We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Far-Field, Multi-Camera, Video-to-Video Face Recognition

Aristodemos Pnevmatikakis and Lazaros Polymenakos
Athens Information Technology
Greece

1. Introduction

Face recognition on still images has been extensively studied. Given sufficient training data (many gallery stills of each person) and/or high resolution images, the 90% recognition barrier can be exceeded, even for hundreds of different people to be recognized (Phillips et al., 2006). Face recognition on video streams has only recently begun to receive attention (Weng et al., 2000; Li et al., 2001; Gorodnichy, 2003; Lee et al., 2003; Liu and Chen, 2003; Raytchev and Murase, 2003; Aggarval et al., 2004; Xie et al., 2004; Stergiou et al., 2006). Video-to-video face recognition refers to the problem of training and testing face recognition systems using video streams. Usually these video streams are near-field, where the person to be recognized occupies most of the frame. They are also constrained in the sense that the person looks mainly at the camera. Typical such video streams originate from video-calls and news narration, where a person's head and upper torso is visible.

A much more interesting application domain is that of the far-field unconstrained video streams. In such streams the people are far from the camera, which is typically mounted on a room corner near the ceiling. VGA-resolution cameras in such a setup can easily lead to quite small faces – down to less than ten pixels between the eyes (Stiefelhagen et al., 2007), contrasted to over two hundred pixels in many of the latest face recognition evaluations (Phillips et al., 2006). Also, the people go about their business, almost never facing the camera directly. As a result, faces undergo large pose, expression and lighting variations. Part of the problem is alleviated by the use of multiple cameras; getting approximately frontal faces is more probable with four cameras at the corners of a room than with a single one. The problem is further alleviated by the fact that the goal is not to derive a person's identity from a single frame, but rather from some video duration. Faces to be recognized are collected from a number of frames; the person identity is then established based on that collection of faces.

Far-field unconstrained video-to-video face recognition needs to address the following challenges:

- Detection, tracking and segmentation of the faces from the video streams, both for system training and recognition.
- Selection of the most suitable faces to train the system and to base the recognition upon.
- The face recognition algorithm needs to cope with very small faces, with unconstrained pose, expression and illumination, and also with inaccurate face framing.
- Fusion of the individual decisions on faces, to provide the identity of the person given some time interval.

Source: Face Recognition, Book edited by: Kresimir Delac and Mislav Grgic, ISBN 978-3-902613-03-5, pp.558, I-Tech, Vienna, Austria, June 2007

In section 2 of this chapter we will present the state-of-the-art in video-to-video face recognition, mostly near-field with people moving towards the camera. In section 3 we will address all the before-mentioned challenges of video-to-video face recognition, by analyzing the tradeoffs of different face segmentation approaches, face recognition methods and decision fusion strategies. We will base our analysis on a publicly available database of videos, built by the partners of the CHIL project (Waibel et al., 2004) and already used in the CLEAR 2006 evaluations (Stiefelhagen et al., 2007). This database offers recordings at five different sites, 26 individuals, two different gallery video lengths and four different probe video lengths.

2. Algorithms and databases for video-to-video face recognition

Video-to-video face recognition is split into two tasks. Firstly stills containing faces are extracted from the gallery and probe videos, generating the gallery and probe stills. Then, traditional still-to-still face recognition is applied, with one addition: the goal is the recognition of a person throughout the complete probe video, i.e. using all the probe stills coming from it. Hence, apart from recognition, the video-to-video face recognition task has some sort of face detection/tracking and utilization of temporal information embedded in it. Even though video-to-video face recognition is a relatively new field, many algorithms can be found in the literature. These algorithms differ on the face detection, the way the face recognizer utilizes temporal information, as well as on the video databases they are tested with.

These algorithms are categorized regarding the way temporal information is used, to report people identities per probe video and not per extracted probe still. There are algorithms based on post-decision fusion (Xie et al., 2004; Stergiou et al., 2006), while others embed the use of temporal information within the face recognizer (Weng et al., 2000; Li et al., 2001; Lee et al., 2003; Liu and Chen, 2003; Raytchev and Murase, 2003; Aggarval et al., 2004). An exception to this categorization can be found in (Gorodnichy, 2003), where temporal information is only utilized in face detection, to provide the best still to attempt recognition. The subjects are approaching the camera, allowing for a coarse-to-fine face detection scheme.

Xie et al. employ post-decision methods (Xie et al., 2004). Their classifier is a polynomial correlation filter bank with non-linear output combination. It operates on faces extracted using template matching in a head region found by motion. Since the videos they employ are near-field, such a detector suffices.

Weng et al. are concerned with the computational burden of training in a batch mode from many and long gallery videos and propose an iterative tree building algorithm for on-line training (Weng et al., 2000). They do not address face detection at all. Their approach falls a bit short of the nearest neighbour classifier and is a good candidate when the amount of data prohibits batch training. Another graph-based approach is (Raytchev and Murase, 2003), where face sequences act as nodes and node attraction and repulsion are defined in the sequence proximity matrix. Two clustering algorithms are introduced that can lead to unsupervised face recognition.

Li et al. utilize a pose estimator to fit a multi-view dynamic face model on the video frames (Li et al., 2001). This gives pose invariant textures. Kernel discriminant analysis of those textures yields identity surfaces. Trajectories are defined on these surfaces using gallery videos, and are compared with those from probe videos for recognition. Lee et al. split the

gallery stills extracted from the videos of each person into pose manifolds (Lee et al., 2003). They then use the temporal information to learn the transition probabilities between those pose manifolds and to handle occlusions. Face detection is again not addressed. They show their approach to be superior to temporal voting across the 20 last extracted probe stills. Unlike other video-to-video face recognition methods, they report performance on a per still, not video probe basis, which does not reflect the goal of such algorithms. Liu and Chen use temporal information in gallery face sequences to train Hidden Markov Models (HMMs) (Liu and Chen, 2003). The probe face sequences are analyzed with each of the trained HMMs, to yield the person identity based on maximum likelihood scores. Face sequences are manually extracted from the videos. They show enhanced performance compared to post decision fusion using voting. Aggarval et al. use temporal information to learn ARMA pose variation models from gallery and probe face sequences (Aggarval et al., 2004). They then employ model matching criteria to associate a gallery model to each probe one. Face detection is again not addressed.

All the above algorithms perform face detection and recognition independently. Zhou et al. on the other hand perform face tracking and recognition jointly in a particle filtering framework by adding an identity variable in the state vector and demanding identity consistency across time. In (Zhou et al., 2003) they show good performance employing the extracted probe stills as appearance models for tracking, while in (Zhou et al., 2004) they improve tracking robustness for moderate pose changes and occlusions using adaptive appearance and state transition models.

The various databases used for video-to-video face recognition are characterized by the number of individuals, the degree of pose and illumination variations, the recording conditions (far, medium or near field), the duration of the gallery and probe videos and the number of probe videos. Some things are common in these databases. The number of different people to be recognized is much smaller than the still-to-still face recognition databases. While in the latest Face Recognition Grand Challenge (Philips et al., 2006) there are thousands of different individuals, all video-to-video face recognition algorithms are tested on video databases of 10 to 33 individuals. The only exception is (Weng et al., 2000), which employs 143 individuals. There is no significant temporal separation between gallery and probe videos; the difficulty of the task stems from the fact that there is action depicted in the videos, that results to gross pose, expression and illumination changes and the lower quality images, as the resolution of the faces is typically much smaller than the one found in still-to-still face recognition databases. Most of the algorithms are tested with videos taken indoors. Exceptions can be found in some experiments of (Zhou et al., 2003) and in (Raytchev and Murase, 2003). In most cases the recording conditions are near-field: The faces occupy a significant part of the image, either during the whole of the video (Weng et al., 2000; Li at al., 2001; Liu and Chen, 2003) or towards the end of it as the people are walking towards the camera (Gorodnichy, 2003; Raytchev and Murase, 2003; Zhou et al., 2003; Xie et al., 2004). The only truly far-field video recordings known to the authors are those collected by the partners of the CHIL project (Waibel et al., 2004) and already used in the Classification of Events, Activities and Relationships (CLEAR 2006) evaluations (Stiefelhagen et al., 2007). Unfortunately, many of the algorithms in the field are only tested on custom built video databases, which are not publicly available, or for which not all the necessary data are reported. Unlike still-to-still face recognition, there have been no evaluations for its video-to-video counterpart. The single exception are the CLEAR 2006 and

the upcoming CLEAR 2007 evaluations (Stiefelhagen et al., 2007), which include a video-to-video face recognition task. Table 1 summarizes the most commonly used and publicly available video databases.

Parameter	MoBo	CLEAR 2006
No. of people	25	26
Camera views	Single, facing person	4, at room corners
Gallery duration	10 sec	15 and 30 sec
Probe duration	10 sec	1, 5, 10 and 20 sec
No. of probe videos	74	613 (1 sec), 411 (5 sec), 289 (10 sec) and 178 (20 sec),
Scenario	Walking on a treadmill	Moving freely: meeting with presentation
Pose, expression	Approximately frontal; always both eyes visible	Any pose, natural talking expression
Illumination	Constant	Changes due to projector beam, overhead lights
Recording conditions	Medium field, 30 to 40 pixels wide faces	Far field, median eye distance 9 pixels

Table 1. Summary of publicly available video databases used for video-to-video face recognition. The frame rate is 30 fps

Note that the pose variations in the CLEAR 2006 database are extreme: some of the shorter videos do not contain any face with both eyes visible. This is alleviated by the use of 4 different camera views: one of the views is bound to capture some frames with faces having both eyes visible. The durations reported in Table 1 for this database are per camera view; there are actually four times as much frames to extract faces from.

While some of the algorithms that jointly utilize temporal information and perform recognition claim better results than post-decision fusion, the latter should not be discounted for two reasons. Firstly, only simple (not weighted) voting is used in these comparisons. Secondly, all these algorithms are based on learning the evolution of a face manifold, as pose, expression and illumination change with time. On the one hand, there can be valid changes in the probe videos not present in the gallery videos. On the other hand, the face manifold depends on the appearance of the face, which is not only dependant on pose, expression and illumination, but also on face detection accuracy. The randomness of face detection errors leads to greater face manifold spreading with random transitions. Attempting to learn such random transitions just overfits the classifier on the gallery data.

The effect of these errors is even more pronounced on far-field viewing conditions and unconstrained people movement, where face detection is much harder. All these algorithms have not been tested on such videos. For this reason, we have chosen the post-decision fusion scheme in (Stergiou et al., 2006) for the far-field, unconstrained video-to-video face recognition system detailed in the next section.

3. Proposed face recognition system

In this section we analyze the different options for video-to-video face recognition using the CLEAR 2006 database. We present different solutions for all the detection and recognition subtasks and we investigate their effect on recognition rate. For the reasons discussed in section 2 we choose a post-decision fusion scheme to utilize the temporal information in the video streams.

3.1 Face detection for gallery and probe generation

The CLEAR 2006 database comes with a set of annotations (Stiefelhagen et al., 2007). The face bounding box is marked every 1sec, while the centers of the eyes every 200ms. The lower frequency of the face annotations is due to the severe difficulty of this kind f annotation. Hence the first option for face detection is to simply use these labels to extract the faces. The labels are linearly interpolated to provide the eyes of the person in each frame. Should two eyes exist, the face is cropped, normalized and added to the probe or gallery. Normalization accounts for face geometry and illumination changes. First the marked eyes are positioned on specific coordinates on a 34 by 42 template that contains mostly the face for approximately frontal views of the people. This is a big template for most of the faces; it is selected to favor upsampling of the small faces to downsampling of the large ones. No deliberate perturbation of the eye positions is carried out to alleviate the effect of eye labeling errors (Lee et al., 2003; Ekenel and Pnevmatikakis, 2006). Such an approach is very important for small galleries, and has been applied in the past on still-to-video face recognition on data similar to those of the CLEAR2006 (Ekenel and Pnevmatikakis, 2006), but the rich gallery of this dataset is enough to randomize the errors and alleviate their effect. Then the intensity is made zero-mean, unit variance. Although more aggressive normalization techniques exist to account for illumination changes (Pnevmatikakis & Polymenakos, 2005), these also degrade performance under pose and expression changes (Pnevmatikakis & Polymenakos, 2005). Hence the mild normalization approach is taken here, to provide some immunity to illumination changes without degrading performance under pose changes too much. The normalized gallery images extracted for one person are shown in Figure 1.

Evidently there are problems with the accuracy of the interpolated labels, or the 200 ms labels themselves, that lead to scaling errors, shifting and rotation of the faces. Such effects can be from minor up to major, leading to image segments that are definitely not faces (end of row four, beginning of row five). Also, there are pose variations, both left-right (even extreme profile with only one eye visible – row five) and up-down. Finally note the large resolution changes; there are faces where details are visible, and others that are a blur due to the upsampling to bring them to a standard size (contrast the level of detail in the two last rows). The gross resolution variation present in the probe videos is apparent in the histogram of eye distances shown in Figure 2.

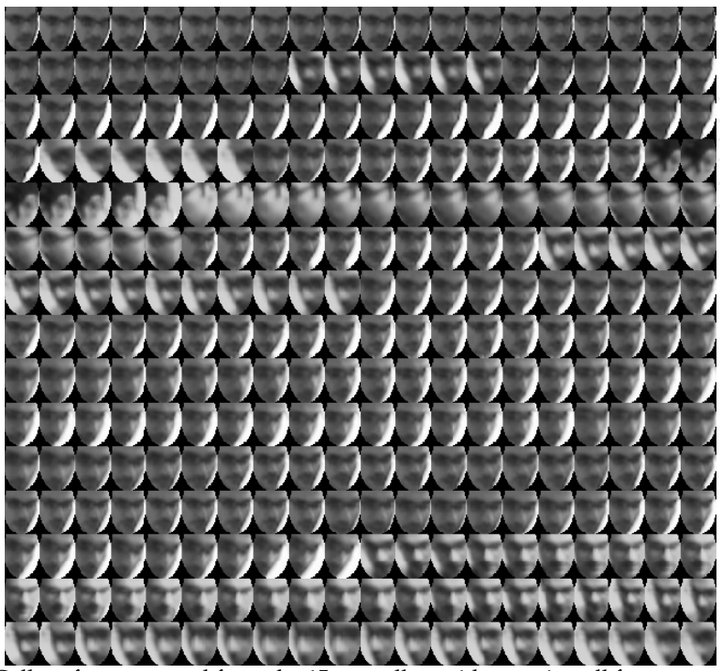


Figure 1. Gallery faces cropped from the 15 sec gallery videos, using all four cameras, for one person

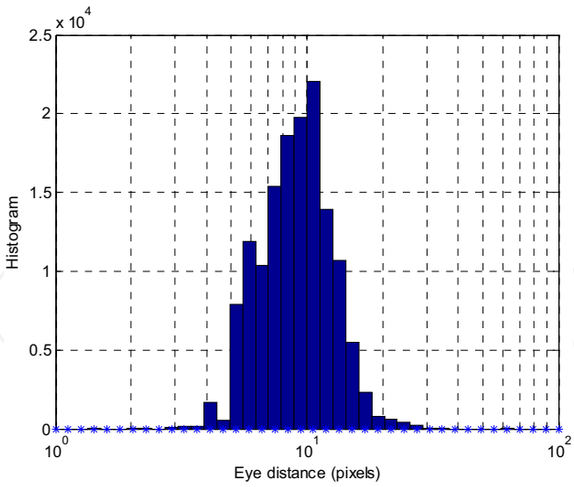


Figure 2. Histogram of the eye distances of the faces segmented from the probe videos using the manual annotations. The video-to-video face recognition system has to cope with eye distances of 4 to 28 pixels

When the view is not approximately frontal, then the template might include other parts of the head, or even background. Such views are not wanted, and some means for automatically discarding them is needed. Note at this point that automatic selection of faces is a prerequisite only for the probe videos. But it is not only cumbersome to manually filter the gallery stills; such a selection can cause mismatches between the automatically selected probe stills and the manually selected gallery stills. For both these reasons an automatic mechanism for the selection of faces is utilized. This mechanism employs a measure of frontality, based on the supplied face bounding boxes and eye positions. Frontal views should have both eyes symmetrically positioned around the vertical face axis. This symmetry is enumerated in the frontality measure. The measure can unfortunately be inaccurate for two reasons. The first has to do with the provided label files: eye positions are provided every 200 ms, while face bounding boxes every 1 sec, causing larger errors due to interpolation. The second reason has to do with the positioning of the head: when it is not upright, then the major axis of the face does not coincide with the central vertical axis of the face bounding box. Nevertheless, employing the proposed frontality measure rids the system from most of the non-frontal faces at the expense of missing some frontal but tilted ones. As for the threshold on frontality, this should not be too strict to diminish the training and testing data. It is set to 0.1 for all training durations and testing durations up to 10 sec. For testing durations of 20 sec, it is doubled, as the abundance of images in this case allows for a stricter threshold. A final problem with the application of the frontality threshold is that there are some testing segments for which both eyes are never visible. This leads to empty segments. These profile faces can in principle be classified by face recognizers trained on profile faces, but such classifiers have not been implemented in the scope of these experiments. The still gallery and probe sets generated using the face annotations are summarized in Table 2.

Face cropping method		Interpolated hand-annotated eye centers				Viola-Jones detector			
Face normalization		De-rotation using the eye centers, scaling to 42 by 34 pixels				No de-rotation, scaling to 48 by 36 pixels			
Gallery stills per person	Length (sec)	15		30		15		30	
	Min	47		56		118		251	
	Average	241		517		428		886	
	Max	613		1213		890		1696	
Probe stills per video	Length (sec)	1	5	10	20	1	5	10	20
	Min	0	0	0	0	1	2	19	81
	Average	16	78	148	301	25	127	226	515
	Max	60	282	479	930	90	348	793	1406
	Empty videos	13%	3.4%	1.7%	1.1%	0	0	0	0

Table 2. Summary of the gallery and probe still sets generated from the CLEAR 2006 videos using either the provided face annotations or the trained cascaded detector

Basing the gallery and probe generation of video-to-video face recognition on annotations is not good practice. Annotations and expensive and inaccurate, both because it is difficult to label facial features on far-field recordings, and because interpolation is needed, as the

frames are annotated sparsely. Also, actual systems have to be fully automatic. Hence a face detector is needed. As multiple people are present in the frames, and the faces are tiny compared to the frame size, the natural choice for a detector is the boosted cascade of simple features (Viola and Jones, 2001). Although many improvements on the original algorithm have been proposed (Li and Zhang, 2004; Schneiderman, 2004), we opted to stick to the original version that uses AdaBoost and its implementation in OpenCV (Bradski, 2005), as this is publicly available. Although a trained cascade of simple classifiers is already provided with OpenCV, it is not suitable for our needs as the faces in our far-field recordings are too small. That detector has very high miss rate. A more suitable detector is thus trained. To do so we use 6,000 positive samples (images with marked faces), 20,000 negative samples (images with no human or animal face present), an aspect ratio of 3/4, minimum feature size 0, 99.9% hit rate, 50% false alarm, tilted features, non-symmetric faces and gentle AdaBoost learning (Bradski, 2005). We run the cascaded classifier on all the frames of the gallery and probe videos, and we collect the faces. Note that due to the existence of many people in the frames, the labels are still needed to tell apart the person under consideration from the other meeting participants. If any detection exists close to the provided face bounding box, then it is selected as the face of interest. The temporally subsampled gallery images for the same person shown in Figure 1 are shown in Figure 3.

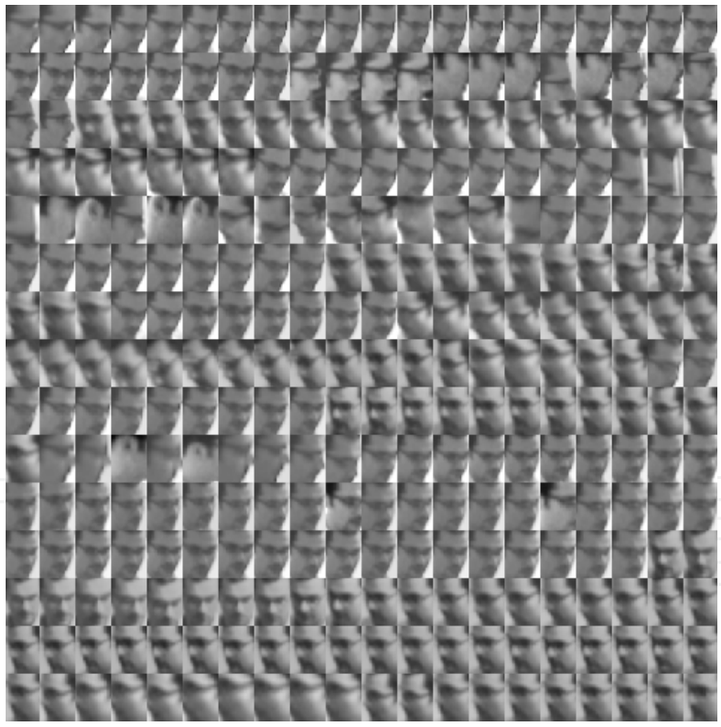


Figure 3. Temporally subsampled gallery faces automatically cropped from the 15 sec gallery videos, using all four cameras, for one person

Comparing the faces in Figure 1 and 3, it is evident that using the automatic detection scheme we get more faces, but less accurately framed than with the face annotations. Also, there is no attempt to geometrically normalize the faces based on the eye positions, nor any filtering of profile faces. The statistics of the automatically extracted gallery and probe stills are also shown in Table 2.

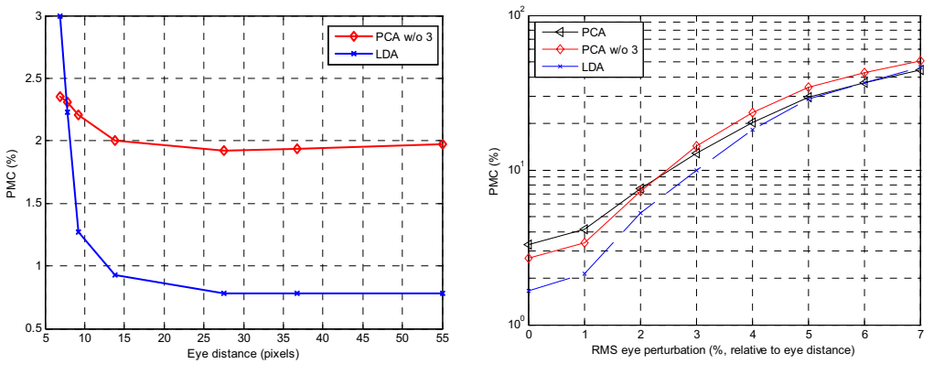
3.2 Classification

For classification the gallery faces are vectorized by rearranging the intensities of their pixels into a vector, e.g. by reading the intensities in a column-wise fashion. The mean vector is subtracted, yielding zero-mean vectors, to be used for the training of the classifiers.

The classifiers employed are of the linear subspace projection family. Both Principal Components Analysis (PCA) (Turk and Pentland, 1991) and Linear Discriminant Analysis (LDA) (Belhumeur et al., 1997) are employed to build unsupervised and supervised projection matrices respectively. PCA aims at transforming the training vectors so that their projections in lower-dimensional spaces has maximum scatter. This guarantees optimality in terms of minimum squared error of the representation of the original vectors in any lower-dimensional space (Duda et al., 2000). The determination of the transformation matrix does not require any class information, hence it is unsupervised. Although the optimality in representation does not offer any guarantee for optimality in classification, the use of PCA has led to the successful Eigenface face recognition method (Turk and Pentland, 1991). The dimension D of the recognition subspace onto which the training vectors are projected is a parameter of the method, to be determined empirically. Suppressing some of the dimensions along which the scatter of the projected vectors is smallest not only increases the speed of the classification, but also seems to be suppressing variability that is irrelevant to the recognition, leading to increased performance. LDA on the other hand aims at maximizing the between-class scatter under the constraint of minimum within-class scatter of the training vectors, effectively minimizing the volume of each class in the recognition space, while maximizing the distance between the classes (Duda et al., 2000). The dimensions of the LDA subspace is $K-1$, where K is the number of classes. The determination o the LDA projection matrix requires class information, hence it is supervised. LDA suffers from ill-training (Martinez and Kak, 2001), when the training vectors do not represent well the scatter of the various classes. Nevertheless, given sufficient training, its use in the Fisherfaces method (Belhumeur et al., 1997) has led to very good results.

LDA is better for large faces with accurate eye labels (Rentzeperis et al., 2006), but PCA is more robust as face size and eye labeling accuracy drop. LDA is robust to illumination changes (Belhumeur et al., 1997). PCA can be made more robust to illumination changes if some of the eigenvectors corresponding to the largest eigenvalues are excluded from the projection matrix, but this reduces the robustness of PCA under eye misalignment errors. At far-field viewing conditions, resolution is low and the accurate determination of the eye position is very difficult, even for human annotators. To demonstrate the difficulties the far-field viewing conditions impose on face recognition, a comparison of the error rate of PCA, PCA without the three eigenvectors correspond=ding to the three largest eigenvalues (PCA w/o 3) and LDA is carried out in Figure 4, for different face resolutions and eye alignment accuracies. Note that the database used for these experiments is not the video database of CLEAR 2006, but HumanScan (Jesorsky et al., 2001) that offers very large faces which can be

decimated to smaller dimensions and the evaluation methodology is the one presented in (Pnevmatikakis and Polymenakos 2005). The probability of misclassification (PMC) increases below 10 pixels of eye distance, even with perfect eye labelling, and LDA can become worse than PCA, even when as many as 10 gallery faces per person are used (Figure 4.a). The PMC degrades even less gracefully when the faces are registered with incorrect eye positions. For 5 gallery faces per person and RMS eye alignment errors greater than 5% of the eye distance, PCA and LDA perform similarly. PCA w/o 3 becomes worse than PCA for eye misalignments larger than 2% of the eye distance



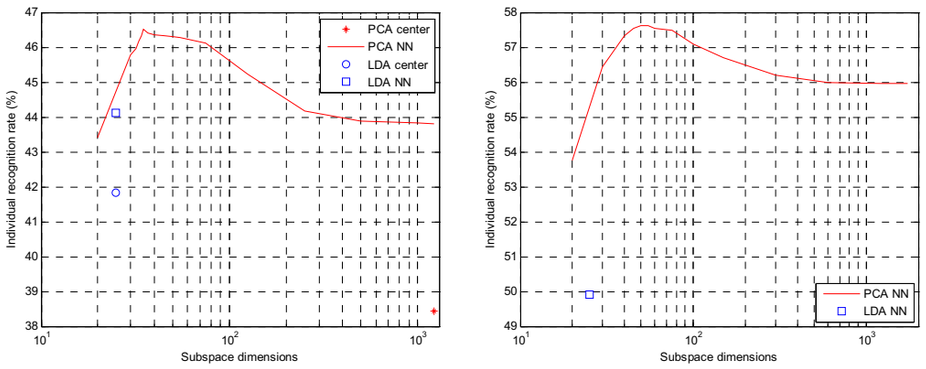
(a) (b)

Figure 4. Effect of far-field viewing conditions on linear subspace projection face recognition. (a) Performance as a function of face resolution. (b) Performance as a function of eye misalignment

It is evident from the above example that the performance of LDA and PCA at the face resolutions and eye misalignments of interest is expected to be very close, but each method performs better under different conditions. When there are many probe images per testing segments, LDA is expected to be a better choice to PCA. The latter is expected to surpass LDA when there are fewer gallery images or more probe images to fuse the individual decisions. Hence both methods are used, and their results are fused, as explained in the next section. A note is due at this point for the application of LDA. Contrary to the Fisherfaces algorithm (Belhumeur et al., 1997), in this case the small sample size problem (Yu and Yang, 2001) does not apply. The number of pixels of the faces is smaller than the available gallery stills, no matter the gallery duration or the face cropping method employed. Hence no PCA step is used, without the need for a direct LDA algorithm (Yu and Yang, 2001).

According to the Eigenfaces (Turk and Pentland, 1991) or Fisherfaces (Belhumeur et al., 1997) methods, the gallery images are represented by their class means after projection to the recognition space. Recognition is based on the distance of a projected gallery face from those means. This is not effective in the case of unconstrained movement of the person, since then the intra-personal variations of the face manifold due to pose variations can be far more pronounced than the extra-personal variations (Li et al., 2001). In this case it is better to use a nearest neighbour classifier. The implication is that all the projected gallery faces have to be kept and compared against every probe projected face.

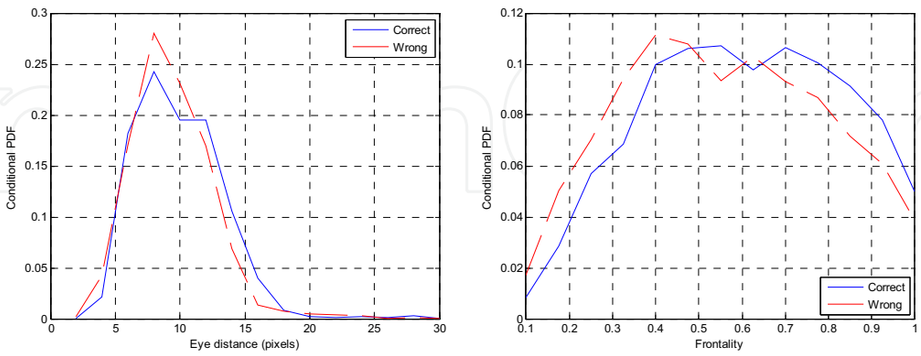
Different distance metrics can be used for classification. When the probe faces are compared to the gallery class centres, then the weighted Euclidian distance is used for PCA projection and the Cosine for LDA projection (Pnevmatikakis & Polymenakos, 2005). When the comparison is against any individual gallery face, then the Euclidian distance is used.
Although the individual recognition rate for each probe face is not the goal of the video-to-video system, it is instructive to report it for the different options of LDA and PCA classifiers. This is done in Figure 5 for the manually cropped faces using the annotations and the automatically cropped faces from the 15 sec long gallery and the 1 sec long probe videos. Obviously, for manual cropping, the best recognition results with PCA (46.5%) are obtained using the nearest neighbour classifier and retaining 35 dimensions in the recognition space. The best individual results with LDA (44.1%) are again obtained using the nearest neighbour classifier. For automatic cropping, the best recognition results with PCA (57.5%) are obtained using the nearest neighbour classifier and retaining 45 dimensions in the recognition space. The higher optimum recognition subspace dimension for this case is justified by the higher maximum recognition subspace dimension due to the increased normalized resolution of the automatically cropped faces. The best individual results with LDA (49.9%) are again obtained using the nearest neighbour classifier, but notice in this case how worse the LDA performance is compared to PCA.



(a) (b)

Figure 5. Individual PMC for the manually cropped faces using the annotations (a) and the automatically cropped faces (b) from the 15 sec long gallery and the 1 sec long probe videos. The effect of projection type (PCA or LDA), classifier (class centre or NN) and recognition space dimension is shown

Finally, the correlation of successful individual recognition to face resolution and frontality is investigated. The probability density functions (PDF) of eye distance and frontality conditioned on correct or wrong recognition results are shown in Figure 6, again for the manually cropped faces using the annotations in the 15 sec long gallery and the 1 sec long probe videos. It can be seen that compared to the PDFs given wrong results, the shift of the PDFs given correct results towards larger eye distances or frontality values is very small. This signifies that the performance of the system does not depend significantly on the pose or the size of the faces.



(a) (b)

Figure 6. Conditional PDFs of eye distance and frontality leading to correct or wrong recognition

3.3 Post-decision fusion

A two-stage fusion scheme is employed, based on the sum rule (Kittler et al., 1998). The first stage performs fusion jointly across time and camera views, while the second stage fuses the results of the two classifiers. The fusion scheme is illustrated in Figure 7.

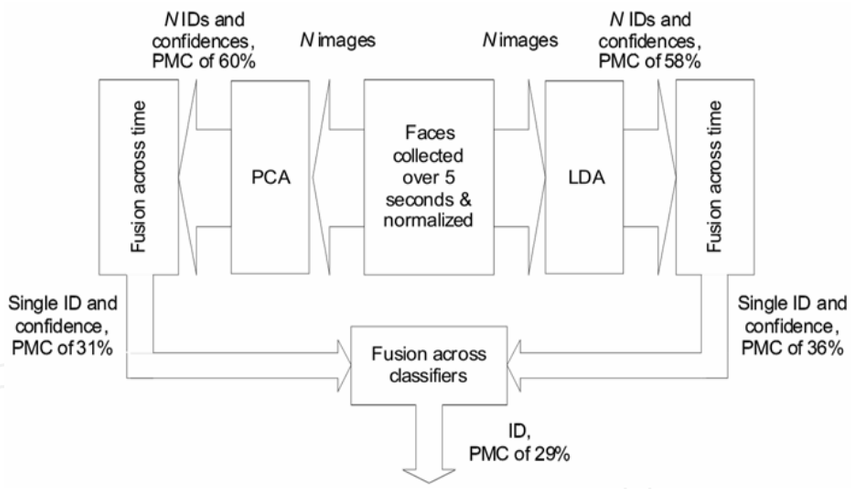


Figure 7. Two-stage fusion scheme. The PMC shown at the various stages of the scheme correspond to the 15 sec gallery videos, face extraction using the provided annotations and classifying the extracted probe stills according to the distance from the gallery class centres

The individual decisions for the probe faces are fused using the sum rule (Kittler et al., 1998). According to the sum rule, each of the decision ID_i of the probe faces in a testing

segment casts a vote that carries a weight w_i. The weights w_i of every decision such as $ID_i = k$ are summed to yield the weights W_k of each class:

$$W_k = \sum_{i:ID_i=k} w_i \tag{1}$$

where $k = 1,\ldots,K$ and K is the number of classes. Then the fused decision based on the N individual identities is:

$$ID^{(N)} = \arg\max_k (W_k) \tag{2}$$

The weight w_i in the sum rule for the i-th decision is the sixth power of the ratio of the second-minimum distance $d_i^{(1)}$ over the minimum distance $d_i^{(1)}$:

$$w_i = \left[\frac{d_i^{(2)}}{d_i^{(1)}}\right]^6 \tag{3}$$

This choice for weight reflects the classification confidence: If the two smallest distances from the class centers are approximately equal, then the selection of the identity leading to the smallest distance is unreliable. In this case the weight is close to unity, weighting down the particular decision. If on the other hand the minimum distance is much smaller than the second-minimum, the decision is heavily weighted as the selection of the identity is reliable. The sixth power allows for a few very confident decisions to be weighted more then many less confident ones. The suitability of the proposed weights is demonstrated in Figure 8, where the conditional cumulative density functions (CDF) of the weights, conditioned on correct or wrong recognition are shown for the manually cropped faces using the annotations and the automatic detection scheme, in the 15 sec long gallery and the 1 sec long probe videos.

It is evident from Figure 8 that the probability of wrong recognition diminishes as the proposed weight increases, hence they can be used in a weighted voting scheme. The fused recognition rate of PCA increases from the 71.7% obtained by majority voting, to 72.8% obtained by using the proposed weighted voting scheme. Also, the weights for the faces cropped using the automatic detection scheme are more suitable than those of the manual: The CDFs given wrong decisions are practically the same, while the CDF given correct decisions for the automatic scheme is shifted to larger weights compared to that for manual cropping. Hence, not only the individual recognition rates for the automatic scheme are higher (see Figure 5), but in addition it is expected that the gain due to fusion will be higher. Indeed, fusing the individual PCA results on the manually cropped probes from the 1 sec long videos, we obtain a recognition rate of 53.8%, with a relative increase from the individual rate of 15.7%. On the other hand, fusing the individual PCA results on the automatically cropped probes, we obtain a recognition rate of 72.8%, with a relative increase from the individual rate of 26.5%.

The decisions $ID^{(\mathrm{PCA})}$ and $ID^{(\mathrm{LDA})}$ of the PCA and the LDA classifiers are again fused using the sum rule to yield the reported identity. For this fusion, the class weights W_k of equation (1) are used instead of the distances in equation (3). Setting:

$$k_1 \equiv [\text{best matching class}] = ID^{(N)}$$
$$k_2 \equiv [\text{second-best matching class}] \tag{4}$$

the weights of the PCA and LDA decisions become:

$$w_i = \frac{W_{k_1}^{(i)}}{W_{k_2}^{(i)}},\ i \in \{\text{PCA, LDA}\} \tag{5}$$

Then the fused PCA/LDA decision is:

$$ID = \begin{cases} ID^{(\text{PCA})} & \text{if } w_{\text{PCA}} \geq w_{\text{LDA}} \\ ID^{(\text{LDA})} & \text{if } w_{\text{PCA}} < w_{\text{LDA}} \end{cases} \tag{6}$$

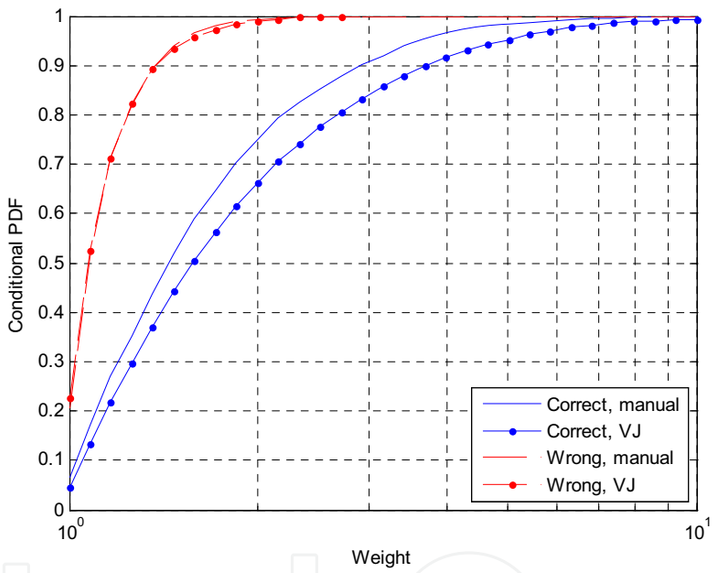


Figure 8. Conditional cumulative density functions of the weights, conditioned on correct or wrong recognition are shown for the manually cropped faces using the annotations and the automatic detection scheme, in the 15 sec long gallery and the 1 sec long probe videos. The weights from the PCA classifier are used

3.4 Performance

The performance of the video-to-video face recognition system described in this section is presented next. This system using the manual annotations for gallery and probe still generation and classification based on the distance from projected gallery class centres has been evaluated in CLEAR 2006. Performance can be significantly boosted using the nearest neighbour classifier, especially for the 30 sec long gallery videos. An even greater

performance boost is achieved by using the automatic face detection scheme. The somehow degraded framing of the faces in some still images thus generated is by far compensated by the larger number of gallery stills available for training and the larger number of probe stills per test, that allow for more efficient post-decision fusion. The recognition rate in the probe videos is presented in Table 3 and Figure 9. For comparison, also the best performance achieved in the CLEAR 2006 evaluations is also included.

Method	15 sec gallery duration				30 sec gallery duration			
	Probe duration (sec)							
	1	5	10	20	1	5	10	20
Annotations, distance from class centres (Man-centre)	49.4	70.3	75.8	79.8	52.7	68.9	73.4	75.3
Annotations, nearest neighbour (Man-NN)	53.8	72.3	78.2	83.1	60.7	79.6	85.5	91.6
Viola-Jones detector, nearest neighbour (VJ-NN)	72.8	86.6	87.9	93.3	79.5	93.47	93.8	97.8
CLEAR-Best	62.3	73.2	79	80.7	71	81.5	83.9	85.2

Table 3. Average recognition rates for the various probe video durations, given any of the two gallery video durations. The first three entries correspond to the different options for the system described in this section, while the last one refers to the best performance reported (across all systems) in the CLEAR 2006 evaluations

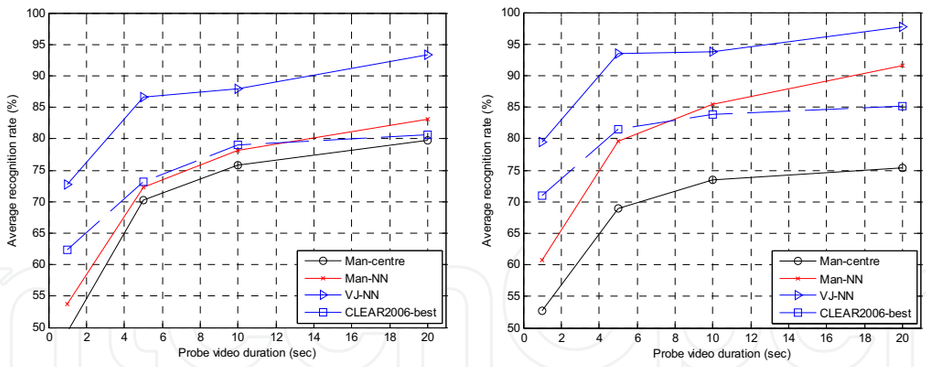


Figure 9. Average recognition rates for the various probe video durations, for (a) 15 sec gallery videos duration and (b) 30 sec gallery videos duration

Next we investigate the effect of the amount of probe faces extracted from the videos and of the weights obtained when the probes are recognized individually on the correct recognition over the complete sequence. Figure 10 depicts the scatter plot of the maximum weight

versus the number of probe faces extracted, for each of the 1 sec long probe videos that lead to correct or wrong recognition.

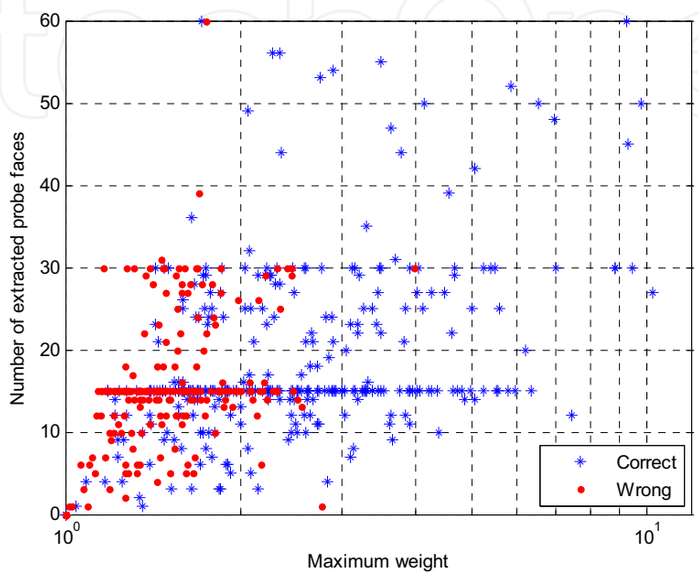


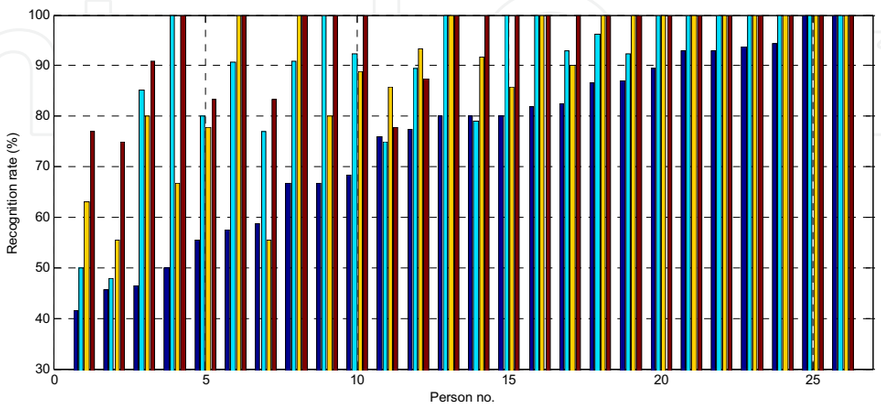
Figure 10. Scatter plot of the maximum weight versus the number of probe faces extracted, for each of the 1 sec long probe videos that lead to correct (asterisks) or wrong (points) recognition

The more probe faces the system extracts and the highest the maximum weight from the individual recognition is, the easiest is the person in the video correctly recognized. For all practical reasons, when there is a weight higher than 2.5 or there are more than 30 extracted probe faces, the person is identified correctly. Given longer probe video durations, these conditions are more likely to be met. Of course this depends on the situation depicted in the video, for example a person looking down all the time.

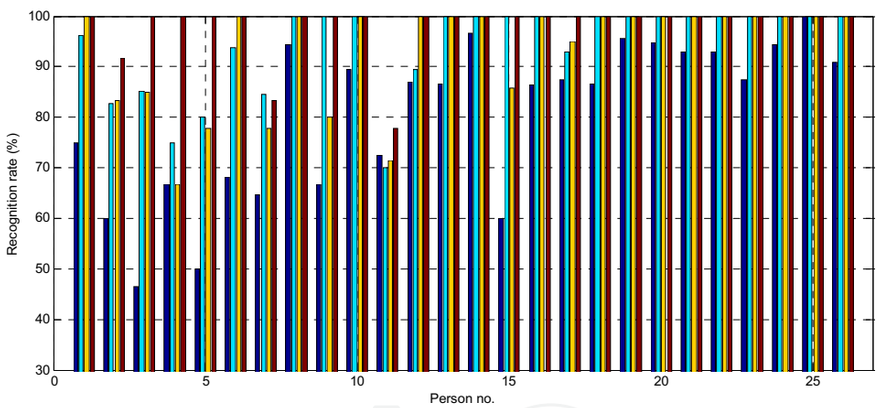
Finally, it is interesting to investigate if some people are harder to recognize than others. The bar graph of Figure 11 depicts the recognition rates for the 26 different people, under the two training and four testing conditions. Some people that are hard to recognize remain so no matter the gallery or probe video lengths. This variation in the performance across different people can not be attributed to the properties of the extracted gallery or probe faces; like their number, eye distance of frontality metric. It is due to the difference in matching between training and testing conditions: Some people act similarly in the gallery and probe videos, hence appearing similar, while others do not.

It is evident from Figure 11 that not always people that are very difficult to recognize in one of the eight training and testing conditions remain difficult in other conditions. This is because the actions of a person in the probe and gallery videos can be more or less matched as those videos change. For example, the most difficult person in the 15 sec gallery video

and 1 sec probe videos, is easier than people 2 and 7 in the 10 sec probe videos, and easier than people 2-6, 8, 10 and 14 in the 30 sec gallery video.



(a)



(b)

Figure 11. Per person recognition rates for the different durations of the probe videos (grouped) and for the 15 sec (a) or 30 sec (b) long gallery videos. The people are sorted by ascending recognition rate for the 1 sec long probe and the 15 sec long gallery videos

Finally, there is a large deviation in recognition performance in the 15 sec gallery video and 1 sec probe videos. This drops somewhat for longer probe and gallery videos. This is demonstrated in Figure 12, where the standard deviation of the recognition rate across the 26 different people is depicted for the four probe video durations and the two gallery video durations. Hence increasing the probe or gallery durations tend to make performance across different people both better and more uniform.

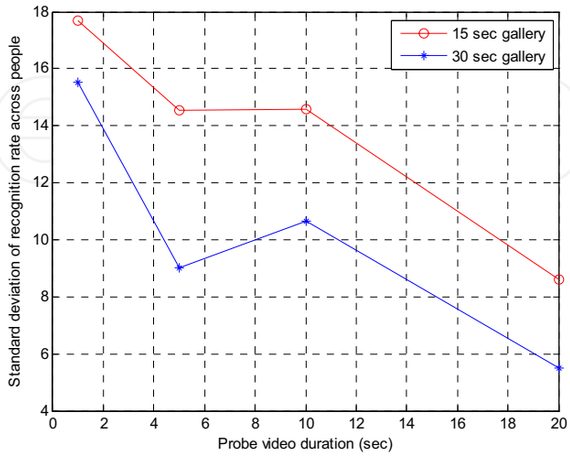


Figure 12. Standard deviation of the recognition rate across the 26 different people for the four probe video durations and the two gallery video durations. Performance across the different people is more uniform as the durations increase

4. Conclusion and possible extensions

In this chapter we have presented the tradeoffs in video-to-video face recognition, applied on far-field, unconstrained recordings. We have demonstrated that given long probe video durations the performance of a system based on a frontal Viola-Jones face detector, linear subspace projection and nearest neighbour classifier more or less solves the problem, with average recognition rates above 95%. In applications where long probe videos are impractical, performance is still low (recognition rates of 74% or 80% for 1 sec probe and 15 sec or 30 sec gallery video durations), especially given that the number of people are limited to the modest number of 26. To further enhance performance, there are some possible system enhancements:

- Multiple face detectors can be trained, including poses other than frontal. Also, face detection can be coupled with a probabilistic tracker based on particle filtering (Zhou et al., 2004) or a deterministic tracker based on colour histograms using CAMShift (Bradski, 1998). This will provide more stills, capturing more pose variations.
- Other distance metrics (weighted Euclidian, cosine) can be used for nearest neighbour classification.
- Modelling of face sequences, similar to the exemplar approach of (Zhou et al., 2003), to automatically detect outliers that are not smooth pose transitions, but rather face detector errors. The cleaner face sequences thus obtained can be used to model pose transitions, allowing more efficient utilization of temporal information than weighted voting (Weng et al., 2000; Li et al., 2001; Lee et al., 2003; Liu and Chen, 2003; Aggarval et al., 2004).

5. Acknowledgements

This work is sponsored by the European Union under the integrated project CHIL, contract number 506909.

6. References

Aggarwal, G.; Roy-Chowdhury, A.K. & Chellappa, R. (2004). A System Identification Approach for Video-based Face Recognition, *Proceedings of International Conference on Pattern Recognition*, Cambridge, UK, Aug. 2004

Belhumeur, P.; Hespanha, J. & Kriegman, D. (1997). Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, 7, 711-720

Bradski, G. (1998). Computer Vision Face Tracking for Use in a Perceptual User Interface. *Intel Technology Journal*, 2

Bradski, G.; Kaehler, A. & Pisarevsky, V. (2005). Learning-Based Computer Vision with Intel's Open Source Computer Vision Library. *Intel Technology Journal*, 9

Duda, R.; Hart, P. & Stork, D. (2000). *Pattern Classification*. Wiley-Interscience, New York

Ekenel, H. & Pnevmatikakis, A. (2006). Video-Based Face Recognition Evaluation in the CHIL Project – Run 1, *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*, pp. 85-90, Southampton, UK, Apr., 2006

Gorodnichi, D. (2003). Facial Recognition in Video. In: *AVBPA 2003, Lecture Notes in Computer Science 2688,* Kittler, J. & Nixon, M.S. (Ed.), 505-514, Springer-Verlag, Berlin Heidelberg

Jesorsky, O.; Kirchberg, K. & Frischholz, R. (2001). Robust Face Detection Using the Hausdorff Distance. In Bigun, J. & Smeraldi, F. (ed.), *Audio and Video based Person Authentication*, 90-95, Springer-Verlag, Berlin Heidelberg

Kittler, J.; Hatef, M.; Duin, R.P.W. & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20, 3, 226–239

Lee, K.-C..; Ho, J.; Yang, M.-H. & Kriegman, D. (2003). Video-based face recognition using probabilistic appearance manifolds. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,* 1, pp. 313-320, Madison, Wisconsin, USA, June 2003

Li, Y.; Gong, S. & Liddell, H. (2001). Video-Based Online Face Recognition Using Identity Surfaces. *Proceedings of IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pp. 40-46, Vancouver, Canada, July 2001

Li, S.-Z. & Zhang, Z.Q. (2004). FloatBoost Learning and Statistical Face Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 26, 9, 1112-1123

Liu, X.. & Chen, T. (2003). Video-based face recognition using adaptive hidden markov models. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1, pp. 340-345, Madison, Wisconsin, USA, June 2003

Martínez , A. & Kak, A. (2001). PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23, 2, 228-233

Phillips, J.; Flynn, P.; Scruggs, T.; Boyer, K. & Worek, W. (2006). Preliminary Face Recognition Grand Challenge Results. *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*, pp. 15-21, Southampton, UK, Apr., 2006

Pnevmatikakis, A. & Polymenakos, L. (2005). A testing methodology for face recognition algorithms. In: *MLMI 2005, Lecture Notes in Computer Science 3869,* Renals, S. & Bengio, S. (Ed.), 218-229, Springer-Verlag, Berlin Heidelberg

Raytchev, B. & Murase, H. (2003). Unsupervised recognition of multi-view face sequences based on pairwise clustering with attraction and repulsion. *Computer Vision and Image Understanding,* 91, 22-52

Rentzeperis, E.; Stergiou, A.; Pnevmatikakis, A. & Polymenakos, L. (2006). Impact of Face Registration Errors on Recognition. *Artificial Intelligence Applications and Innovations*, Peania, Greece, June 2006

Schneiderman, H. (2004). Feature-Centric Evaluation for Efficient Cascaded Object Detection. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, June 2004

Stergiou, A.; Pnevmatikakis, A. & Polymenakos, L. (2007). A Decision Fusion System across Time and Classifiers for Audio-visual Person Identification. In: *CLEAR 2006, Lecture Notes in Computer Science 4122,* Stiefelhagen, R. & Garofolo, J. (Ed.), 218-229, Springer-Verlag, Berlin Heidelberg

Stiefelhagen, R.; Bernardin, K., Bowers, R., Garofolo, J., Mostefa, D. & Soundararajan, P. (2007). The CLEAR 2006 Evaluation. In: *CLEAR 2006, Lecture Notes in Computer Science 4122,* Stiefelhagen, R. & Garofolo, J. (Ed.), 218-229, Springer-Verlag, Berlin Heidelberg

Turk, M. & Pentland, A. (1991). Eigenfaces for Recognition. *J. Cognitive Neuroscience*, 3, 71-86

Viola, P. & Jones, M. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, p. 511, Hawaii, Dec. 2001

Waibel, A.; Steusloff, H. & Stiefelhagen, R. (2004). CHIL: Computers in the Human Interaction Loop, *5th International Workshop on Image Analysis for Multimedia Interactive Services*, Lisboa, Portugal, April 21-23, 2004

Weng, J.; Evans, C.H. & Hwang, W.-S. (2000). An Incremental Learning Method for Face Recognition under Continuous Video Stream. *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*, pp. 251-256, Grenoble, France, March 2000

Xie, C.; Vijaya Kumar, B. V. K.; Palanivel, S. & B. Yegnanarayana (2004). A Still-to-Video Face Verification System Using Advanced Correlation Filters. In: *ICBA 2004, Lecture Notes in Computer Science 3072,* Zhang, D. & Jain, A.K. (Ed.), 102-108, Springer-Verlag, Berlin Heidelberg

Yu, H. & Yang, J. (2001). A direct LDA algorithm for high-dimensional data with application to face recognition. *Pattern Recognition*, 34, 2067–2070

Zhou, S.; Krueger, V. & Chellappa, R. (2003). Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding,* 91, 7, 214-245

Zhou, S.; Chellappa, R. & Moghaddam, B. (2004). Visual tracking and recognition using appearance-adaptive models in particle filters. *IEEE Transactions on Image Processing,* 13, 11, 1491-1506

Face Recognition

Edited by Kresimir Delac and Mislav Grgic

ISBN 978-3-902613-03-5

Hard cover, 558 pages

Publisher I-Tech Education and Publishing

Published online 01, July, 2007

Published in print edition July, 2007

This book will serve as a handbook for students, researchers and practitioners in the area of automatic (computer) face recognition and inspire some future research ideas by identifying potential research directions. The book consists of 28 chapters, each focusing on a certain aspect of the problem. Within every chapter the reader will be given an overview of background information on the subject at hand and in many cases a description of the authors' original proposed solution. The chapters in this book are sorted alphabetically, according to the first author's surname. They should give the reader a general idea where the current research efforts are heading, both within the face recognition area itself and in interdisciplinary approaches.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Aristodemos Pnevmatikakis and Lazaros Polymenakos (2007). Far-Field, Multi-Camera, Video-to-Video Face Recognition, Face Recognition, Kresimir Delac and Mislav Grgic (Ed.), ISBN: 978-3-902613-03-5, InTech, Available from: http://www.intechopen.com/books/face_recognition/far-field__multi-camera__video-to-video_face_recognition

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen