying segments while the rest were labeled as non-buying ones.

Trajectory estimation

The customers' trajectories are classified into buying and non-buying ones. Features such as first, second derivative, Euclidian distance, curvature, and spatial-temporal curvature are extracted and then used to train a HMM classifier. Using a 10-fold cross validation we obtained an accuracy of 70%. Still this approach is not discriminative enough, as it lacks information regarding customer's interaction with products.

Trajectory analysis

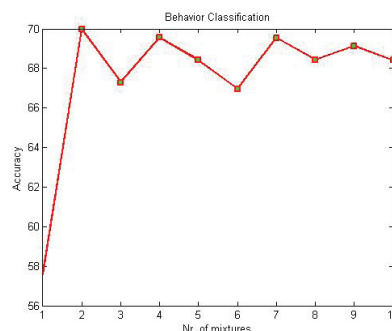
The trajectory analysis module detects segments of interest and includes potential buying segments. The analysis is done using relevant features such as speed, curvature, and duration of staying in a region.

Motion analysis

The motion analysis module is applied to each segment, by estimating optical flow between every two consecutive images in the corresponding human area. Histograms of the motion-related vectors are extracted and fed to a Hidden Markov Model classifier. The results obtained for different HMMs topologies are depicted in the table below.

HMM States	Accuracy
2	67.72%
3	67.73%
4	69.15%
5	69.18%
6	69.96%
7	68.41%
8	69.13%
9	68.07%
10	69.54%

In our experiments we employed different numbers of mixtures. Gaussian Mixture Models (GMM) proved to be better than simple Gaussian models and the optimum number of mixtures for the given problem can be seen in the graph below.



Detection of Human Groups in Videos

Selçuk Sandıkcı, Svitlana Zinger and Peter H.N. de With
Eindhoven University of Technology
P.O. Box 513, 5600 MB, Eindhoven, The Netherlands
Email: {s.sandikci, s.zinger, p.h.n.de.with}@tue.nl

Abstract—In this paper, we propose an approach for detecting and localizing social human groups in videos, which can form a basis for further analysis of groups in general. Our approach is motivated by the collective behavior of individuals which has a fundament in sociological studies. Human groups are discovered by hierarchically clustering trajectories of individuals. A novel similarity function robustly measures the similarity of noisy trajectories. This function is consistent with the typical distances in social group models. We design a detection-based multi-target tracking framework in order to extract trajectories of humans. We have evaluated our approach on several video sequences and achieved acceptable miss rates at reasonable false positive detections per frame.

I. INTRODUCTION

Automatic visual analysis of human groups has significant importance for security surveillance. Humans groups are needed to be detected and located before performing further analysis such as tracking and behavior recognition. In addition, detecting social human groups is beneficial for developing more realistic crowd models and multi-target tracking.

Human groups have been extensively studied by social science researchers. McPhail and Wohlstein [1] state that two people belong to the same group, if they are in close proximity and traveling in the same direction with similar velocities. In our work, human groups are detected by hierarchically clustering trajectories of individuals. We propose a robust trajectory similarity measure which fulfills these requirements to construct the group structure. Accordingly, proposed similarity metric is composed of positional, velocity and directional similarities of trajectories. We design a multi-target tracking system to extract trajectories, which combines human detection with mean-shift tracking [2]. We evaluate and compare our group detection method to earlier work, using video sequences with varying number of groups and difficulty.

II. DESCRIPTION OF GROUP DETECTION ALGORITHM

An overview of our algorithm is depicted in Fig. 1. A video sequence is the input to our algorithm, and detected human groups are the outputs. It includes three principal blocks: *Human Detection*, *Multi-target Tracking* and *Human Group Detection*.

A. Human Detection

We detect humans in each frame using multi-scale Deformable Part Models (DPM) human detector [3] in order to initialize, guide, and terminate individual person trackers. A geometric filtering is carried out to discard detections with inappropriate size.

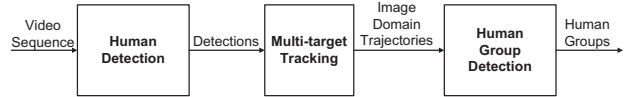


Fig. 1. Overview of our group detection algorithm.

B. Multi-target Tracking

In this block, we track all detected humans to extract image-domain trajectories. Trackers are automatically initialized using human detection results. We terminate a tracker if it is not supported by detections or the average tracking score for the past few frames is lower than a threshold.

We employ color-based mean-shift algorithm to track the targets, where the target appearance is modeled with an RGB color histogram. In order to increase tracking precision during occlusions, a simple inter-object occlusion reasoning method is utilized.

Detections are also used for guiding trackers. In each frame, new detections are matched to existing trackers, using color similarity and bounding box overlap. We linearly blend bounding boxes of matching detection-tracker pairs.

The Double-Exponential Smoothing filter is adopted to smooth and refine the position and the scale of trackers.

C. Human Group Detection

We first project image-domain trajectories onto the ground plane. Then, human groups are detected by hierarchically clustering ground-plane trajectories of individuals.

Our group detection algorithm operates in a sliding-window fashion. We estimate the trajectory similarity using the trajectory samples falling into the temporal interval centered around the current time. The trajectory similarity metric is composed of positional, velocity and directional similarities. The positional similarity of two trajectory sample sets is based on the time-averaged positional distance between them. Similarly, velocity similarity is estimated using difference of their average velocities. Finally, directional similarity is quantified by the inner product of their unit-length direction vectors. The overall trajectory similarity is obtained by multiplying the positional, velocity and directional similarities. However, direction estimation for stationary humans may have significant fluctuations due to tracking inaccuracies. Therefore, we discard the directional similarity for slowly moving humans.

A simple bottom-up clustering algorithm is employed to group trajectories using their pairwise similarities.

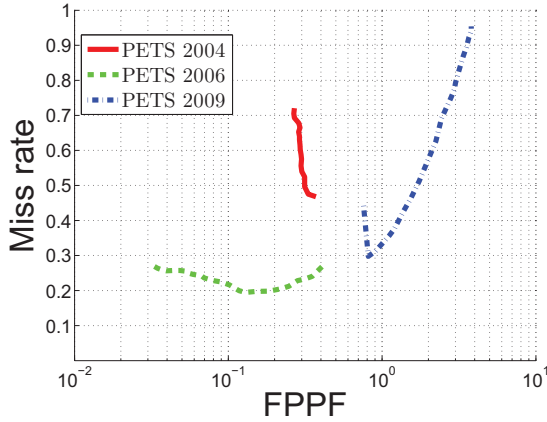


Fig. 2. Quantitative evaluation of our group detection algorithm for varying clustering threshold.

III. EXPERIMENTS AND RESULTS

We evaluate our approach with three sequences from PETS 2004, PETS 2006 and PETS 2009 datasets. Human groups are manually annotated by bounding boxes in each frame.

We detect human groups as described in Section II. A detected group is counted as a true detection only if the detection bounding box significantly overlaps with the ground truth bounding box. The overall performance of the algorithm is measured by evaluating the miss rate and the amount of False Positive detections Per Frame (FPPF). Quantitative evaluation of our approach is shown in Fig. 2. The high miss rate for the PETS 2004 sequence is due to noisy and fragmented trajectories of small-sized humans. In the PETS 2009 sequence, groups are over-divided into smaller groups with increasing clustering threshold, resulting in a steep increase in both the miss rate and FPPF. The strong variation in the curves indicates that the performance of our system depends heavily on the scene contents.

We measure the group detection performance using different trajectory inputs to assess the effect of the tracking. We experiment with trajectories produced by our tracking system with DPM and Histograms of Oriented Gradients (HOG) human detector, and Boosted Particle Filter [4] using DPM. The CLEAR MOT scores of the tested combinations are given in Table I. In Table II, we compare the group detection performance using tested tracking methods. Our DPM-based tracking achieves the lowest miss rate and FPPF value. High false negative and false positive rates of the HOG-based tracking are directly translated into a higher miss rate and FPPF value. A higher number of ID switches of the BPF-DPM combination results in a higher miss rate.

Lastly, we compare our human group detection algorithm with the method of Ge *et al.* [5], which we call hierarchical Clustering based on the symmetric Hausdorff Distance (CHD). CHD uses our DPM-based trajectories as its input. Comparison of group detection performance of our method with CHD is provided in Table III. Our method constantly achieves very

TABLE I
CLEAR MOT RESULTS OF OUR TRACKING SYSTEM USING DPM AND HOG DETECTOR AND BPF ON THE PETS 2006 S7-T6-B SEQUENCE.

Tracking Method	Prec.	Accur.	F. Neg.	F. Pos.	ID Sw.
Ours with DPM	79.6%	88.4%	7.2%	4.3%	7
Ours with HOG	75.0%	64.7%	14.2%	21.0%	8
BPF with DPM	78.5%	88.7%	7.0%	4.2%	21

TABLE II
GROUP DETECTION RESULTS OF DIFFERENT TRACKING METHODS ON THE PETS 2006 SEQUENCE.

Tracking Method	Miss rate	FPPF
Ours with DPM	19.4%	0.129
Ours with HOG	36.3%	0.261
BPF with DPM	30.5%	0.130

TABLE III
COMPARISON OF OUR HUMAN GROUP DETECTION METHOD WITH GROUP DETECTION BASED ON THE HAUSDORFF DISTANCE (CHD).

Dataset	Miss rate	FPPF
PETS 2004 (Ours)	46.7%	0.371
PETS 2004 (CHD)	66.3%	1.106
PETS 2006 (Ours)	19.4%	0.129
PETS 2006 (CHD)	19.1%	0.620
PETS 2009 (Ours)	29.7%	0.813
PETS 2009 (CHD)	26.6%	0.811

similar or lower miss rates and FPPF values compared to the CHD, proving the robustness of our trajectory similarity metric.

IV. CONCLUSION

This paper describes a fully automatic approach for detecting and localizing human groups in videos. The framework can serve as a prior step for high-level group analysis. Our proposed algorithm involves tracking of individuals and grouping their trajectories based on their similarity. The trajectory similarity is estimated by combining positional, velocity and directional similarities. The choice for these features is motivated by sociological studies and appears to be robust for group detection. We have found that extracting accurate and continuous trajectories is the key point for successful detection of human groups. We achieve reasonable miss rates (19.4%, 29.7% and 46.7%) at acceptable false positive detections per frame (FPPF) for challenging video sequences.

REFERENCES

- [1] C. McPhail and R. Wohlstein. Using film to analyze pedestrian behavior. *Sociological Methods and Research*, 10(3): 347-375, 1982.
- [2] D. Comaniciu, V. Ramesh and P. Meer. Kernel-based object tracking. *IEEE T. Pattern Anal. and Machine Intell.*, 35(5):564-575, 2003.
- [3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE T. Pattern Anal. and Machine Intell.*, 32(9): 1627-1645, 2010.
- [4] K. Okuma, A. Taleghani, N. d. Freitas, J. J. Little and D. G. Lowe. A boosted particle filter: multitarget detection and tracking. In *Eur. Conf. Comp. Vision*, 2004.
- [5] W. Ge, R. T. Collins and B. Ruback. Automatically detecting the small group structure of a crowd. In *IEEE Workshop on Applications of Comp. Vis.*, 2009.

Detection of Human Groups in Videos

S. Sandikci, S. Zinger, and P.H.N. de With



1. Introduction

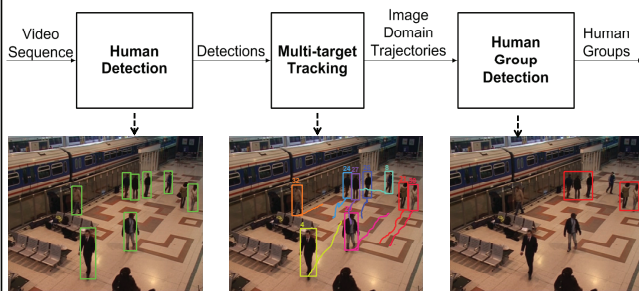
Detection and localization of human groups in videos

- Forms a basis for further group analysis, such as tracking and behavior recognition
- Beneficial for developing realistic crowd models and for multi-target tracking

Our approach is motivated by social group definition of McPhail and Wohlstein (1982): “Group members should be in **close proximity** and should travel in the **same direction** with **similar velocities**”

2. Group Detection Algorithm

Hierarchical clustering of ground-plane trajectories of humans for group detection.



Human Detection:

- Multi-scale Deformable Part Models (DPM) human detector (Felzenszwalb *et al.*, 2010)

Multi-target Tracking:

- Human detection for tracker initialization and termination
- Color-based mean-shift algorithm with simple occlusion handling for tracking humans
- Linear blending of detection results with mean-shift output
- Double exponential smoothing filter for motion prediction and smoothing

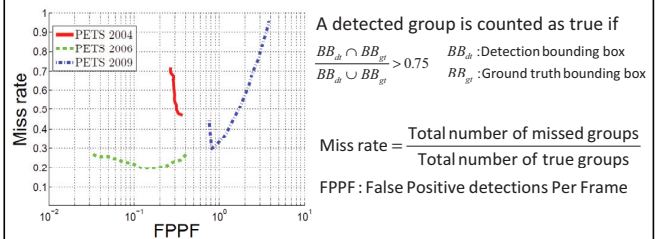
Human Group Detection:

- Mapping of trajectories onto ground-plane to be independent of projective scaling
- A novel trajectory similarity as a combination of **positional**, **velocity** and **directional** similarities, motivated by group definition of McPhail and Wohlstein (1982)
- Group detection by hierarchically clustering trajectories based on their similarity

3. Experiments and Results

Test Set: PETS 2004, PETS 2006 and PETS 2009 datasets

Experiment 1: Group Detection Performance on test sequences



Experiment 2: Influence of Tracking on Group Detection

Performance

CLEAR MOT scores of different tracking methods on PETS 2006						Corresponding Group Detection Performance	
Tracking Method	Prec.	Accur.	F. Neg.	F. Pos.	ID Sw.	Miss rate	FPPF
Ours with DPM	79.6%	88.4%	7.2%	4.3%	7	19.4%	0.129
Ours with HOG	75.0%	64.7%	14.2%	21.0%	8	36.3%	0.261
BPF with DPM	78.5%	88.7%	7.0%	4.2%	21	30.5%	0.130

Experiment 3: Comparison of our human group detection with group detection based on the Hausdorff Distance (CHD) (Ge *et al.*, 2009)

Dataset	Our method		CHD	
	Miss rate	FPPF	Miss rate	FPPF
PETS 2004	46.7%	0.371	66.3%	1.106
PETS 2006	19.4%	0.129	19.1%	0.620
PETS 2009	29.7%	0.813	26.6%	0.811

4. Conclusion

- A fully automatic approach is proposed for detecting and localizing human groups in videos, motivated by sociological studies.
- Our approach is based on hierarchically clustering trajectories of individuals which are similar in position, velocity and direction.
- Extracting accurate and continuous trajectories is the key point for successful group detection.
- Accuracy and stability of multi-target tracking part is planned to be improved by
 - Considering social interactions in the motion model and using more robust appearance models in mean-shift tracking.

Towards Demographic Classification in Unconstrained Environments

Caifeng Shan

Philips Research, High Tech Campus 36
Eindhoven 5656 AE, The Netherlands
Email: caifeng.shan@philips.com

I. INTRODUCTION

Demographic classification, including gender classification and age estimation, has many important applications, for example, visual surveillance, marketing intelligence, intelligent user interface, smart environment, etc. Human faces provide important visual information for gender and age perception, so demographic classification from face images has received much research interest in the last two decades [1], [2].

The face demographic classification system typically consists of two key components: face image representation and classifier design. Different approaches have been exploited for this task. Moghaddam and Yang [1] used raw image pixels with nonlinear Support Vector Machines (SVMs) for gender classification on thumbnail faces; their approach achieves the accuracy of 96.6% on the FERET database. Baluja and Rowley [3] introduced an efficient gender recognition system by boosting pixel comparisons in face images. On the FERET database, their approach matches SVM with 500 comparison operations on images of 20×20 pixels. For age estimation, different facial representations, such as anthropometric models, active appearance models, and age subspace or manifold, have been considered [2]. Regarding estimation techniques, age estimation can be solved as a (multi-class) classification problem or a regression problem [4], and difference classifiers and regression techniques can be adopted [2].

A common problem of the existing studies is face images acquired under controlled conditions are considered, which usually are high-quality frontal faces, occlusion-free, with clean background and limited facial expressions. However, in real-world applications, demographic classification needs to be performed on face images captured in unconstrained environments (see Figure 1 for some examples). As can be expected, there are significant appearance variations in the real-life faces, which include facial expressions, illumination changes, head pose variations, occlusion or make-up, poor image quality, and so on. Therefore, demographic classification in unconstrained environments is much more challenging. However, few studies in the literature have addressed this problem. In this study, we investigate demographic classification on real-life faces.

II. OUR STUDY

A. Our Approach

As an efficient non-parametric method summarizing the local structure of an image, Local Binary Patterns (LBP) has

been widely used for face image analysis [5]. The LBP operator labels the image pixels by thresholding a neighborhood of each pixel with the center value and considering the results as a binary number. The histogram of LBP labels computed over a image region can be used as a texture descriptor. When deriving LBP-based face representation, face images are divided into non-overlapping sub-regions, and the LBP histograms extracted from each sub-region are concatenated into a feature histogram.

In the existing work, LBP histograms are extracted from local facial regions as the region-level description, where the n -bin histogram is utilized as a whole. However, not all bins in the LBP histogram are necessary to contain useful information. Here we adopt Adaboost to learn the discriminative LBP-Histogram (LBPH) bins for demographic classification. Adaboost [6] learns a small number of weak classifiers whose performance is just better than random guessing, and boosts them iteratively into a strong classifier of higher accuracy. Similar to [6], a weak classifier $h_j(x)$ consists of a feature f_j which corresponds to the individual LBPH bin, a threshold θ_j and a parity p_j indicating the direction of the inequality sign:

$$h_j(x) = \begin{cases} 1 & \text{if } p_j f_j(x) \leq p_j \theta_j \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

With the selected features, we adopt SVM for demographic classification, which has been an effective classifier in the existing studies.

B. Data Sets

Dataset 1 — The database collected in [7] consists of 28,231 faces from 5,080 Flickr images. 86% of the faces were detected by face detector. Each face was labeled with the gender and age category, and seven age categories were considered: 0-2, 3-7, 8-12, 13-19, 20-36, 37-65, and 66+. Many faces in the dataset have low resolution. Some example faces are shown in Figure 1.

Dataset 2 — The LFW database [8] contains 13,233 faces of 5,749 subjects collected from the web. All the faces were detected by face detector. We manually labeled the ground truth regarding gender for each face. We chose 7,443 face images (2,943 females and 4,500 males) for gender classification experiments.

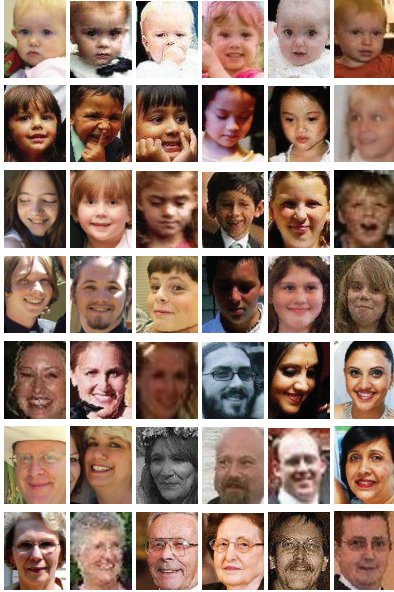


Fig. 1. Examples faces captured in unconstrained environments [7].

Approach	Gender	Age
Appearance [7]	69.6%	38.3%
Appearance + Context [7]	74.1%	42.9%
Gabor + Adaboost	70.2%	43.7%
LBP + Adaboost	71.0%	44.9%
boosted Gabor + SVM	73.3%	48.4%
boosted LBP + SVM	74.9%	50.3%

TABLE I
DEMOGRAPHIC CLASSIFICATION ON THE IMAGE SET USED IN [7].

C. Experiments

Dataset 1 — We first conducted experiments using the same training/testing sets as [7], and show recognition results in Table I. Gabor wavelet features are also compared. It can be seen that our approaches provide better results than the methods in [7], especially for age estimation. It seems that boosted LBP features combined with SVM deliver the best performance for demographic classification.

Considering many faces in the dataset have very low resolution, we also perform experiments after removing faces of very low resolution, i.e., faces with the eye distance less than 24 pixels. The results in Table II show evident improvements compared to Table I; however, the accuracy for age estimation is still around 56%. We show in Table III the confusion matrix of age estimation. It is seen that the age categories 0-2 and 66+ can be better estimated. But, for all other age categories, the accuracies are below 50.0%.

Dataset 2 — As a baseline to compare against, we also applied SVM with raw image pixels, which delivers the best performance on gender classification with faces in controlled environments [1]. The results summarized in Table IV show

Approach	Gender	Age
Gabor + Adaboost	72.5%	48.2%
LBP + Adaboost	73.6%	48.3%
boosted Gabor + SVM	75.7%	52.6%
boosted LBP + SVM	77.4%	55.9%

TABLE II
RESULTS ON THE NEW IMAGE SETS.

	0-2	3-7	8-12	13-19	20-36	37-65	66+
0-2	88.0%	8.0%	2.0%	0	0	1.0%	1.0%
3-7	31.0%	46.0%	16.0%	3.0%	2.0%	2.0%	0
8-12	1.6%	15.6%	37.5%	29.7%	3.1%	6.3%	6.3%
13-19	0	8.0%	18.0%	45.0%	16.0%	12.0%	1.0%
20-36	1.0%	3.0%	0	15.0%	48.0%	23.0%	10.0%
37-65	0	0	2.0%	15.0%	16.0%	41.0%	26.0%
66+	0	1.0%	0	3.0%	4.0%	13.0%	79.0%

TABLE III
CONFUSION MATRIX OF AGE ESTIMATION.

that the boosted LBP features combined with SVM achieve the best performance of 94.44%.

III. CONCLUSION

We investigate demographic classification on faces acquired in unconstrained environments. Our experiments demonstrate that age estimation is a difficult problem.

ACKNOWLEDGMENT

This work was supported by the Visual Context Modeling (ViCoMo) project.

REFERENCES

- [1] B. Moghaddam and M. Yang, "Learning gender with support faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.
- [2] Y. Fu, G. Guo, and T. Huang, "Age synthesis and estimation via faces: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.
- [3] S. Baluja and H. A. Rowley, "Boosting set identification performance," *International Journal of Computer Vision*, 2007.
- [4] G. Guo, Y. Fu, C. Dyer, and T. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Transactions on Image Processing*, 2008.
- [5] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *European Conference on Computer Vision (ECCV)*, 2004, pp. 469–481.
- [6] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001, pp. 511–518.
- [7] A. Gallagher and T. Chen, "Understanding images of groups of people," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 256–263.
- [8] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

Approach			Recognition Rates (%)		
Feature	Dim.	Classifier	Female	Male	Overall
raw pixels	2,944	SVM	86.89	94.13	91.27±1.67
boosted LBP	500	Adaboost	91.13	94.82	93.36±1.49
boosted LBP	500	SVM	91.91	96.09	94.44±1.19

TABLE IV
EXPERIMENTAL RESULTS OF GENDER CLASSIFICATION.

Towards Demographic Classification in Unconstrained Environments

Caifeng Shan

Philips Research, High Tech Campus 36, 5656AE Eindhoven, The Netherlands

caifeng.shan@philips.com

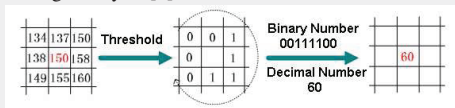
1 Introduction

- Demographic classification, including gender classification and age estimation, has many important applications. Since human faces provide important visual information, demographic classification from face images has received much research interest.
- A common problem of the existing studies is face images acquired under controlled conditions are considered. However, in real-world applications, demographic classification needs to be performed on face images captured in unconstrained environments. Few studies have addressed this problem.

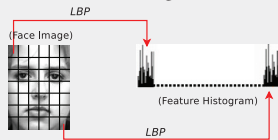
2 Our Approach

Face Image Representation

- As an efficient non-parametric method summarizing the local structure of an image, Local Binary Patterns (LBP) has been widely used for face image analysis [1].



- Face images are normally divided into non-overlapping sub-regions, and the LBP histograms extracted from each sub-region are concatenated into a feature histogram.



- Considering not all bins in the LBP histogram are necessary to be informative, we adopt Adaboost to learn the discriminative LBP-Histogram bins for demographic classification.

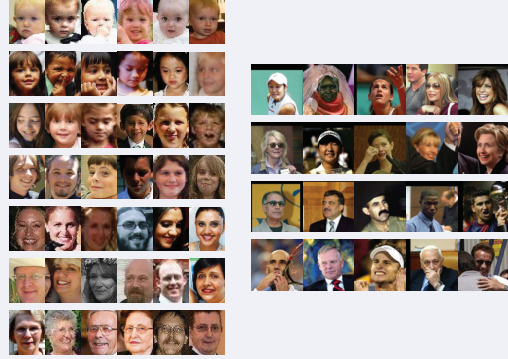
Classifier

- Support Vector Machine (SVM) is adopted as classifier.

3 Data Sets

- Dataset 1** [2] consists of 28,231 faces from 5,080 Flickr images. 86% of the faces were detected by face detector. Each face was labeled with the gender and age category, and seven age categories were considered: 0-2, 3-7, 8-12, 13-19, 20-36, 37-65, and 66+. Many faces have low resolution.
- Dataset 2** (LFW database) [3] contains 13,233 faces of 5,749 subjects collected from the web. All the faces were detected by face detector. We manually labeled the ground truth regarding gender. We chose 7,443 faces (2,943 females and 4,500 males) for gender classification.

4 Experiments



Examples faces captured in unconstrained environments.

- On Dataset 1, we first conduct experiments using the same data as [2], and then perform experiments after removing faces of very low resolution.

Approach	Gender	Age
Appearance [2]	69.6%	38.3% 71.3%
Appearance + Context [2]	74.1%	42.9% 78.1%
Gabor + Adaboost	70.2%	43.7% 80.7%
LBP + Adaboost	71.0%	44.9% 83.0%
boosted Gabor + SVM	73.3%	48.4% 84.4%
boosted LBP + SVM	74.9%	50.3% 87.1%

Demographic classification on the image set used in [2].

Approach	Gender	Age
Gabor + Adaboost	72.5%	48.2% 80.3%
LBP + Adaboost	73.6%	48.3% 83.3%
boosted Gabor + SVM	75.7%	52.6% 83.3%
boosted LBP + SVM	77.4%	55.9% 87.7%

Demographic classification on the new image set.

	0-2	3-7	8-12	13-19	20-36	37-65	66+
0-2	88.0%	8.0%	2.0%	0	0	1.0%	1.0%
3-7	31.0%	46.0%	16.0%	3.0%	2.0%	2.0%	0
8-12	1.6%	15.6%	37.5%	29.7%	3.1%	6.3%	6.3%
13-19	0	8.0%	18.0%	45.0%	16.0%	12.0%	1.0%
20-36	1.0%	3.0%	0	15.0%	48.0%	23.0%	10.0%
37-65	0	0	2.0%	15.0%	16.0%	41.0%	26.0%
66+	0	1.0%	0	3.0%	4.0%	13.0%	79.0%

Confusion matrix of age estimation on the new image set.

- We observe

- Our approaches provide better results than the methods in [2].
- Although performance improved by removing low-resolution faces, the accuracy for age estimation is still around 56%.
- The boosted LBP features combined with SVM always deliver the best performance.
- The age categories 0-2 and 66+ can be better estimated. But, for all other age categories, the accuracies are below 50.0%.

- On Dataset 2, we obtain the accuracy of 94.44% for gender classification.

Approach			Recognition Rates (%)		
Feature	Dim.	Classifier	Female	Male	Overall
raw pixels	2,944	SVM	86.89	94.13	91.27±1.67
boosted LBP	500	Adaboost	91.13	94.82	93.36±1.49
boosted LBP	500	SVM	91.91	96.09	94.44±1.19

Gender classification on Dataset 2.

5 Conclusions

Compared to gender classification (94% on LFW faces, 77% on Flickr faces), age estimation is much more challenging (56% for 7-category estimation on Flickr faces).

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *European Conference on Computer Vision (ECCV)*, pages 469–481, 2004.

[2] A. Gallagher and T. Chen. "Understanding images of groups of people," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 256–263.

[3] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

PHILIPS

Vital Signs Camera

Ingmar van Dijk, Adrienne Heinrich
Philips Research Laboratories, Eindhoven, The Netherlands

Abstract—Measuring the heart rate and respiration is extremely important for healthcare and well-being monitoring. However, current methods are obtrusive and not always reliable. The Vital Signs Camera technology allows monitoring of heart rate and respiration rate remotely using existing video cameras. The Vital Signs camera can be used to monitor heart activity of recovering of high risk cardiac patients in a simple and unobtrusive way in the home. Keeping an eye on sleeping or unattended babies is another possibility.

I. INTRODUCTION

Respiration rate and heartbeat are two prominent vital signs, which provide an early indication of the health condition. Currently, in order to monitor vital signs of a person for lifestyle or health care applications, contact sensors are used. Those sensors are obtrusive, not convenient to wear for a long time, and are not robust to motion. Therefore, we have been developing the technology for monitoring of respiration and heart beat in a non-contact way using a video camera. First, and most importantly, the Vital Signs Camera technology can operate at a distance, measuring subjects without their having to strap on wires or equipment. In early user testing at the hospital in the Netherlands, groups of nurses gave enthusiastic feedback and identified a number of important uses for contact-free technology on general hospital wards. It means, for example, that the same device can be used to perform spot-checks on multiple patients without the risk of spreading infections. It also means that patients need not be disturbed when measurements are taken, and the use of infrared even allows for continued monitoring in darkness or shallow light conditions.

II. HEART BEAT MONITORING

One of the most exciting health care applications for the Vital Signs technology is in neonatal care. Premature infants have extremely fragile skin, and are sometimes so small that they fit in the palm of your hand. Even if there were sufficient space to place a sensor on their bodies, doctors prefer not to risk damaging them. In the future, just one contact-free sensor from Philips - perhaps mounted on an incubator or crib - will monitor a range of vital signs in these most vulnerable patients. The technology for monitoring of heart beat is based on the recent discovery that the physiological phenomenon behind traditional h (Photoplethysmography) sensors can be measured from a distance. Skin reflects part of the light shining on it, the rest is absorbed. Absorption is influenced by heart beats: blood absorbs differently from the bloodless skin. Since the heart beat is present in the visible spectrum, even today's web-cams have sufficient color-depth

to extract heart beats from the temporal signal. Moreover, to monitor a heart beat in the dark environment (e.g., during the night) the technology has been investigated to use a near IR light a monochrome camera sensitive to a IR light spectrum. Heartbeat can be measured from any visible skin, for example from the face. One of the possible applications of the vital signs camera technology is unobtrusive monitoring of babies in home or hospital environment. To verify the feasibility of camera-based monitoring of heart rate of a baby, several recordings in ambient daylight and artificial lighting conditions have been made using a colour camera, and in a dark environment using a monochrome camera and IR light source. In the absence of a reference heartbeat signal, the time of robust measurement of a heartbeat signal was measured by defining the deviation range of 20% around a mean HR value. Based on this approach, the coverage ratio of heart rate monitoring is around 80 %. Based on results of measurements, a conclusion can be made that a stable heartbeat signal is detectable in any lighting conditions if the skin of the baby is not fully occluded due to the motion of the baby. Rotations of the head can cause sudden deteriorations of the heartbeat signal. However, shortly after rotation the signal can be recovered again. A much more serious problem for robust continuous monitoring of the heartbeat signal is caused by active motion of the baby, because it results in occlusions of the face skin area (selected for sensing) by moving hands and legs. In such periods, the signal is lost completely. The camera can still register temporal color variations, but they are caused by motion, not by the heartbeat. The measured signal was also lost when the baby was crying, the open mouth of the baby imposes sudden changes in the temporal behaviour of the skin color, sensed from baby's face. Usually, a couple of seconds is required to recover the heartbeat signal after its loss. To verify the robustness of the measurements in periods of sleep of the baby, we selected two non-overlapping regions of the skin for sensing (face and hand), and we measured heartbeat signals from those regions independently. The results of such measurement show that signals sensed from the hand and from the face have a slight phase difference, but provide exactly the same heart rate numbers. No further investigation of correlation of signals sensed from different skin areas has been made. Analysis of video sequences made when the baby was in deep sleep showed that heart rate values are fluctuating within 10 % from a mean value, calculated over the window of 400s or 600s. Similar results are obtained for periods of shallow sleep of the baby, except for motion moments of the baby.

In general, remote monitoring of heart beat signal from a skin using a Vital Signs Camera technology is possible if a subject is not moving and lighting conditions are stable.

III. RESPIRATION MONITORING

Monitoring of respiration using a Vital Signs Camera technology is based on measurement of respiratory motion of a chest or a belly of a person. Robustness of monitoring has been achieved by means of projecting a light pattern on a chest and belly of a person. The vital signs camera derive breathing signals from an analysis of spatial changes of the projected light pattern.

We have applied our method using projected light in daylight to extract the breathing signals. The results are shown in Figure 1, where the chest expansion is plotted over time. Clearly the signal obtained from our method (black) corresponds to the frequency of the reference signal acquired from the chest belt (red). The blue and the green dots correspond to transitions between the inhalation and exhalation phase and between the exhalation and inhalation phase (expiratory pause) respectively. Again a very good match between the reference signal and our method was found. However, there is room for improvement regarding the breathing wave form. Further, future work should focus on improving the breathing waveform. The camera-based algorithm for respiration monitoring has been tested for sleep monitoring applications. Based on the result of experiments, we can conclude that the high respiration rate correspondences between the video-based respiration rates and the reference signals were achieved with an overall performance of 96% and 95% for the abdomen and chest belts respectively, while 'only' a 97% agreement between the two reference signals was obtained. A solid performance of the video-based respiration rate monitoring system throughout all test subjects and challenges such as thick blankets and various body positions is observed. The algorithm underperforms in estimating the correct peak locations. The correspondence between the reference signals and the processed video signal reaches 65%. The algorithm for unobtrusive monitoring of respiration has been tested also for monitoring of breathing of a baby. When the respiration of a 3-month old baby was monitored, we found breathing rates when the baby was performing no or small body part movements in the expected range between 23/min and 31/min. When the baby was monitored different episodes where captured, from regular breathing and no body movements, to irregular breathing with incidental very deep breaths and small, medium, large body movements with and without crying. From the analyzed data, we believe it is possible to discriminate between deep and shallow sleep of a baby and between the different types of motions in order to alert the parents moments before their baby starts crying.

IV. CONCLUSION

The current technology for unobtrusive monitoring of vital signs allow measurement of heart beat and respiration signals remotely. Currently, robust measurement of heart rate and

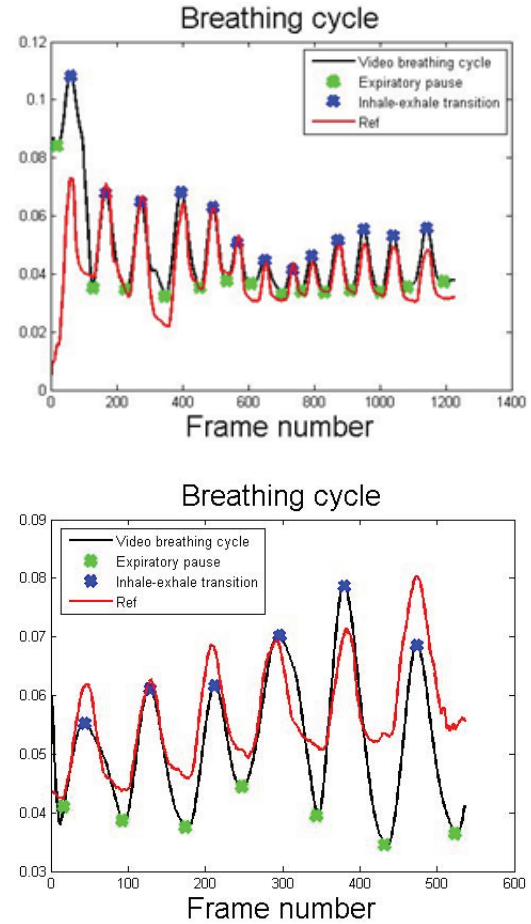


Figure 1. Breathing signals extracted by our camera-based technology (black) and by the chest belt (red) for the front (top) and back (bottom) side recordings.

breathing is possible if a person is not moving and an ambient illumination is static.

Motivation

Customer Benefits

- Enable easy to use & aesthetic relaxing biofeedback and wellness monitoring solutions
- Provide unobtrusive heart rate monitoring for sleep monitoring and cardio-fitness coaching
- Enable true presence detection for advanced user interaction
- Simultaneous unobtrusive monitoring of heart rate, respiration and physical activity

Unique advantage

- Unobtrusive monitoring of several subjects simultaneously at a distance up to 100m.
- Single sensor to detect, recognize persons and measure their physiological conditions.

Solution

General

- (Web)cameras can now measure vital signs like heart rate, breathing, and activity, through advanced motion and color analysis
- Blood volume variations in the skin caused by heart beats can be detected by measuring temporal changes of light reflection from skin
- Physical movements caused by respiration can be detected and measured remotely using a video camera and projected light pattern

Hardware

- Vital Signs Camera uses existing low cost hardware components

Software

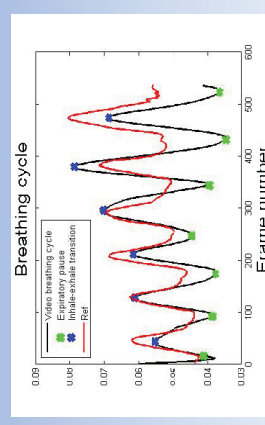
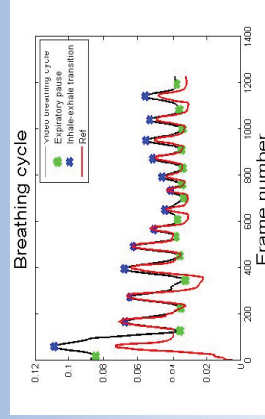
- Software of the algorithms is optimized to run real-time on PC platform

Application areas



Results

- Breathing signals extracted by our camera-based technology (black) and by the chest belt (red) for the front and back side video recordings



Conclusion & future work

- Techniques for unobtrusive monitoring of respiration & heart beat signals have been investigated
- The technology has been tested in lifestyle and healthcare applications
- Future work aimed at further improving performance and robustness

The authors would like to thank Erasmus MC and Philips Digital Pathology Venture for their support. This work was funded by the iCare project and will continue in Cytron-IL.

Subtitle Detection for TV Applications

Bahman Zafarifar, Jingyue Cao, and Peter H. N. de With, IEEE Fellow

Abstract —A method is presented for localization and classification of subtitles in TV videos, based on static-region detection, object-based adaptive temporal filtering and subtitle bounding box classification using text-stroke alignment features¹.

I. INTRODUCTION

Static regions overlaid on video, such as subtitles and Close Captions, require special attention in modern TV signal processing. In advanced Motion-Compensated Picture Rate Conversion (MC-PRC) [1][2], boundaries of such static text overlays are treated specifically, through occlusion detection, separate motion vector determination and pixel interpolation strategy for covering/uncovering occlusion areas, and an additional mechanisms for detection and protection of static (text) overlays by employing a *static-region detector*.

We present a subtitle detection algorithm that builds upon the output of the static region detector of an existing TV video processing chip [3], elevating it from a simple pixel-based still/non-still decision to an *object-based* subtitle/non-subtitle decision. Whereas the existing static region detector generates many false positives, if used to detect subtitles, the proposed system reduces these false positives significantly, at the cost of a slight reduction in true positives.

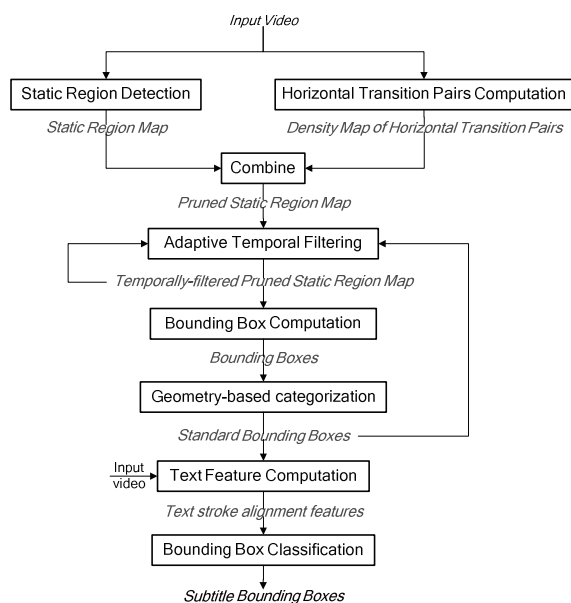


Fig. 1 Overview of the subtitle detection algorithm.

The algorithm provides two levels of information: (1) a set of bounding boxes around blobs of static regions, categorized by their geometry and filling degree of static regions, and (2) bounding-box level subtitle classification, using text-stroke alignment features, which judges whether a bounding box indeed contains subtitle text.

II. ALGORITHM AND RESULTS

The proposed algorithm [4] (Fig. 1) first prunes static regions that are not likely to be part of subtitles, by verifying the density of sharp horizontal luminance transitions. The result of the pruning is that subtitle lines are separated from each other and from the background (Fig. 2). An adaptive temporal filter improves the detection stability by suppressing random variations of the pruned static regions within candidate subtitle bounding boxes, using object-level feedback of candidate bounding box locations to the pixel-level filtering operation. The filter avoids temporal filtering outside these bounding boxes, which offers instantaneous response to (dis)appearing subtitles. Bounding box locations around blobs of static regions are computed using iterative 1D projections. Each bounding box is then categorized into *standard* and *non-standard*, using geometry and filling-degree of the temporally filtered pruned static regions as constraints. Finally, for each *standard* bounding box a pair of horizontal and vertical text-stroke alignment features (Fig. 3) are computed and use to carry out a binary classification of the bounding box into subtitle or non-subtitle classes, based on a trained 2D Gaussian model (Fig. 4).

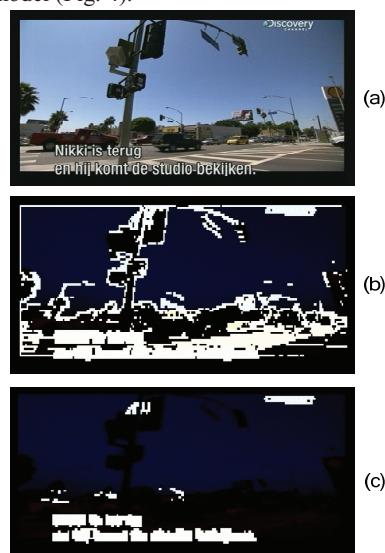


Fig. 2 Pruning the static region detection. (a) Input image. (b) Static region map: nearly all stationary strong edges are detected. (c) Pruned static region map: subtitle lines are separated from the background.

¹ This work has been supported by the NWO Casimir program under grant number 018.002.013, NXP Semiconductors and Trident Microsystems.

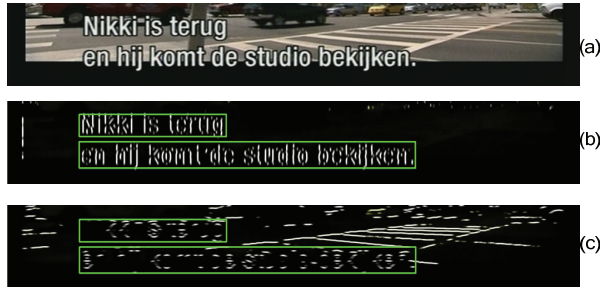


Fig. 3. Computing text-stroke alignment features. (a) Input. (b) Map of vertically aligned strokes. (c) Map of horizontally aligned strokes.

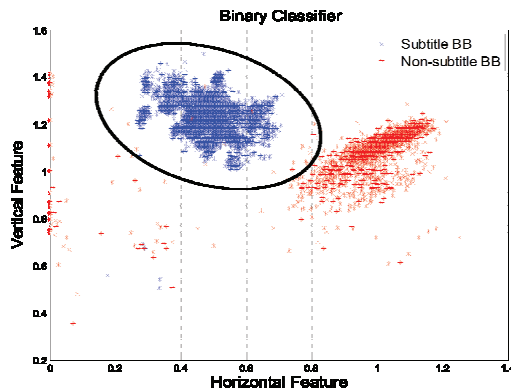


Fig. 4. Distribution of the text-stroke alignment features of the training set. The black ellipse shows the threshold level used for classification to the subtitle class.

TABLE 1—PIXEL-LEVEL EVALUATION RESULTS

	Precision = TP/(TP+FP)	Recall (TPR) = TP/(TP+FN)	FPR = FP/(FP+TN)
Static Region Detector	56%	100.0%	2.24%
Categorized as standard BB	83%	98.5%	0.59%
Classified as subtitle BB	96%	98.0%	0.11%

A test on 5000 video frames (Table 1) showed that the algorithm increases the static region detector's Precision rate from 56% to 96%, which means that among the detected pixels 96% are actually subtitle pixels, versus a slight loss of 2% (100%-98%) in Recall rate (higher erroneous rejection of subtitle pixels).

The proposed system has interesting properties for TV applications in that it forms a causal system where no access to future frames is required, and that it is designed to have an instantaneous response to (dis)appearing subtitles. The majority of operations are suitable for execution on streamed video without the need for large data buffers.

III. POTENTIAL APPLICATION

The proposed subtitle detection system may enable higher levels of TV picture quality enhancement, positioning subtitles at a desired depth in a 3D TV use case, keeping subtitles at a fixed position in image stabilization use case, etc.

We have conducted an experiment to verify the potential for reducing artifacts caused by Motion-Compensated Picture

Rate Conversion (MC-PRC) in the surrounding areas of subtitles. In Section I, we mentioned that static region detection is used to protect static overlays against artifacts. However, the motion vectors in the surrounding of subtitles will still remain disrupted (see Fig. 5(c)), leading to artifacts in these surrounding regions.

The experiment verifies whether exclusion of subtitles in motion estimation (ME) could lead to a more consistent motion vector field in the surrounding areas of subtitles, and thus fewer artifacts. To enable this without modifying the existing ME process, we replaced the static pixels within computed subtitle bounding boxes with the average value of neighboring pixels (in-painting), of which the result is shown in Fig. 5 (b), and performed the MC-PRC on the in-painted sequence. The subtitle pixels are then superimposed on the up-converted sequence. Fig. 5 (d) visualizes the improved motion vector consistency in the surrounding areas of subtitles, as compared to the original motion vectors shown in Fig. 5 (c). Fig. 5 (f) and (g) show the reduction of artifacts as a result of this process, which indicates the potential of the proposed system to enable higher picture quality in MC-PRC.



Fig. 5. Improving picture quality in MC-PRC process. (a) Input. (b) Filled subtitles. (c) MC-PRC on original: disrupted motion vectors in areas surrounding subtitles. (d) Improved motion vectors in these surroundings after performing MC-PRC on the filled version (subtitles are superimposed on the result of MC-PRC). (e), (f), (g) Zoomed versions of (a), (c), (d), resp. Artifacts in (f), shown by the dotted ellipse, are reduced in (g).

REFERENCES

- [1] E.B. Bellers, J. W. van Gorp, J. G. W. M. Janssen, R. Braspenning, R. Wittebrood, "Solving occlusion in Frame-Rate up-Conversion," *Consumer Electronics, IEEE Int. Conf. on, ICCE '07*, pp. 1-2, 2007.
- [2] M. Mertens, and G. de Haan, "Motion Vector field improvement for picture rate conversion with reduced Halo," *Proc. of the SPIE/IST VCIP*, pp. 352-362, 2001.
- [3] PNX5100 video back-end processor, document 10012C, Sept. 2010, <http://www.tridentmicro.com>.
- [4] Bahman Zafarifar, Jingyue Cao, Peter H. N. de With, "Instantaneously Responsive Subtitle Localization and Classification for TV Applications", in *IEEE Transactions in Consumer Electronics*, 2011.

Instantaneously Responsive Subtitle Localization & Classification for TV Applications

Bahman Zafarifar, Jingyue Cao, Peter H. N. de With

P13/5



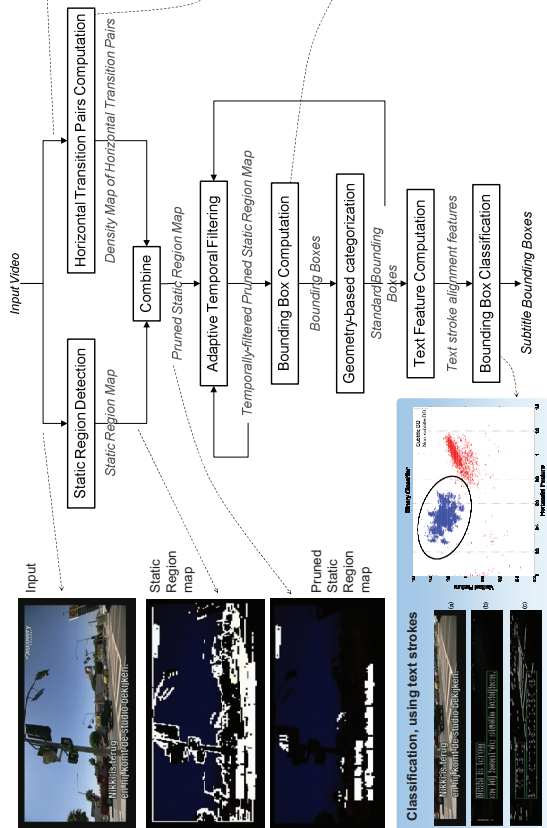
TU/e
technische universiteit
 eindhoven
Casimir grant

Introduction

- Special *static-regions detector* in TVs protects static overlays against artifacts of MC-Frame Rate Conversion.
- Assuming an existing *static-region detector*, we detect subtitle bounding boxes, providing:
 - A set of bounding boxes around blobs of static regions.
 - Bounding-box classification to text / non-text using text stroke alignment features.
- Properties:
 - Instantaneous response (no frame delay).
 - Suited for streamed video processing (low buffering)

Algorithm

1. **Pruning:**
 - Identify steep horizontal luminance transitions.
 - Accept only static regions that have a high density of transition pairs.
2. **Adaptive Temporal Filtering** of Static Regions
 - Subtitles are static. Use this to improve stability.
 - Feed-back bounding box location to temporal filter:
 - Within bounding box: Strong filtering
 - Outside bounding box: No filtering
 - **Instantaneous response** to (dis)appearing subtitles.
3. **Bounding Box Computation**
 - Initial iteration, using fixed thresholds on 1D projections.
 - Refinement iterations using adaptive thresholds.
 - Categorize bounding boxes according to height and filling degree of static regions to *standard* / *non-standard*.
4. **Text Classification**
 - Compute horizontal and vertical text-stroke alignment features.
 - Classify bounding box to *subtitle* / *non-subtitle* using text-stroke features.

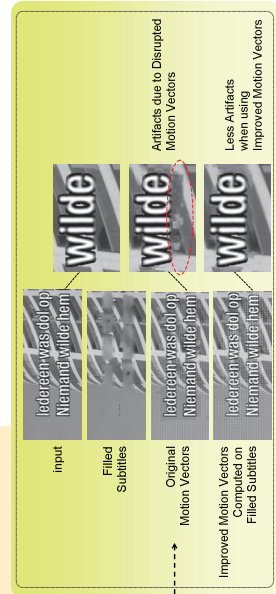
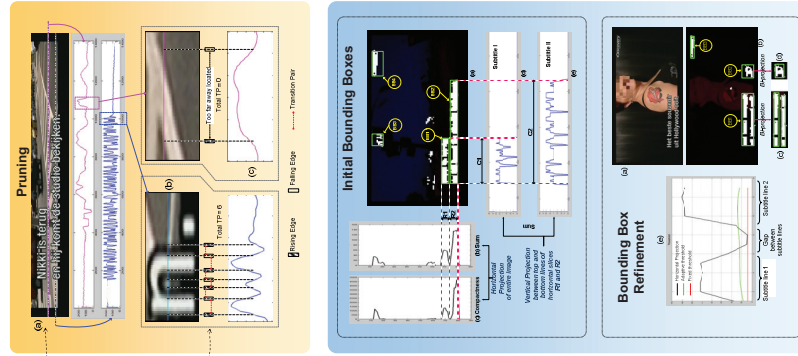


Results

- Pixel-level performance on a set of 5000+ frames of annotated TV video:
PPV: Positive Predictive Value, TPR: True Positive Rate, FPR: False Positive Rate
 - When using **classification by text-stroke features**:
 - 96% PPV, (98.0% TPR at 0.1% FPR).
 - When using **bounding box categorization**:
 - 83% PPV, (98.5% TPR at 0.6% FPR).
 - When using **existing static-region detector**:
 - 0.56% PPV, (100% TPR at 2.2% FPR)

Potential Applications

- Subtitle removal / replacement.
- Better PQ of surrounding text in Frame Rate Conversion.
- 3D positioning of subtitles at desired depth level.
- Keeping subtitles still, in camera stabilization use case.



Context analysis: sky, water and motion

S. Javanbakhti, S. Zinger, J. Han, P. H. N. de With

Video Coding and Architectures Research group, Faculty of Electrical Engineering, Eindhoven University of Technology,
P.O. Box 513, 5600 MB Eindhoven, the Netherlands
Phone: +31 40 247 2540, +31 40 247 3708
{S.Javanbakhti, s.zinger, jg.han, P.H.N.de.With}@tue.nl

Video surveillance has become a widespread technology for increasing security and collecting information on events and behavior of objects. Current research, such as in the European ViCoMo project, has shown that video analysis is an important aid to facilitate decision making for surveillance. Interpreting the events present in the video is a complex task, and the same gesture or motion can be understood in several ways depending on the context of the event and/or the scene. Therefore, it is useful to analyze the context of the scene prior to concluding on the semantic meaning of the video content.

In this paper, we present our research on context analysis on video sequences. By context analysis we mean not only determining the general conditions such as daytime or nighttime, indoor or outdoor environments, but also region labeling [1] and motion analysis of the scene. Our current research concentrates on sky and water labeling and on motion analysis for determining the context. Later, this can be extended with regions such as roads, greenery, buildings, etc.

The techniques applied in this work are capable of robustly detecting various sky and water appearances, even when the color information is very poor. Surveillance videos of the outdoor environments in the Netherlands often lack color because of little sunlight even during daytime. Hence, it is important to explore sky and water detection without relying much on the characteristic blue color features. Context analysis based on motion is another goal of this work which can be used to annotate roads and to restrict the computationally heavy search for moving object to the areas where the motion is detected.

The sky detection algorithm [3] that we apply has two phases: (1) *training phase* which defines the color model, texture properties at multi-resolution and vertical position, (2) *detecting phase* which adapts the color model, vertical position and calculates texture properties. The water detection algorithm that we present in this paper also consists of two parts: (1) graph-based image segmentation, which generates initial regions, and (2) SVM-based region recognition. Normalized RGB color information is used as a feature for SVM. We have evaluated the entropy of the pixel as an additional metric but not much improvement has been achieved. We also consider the location of the pixels, but this feature reduces the flexibility.

Experiments based on the above detection techniques show that we achieve results comparable with other state-of-the-art techniques for sky and water detection, although in our case

the color information is poor. To evaluate results, we use the Coverability Rate (CR) which measures how much of the true sky or water is detected by the algorithm. The obtained average of CR for water detection is about 96.6% and for sky detection it is about 98%.

To analyze motion, we apply a heat map. A *heat map* is a 2D histogram indicating main regions of motion activity [4]. This information provides us context for identifying regions in the scene where the motion happens and it can, for example, delimit the area for searching moving objects. As a result, less video processing will be needed when this context is used. We also identify the direction of the movement in the scene using optical flow [4], [5]. We expect that this context, combined with a heat map, can be applied for change detection.

REFERENCES

- [1] J. Fan, Y. Gao and H. Luo. Multi-Level annotation of natural scenes using dominant image components and semantic concepts. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, October 2004, New York, USA.
- [2] M. Iqbal, O. Morel and F. Meriaudeau. A survey of outdoor water detection. In *The 5th International Conference on Information & Communication Technology and Systems (ICTS)*, Indonesia, 2009.
- [3] B. Zafarifar and P. H. N. de With. Adaptive modeling of sky for video processing and coding applications. In *WTC*, 2006.
- [4] N. Ihaddadene and C. Djeraba. Real-time crowd motion analysis. In *19th International Conference on Pattern Recognition*, December, 2008, Tampa, Florida - USA.
- [5] Tutorial on Optical Flow, University of Manchester, <http://www.tina-vision.net/docs/memos/2004-012.pdf>, last access 14 Jan 2011.

Context analysis: sky, water and motion

S. Javanbakhti, S. Zinger, P. H.N. de With

Department of Electrical Engineering, Eindhoven University of Technology, The Netherlands

Introduction

Importance: Context analysis of digital images and video sequences is widely used, with applications ranging from high-level image understanding and semantic-driven image- and video retrieval, to pixel-level applications like object recognition and local picture quality improvement. Our current research concentrates on sky and water labeling and on motion analysis for determining the context.

Problem: Surveillance videos of the outdoor environments in the Netherlands often lack color because of little sunlight even during daytime.

Solution: exploring sky and water detection without relying much on the characteristic blue color features.

Approach

Sky detection: The algorithm [1] that we apply has two phases:

- 1) training phase which defines the color model, texture properties at multi-resolution and vertical position,
- 2) detecting phase which adapts the color model, vertical position and calculates texture properties.

Water detection: The algorithm also consists of two parts:

- 1) graph-based image segmentation, which generates initial regions,
- 2) SVM-based region recognition. Normalized RGB color information is used as a feature for SVM.

Motion detection: To analyze motion, we apply a heat map. A heat map is a 2D histogram indicating main regions of motion activity [2]. We also identify the direction of the movement in the scene using optical flow [2].

Experimental results

To evaluate results, we use the Coverability Rate (CR) which measures how much of the true sky or water is detected by the algorithm. We have applied the proposed algorithms for sky and water detection on more than 15 images of Watervisie data set. The obtained average of CR for water detection is about 96.6% and for sky detection it is about 98%. Fig.1 (left) shows the original image of our data set which is one frame of video sequence.

As we can see this data set does not provide sufficient information in terms of color. Fig.1 (middle) shows probability of the sky detection algorithm on this image and Fig.1 (right) shows result of applying threshold.

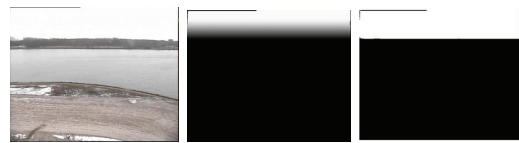


Fig.1. Original image (left). GT of sky (between). Result of the sky detection (right)

Fig.2 shows the original image with result of the water detection algorithm

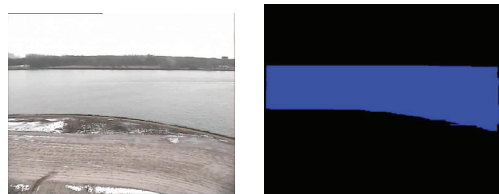


Fig.2. Original image (left). Result of water detection algorithm on image(right)

Heat map of our video sequence is shown in Fig. 3 (left). Fig.3 (right) shows result of motion detection.



Fig.3. Heat map (red indicates regions with highest movement) (left). Optic flow LK (arrow shows the direction of ship's movement) (right).

Conclusion

The proposed sky and water detection algorithms increase correct detection/rejection rates in data sets with insufficient information in terms of color. A heat map can delimit the area for searching moving objects. As a result, less video processing will be needed when this context is used and with combining with motion vector, can be applied for change detection.

References

- [1] B. Zafarifar and P. H. N. de With, "Adaptive Modeling of Sky for Video Processing and Coding Applications", WIC, 2006.
- [2] N. Ihaddadene, "Real-time Crowd Motion Analysis", 19th International Conference on Pattern Recognition, USA, 2008.

Seeing the user: CV in support of self adaptive user interfaces

Hester Bruikman, Hao Wang, and Roy van de Korput

Philips Consumer Lifestyle, High Tech Campus 37, 5656 AE, Eindhoven, The Netherlands
E-mail: {Hester.Bruikman, Hao.Wang, Roy.van.de.Korput}@philips.com

I. INTRODUCTION

When designing and developing user interfaces a cost-benefit trade off is made in deciding whether or not to add accessibility features. The MyUI project proposes the concept of self adaptive interfaces to optimize this trade off. By offering design pattern repositories that contain user interface designs that are suitable for users with special needs, generic user interfaces are to self adapt following user and context models updated in real time.

To create real time user and context models, unobtrusive and easy implementable sensors need to be in place. These sensors need to identify the actual effect of disabilities on user system interaction performance and user experience. For some user model variables computer vision applications are logical software sensors given that they are unobtrusive, and easy implementable.

While accessibility has been a widely addressed topic in the domain of web and PC applications, the need to address accessibility in interactive television applications has only recently emerged with the introduction of interactive digital television (iDTV) and Internet TV (a.k.a.: Connected TV or Smart TV). The television is a familiar, centrally located device that poses special benefits for those with disabilities who have trouble accessing PC's. In this context, the MyUI project focuses on developing self adaptive user interfaces for Internet TV services.

II. FACE DETECTION AND EYE TRACKING APPROACH

An early sensor that is developed in the MyUI project is a sensor to measure the actual user profile in terms of visibility. Following known data about a users' visual deficiency we infer to what extend this deficiency negatively affects actual usage of an Internet TV service and update the user profile accordingly. The system design provides a robust implementation to detect visual difficulties using inexpensive webcams. Here a lean-forward gesture is defined as a measure of visual difficulties which relates to real-time relative distance detection. Distance between two eyes is applied as a cue to determine user-to-camera distance. The system flow diagram is depicted in Figure 1.

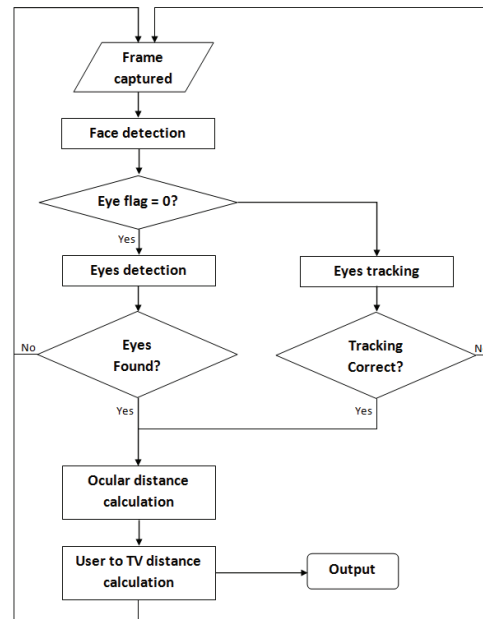


Figure 1 System flow of distance detection

When a captured frame comes in, the first step is detecting the face region, that is Region of Interest (ROI), and eyes' locations. Meanwhile, an eye flag is checked whether the eyes have yet been found or not. If eyes are not located yet, the eyes detection process is in operation on the ROI on the basis of a prior detected face. In the case that eyes locations are known from previous frames, a fast tracking algorithm substitutes detection. Once eyes are correctly positioned either by detection or tracking, the Euclidean distance between eyes, also named ocular distance, is calculated. Comparing the ocular distance with a calibrated distance, the absolute distance between user and camera can be determined. Then a lean-forward or -backward gesture, i.e., relative distance, can be easily detected by the distance changes during a certain period.

III. ALGORITHMS USED

As described above, the system consists of two aspects to detect a lean-forward gesture: detection and tracking. For detection, one face detector and one eye detector are adopted, since faces and eyes have different features. Both of these detectors are on the basis of a cascade scheme consisting of a set of simple Haar-like features based AdaBoost classifiers. These detectors are built by previous training with a representative set of positive and negative examples, where

positive examples are images with a face or eye and negative examples are images randomly selected without face and eye. Using patch templates with different scales, the detectors can rapidly and robustly detect a face or eye from the images.

After detection, a pyramid Lucas-Kanade feature tracker is used to track the eyes. The goal of feature tracking is to find out the optical flow vector \mathbf{d} , for a given point \mathbf{u} in image I , and its corresponding location $\mathbf{v}=\mathbf{u}+\mathbf{d}$ in image J . The optical flow \mathbf{d} is defined as being the vector that minimized the residual function ϵ defined as:

$$\epsilon(\mathbf{d}) = \sum_{x=u_x-w_x}^{u_x+w_x} \sum_{y=u_y-w_y}^{u_y+w_y} (I(x, y) - J(x + d_x, y + d_y))^2,$$

where an integration window of size $(2w_x + 1) \times (2w_y + 1)$ is measured.

In order to handle large motions while keeping high accuracy, the hierarchy structure is introduced. The overall pyramidal tracking algorithm works as follows: first, the optical flow is computed at the deepest pyramid level L_m . Then, the result of the computation is propagated to the upper level L_{m-1} in a form of an initial guess for the pixel displacement (at level L_{m-1}). Given that initial guess, the refined optical flow is computed at level L_{m-1} , and the result is propagated to level L_{m-2} and so up to the level 0 (the original image).

IV. INITIAL DEMONSTRATOR

The initial demonstrator that has been developed within the project includes in terms of hardware: a PC running windows, an RFID reader, Philips Net TV, a basic webcam, and a remote control. In terms of software the system consists of three main components:

- Client (including camera Device Driver (CVA), Event Processor and a Browser)
- Context Manager (including a Sensor Data Manager, and application and context data storage and management components), and
- Middleware application following a Model View Controller approach (including a Design Patterns Repository, a views generator and an Adaptation Engine)

If a measured distance differences X cm between frames in a Y time frame (= a lean-forward or lean-backward gesture), an event is sent to an Event Processor. The Event Processor then decides where the event needs to go, in this case to a Sensor Data Manager. The Sensor Data Manager uses this event to update the User Profile accordingly. When a given user has a visual deficiency that translates to z on a scale of $-x$ to $+x$, this score is updated following a lean-forward event with $+1$. Subsequently, a Middleware layer associates the visual ability value to a user interface design pattern that is suitable for users that score $z+1$ on visual ability. A suitable view to an Internet TV service is then generated and is sent back to the client, which is a Philips Net TV browser.

The sample Internet TV service that is used is an e-mail application. The view adapts in fontsize and consequently number of items on the screen. This early demonstrator is used to convey the concept of self adaptive interfaces.

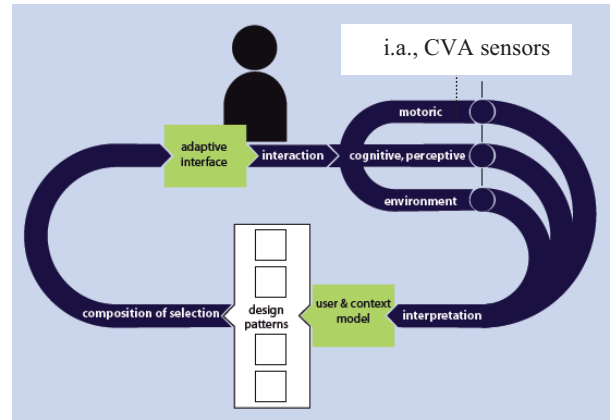


Figure 2 MyUI self adaptive user interface concept

V. FUTURE WORK AND CHALLENGES

Future work includes an extension of the CV sensors set. Current research is aimed at measuring users' attention, given that particularly elderly and stroke patients (i.e., main target users within the MyUI project) have difficulties in retaining attention. We consider head pose (i.e., horizontal head orientation) as a measure of *visual* attention. To infer what effect visual attention has on the actual user interaction, CV sensors need to be combined with known user profile data and other user interaction data. For example, when taking response time of remote control input as a performance measure of how suitable this device is, attention measures can contextualize performance results.

In the remaining of the MyUI project we look into using CVA as more conscious input mechanisms. Salient and repetitive lean-forward movements can indicate visual difficulties, but also a preference for bigger fonts. Conscious user behavior will need to be distinguished from deficiency sensing that is currently implemented. Aside from this gesture control is a logical addition in supporting disabled users.

In developing other sensors, albeit to detect input or user profile and context related data, optimizing applications for use in a living room with a –rule of thumb– 10 feet position to screen poses challenges on detection accuracy. Since, even though bodily gestures are very much apparent in social communication, in private situations these movements are much less distinct. These challenges remain not only for self adaptive user interfaces, but for any CVA that is used to improve (Internet) TV user experience.

VI. ACKNOWLEDGEMENTS

The MyUI project is funded under the EU 7th Framework Programme. We would like to acknowledge all partners working on this project. In addition, we would like to acknowledge colleagues who have made available their expert knowledge, namely Nadejda Roubtsova, Chris Varekamp and Gerard de Haan.

PHILIPS

Seeing the user: CV in support of self adaptive user interfaces

MyUI is an EU R&D project that aims to develop a design pattern repository and accompanying logic, for generic user interfaces to self adapt following user and context models updated in real time.

To create real time user and context models, unobtrusive and easy implementable sensors need to be in place. These sensors need to identify the actual effect of disabilities on user system interaction performance and user experience. For some user model variables computer vision applications are logical software sensors given that they are unobtrusive, and easy implementable.

The MyUI project focuses on developing self adaptive user interfaces for Internet TV services.

An early sensor that is developed in the MyUI project is a sensor to measure the actual user profile in terms of visibility. Following known data about a users' visual deficiency we infer to what extend this deficiency negatively affects actual usage of an Internet TV service and update the user profile accordingly. The system design provides a robust implementation to detect visual difficulties using inexpensive webcams. Here a lean-forward gesture is defined as a measure of visual difficulties which relates to real-time relative distance detection. Distance between two eyes is applied as a cue to determine user-to-camera distance.

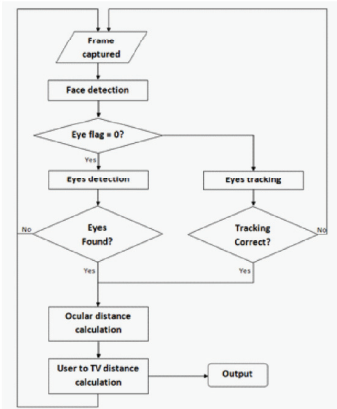


Figure 2. System flow of distance detection

The initial demonstrator that has been developed within the project includes in terms of hardware: a PC running windows, an RFID reader, Philips NetTV, a basic webcam, and a remote control.

If a measured distance differences X cm between frames in a Y time frame (= a lean-forward or lean-backward gesture), an event is sent to an Event Processor. The Event Processor then decides where the event needs to go, in this case to a Sensor Data Manager. The Sensor Data Manager uses this event to update the User Profile accordingly. When a given user has a visual deficiency that translates to z on a scale of -x to +x, this score is updated following a lean-forward event with +1. Subsequently, a Middleware layer associates the visual ability value to a user interface design pattern that is suitable for users that score z+1 on visual ability. A suitable view to an Internet TV service is then generated and is send back to the client, which is a Philips NetTV browser.

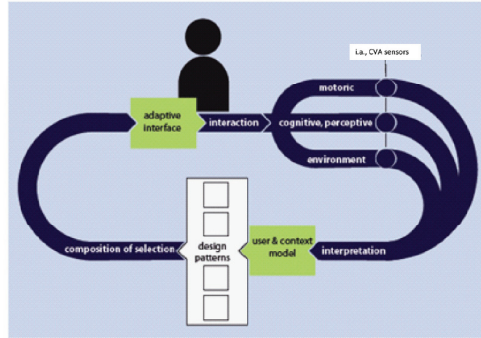


Figure 1. How CV software sensors are used in support of self adaptive interfaces

Both, the face detector and the eye detector are on the basis of a cascade scheme consisting of a set of simple Haar-like features based Ada-Boost classifiers.

After detection, a pyramid Lucas-Kanade feature tracker is used to track the eyes. The goal of feature tracking is to find out the optical flow vector d , for a given point u in image I , and its corresponding location $v=u+d$ in image J . The optical flow d is defined as being the vector that minimized the residual function ϵ defined as:

$$\epsilon(d) = \sum_{x=u_x-w_x}^{u_x+w_x} \sum_{y=u_y-w_y}^{u_y+w_y} (I(x, y) - J(x + d_x, y + d_y))^2$$

where an integration window of size $(2w_x + 1) \times (2w_y + 1)$ is measured.

In order to handle large motions while keeping high accuracy, the hierarchy structure is introduced. The optical flow is computed at the deepest pyramid level L_m . Then, the result of the computation is propagated to the upper level L_{m-1} in a form of an initial guess for the pixel displacement (at level L_{m-1}). Given that initial guess, the refine optical flow is computed at level L_{m-1} , and the result is propagated to level L_{m-2} and so up to the level 0 (the original image).

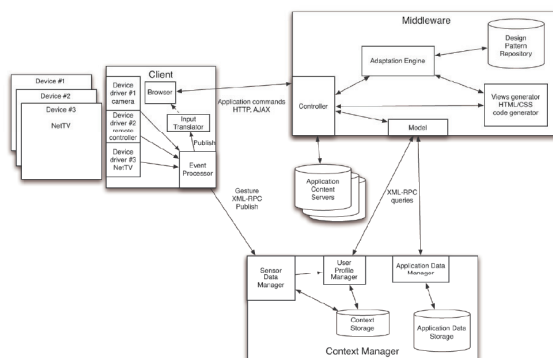


Figure 3. System architecture of MyUI user and context modelling and self adaptive interface generation

Hester Bruikman, Hao Wang, and Roy van de Korput
Philips Consumer Lifestyle, High Tech Campus 37, 5656 AE, Eindhoven, The Netherlands
E-mail: {Hester.Bruikman, Hao.Wang, Roy.van.de.Korput}@philips.com

Towards Multi-View Ship Detection for Maritime Surveillance

Rob Wijnhoven
ViNotion
P.O. Box 2346, 5600 CH Eindhoven
The Netherlands
rob.wijnhoven@vinotion.nl

Kris van Rens
Vinotion
P.O. Box 2346, 5600 CH Eindhoven
The Netherlands
kris.van.rens@vinotion.nl

I. INTRODUCTION

To aid traffic operators, maritime traffic control systems overlay the position of ships on a map of the area. Typically, the detection and tracking of ships are generated using radar technology. Video cameras are used as visual verification tools for the traffic guidance operators. Although radar technology is mature and gives highly accurate detection results, interference in the radar signal from clutter makes object detection more difficult. Furthermore, small- and non-metal ships are not detected. Next to these technical difficulties, the application of radar technology is quite expensive.

Within the Dutch WATERVisie project¹, we aim at developing a system for real-time ship detection and tracking by using commercially available Pan-Tilt-Zoom (PTZ) camera systems. The new system should reduce or remove the shortcomings of the current radar-based system. Generated detection information will be integrated with the existing traffic waterway analysis system. An graphical representation of the system is shown in Figure 1.

Typical object detection for video surveillance applies background modeling techniques and classifies each pixel of the image into background or object. These methods have difficulty dealing with changing light conditions. Randomly moving objects like flags or water result in many false detections. Because the speed of ships in harbors is typically low, continuous monitoring of the complete area is too costly. Scanning the area with a single pan-tilt-zoom camera is intrinsically sufficient for regularly sampling the ships locations. However, camera movement poses a new problem for the detection algorithm in such a system.

Background segmentation cannot be used, as ships should be detected for any position of the camera. Therefore, we propose the use of shape-based detections. In an initial experiment, the authors have evaluated the use of shape-based ship detection [1]. The camera was fixed to a viewpoint perpendicular to the waterway, thereby limiting the monitored ships to sideviews only. A dataset of 150 ships was created from three days of video data, captured at the "Botlek",

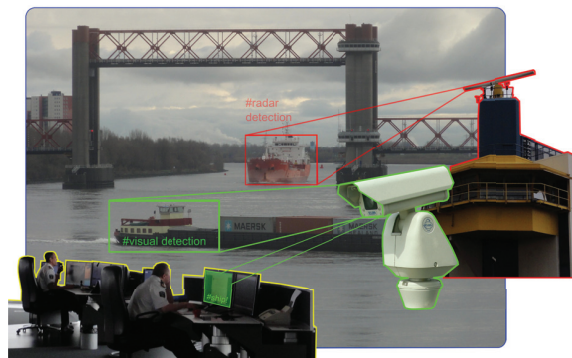


Fig. 1. Visual detections in addition to existing radar detections.

an industrial area in Rotterdam, The Netherlands. Because of the high variation in ship appearance, the stern (back) of the ships was annotated (see Figure 2). Analyzing the detection performance shows that the detection of ships given limited viewpoint variation is very good (see Figures 3 and 4). Although the viewpoint of this dataset is fixed, the visual variation of ships is very high. However, if we want to detect ships over the full range of viewpoints, a single viewpoint detector is insufficient.

II. TOWARDS MULTI-VIEW DETECTION

Solving the multi-view detection problem can be done by dividing the total problem in smaller sub-problems. The total variation in viewpoint can be divided in smaller subsets that each cover a range of viewpoints (e.g. only frontal or sideviews of ships). However, the selection of the number of sub-classes and the division of ship images in subclasses is challenging.

A. Multi-View Detection Algorithm

Manual annotation into subclasses is very time-consuming and can result in a non-optimal division of the samples with respect to detection performance. Therefore, unsupervised sub-categorization is preferred. *k*-means clustering could be used to divide the training set, but includes background information (around the ships) in the clustering process. Approaches that do use this background information in the sub-categorization process are e.g. Cluster Boosted Tree [2] and the algorithm by Kuo and Nevatia [3]. Most algorithms assume a fixed number

¹The WATERVisie project is a cooperation between HITT, Vector Fabrics, Eindhoven University of Technology, ViNotion and Havenbedrijf Rotterdam (HbR) and is supported by Point One, an innovation program of the Dutch Ministry of Economic affairs.

of final subclasses and/or do not consider merging or splitting of classes.

Because the amount of background information is limited for ship images (typically, background is water), the use of k -means clustering can be used. Although limited training images are available, first experiments show that the grouping of ship images of multiple viewpoints using k -means clustering shows an improvement in the detection performance. As future work, we will investigate more advanced methods that take the background information into account.

B. Multi-Viewpoint Dataset Generation

Apart from algorithmic challenges, there are several practical constraints, when applying in a maritime surveillance system. Most difficult is the generation of sufficient training data. Although public datasets are available for some object classes (*e.g.* cars, persons, faces), the problem of ship detection is an unexplored area, requiring the generation of a new dataset. Recording several days of video at a maritime location gives sufficient training information, but manual annotation of the data takes too much time. Moreover, not all variation in both ship appearance and viewpoints might be contained in this limited timespan (*e.g.* cruise ships typically only appear every few weeks/months).



Fig. 2. Large variation in ship training images for single viewpoint only.



Fig. 3. Ship detections (red: annotations, green: detections).



Fig. 4. Ship detections (red: annotations, green: detections).

We propose to extend the traditional continuous recording with a motion-detection filter, to limit the amount of video data to be recorded, keeping the amount of human annotation effort to a minimum. By using a PTZ camera that covers a large range of positions, but remains stationary at each position for some timespace, the motion-detection algorithm can be used for different backgrounds. This hybrid solution uses both the advantages of the moving PTZ camera, and the motion

information from background modeling using a static camera to generate training data.

In addition to using motion filtering as pre-selection, the real video data can be extended with virtual data from 3D models of ships. Yu *et al.* [4] have shown that the addition of synthetic images enhances the performance of a pedestrian detection problem. Liebelt and Schmid [5] use a database of 3D car models for additional training data. Although the use of 3D information is very useful because full viewpoint information is available (see Figure 5), the generation of 3D models is very time consuming.

III. CONCLUSIONS

In this paper, we have analyzed the problems of multi-view ship detection. Previous experiments have shown that the detection of ships with limited variation in viewpoint is possible and is very robust to changes in the ship appearance. We propose to extend the single detector to multiple independent detectors that each cover a range of viewpoints. First experiments show an increase in detection performance, showing the feasibility of the approach. As future work, we will investigate the selection of the viewpoint ranges for the separate detectors and the corresponding training samples. To generate a real dataset from large amounts of video data we have proposed a motion-based filter to keep the amount of human annotation effort to a minimum. In addition, to extend the training images of a multi-viewpoint ship detector, 3D models can be used.

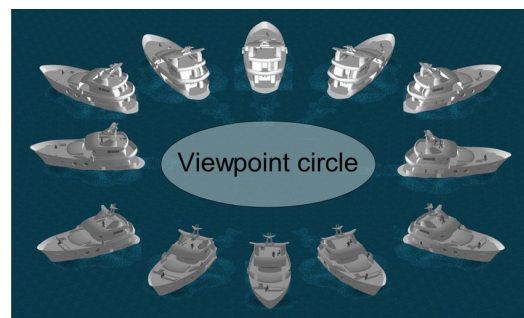


Fig. 5. Synthetic 3D models can be used to generate additional training data for ship detection. Note the large change in appearance over the full 360-degree viewing circle.

REFERENCES

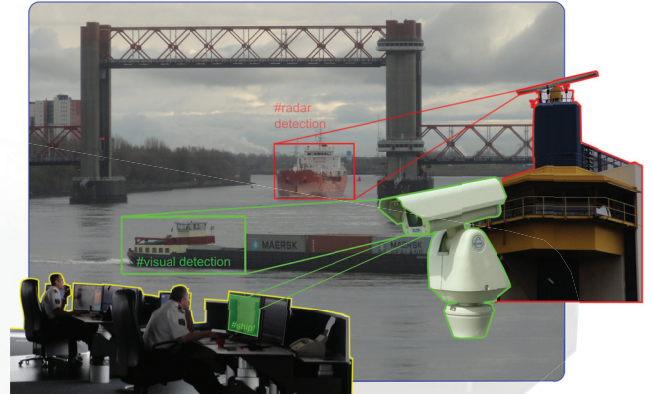
- [1] R. Wijnhoven, K. van Rens, E. G. T. Jaspers, and P. H. N. de With, "Online learning for ship detection in maritime surveillance," in *Proc. of 31th Symposium on Information Theory in the Benelux (WIC)*, May 2010, pp. 73–80.
- [2] B. Wu and N. Ram, "Cluster Boosted Tree Classifier for Multi-View, Multi-Pose Object Detection," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2007, pp. 1303–1310.
- [3] C.-H. Kuo and R. Nevatia, "Robust multi-view car detection using unsupervised sub-categorization," in *Workshop on Applications of Computer Vision (WACV)*, 2009.
- [4] J. Yu, D. Farin, C. Krüger, and B. Schiele, "Improving person detection using synthetic training data," in *Proc. International Conference on Image Processing (ICIP)*, September 2010.
- [5] J. Liebelt and C. Schmid, "Multi-view object class detection with a 3d geometric model," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

Towards Multi-View Ship Detection for Maritime Surveillance

Rob Wijnhoven & Kris van Rens
ViNotion, Eindhoven, The Netherlands
{rob.wijnhoven,kris.van.rens}@vinotion.nl

Introduction

- Traffic control, Rotterdam Harbor: Detect ships and their locations
- Radar technology: interference, misses small ships
- WATERVisie Project: use moving PTZ camera to detect ships



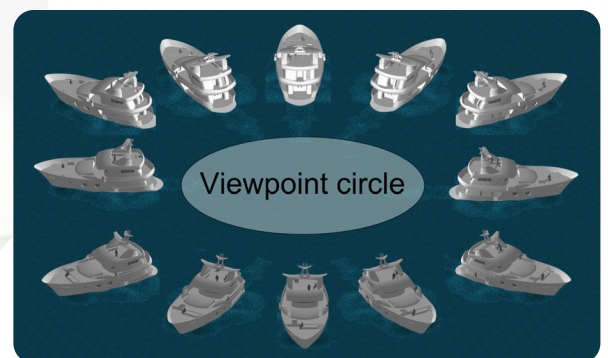
Ship Detection

- Background modeling cannot be used (moving camera)
- Propose shape-based ship detection
- Experiments show feasibility, given limited ship viewpoint variation



Multi-View Ship Detection

- To cover variation of all viewpoints, we propose multiple detectors
- How to define these detectors (e.g. k-means clustering / hierarchy)
- How to obtain training data (e.g. motion filtering to select images)



Conclusions

- Ship detection feasible for single viewpoint
- Extension to full multi-view requires multiple detectors
- More training image data needed, many practical limitations



