Face Recognition From Video using Active Appearance Model Segmentation

Nathan Faggian, Andrew Paplinski
Clayton School of Information Technology,
Monash University, Victoria, Australia,
{nathanf, app}@mail.csse.monash.edu.au

Tat-Jun Chin*
Institute for Vision Systems Engineering,
Monash University, Victoria, Australia,
tat.chin@eng.monash.edu.au

Abstract

Face recognition from video can be improved if good face segmentation of the subject under test is achieved. Many video based face recognition rely on simple background modeling and coarse alignment strategies for segmentation. This work presents a face recognition from video framework based on using Active Appearance Models (AAM) to achieve accurate face segmentation and consistent shape free representation across a video sequence. The segmentation provided by the AAM can be effectively normalized (morphed) to a mean shape. The resulting subimage can then be delivered to conventional face recognition from video algorithms for robust classification. We present preliminary results on a dataset of 17 individuals and outline the problems encountered in this approach.

1. Introduction

Traditionally, the problem of face recognition from video was approached from the still image perspective i.e. find an instance of a face in the video stream that is suitable under certain quality measures (e.g. pose, illumination etc.) and apply conventional still image-based recognition algorithms such as Eigenfaces or Fisherfaces. By incorporating AAM segmentation, these methods stand to benefit from the vast amounts of information available in a video sequence since non-frontal faces can be morphed into approximately frontal faces. Hence, a voting based technique (refer to [13]) would be more robust when more face images which are previously unusable due to unsuitable pose can contribute towards determining the identity of a face.

Recently, probabilistic approaches have gained popularity in video-based face recognition. For a recent survey, see [5]. Most of these approaches essentially formulate face tracking and classification in the same probability framework. By performing fast face detection followed by AAM segmentation, we can decouple tracking from recognition. This can potentially improve the overall performance since tracking/segmentation and recognition are delegated to more specialized algorithms. However, this "divide-and-conquer" approach does not necessarily mean that we have to resort to still-image-based methods for classification. Methods such as [12, 1] which distinguish face image sets (from video sequences) can capitalize on accurate face segmentation to arrive at better performance.

Edwards et al [7] presented the Active Appearance Model (or AAM) as a method to model objects in images. It is a modeling by synthesis approach to image analysis and is a popular technique that has been broadly used in the field of computer vision; specifically in the domain of facial modeling. An AAM represents an encoding of both the shape and texture information of the object, where the goal is to be able to estimate any valid instance of the object using PCA based parametric models:

$$\widehat{\mathrm{s}} = \bar{\mathrm{s}} + \mathrm{S}\alpha \ , \quad \widehat{\mathrm{t}} = \bar{\mathrm{t}} + \mathrm{T}\beta \ . \tag{1}$$

The shape and texture models encode the modes of variation that the hand-labeled training samples provide. A new shape $\widehat{\mathrm{s}}$ or texture $\widehat{\mathrm{t}}$ can be constructed as a linear combination (α or β) of the principle components (column space) of the measurement matrices for shape and texture.

2. AAM Fitting

Fitting the texture and shape models for an AAM is a complex nonlinear problem; because pixels and their locations in images are generally not related [2]. Fitting an AAM is much like the nonlinear optimization applied to image alignment. In fact; image alignment algorithms can be directly applied to the fitting of an AAM. In this case the AAM (shape component) is represented as a special transform, termed a piecewise affine transform. Image alignment is the process of aligning an image and a constant template

*Tat-Jun Chin is a recipient of the Australia-Asia Awards 2004 conferred by the Department of Education, Science and Training (DEST) of the Australian Government.

Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

through a nonlinear optimization process, this boils down to an application of Gauss-Newton [4] and was first proposed by Lucas and Kanade [10] to solve affine relationships. In the context of AAMs a search for a constant template is akin to a search for the shape parameters that minimize the difference between a rendered AAM ($\mathbf{I}(\alpha, \beta)$) and the image ($I$), optionally the texture parameters can also be searched for. Fortunately the common methods for image alignment can be applied to the special piecewise affine transform that is used in AAMs; the quickest being Inverse Compositional Image Alignment (or ICIA).

2.1. ICIA Fitting

ICIA is a fitting method that is based on the earlier forwards additive method [10] although the roles of the template and the image is reversed. By reversing the role of the template and the image for Gauss-Newton image alignment, a large amount of the computation (specifically related to the Jacobian) is constant, and thus precomputed. This results in a very efficient and different alignment algorithm. ICIA differs from the earlier additive method in three important ways: 1) ICIA is a compositional method, the current parameter estimate is multiplied with the update, 2) It is a search for parameters in the inverse direction, since the template and image roles are reversed, 3) The majority of its computation degenerates to simple linear operators (matrix multiplications, additions and subtractions). When applied to AAMs the ICIA optimizes for two components that define the shape of the AAM: 1) The local vertex transforming warp, α, 2) The global transforming warp, γ. This is parameterized in the following manner:

$$\mathrm{W}(\bar{\mathrm{s}}; \alpha) \neq \widehat{\mathrm{s}} = \bar{\mathrm{s}} + \mathrm{S}\alpha \ , \quad \mathrm{N}(\bar{\mathrm{s}}; \gamma) \neq \widehat{\mathrm{s}} = \bar{\mathrm{s}} + \mathrm{S}^{*}\gamma \ . \quad (2)$$

It is important to note that when using the "projected out" method for AAM fitting described by Matthews and Baker [2, 3], it is only necessary to fit with respect to shape, and once shape is determined solving the texture becomes a trivial projection into the shape-free texture space.

We implemented ICIA the fast AAM fitting method of Baker and Matthews [2, 3], using the same additions as Faggian [8], 1) Multi-resolution fitting was implemented, 2) QR decomposition was used to improve the shape parameter estimate, 3) The Levernberg-Marquadt method was applied to improve convergence rates.

2.2. Template drift

Template drift is a common problem in image based tracking. In the context of AAMs template drift occurs when errors of identity and camera parameters in an incorrect AAM fitting propagates to the next fitting stage. Over time the error in estimates increases and the AAM looses track of the object.

In our work we experimented with two possible methods for avoiding template drift. The first was to treat the problem as a recursive filtering problem by using AAM segmentation to determine eye centroids. This is then used as the feature vector x input to a kalman-filter [9], where the assumption is made that bad fittings of the AAM will result in eye locations which vary greatly from their true position. When the difference (Euclidean norm) between the kalman estimate z and the measurement x is too large then the system is re-initialized using the kalman estimate to align the model. The second and more simplistic approach was simply to use eye detections every 10 video frames to re-initialize the model. At this stage this was a set of hand-labeled eye locations, but other eye-detection algorithms can certainly be applied. During the segmentation process the second method was found to be more stable.

3. Face Recognition from Video

To highlight the benefits of AAM segmentation, we implemented two face recognition from video approaches. We first employ the Eigenfaces method in the context of video based recognition which is described as follows. Assume we have a frontal face image database $\mathrm{X} = \{\mathrm{x}_1, \cdots, \mathrm{x}_\mathrm{n}\}$ of m different faces, with $\mathrm{Y} = \{\mathrm{y}_1, \cdots, \mathrm{y}_\mathrm{n}\} \in \{1, \cdots, \mathrm{m}\}$ denoting the respective classes of the elements of X. PCA is performed on X to obtain a low dimensional eigenspace $\mathcal{Z}$ with orthogonal basis Z which maximizes the variance of X. Using AAM segmentation on a test video (with only one face present), a set of v test images $\mathrm{T} = \{\mathrm{t}_1, \cdots, \mathrm{t}_\mathrm{v}\}$ is acquired. The segmented faces are morphed into the mean shape and resized to have the same size as the elements of X. Both T and X are then projected onto $\mathcal{Z}$ via

$$\mathrm{T}_\mathrm{Z} = \mathrm{Z}^\mathrm{T}(\mathrm{T} - \bar{\mathrm{x}}) \ , \quad \mathrm{X}_\mathrm{Z} = \mathrm{Z}^\mathrm{T}(\mathrm{X} - \bar{\mathrm{x}}) \ , \quad (3)$$

where $\bar{\mathrm{x}}$ denotes the mean of X. Classification can be performed within $\mathcal{Z}$ using the nearest neighbour rule with T_Z as inputs and X_Z as class prototypes. Given the individual decisions as $\{\mathrm{c}_1, \cdots, \mathrm{c}_\mathrm{v}\}$, by simple majority voting the overall decision for the test video sequence is

$$\mathrm{C} = \arg\min_{\mathrm{y}} \{ \sum_{\mathrm{i}=1}^{\mathrm{v}} |\mathrm{c}_\mathrm{i} - \mathrm{y}| \, , \ \forall \ \mathrm{y} \in \mathrm{Y} \} \ . \quad (4)$$

A confidence rating for each possible outcome y can be computed as

$$\mathrm{R}_\mathrm{y} = \frac{1}{\mathrm{v}} \sum_{\mathrm{i}=1}^{\mathrm{v}} \delta(\mathrm{c}_\mathrm{i} - \mathrm{y}) \ , \quad (5)$$

Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

with $\delta(\mathrm{x}) = 0$ for $\mathrm{x} \neq 0$ and $\delta(\mathrm{x}) = 1$ for $\mathrm{x} = 0$. $\mathrm{R_y} = 1$ and $\mathrm{R_y} = 0$ signifies absolute confidence and no confidence respectively for class y. Alternatively, a more sophisticated probabilistic voting [11] can be applied.

Secondly, we consider the Mutual Subspace Method (MSM) of [12] for face recognition from video. We start by acquiring sets of face images $\{\mathrm{X_1}, \cdots, \mathrm{X_n}\}$ (not necessarily frontally posed) segmented from video sequences, with the faces in each set bearing the same identity. Face recognition is performed on the premise that different faces occupy different subspaces of the image space. To estimate an r-dimensional face-specific subspace $\mathcal{F}_\mathrm{i}$ of $\mathrm{X_i}$, a rank-r Singular Value Factorization can be invoked on $\mathrm{X_i}$,

$$\mathrm{X_i^r} = \mathrm{U_i^r}\Sigma_\mathrm{i}^\mathrm{r}(\mathrm{V_i^r})^\mathrm{T} \ , \tag{6}$$

with $\mathrm{X_i^r}$ being the rank-r approximation of $\mathrm{X_i}$ and $\mathrm{U_i^r}$ being the basis of the face subspace required. Given another face subspace $\mathcal{F}_\mathrm{j}$ of $\mathrm{X_j}$ with basis $\mathrm{U_j^r}$, the distance between these subspaces can be used for classification. First, we compute the matrix $\mathrm{K} = (\mathrm{U_i^r})^\mathrm{T}\mathrm{U_j^r}$. By invoking the SVD, we obtain the principal angles between the subspaces:

$$\mathrm{K} = \mathrm{U_K}\Sigma_\mathrm{K}(\mathrm{V_K})^\mathrm{T} \tag{7}$$

with $\Sigma_\mathrm{K} = \mathrm{diag}\{\cos\theta_1, \cdots, \cos\theta_\mathrm{r}\}$, with $\{\theta_1, \cdots, \theta_\mathrm{r}\}$ being the principal angles. Functions on principal angles, for example

$$\mathrm{d}(\mathcal{F}_\mathrm{i}, \mathcal{F}_\mathrm{j}) = \min\{\theta_1, \cdots, \theta_\mathrm{r}\} \ , \tag{8}$$

can be used as distance measures for classification. AAM can help in improving robustness by allowing accurate segmentation of faces unachievable by using just face detectors. By incorporating only relevant pixels of faces, face subspaces can be estimated without being biased by unnecessary information from the hair, clothing and background. To build the face subspaces on-the-fly, incremental SVD can be applied. Refer to [6] for more details.

4. Video Database

We collected a 17-subject video database for experiments. Each video contains 1000 frames and was obtained using a high-resolution firewire camera at VGA resolution. At the beginning, each subject was asked to carry out a predefined set of actions that put them through a number of varied poses and expressions. Subsequently the subjects were not given any directions and were allowed to do what they liked in front of the camera until recording ended. Approximately 30% of faces in the videos are frontally posed. Several examples of the videos are shown in Figure 1. Each 1000-frame video was then partitioned into contiguous 100-frame sequences of which the first was used for constructing AAMs and training face video classifiers. The rest of the 100-frame sequences were used for tests.

5. Results

A set of 17 person-specific AAMs where constructed to segment the faces from the video sequences. These were constructed using only 8 images (Figure 1) from the training sequences. As a benchmark for the proposed face recognition from video using AAM segmentation, we used the AdaBoost face detector as an alternative segmentation method (see Figure 3). In each case there is possibility for segmentation error, in the case of slight pose changes the face detector can fail, and in the case of extreme pose changes the AAM can also fail. This is illustrated by Figures 2(a),2(b) and 2(c). AdaBoost managed to segment roughly 65% of the faces available in all the video sequences, while AAM managed approximately 80%. By using the face segmentations of the training sequences, Eigenface and MSM classifiers were trained. The Eigenfaces were obtained by invoking a PCA on a database of segmented frontal face images of the subjects (10 per training sequence). For the MSM, 25-dimensional face-specific subspaces were estimated from all the segmented images in the training sequence. The test sequences were then evaluated using the approaches detailed in Section 3. In total, 153 (17×9) tests are available using our database. Table 1 presents the results obtained.

Method	Segmentation	Correct (%)	Undecided (%)
EF	AB	62.75 (96/153)	35.95 (55/153)
MSM	AB	84.97 (130/153)	11.76 (18/153)
EF	AAM	92.16 (141/153)	3.27 (5/153)
MSM	AAM	94.77 (145/153)	1.96 (3/153)

Table 1. Face recognition from video results. EF = Eigenface, AB = AdaBoost.

When the best match for a video sequence failed to attain a pre-determined confidence measure, the system was instructed to decline classification and the label "Undecided" was used. For the Eigenface method, the threshold was at least 80% confidence rating (see Section 3). For the MSM, the threshold was set at 0.15 of the distance measure defined by Equation (8), and the closest subspace to the test subspace must have a distance below this value to be considered a match.

6. Conclusion

AAM segmentation improves the performance of face recognition from video in the two methods we have implemented. In the case of the EigenFace method, an improvement of 29.41% in terms of correct matches is achieved, because of the normalization of shape. In the case of the

Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

(a) Ideal AAM segmentation. (b) Face detection failure. (c) AAM segmentation failure.

Figure 2. Examples for AAM segmentations.

Figure 1. Several example frames of our video database.

Figure 3. Row 1: Face segmented by AdaBoost. Row 2: Face segmented and normalized by AAM. Row 3: Face segmented by AAM but not normalized.

MSM, by removing significant background clutter the subspace estimations are less effected by noise and an improvement of 9.8% in terms of correct matches is achieved. In addition, our experiments confirmed that the MSM is better than Eigenface in the context of face recognition from video.

7. Acknowledgments

We would like to express our appreciation to the video database subjects for their kind contribution and the Australian Research Council for its continued funding.

References

[1] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *CVPR*, 2005.

[2] S. Baker and I. Matthews. Lucas-kanade 20 years on; a unifying framework: Part1. Technical Report 16, Robotics Institute, Carnegie Mellon University, 2002.

[3] S. Baker and I. Matthews. Lucas-kanade 20 years on; a unifying framework: Part2. Technical Report 01, Robotics Institute, Carnegie Mellon University, 2003.

[4] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[5] R. Chellappa and S. Zhou. *Handbook of face recognition*, chapter 9: Face tracking and recognition from video. Springer, 2005.

[6] T.-J. Chin, J. U, K. Schindler, and D. Suter. Face recognition from video by matching image sets. In *Digital Image Computing, Techniques and Applications*, 2005.

[7] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *Proc. European Conference on Computer Vision*, volume 2, pages 484–498. Springer, 1998.

[8] N. Faggian, S. Romdhani, J. Sherrah, and A. Paplinski. Color active appearance model analysis using a 3D morphable model. In *Digital Image Computing: Techniques and Applications*, December 2005.

[9] E. Kalman, Rudolph. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[10] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[11] S. McKenna and S. Gong. *Recognising moving faces*, pages 578–588. SpringerVerlag, 1998.

[12] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *International Conference on Automatic Face and Gesture Recognition*, pages 318–323, 1998.

[13] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld. Face recognition: a literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.

Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE