Video-based Face Recognition on Real-World Data

Johannes Stallkamp Hazım K. Ekenel Rainer Stiefelhagen
Interactive System Labs
University of Karlsruhe, Germany
{jstallkamp,ekenel,stiefel}@ira.uka.de

Abstract

In this paper, we present the classification sub-system of a real-time video-based face identification system which recognizes people entering through the door of a laboratory. Since the subjects are not asked to cooperate with the system but are allowed to behave naturally, this application scenario poses many challenges. Continuous, uncontrolled variations of facial appearance due to illumination, pose, expression, and occlusion need to be handled to allow for successful recognition. Faces are classified by a local appearance-based face recognition algorithm. The obtained confidence scores from each classification are progressively combined to provide the identity estimate of the entire sequence. We introduce three different measures to weight the contribution of each individual frame to the overall classification decision. They are distance-to-model (DTM), distance-to-second-closest (DT2ND), and their combination. Both a k-nearest neighbor approach and a set of Gaussian mixtures are evaluated to produce individual frame scores. We have conducted closed set and open set identification experiments on a database of 41 subjects. The experimental results show that the proposed system is able to reach high correct recognition rates in a difficult scenario.

1. Introduction

Perceptual computer systems which are developed in order to create smart environments (homes, cars) need to follow human interaction patterns in order to allow comfortable usage. These environments are supposed to give benefits to the users and support their interaction among each other. The users must be in focus of attention, not the computer system which supplies certain functionality. In such environments, it is essential that the computer system is able to identify the people it is dealing with.

A feasible approach to biometric identification is the use of facial features. Alongside speech, it is a very natural approach and mimics human recognition. The nature of the cue inherently allows for unobtrusive, interaction-free recognition, as the visibility of the face does not need any specific action. This property is indispensable in smart environments, as the necessity of indiviuals to explicitly cooperate with the system (*e.g.* by putting a finger on a fingerprint scanner) violates the central idea of such environments, since it interrupts a user's actions and moves the computer system into the focus of attention.

1.1. Previous work

Numerous approaches have been developed to recognize faces. While the main focus was on image-based methods in the beginning [17, 2, 7], it shifted more and more towards video-based approaches in the last years. These are developed in order to overcome shortcomings of image-based recognizers like sensitivity to low resolution, pose variations and partial occlusion.

Zhou *et al*. [19] use sequence importance sampling (SIS) to propagate a joint posterior probability distribution of identity and motion over time to do tracking and recognition of a person simultaneously. To overcome continous changes of head pose and facial expressions, Lee *et al*. [10] represent the appearance of a person by the means of pose manifolds which are connected by transition probabilities. In order to model person-specific appearance and dynamics, Liu and Chen [12] train individual hidden Markov models (HMM) on eigenface image sequences. Confident classifications are used to adapt these models.

The problem with models that are based on probability distributions is that they make strong assumptions about underlying distributions in the training set. A model may implicitly learn dependencies which are not characteristic for the data, if the training set turns out not to be representative. The counterpart are exemplar-based approaches which generally do not assume an underlying distribution and are, thus, less affected by non-representative training data.

978-1-4244-1631-8/07/$25.00 ©2007 IEEE

A large variety of head pose and illumination variations, as well as occlusion, is encountered in feature films. Arandjelovic and Zisserman [1] built a system to retrieve all faces from a film that match one or multiple query images. The appearance-based approach uses a modified Euclidean distance for classification. Instead of doing frame-based retrieval, Sivic *et al*. [14] group all face views of a person within the same shot into a face-track, represented as a histogram. Given a query image in one of the scenes, the corresponding face-track is determined. All matching face-tracks are retrieved from the whole film by means of a chi-square goodness-of-fit test..

Face recognition systems that are to be deployed in a real-life scenario, usually encounter the problem that they are confronted with unknown people. Li and Wechsler [11] make use of transduction to derive a rejection criterion. The k-nearest neighbors of a test sample are iteratively misclassifed to determine an error distribution. If classification of the test sample as any of the classes does not yield a credibility sufficiently different from this distribution, it is rejected, otherwise it is classified.

1.2. Motivation

Our central goal is to build a real-time capable face recognition system (FRS) for uncontrolled environments. It is supposed to handle robustly real-life situations with all the challenges they bring along that make the task harder. The central challenge is to achieve unobtrusive recognition, *i.e*. to create a system that operates in the background and does not need specific user interaction. This is essential to grant users the freedom to behave naturally. As a consequence of this freedom, difficulties arise from varying pose, like out-of-plane rotations, and different facial expressions. Accessories and facial hair can cause partial occlusions. Daylight leads to very different illumination depending on the time of day, time of year and weather conditions. In spite of this hardly controllable natural influences, even the artificial light sources are withdrawn from the system's control if unobtrusive recognition as postulated above is to be implemented. Since the users, *i.e*. the persons to be recognized, are not supposed to be restrained by the system, they are free to switch on and off any light sources that might be available. This leads to a wide variety of illumination configurations in terms of light intensity, direction and even color.

In the given scenario, a camera opposite the entrance door to a computer laboratory records people entering the room. As outlined above, they are not explicitly cooperating, and recording conditions can vary largely. Some example views are shown in Figure 1[1].

[1]A demo video can be found at `http://isl.ira.uka.de/~ekenel/door\_monitoring.mpg`

Figure 1. Exemplary recognition situations showing a variety of different lighting, pose and occlusion conditions. No individual explicitly looks into the camera.

1.3. Our approach

The scenario outlined above leads to large variations of "quality" between frames. As we will show, a single frame might not be distinctive enough to allow a clear classification, but it usually still holds enough information that a classifier can get *close* to the correct decision. Therefore, we use a video-based approach to face identification by progressively combining the individual frame-based classification results to one score per sequence.

The different variations of facial appearance entail that some frames are more ambiguous than others. Therefore, two main observations are exploited to derive two different schemes to weight the contribution of each individual frame to the overall classification result. The first, *distance-to-model (DTM)*, takes into account how similar a test sample is to the representatives of the training set. Test samples that are very different from the training data will generally produce larger distances than more similar data. Consequently, they are more likely to cause a misclassification. DTM is used to reduce the impact of these samples on the final score. The second weighting scheme, *distance-to-second-closest (DT2ND)*, reduces the impact of frames which deliver ambiguous classification results. It is based on the same idea as in [13] that reliable matching requires the best match to be significantly better then the second-best match (speaking of classes). Other than Lowe [13], we do not discard "bad" matches (in terms of distance or score) but only reduce their contribution. As we will show, both DTM and DT2ND have different positive effects on the classification results and the combination of the two schemes can improve the performance further.

Individual frame scores are generated with a k-nearest neighbor classifier and a set of Gaussian mixture models. Although the weighting functions have to be computed sep-

Figure 2. Sample aligned face images

arately for both classifiers due to different classifier output, their structure is the same.

The remainder of this paper is organized as follows. Details about the feature extraction and classification steps, including the introduction of the weighting schemes, are given in Section 2. Our approach is evaluated in Section 3. Section 4 finishes the paper with a conclusion and future directions.

2. Face recognition

Faces in the input sequences are detected, tracked and registered as described in [16]. Figure 2 shows some sample images of aligned faces. They suffer from strong variations in illumination, pose, expression and resolution. Furthermore, some of them are blurry due to interlacing effects and movement of the subject. Although there are no occlusion examples in Figure 2, they can be anticipated from the samples in Figure 1.

Feature extraction follows the approach in [5] which applies block-based DCT to non-overlapping blocks of size 8×8 pixels. DCT was chosen for its compact representation of the input signal and the data-independency of its basis functions. Besides, its fast computation helps processing in real-time applications. The first five AC coefficients in each block are selected in order to create compact local feature vectors. The DC coefficient is discarded for illumination normalization as suggested in [5]. Furthermore, robustness against illumination variations is increased by normalizing the local feature vectors to unit norm [6]. This reduces illumination effects (especially illumination differences with a gradient pattern) while keeping the essential frequency information. Concatenation yields a global feature vector. Both a discriminative and a generative approach are followed to classifiy the so-achieved feature vectors. With both approaches, individual models are derived for each person. The granularity of these models depends on the respective amount of available training data. This accounts for the fact that the real-life data collection setting leads to largely varying amounts of data among the different persons. Thus, the more often the system encounters a certain individual, the more detailed this individual's model will be, as more variation can be captured.

2.1. K-nearest neighbors model

A major advantage of discriminative approaches like K-nearest neighbors (KNN) is that they do not make an assumption about the distribution of the underlying data. This allows to build meaningful models with less data than would be necessary to train high-dimensional generative models like Gaussian mixtures. To determine the nearest neighbors, the L_1 norm is employed as distance measure $d(\cdot,\cdot)$, as it was shown to perform best among several popular distance metrics in [6]. The k closest neighbors S_i, $i = 1, 2, \ldots, k$ of a test vector x are selected with score $s_i = d(x, S_i)$. Because the distances and, thus, the resulting scores can differ largely between frames, they need to be normalized. This is achieved with linear *min-max normalization* [15],

$$s'_i = 1 - \frac{s_i - s_{\min}}{s_{\max} - s_{\min}} \qquad i = 1, 2, \ldots, k \tag{1}$$

which maps the scores to $[0, 1]$. To have equal contribution of each frame, these scores are re-normalized to $\sum_{i=1}^{k} s'_i = 1$. Of course, among the k closest representatives, there can be several ones from the same class. Since some people have far fewer representatives than others, care must be taken that their scores are not dominated by those. Individual scores are selected by a simple max-rule [9], which only selects the maximum score for each class.

2.2. Gaussian mixture model

Although generative models, in our case Gaussian mixture models (GMM), usually require more training data than discriminative ones, they allow to model the data with probability density functions (pdf), and, as a consequence, the computation of conditional pdfs.

The Gaussian mixture model approach trains one GMM per class using an expectation-maximization algorithm [3, 18]. Likewise the KNN model, the number of components per mixture depends on the number of training samples available for a person. At runtime, person x is classified as one of the N registered individuals in a maximum log-likelihood manner using

$$\arg\max_{i\in N} \log P(x|i) = \arg\max_{i\in N} \log \sum_{j=1}^{k_i} \alpha_{ij} \cdot \mathcal{N}(x; \mu_{ij}, \Sigma_{ij}) \tag{2}$$

where k_i denotes the number of modes per person, α_{ij} the mixing parameters, and μ_{ij} and Σ_{ij} the mean and the variance of the j's component of person i's model, respectively. To keep the computational effort within reasonable bounds, only a diagonal rather than the full covariance matrix is used.

For the weighting scheme approach, a distance measure in terms of a similarity measure is necessary. It is defined as $d(x,i) = |\log P(x|i)|, i \in N$ (see Equation (2)).

2.3. Temporal fusion: Weighting schemes

A sum-rule [9] decision fusion scheme is employed to take advantage of all frames in a video sequence to decide on the identity of a subject. Two baseline performances are determined. First, every single frame is evaluated individually to be able to evaluate the improvement contributed by video-based classification. Second, the baseline video-based recognition performance is determined by simply adding the scores of all frames. The decision is then made based on the min-max-normalized final score. Due to the real-life quality of the data, however, not all frames are suitable to classify the subject. Low resolution, large occlusion and faulty alignment are examples of negative influences on frame quality. Besides, certain views of a person may simply not be captured by the model due to little training data or due to training data that contains too little variation. Two important observations have been made which are exploited in order to reduce the impact of ambiguous frames.

First, for wrong classifications, the distance to the closest representative is, on average, larger than for correct ones. Moreover, badly aligned frames result in larger distances as well. To account for this, we introduce the weighting scheme *distance-to-model (DTM)*. The frames f_i, $i = 1, 2, \ldots$, are weighted with respect to the closest representative c with

$$w_{\text{DTM}}(f_i) = \begin{cases} 1 & \text{if } d(f_i, c) < \mu \\ e^{-\frac{d(f_i,c)-\mu}{2\sigma^2}} & \text{otherwise} \end{cases} \quad (3)$$

This weighting function is chosen according to the observed distribution of frame distances $d(f_i, c_{f_i,\text{correct}})$, the distances of all frames f_i to the closest representative $c_{f_i,\text{correct}}$ of the corresponding correct class. The distribution, determined on a parameter estimation set, resembles a normal distribution $\mathcal{N}(\cdot;\mu,\sigma^2)$ (in the GMM case, there is an additional peak close to zero). To increase robustness against outliers, μ is chosen as sample median and σ as median absolute deviation (MAD)[8]. Example distributions and weight functions are shown in Figure 3. Using the weight function w_{DTM}, the influence of frames which are not sufficiently close to the model is reduced.

The second observation is that, in case of misclassification of frame f_i, the difference of the distances $\Delta(f_i)$ to the closest and the second closest representatives is generally smaller than in the correct case. The distribution of these distances follows approximately an exponential distribution

$$\varepsilon(x;\lambda) = \lambda \mathrm{e}^{-\lambda x} \quad (4)$$

The weights are then computed as the cumulative distribution function of $\varepsilon(\cdot)$

$$w_{\text{DT2ND}}(f_i) = \mathcal{E}(\Delta(f_i)) = 1 - \mathrm{e}^{-\lambda\Delta(f_i)} \quad (5)$$

Again, example distributions and weight functions are shown in Figure 4.

This weighting scheme will be referred to as *distance-to-second-closest (DT2ND)*.

In addition to individual weighting schemes, a joint weighting scheme employs the product of w_{DTM} and w_{DT2ND} to weight the frames.

3. Experiments

We allow the subjects who are to be identified to behave naturally without the need to cooperate with the recognition system. For this reason, the input data will cover a much larger bandwidth of appearance variations in terms of illumination, pose and occlusion than publicly available databases, which furthermore usually consist of still images. Consequently, we needed to collect our own data.

Our database consists of 2,292 video sequences (205,480 frames) of 41 subjects recorded during 6 months. It is chronologically divided into three sets for training, parameter estimation and testing. The separation by date accounts for the fact that people change appearance over time. Besides the fact that *unaware* recording of large amounts of video data is not trivial, we think that 41 subjects consitute a solid base for smart environments. In those, it is important to cover the "core" set of people, *e.g.* the inhabitants of a smart house.

Face images were automatically extracted from training sequences using the registration process outlined in [16]. Training data is augmented with virtual samples at different scales and slight rotations to increase the amount of training data and reduce effects of registration errors. The augmentation leads to a very large training set which, consequently, would slow down the KNN approach. Since many samples from consecutive frames are very similar, k-means is applied to select representative exemplars. The clustering is performed for each person individually.

For the KNN approach and this database, the $k = 10$ closest neighbors were selected. This is a good trade-off between recognition result and — important for real-time systems — processing time.

Results on a different data set can be found in the CLEAR evaluation 2007 [4] in which this system performed very well.

3.1. Closed set identification

For closed-set identification, the system is only confronted with subjects that are registered in the database. The system has to classify each subject as one of the possible

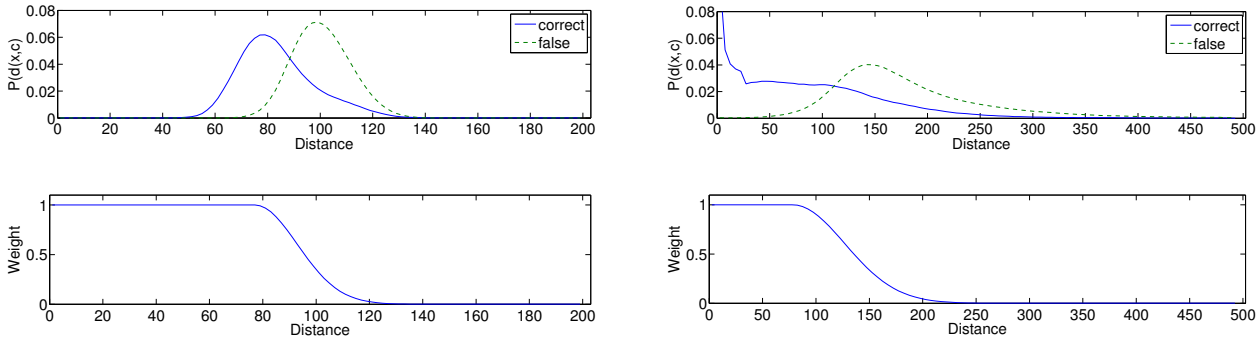

Figure 3. DTM weight functions for (left) KNN ($k = 10$) and (right) GMM approach. (top) Distribution of the distances to the closest representative of the correct class (blue, solid) and to all other classes (green, dashed) and (bottom) the actual weight function.

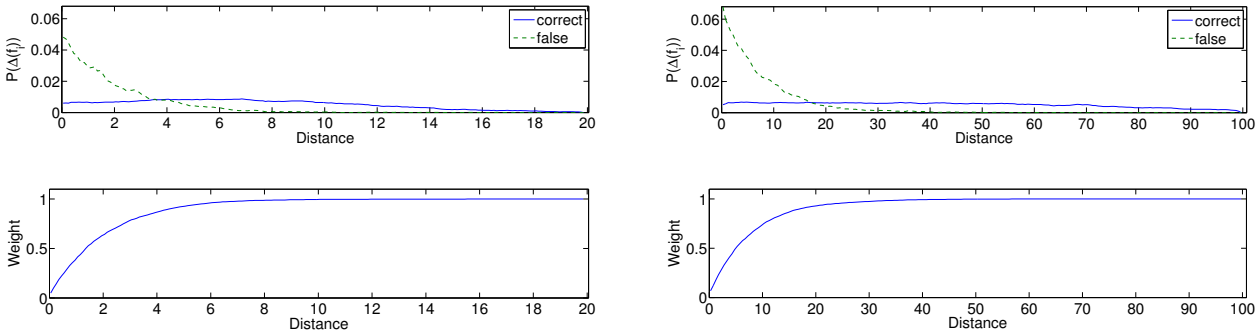

Figure 4. DT2ND weight function for (left) KNN ($k = 10$) and (right) GMM approach. (top) Distribution of distances between closest and second closest representatives for correct (blue, solid) and false classifications (green, dashed) and (bottom) the actual weight function.

Table 1. Closed-set correct classification rates.

KNN	CCR (%)	GMM	CCR (%)
Frame-based	68.4	Frame-based	62.7
Uniform	90.0	Uniform	86.7
DTM	92.0	DTM	90.6
DT2ND	91.3	DT2ND	89.1
Combined	**92.5**	**Combined**	**91.8**

classes. The performance is measured as *correct classification rate (CCR)*, the percentage of correctly classified sequences in the test set. As baseline performance, *every* single frame is evaluated *individually*. The results are given in Table 1. *Uniform* denotes no weighting and therefore equal contribution of each frame, *combined* the combination of DTM and DT2ND by weight multiplication.

Video-based evaluation outperforms frame-based evaluation since the increased amount of available data helps to resolve some ambiguities. Obviously, both weighting schemes improve the classification performance over uniform weighting. The combination takes advantage of both to perform even better. This is true both for the KNN and the GMM approach. The increase is slightly larger for DTM, as it assigns smaller weights to frames that are not similar enough to the representatives of the training set. DT2ND, in contrast, reduces the impact of ambiguous frames, *i.e.* frames which yield similar scores for the top two candidates, albeit how well the "face" is modeled. In fact, a badly aligned image can lead to a distinct score, but it is likely to have a large distance to the model. DTM is able to handle this case, DT2ND is not. Nevertheless, reduction of ambiguity leads to better performance over uniform weighting as well.

As far as the different models are concerned, the discriminative approaches perform slightly better than the generative ones. Since parametric models like GMMs need more training data with increasing dimensionality, this is possibly caused by insufficient training data for some individuals which can prevent derivation of meaningful models. Besides, the number of mixture components might not be sufficient to approximate the underlying probability distribution. The discriminative models are less affected by little training data, as they classify new data only based on existing data, without making any assumptions about its distribution.

To investigate the robustness of the results, it is worth looking at the results including rank-2 and rank-3 classifications, *i.e.* cases in which the correct identity is among the

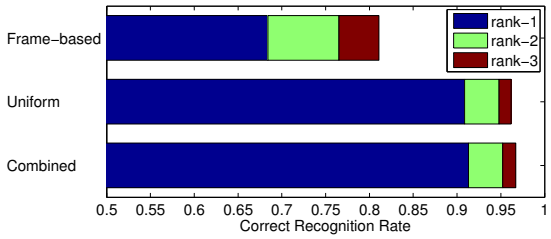

Figure 5. Correct recognition rate by rank for the KNN models.

Table 2. Correct recognition rate by rank for the KNN models.

	Frame-based	Uniform	Combined
rank-1	68.4	90.9	92.5
rank-2	76.5	94.8	95.6
rank-3	81.1	96.2	96.7

best two or three hypotheses. Representatively, results are reported for the KNN approach only. As clearly depicted in Figure 5 and Table 2, the frame-based approach often gets close to the correct decision. However, it has to decide on the identity even in the case that the single feature vector is of questionable quality. The approach lacks an opportunity to support or discard the hypothesis using additional data as done by the sequence-based methods. These are able to exploit the temporal dependency of consecutive frames and to promote the rank-2 and rank-3 classifications of the frame-based models to first place. Since many frames contribute to the decision, the overall performance improvement is larger than the difference between the correct and rank-3 classifications in the frame-based approach. The rank analysis supports the decision of choosing $k = 10$ neighbors, as even in the frame-based approach, the correct result is already among the top three matches in $81.1\,\%$ of the cases.

The more frames can be evaluated, the more likely it is to obtain a correct result. This gets confirmed by the observation that the average length of correctly classified sequences is larger — 39 frames — than that of misclassified ones with 28 frames as depicted in Figure 6.

To justify the increased training efforts caused by the larger training set size, an experiment was conducted to compare the recognition performance using augmented and unaugmented training data. The comparison can only cover the KNN models as it is not possible to train appropriate GMMs due to the fact that many individuals have fewer images in the training set than the feature vector's dimensionality. As listed in Table 3, recognition performance increases significantly in all three KNN cases. This shows that the data augmentation is well worth the increased memory and time resources.

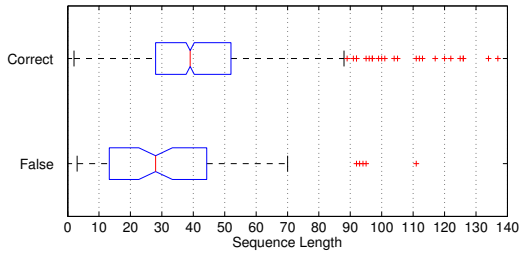

Figure 6. Box plot showing the distribution of sequence lengths for correct and false classifications. The box denotes the median as well as the lower and upper quartiles whereas the "whiskers" denote the range of values. Outliers are marked by crosses. The plot is based on results achieved with the combined approach.

Table 3. Influence of data set augmentation with virtual samples. All results improve significantly. Significance was computed with crosstabulation.

	Frame-based	Uniform	Combined
unaugmented	56.6 %	87.6 %	88.2 %
augmented	68.4 %	90.9 %	92.5 %
p-value	0.00	0.02	0.00
significantly better	✓	✓	✓

3.2. Open set identification

This task extends the previous one by the difficulty that unknown people, *i.e.* persons which are not registered in the database, can be encountered. Therefore, prior to classification as one of the possible identities, the system has to decide whether a person is known or unknown. Impostors are to be rejected, while genuine members of the database need to be accepted and classified correctly. To model this task with the existing data set, the system is trained in a leave-one-out manner. One person at a time is removed from the database and is presented to the system as impostor during the subsequent evaluation on all sequences. This process is repeated N times, so that each person takes the impostor role once. The acceptance-rejection criterion is a threshold on the confidence of the classification, which is a value between 0 and 1. If the confidence is too low, the person is rejected.

A measure of confidence of the classification is derived by min-max normalization (see Equation (1)) of the accumulated scores at the end of the sequence. The frame-based scores are already normalized and can serve as confidence measure without further processing.

Compared to closed-set identification, two more error types can occur. Additional to false classifications, the system can erroneously either reject genuine identities or accept impostors. All three errors have to be traded-off against each other as it is not possible to minimize them at the same

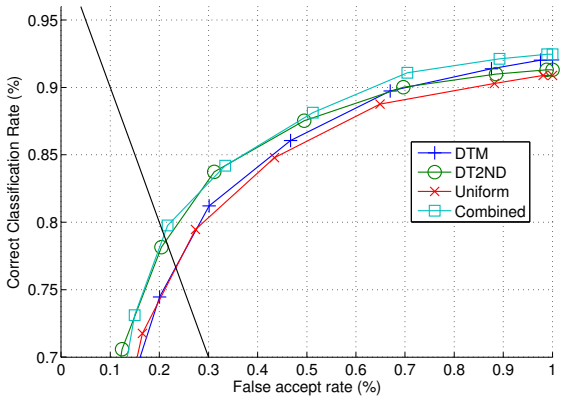

Figure 7. ROC curve of of open-set recognition results for the KNN approach. The black line denotes equal error.

Table 4. Open-set equal error rates.

KNN	EER
Frame-based	50.0 %
Uniform	23.4 %
DTM	23.3 %
DT2ND	21.3 %
Combined	**21.0 %**

GMM	EER
Frame-based	43.7 %
Uniform	23.0 %
DTM	20.2 %
DT2ND	20.5 %
Combined	**18.0 %**

time. For this reason, a different performance measure is necessary. The employed *equal error rate (EER)* denotes the minimum combined error rate. It is reached when

$$\mathrm{FAR} = \mathrm{FRR} + \mathrm{FCR} \tag{6}$$

i.e. when the *false acceptance rate (FAR)* among the impostors is equal to the sum of the *false rejection rate (FRR)* and the *false classification rate (FCR)* among the registered persons. The rates are defined as

$$\mathrm{FAR} = \frac{n_{i,\mathrm{accepted}}}{n_i} \tag{7}$$

$$\mathrm{FRR} = \frac{n_{g,\mathrm{rejected}}}{n_g} \tag{8}$$

$$\mathrm{FCR} = \frac{n_{g,\mathrm{misclassified}}}{n_{g,\mathrm{accepted}}} \tag{9}$$

where n denotes number of frames or sequences and the subscripts g and i denote genuine or impostor samples, respectively. Please note that FCR is the false classification rate conditional on acceptance as "known" individual.

For the KNN approach, it can be seen from the results in Figure 7 that uniform and combined weighting form a lower and upper bound for the individual weighting schemes. The two weighting schemes affect different parts of the ROC curve. The DTM scheme improves the recognition rate for high false acceptance rates. A FAR of 100 percent

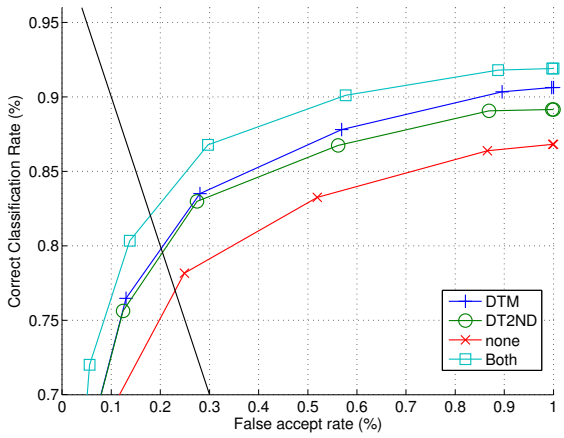

Figure 8. ROC curve of of open-set recognition results for the GMM approach. The black line denotes equal error.

is equivalent to closed-set identification in terms of the ROC curve because the CCR can only be computed over genuine samples. This confirms the results from the closed-set experiment. DTM reduces the false classification rate, but it is not able to discriminate between known and unknown persons as the feature vector of an impostor can be indeed very similar to the training representatives. Therefore, as can be seen from Table 4, the EER is approximately the same than in the unweighted case.

Genuine identities, however, usually have smaller distances to one single class representative than to all other classes in the model, while impostors are similarly close to multiple classes (*c.f.* Figure 4). This ambiguity is exploited by the DT2ND weighting scheme to identify impostors. Their scores are reduced, leading to smaller confidence values, which in turn result in better rejection. The same threshold causes rejection of more impostors than in the unweighted case or, to put it the other way round, the threshold can be reduced causing fewer false rejections. As a consequence, the EER is reduced in open-set identification.

Looking at Figure 8, the observations are similar if the GMM output is used as similarity measure: Combined and uniform weighting represent upper and lower bounds on the individual weighting schemes and DTM outperforms DT2ND for high FARs (*i.e.* getting closer to closed-set identification). However, in contrast to the results above, DT2ND alone does not perform better than DTM in terms of EER. In combination with DTM, however, there is a performance improvement observable that does not exist in Figure 7 (where the *Combined* ROC merely represents the best out of DTM and DT2ND). The combination leads to the lowest EER of all runs, as can be seen from Table 4.

Since the ROC curves show the CCR depending on the FAR, the question arises to which extent each of the other

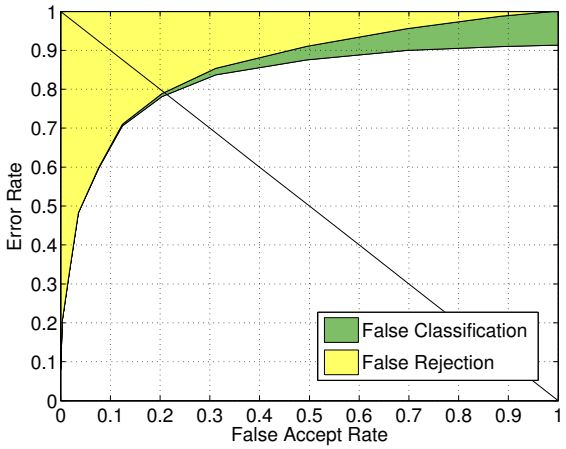

Figure 9. Analysis of the contribution of FRR and FCR to the overall error rate of Combined-KNN. The black line denotes equal error.

two types of error, false rejection and false classification, impair classification performance. To investigate this, Figure 9 plots FRR and FCR separately for Combined-KNN. The lower bound corresponds to the CCR as depicted in Figure 7. At the point of equal error, there remains only a minimal FCR of about 1 percent while the major part of about 20 percent is caused by false rejection of genuine identities. This reflects the difficulty of separating impostors and genuines as they represent arbitrary subsets of all possible faces.

4. Conclusion

In this study, a video-based approach to face recognition is presented which is able to process real-world data, comprising large variations of a subject's visual appearance.

Three weighting schemes have been introduced to weight the contribution of individual frames in order to improve the classification performance. The first, DTM, takes into account how similar a test sample is to the representatives of the training set and therefore reduces the negative impact of unfamiliar data (*e.g.* due to bad recording conditions or faulty face registration). The second, DT2ND, reduces the influence of frames which deliver ambiguous classification results. At least in the KNN case, it is able to reduce the equal-error rate in open-set identification. Finally, the combination of DTM and DT2ND was shown to join the benefits of both.

Extensive experiments on a large video database of 41 non-cooperative subjects have been conducted for both closed-set and open-set identification tasks. The results show that the combination of video-based face recognition with a local appearance-based approach is able to handle a large number of variations of facial appearance caused by illumination, pose, expression, and occlusion.

In the future, detection of impostors in the open-set identification task needs further investigation. However, it has to be kept in mind that impostor detection is a non-trivial task because it requires the separation of an arbitrary subset of all possible faces from the rest.

Acknowledgements

This work is sponsored by the European Union under the Integrated Project CHIL, contract number 506909, and under the FP6-2004-ACC-SSA-2016684 SPICE project.

References

[1] O. Arandjelovic and A. Zisserman. Automatic face recognition for film character retrieval in feature-length films. In *CVPR*, pages 860–867, Washington, DC, USA, 2005.

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class-specific linear projection. *IEEE Trans. PAMI*, 19(7):711–720, July 1997.

[3] A. Dempster, N. Laird, and D. Rubin. Maximum-likelihood from incomplete data via EM algorithm. *Journal Royal Statistical Society, Series B*, 39:1–38, 1977.

[4] H. K. Ekenel, Q. Jin, and M. Fischer. ISL person identification systems in the CLEAR 2007 evaluations. 2007. `http://www.clear-evaluation.org`.

[5] H. K. Ekenel and R. Stiefelhagen. Local appearance-based face recognition using discrete cosine transform. In *EUSIPCO*, 2005.

[6] H. K. Ekenel and R. Stiefelhagen. Analysis of local appearance-based face recognition: Effects of feature selection and feature normalization. *CVPRW*, page 34, 2006.

[7] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. PAMI*, 23(6):643–660, 2001.

[8] P. J. Huber. *Robust Statistics*. Wiley, 1981.

[9] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Trans. PAMI*, 20(3):226–239, 1998.

[10] K. C. Lee, J. Ho, M. H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *CVPR*, pages I: 313–320. IEEE Computer Society, 2003.

[11] F. Li and H. Wechsler. Open set face recognition using transduction. *IEEE Trans. PAMI*, 27(11):1686–1697, 2005.

[12] X. Liu and T. Chen. Video-based face recognition using adaptive hidden markov models. In *CVPR*, pages I: 340–345. IEEE Computer Society, 2003.

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[14] J. Sivic, M. Everingham, and A. Zisserman. Person spotting: Video shot retrieval for face sets. In *CIVR*, volume 3568, pages 226–236. Springer, July 2005.

[15] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. K. Jain. Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems. *PAMI*, 27(3):450–455, 2005.

[16] J. Stallkamp. Video-based face recognition using local appearance-based models. Master's thesis, Interactive System Labs (ISL), University of Karlsruhe, Germany, Nov. 2006.

[17] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cogn. Neurosci.*, 3(1):71–86, 1991.

[18] G. Welch and G. Bishop. An introduction to the Kalman filter. In *SIGGRAPH, Course 08*, 2001.

[19] S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. *CVIU*, 91(1-2):214–245, Feb. 2003.