Monocular and Stereo Methods for AAM Learning from Video

Jason Saragih[1] and Roland Goecke[1,2*]
[1]Research School of Information Sciences and Engineering, Australian National University
[2]National ICT Australia, Canberra Research Laboratory
Canberra, Australia
Email: jason.saragih@rsise.anu.edu.au, roland.goecke@anu.edu.au

Abstract

The active appearance model (AAM) is a powerful method for modeling deformable visual objects. One of the major drawbacks of the AAM is that it requires a training set of pseudo-dense correspondences over the whole database. In this work, we investigate the utility of stereo constraints for automatic model building from video. First, we propose a new method for automatic correspondence finding in monocular images which is based on an adaptive template tracking paradigm. We then extend this method to take the scene geometry into account, proposing three approaches, each accounting for the availability of the fundamental matrix and calibration parameters or the lack thereof. The performance of the monocular method was first evaluated on a pre-annotated database of a talking face. We then compared the monocular method against its three stereo extensions using a stereo database.

1. Introduction

The active appearance model (AAM) [9] is a powerful method for modeling deformable visual objects, coupling a compact parametric representation with an efficient generative alignment method. As such, the method has found applications in many image modeling, alignment and tracking problems, for example [16, 21, 23].

To build the AAM's generative model, a pseudo-dense set of annotations is required for each training image. This is a major drawback of the method as manual annotations of large databases is both tedious and error prone. Although there are a number of methods for automatic correspondence finding, most are lacking in two respects. Firstly, the groupwise methods that are commonly employed usually ignore the sequential nature of images in video. Secondly, despite generally modeling a non-rigid 3D object, no constraint on 3D geometry is enforced, which is the consequence of using monocular images.

In this paper, we present contributions on two fronts. Firstly, we demonstrate that embedding automatic correspondence finding within an adaptive tracking paradigm can yield good results. Secondly, we extend this method to account for the epipolar geometry which must be adhered to by stereo images. We begin with an overview of related work in Section 2. The adaptive template paradigm for correspondence finding is presented in Section 3. Extensions of this method to stereo sequences are then presented in Section 4. In Section 5, we describe our experiments and analyze the results. We conclude in Section 6 with a summary of our results and directions of future work.

2. Related Work

There has been a significant amount of research over the years to automatically find pseudo-dense correspondences across images of the same class for building AAMs. These methods can be broadly categorized into either feature or image based approaches.

Feature based approaches, for example [6, 13], find correspondences between salient features in the image by examining the local structure of the features. The advantage of this approach is that feature comparisons and calculations are relatively cheap. The downside however is twofold. Firstly, there may be insufficient salient features in the object to build a good appearance model. Secondly, as the feature comparisons generally consider only local image structure, the global image structure on which the AAM is then modeled is ignored, and hence, the model built using annotations found in this manner may be suboptimal.

Image based methods, for example [2, 7], find dense correspondences across images by learning a warping function which minimizes some type of error measure between the intensities of the images. The main advantage of this approach is that the global structure of the image is taken into

*National ICT Australia is funded by the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council.

1-4244-1180-7/07/$25.00 ©2007 IEEE

account, better mimicking the AAM for which the correspondences will be used later. The main drawback of this approach is that to accurately represent the shape variations of the visual object, the warping function will generally need to be parametrized using a large number of landmarks. This, in turn, leads to a very large optimization problem which is slow to optimize and prone to terminating in local minima.

There has been comparatively little research in building AAMs from image sequences. In [24], the authors track the most salient features independently throughout a sequence. Although this method can be computationally cheap, it is essentially a feature based approach, suffering from the drawbacks described above. Even sparser is the attempt to incorporate scene geometry into the model building process. The utility of such an approach has been investigated in model fitting, however. In [14], for example, the fidelity of AAM fitting was improved by simultaneously fitting to a number of images of the same scene using independent AAMs linked by a pre-learnt 3D shape model.

3. Adaptive Template Tracking

In video databases, it is advantageous to assume that the change in the objects appearance between consecutive frames is small. As such, a deformable template annotated in the first frame can be propagated through the sequence by sequentially finding the small perturbations in the warping and lighting function which brings each frame to the next one in the sequence.

There are difficulties here, however, regarding how to treat the changes in the object's appearance throughout the sequence. Although the object's appearance in the previous frame is usually a good approximation to that of the current frame, small misalignments in the fitting process can accumulate throughout the sequence, leading to the drifting phenomenon. There are a number of approaches to the template update problem which minimize drifting, for example [19, 20, 25]. In this work, we follow the approach in [19], where the object's texture is modeled as a weighted sum of an initial template and the texture from the most recent image:

$$\mathbf{T}(\mathbf{x}) = \gamma \mathbf{T}_0(\mathbf{x}) + (1-\gamma)\mathbf{T}_{t-1}(\mathbf{x}). \tag{1}$$

The parameter $\gamma \in (0 \dots 1)$ is a grounding factor which reduces drifting whist allowing the template to adapt to the current object's texture.

Apart from this template adaptation process, we also account for global changes in lighting between the template and the object in the current frame using the linear model:

$$\mathbf{t}(\mathbf{x};\mathbf{q}) = \alpha \mathbf{T}(\mathbf{x}) + \beta, \tag{2}$$

where $\mathbf{t}(\mathbf{x};\mathbf{q})$ is the texture model to be fitted to the current image and $\mathbf{q} = (\alpha, \beta)$ are the global lighting parameters.

To fit the template to the current image, we minimize the following cost function:

$$C = C_D + w_S C_S, \tag{3}$$

where C_D is a data term which measures the similarity between the warped image and the template and C_S is a smoothing term which penalizes complex deformations induced by the warping function. The smoothing weight w_S regularizes the trade off between these two terms in the total cost.

3.1. The Data Term

To account for differences in appearance between the template and the image, the data term is chosen to be a robust penalization of the pixel-by-pixel difference:

$$C_D = \sum_{\mathbf{x}\in\Omega} \rho\left(E(\mathbf{x});\sigma\right), \tag{4}$$

where Ω is the template's spatial domain and

$$E(\mathbf{x}) = \mathbf{t}(\mathbf{x};\mathbf{q}) - I(\mathbf{W}(\mathbf{x};\mathbf{p})) \tag{5}$$

is the texture residual at pixel $\mathbf{x}$. Here, $\mathbf{I}(\mathbf{W}(\mathbf{x},\mathbf{p}))$ is the texture of image $\mathbf{I}$ warped back to the reference frame at location $\mathbf{x}$ using the warp $\mathbf{W}(\mathbf{x};\mathbf{p})$, parameterized by a set of landmarks $\mathbf{p} = [\mathbf{x}_1;\dots;\mathbf{x}_n]$. To simplify notation in the following, a pixel location in the reference frame warped onto the image will be written as:

$$\vec{\mathbf{x}} = \mathbf{W}(\mathbf{x};\mathbf{p}) \tag{6}$$

As a robust penalizer, we use the *Geman-McClure* function:

$$\rho(r;\sigma) = \frac{r^2}{\sigma^2 + r^2}, \tag{7}$$

which has been used extensively for optical flow estimation [3, 4]. For the choice of the scaling parameter σ, we follow the work in [22] and assume a contaminated Gaussian distribution for the residuals. With this, the authors derive σ from the median of the absolute residuals:

$$\sigma = 1.4826\,\mathrm{med}\left(|E(\mathbf{x})|\right) \tag{8}$$

which they claim to tolerate almost 50% of outliers when the assumption holds.

As the relationship between the parameters $\mathbf{p}$ and the image pixels is generally nonlinear, minimizing Equation (3) requires a general-purpose nonlinear optimizer. Since we assume the object's appearance changes only by a small amount between frames, the estimate of the warp and lighting parameters from the previous frame is assumed to lie within the quadratic region of the cost function of the current frame. As such, we use the Gauss-Newton method,

which can be expected to exhibit superlinear if not quadratic convergence baring the assumptions hold [10]. To allow the use of the robust error function in the Gauss-Newton method, we follow the approach in [1] and replace the data term in Equation (4) with:

$$C_D = \sum_{\mathbf{x}\in\Omega} \varrho\left(E(\mathbf{x})^2; \sigma\right) \tag{9}$$

and the robust error function with:

$$\varrho(r;\sigma) = \frac{r}{\sigma^2 + r}. \tag{10}$$

This requires only that the error function is symmetric, which is satisfied by the Geman McClure function.

With this formulation, the gradient and Gauss-Newton Hessian of the data term are given by:

$$\mathbf{g}_D = \sum_{\mathbf{x}\in\Omega} \varrho'(E(\mathbf{x})^2)\mathbf{J}_D(\mathbf{x})^T E(\mathbf{x}) \tag{11}$$

$$\mathbf{H}_D = \sum_{\mathbf{x}\in\Omega} \varrho'(E(\mathbf{x})^2)\mathbf{J}_D(\mathbf{x})^T \mathbf{J}_D(\mathbf{x}), \tag{12}$$

where $\varrho'(E(\mathbf{x})^2)$ is the derivative of the reformulated robust error function and

$$\mathbf{J}_D(\mathbf{x}) = \left[-\nabla I(\vec{\mathbf{x}})\frac{\partial \vec{\mathbf{x}}}{\partial \mathbf{p}}, \mathbf{T}(\mathbf{x}), 1\right] \tag{13}$$

is the Jacobian at $\mathbf{x}$. Here, $\nabla\mathbf{I}$ is the spatial gradient of the current image, $\frac{\partial \vec{\mathbf{x}}}{\partial \mathbf{p}}$ is the derivative of the warped pixel locations with respect to the landmarks, and the last two columns are the components of the Jacobian pertaining to the global lighting and bias $\mathbf{q}$ (see Equations (2) and (5)).

Following [8], we use a piecewise affine function for the warping where the landmarks are triangulated and the motion of pixels within the same triangle are assumed to exhibit affine motion with the same parameters. This warping function gives significant computational savings over other methods such B-splines or thin plate splines. Since the warped location of each pixel depends only on the landmarks making up the triangle containing it, only entries of the gradient and Hessian pertaining to these landmarks need to be updated in each component of the sum's in equations (11) and (12).

3.2. The Smoothing Term

Without any constraint on landmark deformation, the optimization is likely to terminate in a local minima. The purpose of the smoothing term C_S in Equation (3) is, hence, to penalize complex deformations. Taking inspiration from variational optical flow estimation [5], we penalize the *differences* between landmark deformations as follows:

$$C_S = \sum_{i,j}^{n-1} \kappa_{ij} \|(\mathbf{x}_i - \mathbf{x}_i^0) - (\mathbf{x}_j - \mathbf{x}_j^0)\|^2, \tag{14}$$

where $\mathbf{x}_i^0$ is the pixel location in the image from which the deformation of landmark $\mathbf{x}_i$ is measured, and

$$\kappa_{ij} = \frac{\exp\left(-\frac{\|\mathbf{x}_i^0 - \mathbf{x}_j^0\|^2}{2\nu_s^2}\right)}{\sum_j^{n-1}\exp\left(-\frac{\|\mathbf{x}_i^0 - \mathbf{x}_j^0\|^2}{2\nu_s^2}\right)} \tag{15}$$

is a weighting term which encourages points which are close to each other to deform in a fashion more similar than those which are further apart. Also, rather than measuring deformation from the shape in the reference frame, the anisotropic deformation of a linear object is better accounted for by measuring deformation from the previous frame.

As the smoothing term is quadratic in the landmark locations, the Gauss-Newton Hessian of this term is fixed and can be precalculated as follows:

$$\mathbf{H}_S = \sum_{i,j}^{n-1} \kappa_{ij}\left[\mathbf{J}_x(i,j)^T\mathbf{J}_x(i,j) + \mathbf{J}_y(i,j)^T\mathbf{J}_y(i,j)\right] \tag{16}$$

where the k^{th} entry of the smoothing term's Jacobian in the x-direction is given by:

$$\mathbf{J}_x(i,j)^k = \begin{cases} 1 & \text{if } k = 2i \\ -1 & \text{if } k = 2j \\ 0 & \text{otherwise} \end{cases}, \tag{17}$$

and similarly for $\mathbf{J}_y$. With this, the gradient is as follows:

$$\mathbf{g}_S = \sum_{i,j}^{n-1} \kappa_{ij}\left[\mathbf{J}_x(i,j), \mathbf{J}_y(i,j)\right](\mathbf{x}_i - \mathbf{x}_j - \mathbf{x}_i^0 + \mathbf{x}_j^0) \tag{18}$$

Finally, using equations (11), (12), (16) and (18) for the expressions of the gradient and Hessians, the Gauss-Newton parameter updates are given by:

$$[\Delta\mathbf{p}; \Delta\mathbf{q}] = -\mathbf{H}^{-1}\mathbf{g}, \tag{19}$$

with

$$\mathbf{g} = \mathbf{g}_D + w_S\left[\mathbf{g}_S; \mathbf{0}\right] \tag{20}$$

$$\mathbf{H} = \mathbf{H}_D + w_S\begin{bmatrix}\mathbf{H}_S & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}, \tag{21}$$

where the zeros result because the smoothing term is not dependent on the lighting parameters.

4. Tracking with Stereo Constraints

When the database consists of stereo sequences, we can further constrain the tracking process using the scene's geometry. Here, we consider three cases. In the first case, we assume that nothing is known about the geometry of the

scene. Secondly, we assume that the fundamental matrix relating the stereo pairs is known. Finally, we assume that we also know the intrinsic camera parameters. In all these cases, we assume that an initial estimate of a landmark correspondences in the stereo pair of the first frame is available.

4.1. Initial Correspondence Refinement

The initial estimate of the landmark correspondences are obtained from a manual annotation of the stereo pair of the first frame in all sequences. This process is usually error prone as it depends on a subjective decision about locations which are most similar. In this section, we describe a method used to refine the initial landmark estimates such that they better correspond. To this end, we propose minimizing the following cost function:

$$C = C_C + w_S C_S + w_E C_E, \tag{22}$$

where C_C is a data term, C_S is a smoothing term and C_E is an epipolar constraint. We assume that the landmarks in one of the stereo pairs are fixed and optimize over the landmarks in the other image only, using the Gauss-Newton method.

For the data term, we utilize the color constancy assumption, commonly used in stereo matching:

$$C_C = \sum_{\mathbf{x}\in\Omega} \varrho\left(\left[\alpha_C \mathbf{I}_1(\mathbf{x}) + \beta_C - \mathbf{I}_2(\vec{\mathbf{x}})\right]^2; \sigma\right), \tag{23}$$

where ϱ is given in Equation (10), α_C and β_C are global lighting parameters, $\mathbf{I}_1(\mathbf{x})$ is the texture of the fixed image at location $\mathbf{x}$ and $\mathbf{I}_2(\vec{\mathbf{x}})$ is its stereo pair at the warped pixel location. The robust penalization is required to account for texture differences which arise from the warping process and differences in noise between the images. The Jacobian of the color constancy term takes the following form:

$$\mathbf{J}_C(\mathbf{x}) = \left[-\nabla \mathbf{I}_2(\vec{\mathbf{x}}) \frac{\partial \vec{\mathbf{x}}}{\partial \mathbf{p}}, \mathbf{I}_1(\mathbf{x}), 1\right]. \tag{24}$$

The gradient and Gauss-Newton Hessian then take the same form as Equations (11) and (12), substituting $\mathbf{J}_D$ for $\mathbf{J}_C$.

The deformation complexity between the stereo pairs is penalized as described in Section 3.2. The only difference here is that the reference landmarks from which deformation is computed are the landmarks in the fixed image rather than the previous image in the sequence.

To encourage the landmarks to adhere to scene geometry, the following epipolar constraint is used:

$$C_E = \sum_{i}^{n-1} \left(\tilde{\mathbf{x}}_i^T \mathbf{F} \tilde{\mathbf{x}}_i^0\right)^2, \tag{25}$$

where $\mathbf{F}$ is the fundamental matrix, $\tilde{\mathbf{x}}_i$ is the homogeneous point of the i^{th} landmark $[x_i; y_i; 1]$ and $\tilde{\mathbf{x}}_i^0$ is that of the corresponding point in the fixed image. The squared term is required to ensure the cost of this constraint is positive. This is effectively a soft constraint on the epipolar geometry of the landmarks. As the weight w_E tends to infinity, this term becomes a hard constraint. In practice, w_E is chosen to be a suitably large constant.

If the fundamental matrix relating the stereo pairs is known, the gradient and Gauss-Newton Hessian take on particularly simple forms. As this constraint is quadratic in $\mathbf{x}_i$, the Gauss-Newton Hessian of this term is fixed:

$$\mathbf{H}_E = \sum_{i}^{n-1} \mathbf{J}_E(i)^T \mathbf{J}_E(i), \tag{26}$$

where the k^{th} entry of the epipolar Jacobian is given by

$$\mathbf{J}_E(i)^k = \begin{cases} \mathbf{F}_0 \tilde{\mathbf{x}}_i^0 & \text{if } k = 2i \\ \mathbf{F}_1 \tilde{\mathbf{x}}_i^0 & \text{if } k = 2i+1 \\ 0 & \text{otherwise} \end{cases}, \tag{27}$$

with $\mathbf{F}_l$ denoting the l^{th} row of $\mathbf{F}$. The gradient of this constraint is then:

$$\mathbf{g}_E = \sum_{i}^{n-1} \left(\tilde{\mathbf{x}}_i^T \mathbf{F} \tilde{\mathbf{x}}_i^0\right) \mathbf{J}_E(i)^T \tag{28}$$

If $\mathbf{F}$ is unknown, we must optimize both over the landmarks as well as the fundamental matrix. However, optimizing in the space of fundamental matrices is non-trivial. Nonetheless, for a given set of landmarks in both images, an estimate of $\mathbf{F}$ can be easily obtained using the 8-point algorithm [18]. As such, the derivative of the fundamental matrix with respect to the landmark locations can be found through finite differences:

$$\frac{\partial \mathbf{F}}{\partial x_i} = \frac{\psi(\ldots, x_i + \delta, \ldots) - \psi(\ldots, x_i - \delta, \ldots)}{2\delta}, \tag{29}$$

where δ is a suitably small constant and ψ is the 8-point algorithm, taking as its input a set of corresponding landmarks and returning a fundamental matrix. With this, the k^{th} entry of the epipolar constraints' Jacobian is now given by:

$$\hat{\mathbf{J}}_E(i)^k = \tilde{\mathbf{x}}_k^T \frac{\partial \mathbf{F}}{\partial \mathbf{p}(k)} \tilde{\mathbf{x}}_k^0 + \mathbf{J}_E(i)^k. \tag{30}$$

The form of the gradient and Gauss-Newton Hessian are still those given in Equations (28) and (26), but now the Hessian is no longer fixed.

With the forms given above for the gradient and Hessians of each term, the gradient and Hessian of the cost function in Equation (22) can be computed by a weighted sum of the gradient and Hessians of the individual terms, with their entries reordered appropriately.

4.2. Unknown Fundamental Matrix

There may be scenarios where the fundamental matrix relating the stereo images in the sequence is not available. This may be the case if the cameras are moving or the intrinsic parameters change throughout the sequence, for example when zooming. For this, we propose minimizing the following cost function:

$$C = C_{D_1} + C_{D_2} + w_C C_C + w_S (C_{S_1} + C_{S_2}) + w_E C_E \quad (31)$$

for every frame in the sequence, with respect to landmarks in both images, $\mathbf{p}_1$ and $\mathbf{p}_2$, as well as six global lighting parameters. Here, C_{D_1} and C_{D_2} take the form of the data term in Equation (9), where the templates and landmarks used in Equation (5) are unique to each of the images in the stereo pair. C_{S_1} and C_{S_2} are smoothing terms given in Equation (14), one for each set of landmarks.

The term C_C is a color constancy term similar to that in Equation (23), however both images are now warped to the reference frame:

$$C_C = \sum_{\mathbf{x} \in \Omega} \rho \left(\left[\alpha_C \mathbf{I}_1(\vec{\mathbf{u}}) + \beta_C - \mathbf{I}_2(\vec{\mathbf{v}}) \right]^2 ; \sigma \right), \quad (32)$$

where $\vec{\mathbf{u}}$ are the warped pixel locations as given in Equation (6) with the warp parameterized by $\mathbf{p}_1 = [\mathbf{u}_1; \ldots; \mathbf{u}_n]$, the landmark locations in the first image of the stereo pair. Similarly, $\vec{\mathbf{v}}$ is that of the second image. The effect of this is that the Jacobian of this term takes a slightly different form from that in Equation (24):

$$\mathbf{J}_C(\mathbf{x}) = \left[\nabla \mathbf{I}_1(\vec{\mathbf{u}}) \frac{\partial \vec{\mathbf{u}}}{\partial \mathbf{p}_1} - \nabla \mathbf{I}_2(\vec{\mathbf{v}}) \frac{\partial \vec{\mathbf{v}}}{\partial \mathbf{p}_2}, \mathbf{I}_1(\vec{\mathbf{u}}), 1 \right]. \quad (33)$$

The term C_E is an epipolar constraint similar to that in Equation (25), however, now the landmarks in both images are variables:

$$C_E = \sum_{i}^{n-1} \left(\tilde{\mathbf{v}}_i^T \mathbf{F} \tilde{\mathbf{u}}_i \right)^2, \quad (34)$$

where $\tilde{\mathbf{u}}_i$ is the homogeneous coordinate of $\mathbf{u}_i$ and similarly for $\tilde{\mathbf{v}}_i$. As described in Section 4.1, for the case where $\mathbf{F}$ is unknown, we can find the derivative of this matrix with respect to the landmarks using finite differences (see Equation (29)). The Jacobian of this term is now twice as long, as we need to optimize over landmarks in both images, and takes the form:

$$\mathbf{J}_E(i)^k = \begin{cases} M_k + \mathbf{F}_0 \tilde{\mathbf{u}}_i & \text{if } k = 2i \\ M_k + \mathbf{F}_1 \tilde{\mathbf{u}}_i & \text{if } k = 2i+1 \\ M_k + \tilde{\mathbf{v}}_i{}^T \mathbf{F}^0 & \text{if } k = 2n+2i \\ M_k + \tilde{\mathbf{v}}_i^T \mathbf{F}^1 & \text{if } k = 2n+2i+1 \\ M_k & \text{otherwise} \end{cases}, \quad (35)$$

where $\mathbf{F}_l$ denotes the l^{th} row and $\mathbf{F}^l$ the l^{th} column of $\mathbf{F}$ and

$$M_k = \tilde{\mathbf{v}}_k^T \frac{\partial \mathbf{F}}{\partial \mathbf{p}(k)} \tilde{\mathbf{u}}_k. \quad (36)$$

4.3. Unknown Camera Calibration

When the fundamental matrix throughout the sequence is known, we can use exactly the same cost function as that in Section 4.2. The only difference here is in the evaluation of the gradient and Hessian of the epipolar constraint. Since the fundamental matrix is known, the k^{th} entry of the epipolar Jacobian in Equation (35) becomes:

$$\hat{\mathbf{J}}_E(i)^k = \mathbf{J}_E(i)^k - M_k \quad (37)$$

4.4. Known Scene Geometry

When the camera calibrations as well as the fundamental matrix are known, the projection matrices for each camera can be found [12]. With these, a set of 3D landmarks $\mathbf{p}^{3\text{D}}$ can be found from the refined 2D landmarks (see Section 4.1). Propagating these 3D landmarks throughout the sequence is advantageous for a number of reasons. Firstly, optimization needs only be performed over 3D landmarks rather than two sets of 2D landmarks, leading to a more efficient fitting process. Secondly, the smoothing term in Equation (14) is invariant to translation, which, when extended to 3D, has the effect that it is scale invariant in the image plane as well. Finally, as the geometry of the scene is known, we can do away with the epipolar constraints used in the methods described in Sections 4.2 and 4.3.

In this scenario, we propose minimizing the following cost function:

$$C = C_{D_1} + C_{D_2} + w_C C_C + w_S C_S \quad (38)$$

where C_{D_1} and C_{D_2} are data terms defined in Equation (9) for each image, C_C is the color constancy term in Equation (32) and C_S is a smoothness term as defined in Equation (14). The difference between these terms and the ones described previously is that here they depend on the 3D landmarks of the model, rather than the two sets of 2D landmarks. For example, the Jacobian of the data term in Equation (13) is now given by:

$$\mathbf{J}_D(\mathbf{x}) = \left[-\nabla I(\vec{\mathbf{x}}) \frac{\partial \vec{\mathbf{x}}}{\partial \mathbf{p}^{2\text{D}}} \frac{\partial \mathbf{p}^{2\text{D}}}{\partial \mathbf{p}^{3\text{D}}}, \mathbf{T}(\mathbf{x}), 1 \right], \quad (39)$$

where $\frac{\partial \mathbf{p}^{2\text{D}}}{\partial \mathbf{p}^{3\text{D}}}$ is the derivative of the 2D landmarks with respect to the 3D landmarks.

5. Experiments

To motivate our approach, we conducted two sets of experiments. The first is an evaluation of the monocular tracking method described in Section 3. For this, we used the

FGNet talking face database[1]. In the second set of experiments, we evaluate the utility of adding scene geometry into the fitting process. For this, we used the stereo database AVOZES [11]. In all experiments we set $w_C = 1$, $\nu = 10$, and w_E one order of magnitude larger than the typical total pixel error.

5.1. Monocular Method

The FGNet talking face database is a sequence of 5000 annotated images with 68 landmarks out of which we used the first 1000 to test the method described in Section 3. To evaluate the quality of the found correspondences we built a separate model for shape and texture. The shapes are first registered using Procrustes alignment to remove rigid motion from the data. PCA is applied separately to shape and texture to obtain their modes and magnitude of variation. Assuming the distributions of both shape and texture follow that of a degenerate Gaussian, the quality of the model can be assessed through the *compactness* of the resulting distribution [15]. This is approximated by the volume of covariance:

$$Q = \sum_{i=1}^{N} \lambda_i, \tag{40}$$

where λ_i is the eigenvalue of the i^{th} mode of variation. To avoid discarding different amounts of energy as noise in different trials, we use all non-zero eigenvalues. Unlike that suggested in [15], however, we argue that the quality of a model cannot be assessed through the shape or texture compactness independently. This is because the parameter settings at which Q_s, the shape compactness, is optimal generally disagrees with that of Q_t, the texture compactness.

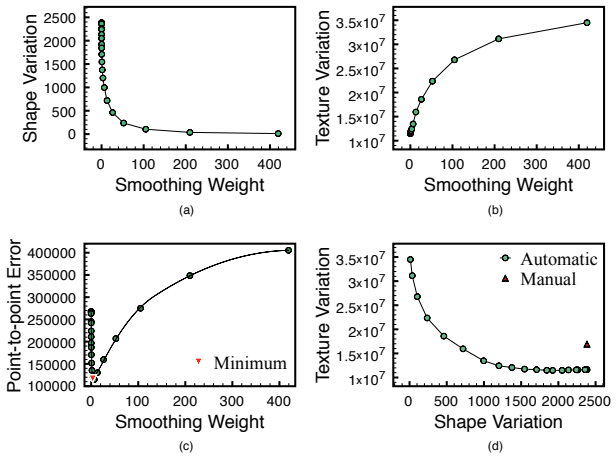

Figure 1. Monocular Tracking results:(a). Q_s vs. w_S (b). Q_t vs. w_S (c). Total point-to-point error vs. w_S (d). Q_s vs. Q_t.

[1] http://www.isbe.man.ac.uk/~bim/data/talking_face/talking_face.html

As such, we ran the monocular tracking algorithm on the database at a number of different settings of the smoothing weight w_S, the results of which are shown in Figure 1. Plots (a) and (b) illustrate the effect of different smoothing weights on Q_s and Q_t, respectively. As expected, as w_S is increased, Q_s improves but Q_t degrades. Plot (c) illustrates the effect of w_S on the total point-to-point error in all images between the supplied annotations with those obtained using our method. The optimal setting of the smoothing weight, with respect to the supplied annotations, is clearly visible as a minimum at $w_S = 23.84$. Increasing w_S beyond this value over-constrains the shape, leading to increasing point-to-point error. The relationship between Q_s and Q_t is illustrated in plot (d). It is worth noting here that models built using some settings of w_S actually exhibit better compactness compared to that built using the supplied annotations, confirming our observation that manual annotation is subjective and prone to error.

5.2. Stereo Methods

The AVOZES database consists of stereo sequences of 20 speakers uttering a variety of phrases, with each sequence ranging from 90 to 120 frames in length. For experiments in this section, we used four female and four male subjects from the continuous speech section of the database. The intrinsic and extrinsic camera parameters are supplied with the database, allowing the fundamental and projection matrices to be calculated. Annotations are not available for this database however, and hence we use it in a comparative setting between the monocular method and the three stereo methods proposed in Section 4. In the following, we will refer to the method in Section 3 as Method 1 and the three stereo extensions in Sections 4.2, 4.3, and 4.4 as Method 2,3 and 4 respectively.

We manually annotated one stereo pair in every sequence with 83 landmarks. Using this as an initial correspondence estimate, we refined the landmark locations in the right image using the method described in Section 4.1, both using the provided fundamental matrix and assuming it is unknown. The 3D reconstruction of one of the subjects in the database is shown in Figure 2. As expected, the reconstruction using the manual annotations without refinement yielded a poor reconstruction. More surprisingly however, the reconstruction using landmarks refined with the supplied fundamental matrix also resulted in a poor reconstruction. In fact, using landmarks refined with an unknown fundamental matrix resulted in a much more qualitatively pleasing reconstruction. On closer inspection, we found that the fundamental matrix built from the supplied intrinsic and extrinsic camera parameters is erroneous. To see this, in Figure 3 the epipolar line of a landmark in the left image is drawn on the right image. Notice that the line does not pass through the same physical point in the right image (outer

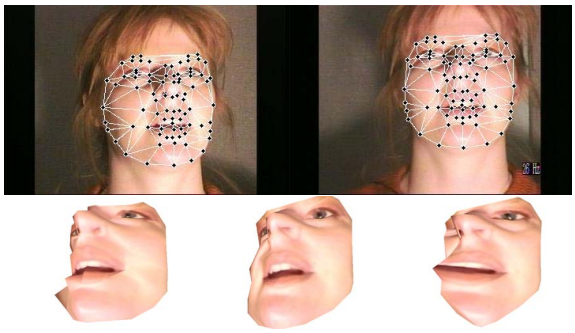

Figure 2. 3D Reconstruction of the reference image of subject f5 in AVOZES. Top row: stereo pair with manual annotations and triangulation shown. Bottom row (left to right): 3D reconstruction using manual annotations, refined annotations with unknown $\mathbf{F}$, refined with $\mathbf{F}$ from supplied camera parameters.

Figure 3. Epipolar error of supplied camera parameters. Left: Image with fixed landmark. Right: Epipolar line of point in left image.

corner of right eye). Although only of the order of two to three pixels, the error here has the effect that the landmarks are constrained to move along epipolar lines which do not contain the same image structures. This in-turn yields the color constancy assumption, used to drive the refinement process, invalid.

Nonetheless, we ran the monocular tracking method and its three stereo extensions on the eight sequences from the database at different settings of w_S. To allow direct comparison between results, the initial correspondence in the first frame for all methods is set to the landmarks refined with unknown $\mathbf{F}$. The monocular method was run independently on each of the streams in the stereo sequences. The quality of the model built using correspondences from each trial is calculated as described in Section 5.1 using images in both streams of the stereo sequences. Plots of Q_s against Q_t for all sequences are shown in Figure 4.

Methods 3 and 4, which utilized the supplied camera parameters, gave inconsistent results, outperforming Method 1 for some settings of w_S, for example in plot (e) of Figure 4, but inconclusive over the whole set of experiments. However, Method 2 consistently outperformed Method 1 in almost all trials, in some cases by a significant margin (see plots (b) and (d) in Figure 4). From this we conclude that incorporating scene geometry does provide a useful constraint for automatic correspondence finding in sequences, however it is sensitive to the accuracy of the assumed scene geometry.

6. Conclusion

We have presented a new method for automatic correspondence finding in image sequences by utilizing a tracking perspective. Three extensions to stereo sequences were also presented accounting for different stereo scenarios. Preliminary results show that the tracking paradigm works well in the monocular setting. We also found that epipolar constraints are capable of improving results further, however, they are sensitive to the accuracy of the camera parameters used.

There are a number of possibilities to improve on the current results. Temporal filtering of the landmark motion may improve the results. Also, utilizing an incremental linear model learning method, for example [17], rather than a fixed adaptive template may also improve the fitting process as variations in directions previously seen in the sequence can be accounted for by the texture model.

References

[1] S. Baker, R. Gross, and I. Matthews. Lucas-Kanade 20 Years On: A Unifying Framework: Part 3. Technical report, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, November 2003.

[2] S. Baker, I. Matthews, and J. Schneider. Automatic Construction of Active Appearance Models as an Image Coding Problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(10):1380–1384, October 2004.

[3] M. Black and P. Anandan. The Robust Estimation of Multiple Motions: Affine and Piecewise-smooth Flow Fields. Technical report, Xerox PARC, Dec 1993.

[4] A. Blake, M. Isard, and D. Reynard. Learning to Track Curves in Motion. In *IEEE Conf. Decision Theory and Control*, pages 3788–3793, 1994.

[5] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High Accuracy Optical Flow Estimation Based on Theory of Warping. In *8th European Conference on Computer Vision*, volume 4, pages 25–36, Prague, Czech Republic, May 2004. Springer-Verlag.

[6] H. Chui, L. Win, R. Schultz, J. S. Duncan, and A. Rangarajan. A Unified Non-rigid Feature Registration Method for Brain Mapping. *Medical Image Analysis*, 7(2):113–130, June 2003.

[7] T. F. Cootes, S. Marsland, C. J. Twining, K. Smith, and C. J. Taylor. Groupwise Diffeomorphic Non-rigid Registration for Automatic Model Building. In *European Conference on Computer Vision*, pages 316–327, 2004.

[8] T. F. Cootes, C. J. Twining, V. Petrovic, R. Schestowitz, and C. J. Taylor. Groupwise Construction of Appearance Models Using Piece-wise Affine Deformations. In *British Machine Vision Conference*, volume 2, pages 879–888, 2005.

[9] G. Edwards, C. J. . Taylor, and T. F. Cootes. Interpreting Face Images Using Active Appearance Models. In *IEEE Int. Conf. Automatic Face and Gesture Recognition*, pages 300–305, 1998.

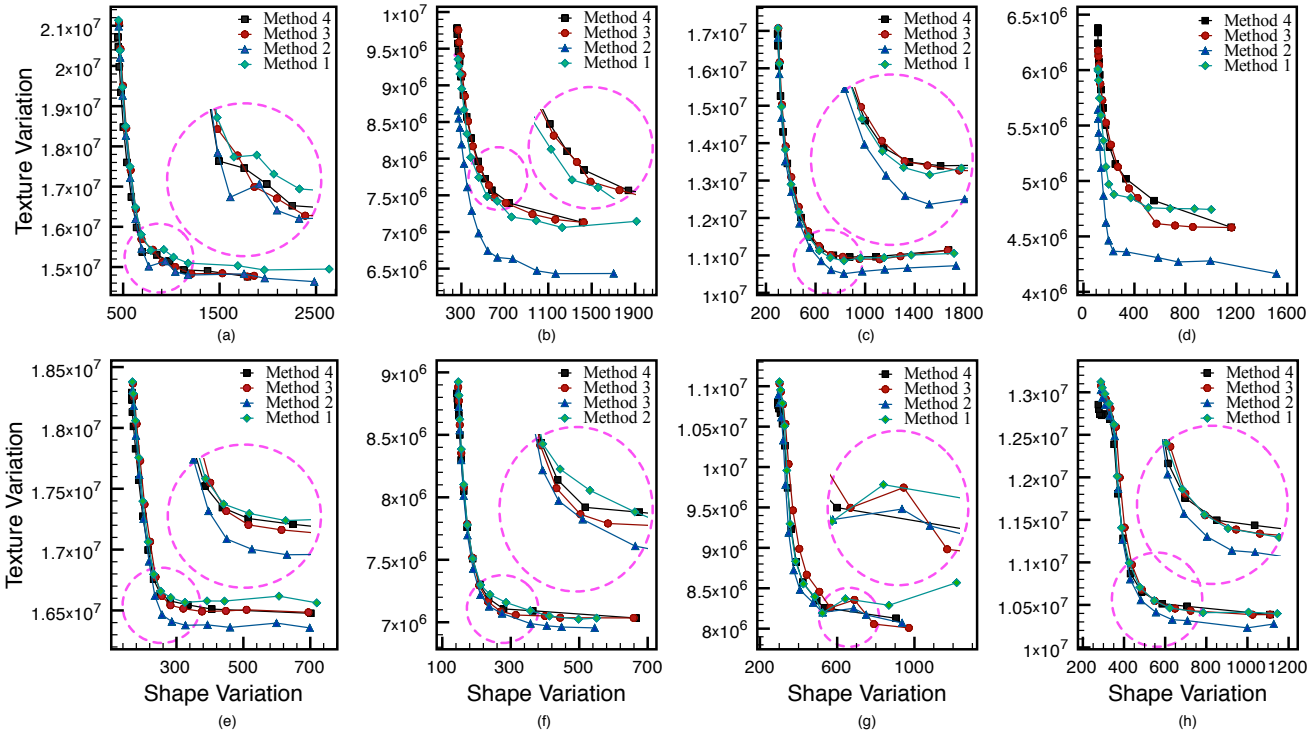

Figure 4. Comparisons of four methods for model building on eight AVOZES sequences, plotting Q_s against Q_t for a number of settings of w_S. From (a) to (h): subject code f3, f5, f9, f10, m1, m3, m4, m7. Upscaled sections of the graphs are shown for clarity.

[10] P. E. Frandsen, K. Jonasson, H. B. Nielsen, and O. Tingleff. *Unconstrained Optimization, 3rd Ed.* IMM, DTU, 2004.

[11] R. Goecke and J. B. Millar. The Audio-Video Australian English Speech Data Corpus AVOZES. In *8th International Conference on Spoken Language Processing INTERSPEECH 2004 - ICSLP*, volume III, pages 2525–2528, Jeju, Korea, October 2004. ISCA.

[12] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[13] A. Hill and C. J. Taylor. A Method of Non-rigid Correspondence for Automatic Landmark Identification. In *7th British Machine Vision Conference*, volume 2, pages 323–332, September 1996.

[14] C. Hu, J. Xiao, I. Matthews, S. Baker, J. Cohn, and T. Kanade. Fitting a Single Active Appearance Model Simultaneously to Multiple Images. In *Proceedings of the British Machine Vision Conference*, September 2004.

[15] T. Jebara. Images as Bags of Pixels. In *International Conference on Computer Vision*, pages 265–272, 2003.

[16] T. Lehn-Schiøler, L. K. Hansen, and J. Larsen. Mapping from Speech to Images Using Continuous State Space Models. In *Lecture Notes in Computer Science*, volume 3361, pages 136 – 145. Springer, Jan 2005.

[17] Y. Li. On incremental and Robust Subspace Learning. *Pattern Recognition*, 37(7):1509–1518, 2004.

[18] H. C. Longuet-Higgins. A Computer Algorithm for Reconstructing a Scene from Two Projections. *Nature*, 293:133–135, September 1981.

[19] G. Loy, R. Goecke, S. Rougeaux, and A. Zelinsky. Stereo 3D Lip Tracking. In *6th Int. Conf. on Control, Automation, Robotics and Vision ICARCV2000*, Singapore, December 2000.

[20] I. Matthews, T. Ishikawa, and S. Baker. The Template Update Problem. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(6):810–815, 2004.

[21] P. Mittrapiyanuruk, G. N. DeSouza, and A. C. Kak. Accurate 3D Tracking of Rigid Objects with Occlusion Using Active Appearance Models. In *WACV/MOTION*, pages 90–95, 2005.

[22] H. S. Sawhney and S. Ayer. Compact Representation of Videos through Dominant and Multiple Motion Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(8):814–830, 1996.

[23] M. B. Stegmann and H. B. Larsson. Fast Registration of Cardiac Perfusion MRI. In *International Society of Magnetic Resonance In Medicine*, page 702, Toronto, Canada, 2003.

[24] K. N. Walker, T. F. Cootes, and C. J. Taylor. Automatically Building Appearance Models from Image Sequences Using Salient Features. *Image and Vision Computing*, 20:435–440, 2002.

[25] Y. Zhong, A. K. Jain, and M. P. Dubuisson-Jolly. Object Tracking Using Deformable Templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(5):544–549, May 2000.