Robust Open-Set Face Recognition for Small-Scale Convenience Applications

Hua Gao, Hazım Kemal Ekenel, Rainer Stiefelhagen

Institute for Anthropomatics
Karlsruhe Institute of Technology
Karlsruhe, Germany
Email: {hua.gao, ekenel, rainer.stiefelhagen}@kit.edu

Abstract. In this paper, a robust real-world video based open-set face recognition system is presented. This system is designed for general small-scale convenience applications, which can be used for providing customized services. In the developed prototype, the system identifies a person in question and conveys customized information according to the identity. Since it does not require any cooperation of the users, the robustness of the system can be easily affected by the confounding factors. To overcome the pose problem, we generated frontal view faces with a tracked 2D face model. We also employed a distance metric to assess the quality of face model tracking. A local appearance-based face representation was used to make the system robust against local appearance variations. We evaluated the system's performance on a face database which was collected in front of an office. The experimental results on this database show that the developed system is able to operate robustly under real-world conditions.

1 INTRODUCTION

There is a large demand for building robust access control systems for the safety and convenience of modern society. Among different biometric identification methods, face recognition is a less obtrusive technique which does not require too much cooperation of the users. Due to the absence of user cooperation and due to uncontrolled conditions, building a robust real-world face recognition system is still a challenging task and has attracted broad interest. In real-world systems, there are many sources of variabilities in facial appearance which may significantly degenerate the performance of the system. The major four confounding factors are pose, illumination, expression, and partial occlusion (i.e. glasses or facial hair). In many previous systems, numerous approaches have been proposed to deal with one or two specific aspects of variations [1–3].

The active appearance model (AAM) [4] was proposed as a 2D deformable face model for modeling pose changes, facial expression and illumination variations. The shape of the face model is optimized by minimizing the texture representation error. Once the model is fitted on an input image, the corresponding model parameter vector can be used as a feature vector. In [5], linear discriminant analysis (LDA) was utilized to construct a discriminant subspace for face identification. Later in [6], the authors found that the shape parameters of AAMs can be used to estimate the pose angle. A frontal view of an input face image can be synthesized by configuring the shape parameters that control the pose. Face recognition with this pose correction method was evaluated by Guillemaut et al. in [7]. However, they corrected the appearance of the rotated faces with some nonlinear warping techniques instead of synthesis. This texture warping approach has the advantage of preserving the textural information such as moles and freckles contained in the original image, which will be lost in the synthesis-based approach in which small local details may not be modeled.

In this paper, we present a face recognition system which is robust against the aforementioned confounding factors. Three key techniques are employed to achieve the goal of this system: 1) A generic AAM is used to track a set of facial landmarks on the face images. With these facial landmarks, we generate shape-free frontal view faces with a nonlinear piecewise affine warping. The variations in pose and expression are normalized to a canonical shape. 2) A distance metric is employed to assess the quality of AAM fitting. According to this distance metric, we filtered out some frames where the model fitting failed. 3) A local appearance-based face representation is used for face recognition. This representation is considered to be invariant to local appearance changes such as expression, partial occlusion, and illumination changes [8]. Experiments in [9] showed that this approach is also robust against the errors introduced in the face model fitting.

The presented open-set face recognition system is suitable for small-scale convenience applications, which can be easily customized for a small group of people such as family members or laboratory members. The system identifies a person in question and conveys customized information or provides personalized services according to the identity of the person. An example system can be a smart TV set, which is able to show personalized TV programs according to the identity of the person in front of the television. It can also be integrated in a smart household robot so that it can identify the family members and customize the dialogue. The prototype application for this presented system is a visitor interface. The system is mounted in front of an office. A welcome message is displayed on the screen. When a visitor appears in front of the system before knocking on the door, the system identifies the visitor unobtrusively without any special cooperation. According to the identity of the person, the system customizes the information that it conveys about the host. For example, if the visitor is unknown, the system only displays availability information about the host. On the other hand, if the person is known, depending on the identity of the person, more detailed information about the host's status is displayed.

The remainder of this paper is organized as follows. In Section 2, we describe the implementation details for building a robust open-set face recognition system. We present the evaluation procedure and discuss the experimental results in Section 3 and give concluding remarks in Section 4.

2 METHODOLOGY

This section explains the processing steps of the developed robust open-set face recognition system.

2.1 Active Appearance Models and Model Fitting

The AAM is a generative parametric model which utilizes both shape and appearance information to represent a certain object such as the human face. A shape in AAM is defined as a set of normalized 2D facial landmarks. An instance of the linear shape model can be represented as $s = s_0 + \sum_{i=1}^{n} p_i s_i$, where s_0 is the mean shape, s_i is the i^{th} shape basis, and $\mathbf{p} = [p_1, p_2, \ldots, p_n]$ are the shape parameters. The appearance model is defined inside the mean shape, which explains the variations in appearance caused by changes in illumination, identity, and expression, etc. It represents an instance appearance as $A = A_0 + \sum_{i=1}^{m} \lambda_i A_i$, where A_0 is the mean appearance, A_i is the i^{th} appearance basis, and $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_m]$ are the appearance parameters.

Given an input facial image $I(\mathbf{x})$, the goal of fitting an AAM is to find the optimal model parameters such that the synthesized model appearance is as similar to the image observation as possible. This leads to a minimization of a cost function defined as:

$$E = \sum_{\mathbf{x} \in s_0} \left[I(\mathbf{W}(\mathbf{x}; \mathbf{p})) - A(\mathbf{x}, \lambda) \right]^2, \tag{1}$$

where $I(\mathbf{W}(\mathbf{x}; \mathbf{p}))$ is the warped input facial image, and $A(\mathbf{x}, \lambda)$ is the synthesized appearance instance. The minimization problem is usually solved by gradient descent methods, which iteratively estimate the incremental update of the shape parameter $\varDelta\mathbf{p}$ and the appearance parameter $\varDelta\lambda$, and update the current model parameters respectively. The Inverse Compositional (IC) and Simultaneously Inverse Compositional (SIC) methods proposed by Baker and Mathews [10] formulated the problem in a more efficient way, where the role of the appearance template and the input image is inversed when computing $\varDelta\mathbf{p}$. The shape is updated by composing an inverse incremental warping $\varDelta W(\mathbf{x}; \mathbf{p})$ which is estimated where $\mathbf{p} = 0$. This framework enables the time-consuming steps of parameter estimation to be pre-computed outside of the iterations. We implemented the SIC fitting algorithm for its better generalization ability to unseen data [11].

It is known that the gradient-descent-based optimization problem usually requires a reasonable initialization. A poor initialization may cause the search to get stuck in a local minimum. We used the responses of a face and eye detector based on Viola & Jones' approach [12] to initialize the face shape with a 2D similarity transformation. This initialization usually suffices for fitting frontal faces. However, when fitting semi-profile faces, part of the initialized shape does not cover the face. Thus the optimization can be affected by the included background pixels. To avoid bad initialization, we adopted a two stage progressive model fitting as used in [9].

2.2 Fitting While Tracking

In the context of tracking AAMs in video sequences, the model parameter $\{\mathbf{p_t}, \lambda_\mathbf{t}\}$ at time t can be initialized with the successfully optimized parameter $\{\mathbf{p_{t-1}}, \lambda_\mathbf{t-1}\}$ at time $t-1$. To improve AAM fitting in video sequences, Liu [13] extended the SIC algorithm to enforce the frame-to-frame registration across video sequences. The proposed algorithm assumes that the warped images will not change over a few frames. This assumption is considered as a prior to the cost function of the SIC, which constrains the parameter searching to the right direction.

The cost function is then reformulated as follows:

$$\sum_{\mathbf{x}} \left[I_t(\mathbf{W}(\mathbf{x};\mathbf{p})) - A(\mathbf{x},\lambda)\right]^2 + k \sum_{\mathbf{x}} \left[I_t(\mathbf{W}(\mathbf{x};\mathbf{p})) - M_t(\mathbf{x})\right]^2, \tag{2}$$

where the first term is the fitting goal of the SIC algorithm. The prior term is defined as the sum of squared error between the current warped image $I_t(\mathbf{W}(\mathbf{x};\mathbf{p}))$ and the appearance from previous frames, $M_t(\mathbf{x})$. For simplicity, we define $M_t(\mathbf{x})$ as the warped image of the video frame at time $t-1$:

$$M_t(\mathbf{x}) = I_{t-1}(\mathbf{W}(\mathbf{x};\mathbf{p}_{t-1})). \tag{3}$$

The benefit of this term is clear; it presents the specific appearance information of the subject being fitted, which may not be modeled by the generic face models. This information can compensate the mismatch between the face models and the input images being fitted.

2.3 Tracking Quality Assessment

A simple way to verify the result of the fitting is to check the residual error, which is also been considered as the stop criterion for the fitting algorithm. The residual error indicates the reconstruction error of the eigenspace decomposition and the measure is referred to as the "distance from feature space" (DFFS) in the context of "eigenfaces". However, the error of the AAM fitting is composed of the reconstruction error and the search error. The residual error alone is not able to assess the quality of the fitting results. Eigenfaces, especially higher-order ones, can be linearly combined to form images which do not resemble faces at all. In this sense, the coefficients of the eigenfaces should also be taken into consideration. For this purpose we employed a modified DFFS definition which was introduced in [14]:

$$DFFS(\lambda_1, \ldots, \lambda_m, \varepsilon) = K \times (\sum_{i=1}^{m}\{\frac{\lambda_i^2}{\sigma_i^2}\} + \frac{\varepsilon^2}{\sigma_{residue}^2}). \tag{4}$$

Here $\varepsilon = \mathbf{I_x} - (A_0 + \sum_{i=1}^m \lambda_i A_i)$ is the residue, and $\sigma_i = \max_{t\in T}|\lambda_{t,i}|$, $\sigma_{residue} = \max_{t\in T}|\varepsilon_t|$, which correspond to the worst outliers of the weights and residue in the training set T. Note that K is an arbitrary constant used as a scale factor.

The modified DFFS value can be used to assess the quality of the AAM fitting result since it yields low values for good face model fitting and high values for poor fitting. It is not exactly zero for a perfect fitting since only the mean face is located precisely in the center of the cloud representing the distribution. All the other face instances have a distance from the center of the cloud and thus have a non-zero value. If the value is larger than a threshold τ, the AAM tracking will be re-initialized.

2.4 Pose Normalization

The most straightforward method to normalize the pose of a face image is the piecewise affine warping which is also used in the fitting algorithm for sampling the texture inside the face mesh. The warp is realized by mapping the pixels in each fitted triangular mesh s to the base mesh s_0. For each pixel $\mathbf{x} = (x, y)^T$ in a triangle in the base mesh s_0, it can find a unique pixel $\mathbf{W}(\mathbf{x}; \mathbf{p}) = \mathbf{x}' = (x', y')^T$ in the corresponding triangle in the mesh s with bilinear interpolation. The implementation for the piecewise affine warp is detailed in [10]. Another nonlinear warping technique based on thin-plate splines (TPS) was also studied in [9]. However, the recognition based on piecewise affine warping outperforms TPS despite its simplicity.

2.5 Open-set Face Recognition

Face Representation The pose normalized facial images are masked with the AAM mean shape. All salient facial features are warped to the canonical locations. However, feature points around the chin area might be misaligned, which may create strong edges. As demonstrated in [15], the chin area does not contribute too much discriminative information compared to other facial features. For this reason, we cropped the chin area in the pose normalized facial image. Following the approach in [8], we scaled the cropped images to 64×64 pixels size and then divided them into 64 non-overlapped blocks of 8×8 pixels size. On each local block, the discrete cosine transform (DCT) is performed. The obtained DCT coefficients are ordered using a zig-zag scanning. The first component is skipped because it represents the average pixel intensity of the entire block. The following five low frequency coefficients are retained which yields a five dimensional local feature vector. This local feature vector is then normalized to unit norm. Finally, the feature vectors extracted from each block are concatenated to construct the global feature vector. For details of the algorithm please see [8].

Classification Open set face recognition is different from the traditional face identification in that it also involves rejection of impostors in addition to identify accepted genuine members that are enrolled in the database. We formulate the open-set face recognition as a multiple verification problem as proposed in [16]. Given a claimed identity, the result of an identity verification is whether the claimed identity is accepted or rejected. Given a number of positive and negative samples it is possible to train a classifier that models the distribution of faces for both cases. Based on this idea, we trained an identity verifier for every one of the n known subjects in the gallery using support vector machines (SVMs). Once a new probe is presented to the system, it is checked against all classifiers; if all of them reject, the person is reported as unknown; if one accepts, the person is accepted with that identity; if more than a single verifier accepts, the identity with the highest score wins. Scores are linearly proportional to the distance to the hyperplane of the corresponding SVM classifier.

Temporal fusion Since a person's identity does not change during the video capture, we can enforce temporal consistency. In order to make it possible to revise a preliminary decision later on, instead of relying on a single classification result for every frame an n-best list is used. N-best lists store the first n highest ranked results. We choose $n = 3$ in this work. For each hypothesis a cumulated score is stored that develops over time. If the face model tracking fails over multiple frames, the cumulated scores are reset assuming that the person has left or is not facing the camera. Resetting scores allows the whole process to restart once a face is located and the face model tracking is successfully initialized.

3 EXPERIMENTS

To evaluate the performance of the presented system, we carried out experiments on the database collected in [16]. Totally 54 people were recorded in front of an office, with natural or artificial lighting conditions depending on the time of the day. These recordings were split into two groups, a group of known people and a group of unknown people. Different sets of data were used for training and testing. Known people's recordings were split into training and testing sessions which do not overlap. Unknown subjects used for training are different from those used for testing. Fig. 1 depicts some example frames in the database. The recorded subjects were free to move, even out the sight of the camera. The depicted example frames show different recording conditions as well as subjects with various poses.

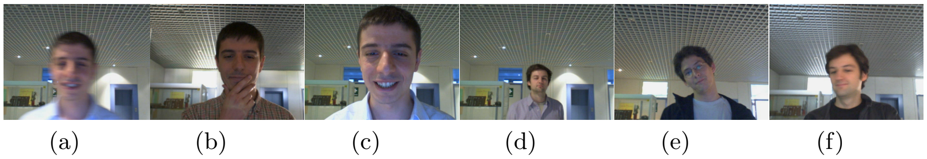

(a) (b) (c) (d) (e) (f)

Fig. 1. Recordings from the data set, different illumination and face sizes. (a) Artificial light, motion blur. (b) Day light, dark, partial occluded. (c) Artificial light, bright, near. (d) Artificial light, far away. (e) Head rotate in plane. (f) Head rotate in depth.

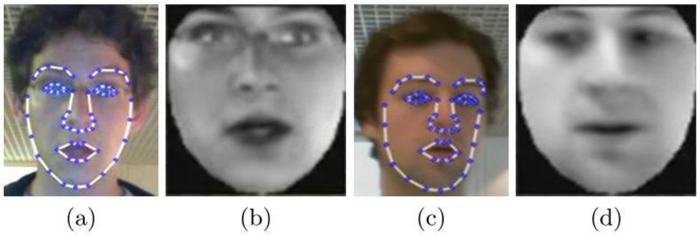


Fig. 2. Pose normalization. (a),(c) Face images overlaid with tracked AAM shape. (b),(d) Pose normalized face images.

3.1 Performance Measure for Open-set Face Recognition

In closed-set face recognition, the false classification rate (FCR) is a common measure for evaluating the performance of the system. However, in open-set face recognition, there are two more types of error which can occur. The impostors can be erroneously accepted by the system (false accept rate (FAR)), while the genuine members can also be rejected (false rejection rate (FRR)). All the three errors have to be traded-off against each other, as it is not possible to minimize them at the same time. The equal error rate (EER = FAR = FCR+FRR) performance measure is employed to trade off against the three error terms. A system with a lower EER is considered to be more robust and accurate.

SVMs automatically minimize the overall error and try to find the global minimum. To fine tune the system performance to equalize the accept error and reject error, the receiver operating characteristic (ROC) curve for SVM-based classification was created by using a parameterized decision surface. The decision hyperplane $\{x \in S : wx + b = 0, (w, b) \in S \times R\}$ is modified to $wx + b = \Delta$, where Δ allows to adjust the FAR and the CCR (correct classification rate) accordingly. A polynomial kernel with degree 2 is used as in [16].

3.2 Performance Comparison

Table 1 lists the data configuration for training and testing. The unknown training data was down-sampled so that the total number of frames from unknown subjects and each known subject is balanced. Note that only the frames fitted under a certain threshold of the modified DFFS value were accepted for training and testing. The threshold was selected empirically so that it discards all possible failed fittings. Here we choose the threshold $\tau = 10.0$. After face pose normalization through the known training sequences, we obtained approximately 600 known training samples for each subject. For the unknown training, 25 sessions of different subject were used, each session was under-sampled to 30 frames. Sample pose normalized face images are plotted in Fig. 2.

We first started with frame-based classification. The results are listed on the first row of Table 2 where $\Delta = -0.108$ and EER = 3.0%. The ROC curve which plots the correct classification rate against the FAR is illustrated in Fig. 3(a).

Training data		
Known	5 subjects	4 sessions and ≈ 600 frames per person
Unknown	25 subjects	1 session, 30 frames per subject
Testing data		
Known	5 subjects	3-7 sessions per person
Unknown	20 subjects	1 session per person

Table 1. Data set for open set experiments

Another frame-based test was also carried out on the same data set to verify the effectiveness of the pose correction. Instead of synthesizing a face with a piecewise affine warp after AAM fitting, a simple Euclidean transformation was performed according to the eye center coordinates in the fitted AAM shape. The corresponding results are listed on the third row of Table 2 where EER = 4.0% and $\mathit{\Delta} = -0.167$. The ROC curve for this test is depicted in Fig. 3(b).

Alignment	Classification	CCR	FRR	FAR	CRR	FCR
AAM	Frame-based	97.1%	2.7%	3.0%	97.0%	0.2%
	Progressive-score	100.0%	0.0%	0.0%	100.0%	0.0%
Euclidean trans.	Frame-based	95.8%	4.1%	4.0%	96.0%	0.1%
	Progressive-score	99.7%	0.3%	0.4%	99.6%	0.0%

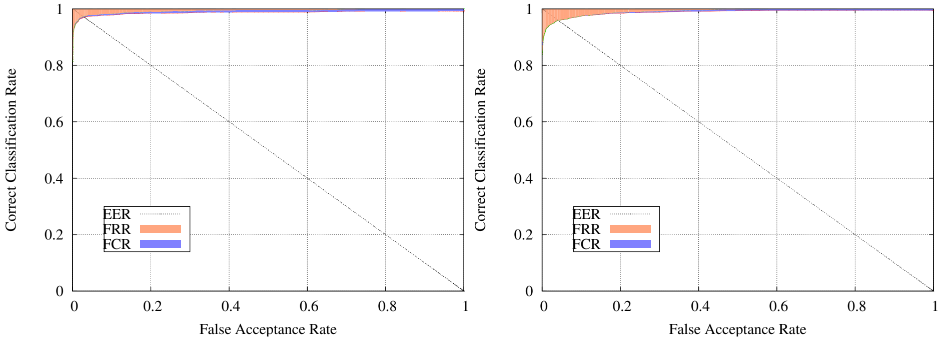
Table 2. Classification results with AAM face warping vs. simple alignment

The frame-based face recognition in video sequences makes a decision on every single frame. The results, therefore, return some insight on the general performance of the registration and classification scheme employed. The EER obtained with pose normalization is 1.0% lower than the one obtained without pose normalization. This means that the AAM-based pose normalization improves the robustness of the system against pose variations. It prevents the system from accepting unknown identities with similar poses to the enrolled person or rejecting genuine members with unmatched poses. The frame filtering with the modified DFFS metric also improves the robustness of the system as the frames with bad alignment are discarded. Compared to the results reported in [16], the equal error rate decreases by 4.5% even without pose normalization.

The temporal fusion-based classification was also applied as a progressive-score-based approach by accumulating frame scores over time. This can be thought of as classifying every frame as if it were the end of a sequence and taking the final score. The results with progressive-score-based classification are listed on the second and fourth row of Table 2, respectively for the two registration approaches.

Observing the results based on the progressive-score-based classification, we noticed that the results were improved compared to the frame-based scheme.

The Euclidean transformation-based alignment achieved 0.4% EER, which is already much better than the frame-based classification. With the AAM-based pose normalization the temporal fusion smoothed out all erroneous decisions and the ERR is 0.0%. [1]



(a)

(b)

Fig. 3. (a) Frame-based ROC curve with pose normalization. (b) Frame-based ROC curve without pose normalization.

4 CONCLUSIONS

In this paper, an open-set face recognition system is presented, in which AAM-based pose normalization is employed to improve its robustness. The system operates fully automatically and runs in near real-time (at 15 fps) on a laptop computer with a 2.0GHz Intel Core 2 Duo processor. It has been observed that normalizing the pose changes improves the recognition performance, because the gallery may not contain the corresponding pose for a given probe. The employed distance metric is able to filter out some misaligned frames, which improved the results further. The local appearance-based face representation makes the system invariant to other confounding factors as well as the misalignment errors.

Currently we only evaluated the system with five known subjects. The performance of the system may decrease as the number of known subjects increases. However, for small-scale convenience applications such as the smart visitor interface, the system is able to operate very robustly with moderate number of group members. In the future, more known subjects will be evaluated and the scalability of the system will be analyzed.

[1] A sample video is available under http://cvhci.anthropomatik.kit.edu/∼hgao/dagm-video.zip, which demonstrates the face tracking, pose normalization and open-set face recognition.

Acknowledgments

This work was realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

References

1. Georghiades, A.S., Belhumeur, P.N., Kriegman, D.J.: From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Trans. on PAMI **23**(6) (2001) 643–660
2. Huang, F.J., Zhou, Z., Zhang, H.J., Chen, T.: Pose invariant face recognition. In: Proc. of the IEEE Int'l Conference on Automatic Face and Gesture Recognition. (2000) 245–250
3. Zhao, W.Y., Chellappa, R.: SFS based view synthesis for robust face recognition. In: Proc. of the IEEE Int'l Conference on Automatic Face and Gesture Recognition. (2000) 285–292
4. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Proc. of 5^{th} European Conference on Computer Vision. Volume 2. (1998) 484–498
5. Edwards, G.J., Cootes, T.F., Taylor, C.J.: Face recognition using active appearance models. In: Proc. of 5^{th} European Conference on Computer Vision. (1998) 581–595
6. Cootes, T.F., Wheeler, G.V., Walker, K.N., Taylor, C.J.: View-based active appearance models. Image and Vision Computing **20**(9-10) (2002) 657–664
7. Guillemaut, J., Kittler, J., Sadeghi, M.T., Christmas, W.J.: General pose face recognition using frontal face model. In: 11^{th} Iberoamerican Congress in Pattern Recognition. Volume 4225/2006. (2006) 79–88
8. Ekenel, H.K.: A robust face recognition algorithm for real-world applications. PhD dissertation, University of Karlsruhe (TH) (2009)
9. Gao, H., Ekenel, H.K., Stiefelhagen, R.: Pose normalization for local appearance-based face recognition. In: 3^{rd} Int'l Conference on Biometrics, LNCS (2009) 32–41
10. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision **60**(2) (2004) 135–164
11. Gross, R., Matthews, I., Baker, S.: Generic vs. person specific active appearance models. Image and Vision Computing **23**(11) (2005) 1080–1093
12. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision **57**(2) (2004) 137–154
13. Liu, X., Wheeler, F.W., Tu, P.H.: Improved face model fitting on video sequences. In: Proc. of British Machine Vision Conference (BMVC) 2007. (2007)
14. Jebara, T.S.: 3D pose estimation and normalization for face recognition. B. Thesis, McGill Centre for Intelligent Machines (1996)
15. Ekenel, H.K., Stiefelhagen, R.: Block selection in the local appearance-based face recognition scheme. In: CVPR Biometrics Workshop. (2006)
16. Ekenel, H.K., Szasz-Toth, L., Stiefelhagen, R.: Open-set face recognition-based visitor interface system. In: Proc of 7^{th} Int'l Conf. on Computer Vision System. Volume 5815., LNCS (2009) 43–52