7

Temporal Synchronization and Normalization of Speech Videos for Face Recognition

Usman Saeed[1] and Jean-Luc Dugelay[2]
[1]*Department of Computer Science, COMSATS Institute of Information Technology,*
[2]*Multimedia Communication Department, Eurecom-Sophia Antipolis,*
[1]*Pakistan*
[2]*France*

1. Introduction

Automatic Face Recognition (AFR) is a domain that provides various advantages over other biometrics, such as acceptability and ease of use, but due to the current trends, the identification rates are still low as compared to more traditional biometrics, such as fingerprints. Image based face recognition, was the mainstay of AFR for several decades but quickly gave way to video based AFR with the arrival of inexpensive video cameras and enhanced processing power.

Video based face recognition has several advantages over image based techniques, the two main being, more data for pixel-based techniques, and availability of temporal information. But with these advantages there are some inconveniences also, the foremost being the augmentation of variation. In the classical image based face recognition degraded performance has mostly been attributed to three main sources of variation in the human face, these being pose, illumination and expression. Among these, pose has been quite problematic both in its effects on the recognition results and the difficulty to compensate for it. Techniques that have been studied for handling pose in face recognition can be classified in 3 categories, first are the ones that estimates an explicit 3D model of the face (Blanz & Vetter, 2003) and then use the parameters of the model for pose compensation, second are subspace based such as eigenspace (Matta & Dugelay, 2008) and the third type are those which build separate subspaces for each pose of the face such as view-based eigenspace (Lee & Kriegman, 2005).

Managing illumination variation in videos has been relatively less studied as compared to pose, mostly image based techniques are extended to video. The two classical image based techniques that have been extended for video with relative success are illumination cones (Georghiades et al., 1998) and 3D morphable models (Blanz & Vetter, 2003). Lastly expression invariant face recognition techniques can be divided in two categories, first are based on subspace methods that model the facial deformations (Tsai et al., 2007). Next are techniques that use morphing techniques (Ramachandran et al., 2005), who morph a smiling into a neutral face.

In this chapter we have focus on another mode of variation that has been conveniently neglected by the research community that is caused by speech. The deformation caused by lip motion during speech can be considered a major cause of low recognition results, especially in videos that have been recorded in studio conditions where illumination and pose variations are minimal. In this chapter we present a novel method of handling this variation by using temporal synchronization and normalization based on lip motion.

The chapter is divided into two main parts; in the first part we propose a temporal synchronization method that, given a group of videos for a person repeating the same phrase in all videos, studies the lip motion in one of the videos and selects synchronization frames based on a criterion of significance (optical flow). The next module then compares the motion of these synchronization frames with the rest of the videos and selects frames with similar motion as synchronization frames. For evaluation of our proposed method we use the classical eigenface algorithm to compare synchronization frames extracted from the videos and random frames to observe the improvement in face recognition results. The second part of this chapter consists of a temporal normalization algorithm that takes the synchronization frames from the previous module and normalizes the length of the video by lip morphing. Firstly the videos are divided into segments defined by the location of the synchronization frames. Next the normalization is carried out independently for each segment of the video by first selecting an optimal number of frames for each segment and then adding and removing frames to normalize the length of the video. For evaluation of our normalization algorithm we have devised a spatio-temporal person recognition algorithm using video information. By applying discrete video tomography, our algorithm summarizes the facial dynamics of a sequence into a single image, which is then analyzed by a modified version of the eigenface for improvement in a face recognition scenario.

The rest of the chapter is divided as follows. In Section 2 we elaborate the lip detection method. In Section 3 we give the synchronization method, after that we present the normalization method in Section 4 and in section 5 we give the concluding remarks and future works.

2. Lip detection

In this section we present a lip detection method to extract the outer lip contour that combines edge based and segmentation based algorithms. The results from the two methods are then combined by OR fusion. The novelty lies in the fusion of two methods, which have different characteristics and thus exhibit different type of strengths and weaknesses. The other significance of this study lies in the extensive testing and evaluation of the detection algorithm on a realistic database. Most previous studies either never carried out empirical comparisons to the ground truth or sufficed by using a limited dataset. Some studies (Liew et al., 2003; Guan, 2008) do exist that have presented results on considerably large datasets but these mostly consists of high resolution images with constant lighting conditions. Figure 1 gives an overview of the lip detection algorithm. Given a database image containing a human face the first step is to select the mouth Region of Interest (ROI) using the tracking points provided with the database. The next step involves the detection, where the same ROI is provided to the edge and segmentation based methods. Finally the results from the two methods are fused to obtain the final outer lip contour.

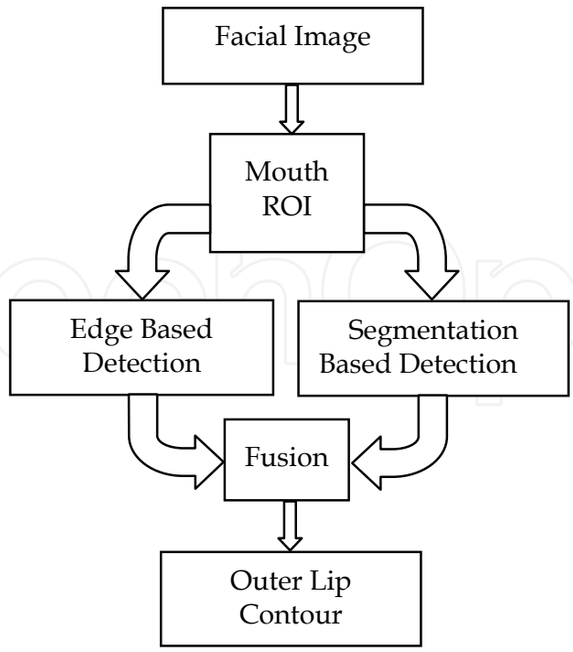


Fig. 1. Overview of lip detection.

2.1 Edge based detection

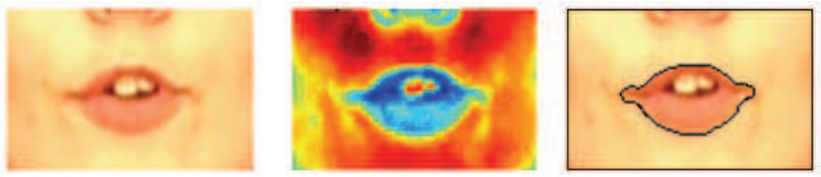
The first algorithm is based on a well accepted edge detection method, it consists of two steps, the first one is a lip enhancing color transform and the second one is edge detection based on active contours. Several color transforms have already been proposed for either enhancing the lip region independently or with respect to the skin. Here, after evaluating several transforms we have selected the color transform (equation 1) proposed by (Canzler & Dziurzyk, 2002). It is based on the principle that blue component has reduced role in lip / skin color discrimination.

$$I = \frac{2G - R - 0.5B}{4} \tag{1}$$

Where R,G,B are the Red, Green and Blue components of the mouth ROI. The next step is the extraction of the outer lip contour, for this we have used active contours (Michael et al., 1987). Active contours (cf. Figure 2) are an edge detection method based on the minimization of an energy associated to the contour. This energy is the sum of internal and external energies; the aim of the internal energy is to maintain the shape as regular and smooth as possible. The most straightforward approach grants high energy to elongated contours (elastic force) and to high curvature contours (rigid force). The external energy models the edge of the object and is supposed to be minimal when the active contours (snake) is at the object boundary. The simplest approach consists of using regularized gradient as the external energy. In our study the contour was initialized as an oval half the size of the ROI with node separation of four pixels.

Since we have applied active contours which have the possibility of detecting multiple objects, on a ROI which may include other features such as the nose tip, jaw line etc. an additional cleanup step needs to be carried out. This consists of selecting the largest detected

object approximately in the middle of the image as the lip and discarding the rest of the detected objects.



a) b) c)

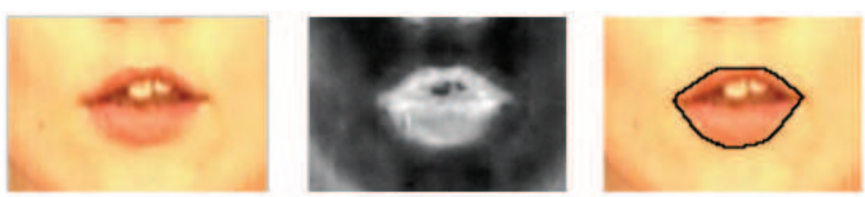
Fig. 2. a) Mouth ROI, b) Color Transform, c) Edge Detection.

2.2 Segmentation based detection

In contrast to the edge based technique the second approach is segmentation based after a color transform in the YIQ domain (cf. Figure 3) . As in the first approach we experimented with several color transform presented in the literature to find the one that is most appropriate for lip segmentation. (Thejaswi & Sengupta, 2008) have presented that skin/lip discrimination can be achieved successfully in the YIQ domain, which firstly de-couples the luminance and chrominance information. They have also suggested that the I channel is most discriminant for skin detection and the Q channel for lip enhancement. Thus we transformed the mouth ROI form RGB to YIQ color space using the equation 2 and retained the Q channel for further processing.

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.595716 & -0.274453 & -0.321263 \\ 0.211456 & -0.522591 & 0.31135 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{2}$$

In classical active contours the external energy is modelled as an edge detector using the gradient of the image, to stop the evolution of the curve on the boundary of the desired object while maintaining smoothness in the curve. This is a major limitation of the active contours as they can only detect objects with reasonably defined edges. Thus for the second method we selected a technique called "active contours without edges" (Chan & Vese, 2001), which models the intensities in different region of the image and uses it as the stopping term in active contours. More precisely this model (Chan & Vese, 2001) is based on Mumford–Shah functional and level sets. In the level set formulation, the problem becomes a mean-curvature flow evolving the active contour, which will stop on the desired boundary. However, the stopping term does not depend on the gradient of the image, as in the classical active contour models, but is instead based on Mumford–Shah functional for segmentation.



a) b) c)

Fig. 3. a) Mouth ROI, b) Color Transform, c) Region Detection

2.3 Error detection and fusion

Lip detection being an intricate problem is prone to errors, especially the lower lip as reported by (Bourel et al., 2000). We faced two types of errors and propose appropriate error detection and correction techniques. The first type of error, which was commonly observed, was caused when the lip was missed altogether and some other feature was selected. This error can easily be detected by applying feature value and locality constraints such as the lip cannot be connected to the ROI's boundary and cannot have an area value less than one-third of the average area value in the entire video sequence. If this error was observed, the detection results were discarded.

The second type occurs when the lip is not detected in its entirety, e.g. missing the lower lip, such errors are difficult to detect thus we proposed to use fusion as a corrective measure, under the assumption that both the detection techniques will not fail simultaneously.

The detection results from the above described methods were then fused using OR logical operator. The outer lip contours are used to create binary masks which describe the interior and the exterior of the outer lip contour. These were then fused using OR Logical Operator defined as

A	B	V
0	0	0
0	1	1
1	0	1
1	1	1

Table 1 presents the commonly observed errors and the effect of OR fusion on the results.

	Type 1 Error	Type 2 Error	No Error
Segmentation Based			
Edge Based			
OR Fusion			

Table 1. Errors and OR Fusion

2.4 Experiments and results

In this section we elaborate the experimental setup and discuss the results obtained. Tests were carried out on Valid Database (Fox et al., 2005) which consists of five recording sessions of 106 subjects using the third utterance. One image was extracted from each of the five videos to create a database of 530 facial images. The reason for selecting one image per video was that the database did not contain any ground truth for lip detection, so ground truth had to be created manually, which is a time consuming task. The images contained both illumination and shape variation; illumination from the fact that they were extracted from all five videos, and shape as they were extracted from random frames of speaker videos.

As already described above the database did not contain any ground truth with respect to the outer lip contour. Thus the ground truth was established manually by a single operator using Adobe Photoshop. The outer lip contour was marked using the magnetic lasso tool which separated the interior and exterior of the outer lip contour by setting the exterior to zero and the interior to one.

To evaluate the lip detection algorithm we used the following two measures proposed by (Guan, 2008), the first measure, equation 3, determines the percentage of overlap (OL) between the segmented lip region A and the ground truth AG. It is defined in Equation 3.

$$OL = \frac{2(A \cap A_G)}{A + A_G} * 100 \tag{3}$$

Using this measure, total agreement will have an overlap of 100%. The second measure, equation 4, is the segmentation error (SE) defined as

$$SE = \frac{OLE + ILE}{2 * TL} * 100 \tag{4}$$

LE (outer lip error) is the number of non-lip pixels being classified as lip pixels and ILE (inner lip error) is the number of lip-pixels classified as non-lip ones. TL denotes the number of lip-pixels in the ground truth. Total agreement will have an SE of 0%.

Initially we calculated the overlap and segmentation errors for edge and segmentation based methods individually, and it was visually observed that edges based method was more accurate but not robust and on several occasions missed almost half of the lip. This can also be observed in the histogram (cf. Figure 4) of segmentation errors; although the majority of lips are detected with 10% or less error but a large number of lip images exhibit approximately 50% of segmentation error. On the other hand segmentation based method was less accurate as majority of lips detected are with 20% error but was quite robust and always succeeded in detecting the lip.

The minimum segmentation, Table 2, error obtained was around 15%, which might seem quite large, but on visual inspection of Figure 4, it is evident that missing the lip corners or including a bit of the skin region can lead to this level of error. Another aspect of the experiment that must be kept in mind is the ground truth. Although every effort was made to establish an ideal ground truth but due to limited time and resources some compromises had to be made. “OR Fusion on 1st Video” are the results that were obtained when OR fusion was applied to only the images from the first video, which are recorded in studio conditions.

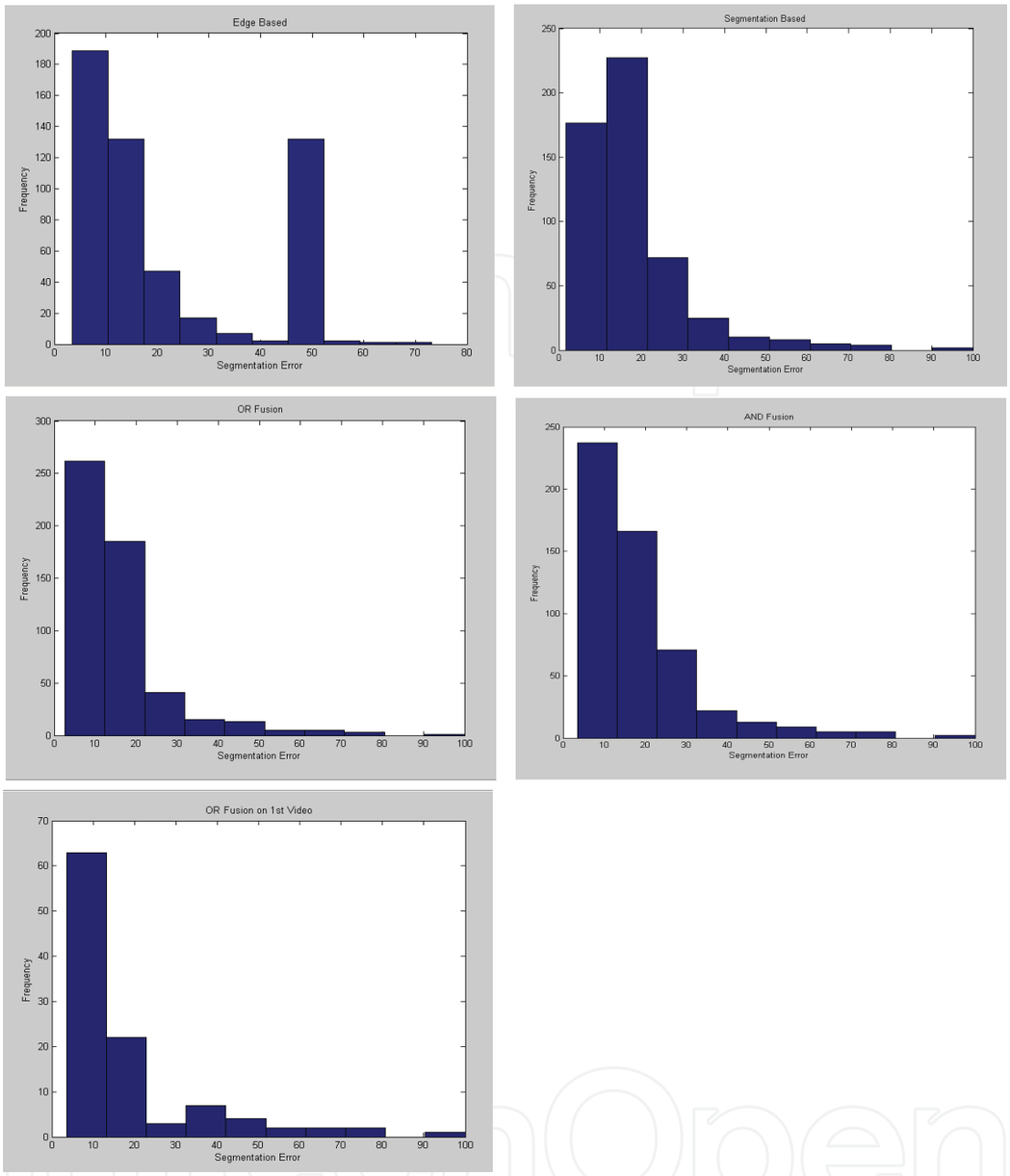


Fig. 4. Histograms for Segmentation Errors

Lip Detection Method	Mean Segmentation Error (SE) %	Mean Overlap (OL) %
Segmentation Based	17.8225	83.6419
Edge Based	22.3665	65.6430
OR Fusion	15.6524	83.9321
AND Fusion	18.4067	84.2452
OR Fusion on 1st Video	13.9964	87.1492

Table 2. Lip detection Results

Fig. 5. Example of Images with 15 % Segmentation Error

3. Synchronization

In this section we propose a temporal synchronization method that, given a group of videos for a person repeating the same phrase in all videos, studies the lip motion in one of the videos and selects synchronization frames based on a criterion of significance (optical flow). The next module then compares the motion of these synchronization frames with the rest of the videos and selects frames with similar motion as synchronization frames. For evaluation of our proposed method we use the classical eigenface algorithm to compare synchronization frames extracted from the videos and random frames to observe the improvement in a face recognition results.

The proposed synchronization method can be divided into two main parts; first is a selection method which selects frames in one of the video that are considered significant, second is a search algorithm in which the synchronization frames selected in the first video are synchronized with the remaining videos.

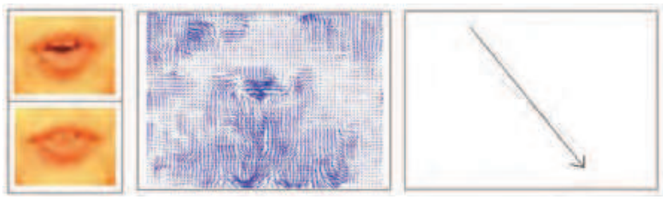
3.1 Synchronization frame selection

The aim of this module is to select synchronization frames from the first video of the group of videos for a specific person. Given a group of videos V_i for the person p, where i is the video index in the group, this module takes the first video V_1 for each person as input and selects synchronization frames SF_1 , that are considered useful for synchronization with the rest of the videos. The criterion for significance is based on amount of lip motion, hence frames that exhibit more lip motion as compared to the frames around them are considered significant. First for the video V_1 the mouth region of interest (ROI) MI_t for each frame t is isolated based on tracking points provided with the database. Then frame by frame optical flow is calculated using the Lucas Kanake method (cf. Figure 6) for the entire video resulting in a matrix of horizontal and vertical motion vectors. As we are interested in a general description of the amount of lip motion in the frame we then calculate the mean of the motion vectors Of_t (cf. Figure 8) for each mouth ROI MI_t.

$$
\begin{aligned}
&\textit{for } t \leftarrow 1 \textit{ to } N-1 \\
&\quad [u_{m,n,t}\ v_{m,n,t}] = LK(MI_t, MI_{t+1}) \\
&\quad Of_t = \sum_{m=1}^{M}\sum_{n=1}^{N}(abs(u_{m,n,t}) + abs(v_{m,n,t})) \\
&\textit{end}
\end{aligned}
$$

Fig. 6. Mean optical flow algorithim

Where T is the number of frames in the video V_i , $LK()$ calculates the Lucas Kanade optical flow. $u_{m,n,t}$ *and* $v_{m,n,t}$ are the horizontal and vertical components of the motion vectors at row m and column n of the frame t.



(a) (b) (c)

Fig. 7. (a) Mouth ROI. (b) *LK* optical flow. (c) Mean vector.

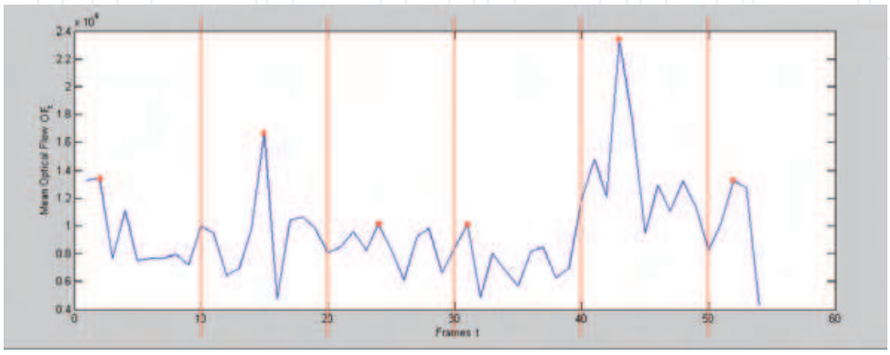


Fig. 8. Mean optical flow Of_t for video

The next step is to select synchronization frames SF_1 based on the mean optical flow Of_t, if we select frames that exhibit maximum lip motion there is a possibility that these frames might lie in close vicinity to each other. Thus we decided to divide the video into predefined segments (cf. Figure 8) and then select the frame with local maxima as synchronization frames.

$$
\begin{aligned}
&for\ t \leftarrow 1\ to\ (N-D)\ with\ incriments\ of\ D \\
&\quad SF_1 = Frame\ with\ value\ (\max(Of_t\ to\ Of_{t+D})) \\
&end \\
&where\ D = \frac{N}{K}
\end{aligned}
$$

Fig. 9. Syncronization frame selecion algorithim

Where T is the total number of frames in the video. K is the number of synchronization frames, its value is predefined and is based on the average temporal length of the videos in the database and will be given in the experiments and results section.

3.2 Synchronization frame matching

In the previous module we have selected synchronization frames from the first video of a person and in this module we try to match these frames with the remaining videos in the group. This module can be broken down into several sub-modules, the first one is a feature extractor where we extracted two features related to lip motion. The second is an alignment algorithm that aligns the extracted lip features before matching, and the last sub-module is a search algorithm that matches the lip features using an adapted mean-square error algorithm. This results in the synchronization frame matrix SF_i for each person.

3.2.1 Feature extraction

In this section we have studied the utility of two mouth features, the first one is quite simply the mouth ROI (MI_t) as used in the previous module, the second is based on lip shape and appearance (LSA_t) and its is based on the outer lip contour extracted in Section 1. Once the outer lip contour is detected the background is then removed and the final feature is obtained as depicted in Figure 9. It contains the shape information in the form of lip contour and the appearance as pixel values inside the outer lip contour. Thus the feature image J may consist of either MI_t or LSA_t.

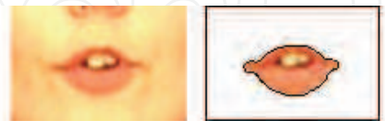


Fig. 10. Lip Feature Image

3.2.2 Alignment

Before the actual matching step, it is imperative that the feature images J (MI_t, LSA_t) are properly aligned, the reason being that some feature images maybe naturally aligned and thus have unfair advantage in matching. The alignment process is based on minimization of mean square error between feature images.

3.2.3 Synchronization frame matching

The last module consists of a search algorithm, which tries to find frames having similar lip motion as synchronization frames selected from the first video in the rest of the videos. The algorithm is based on minimizing the mean square error, adapted for sequences of images.
Let $J_{f(k),i,w}$ be the feature image, where k is the synchronization frame index, $f(k)$ is the location of the synchronization frame in the video, i describes the video number and w the search window, which is fixed to +/-5 frames. Thus the search algorithm tries to find synchronization frames SF_i by matching the current feature image $J_{f(k),1,0}$ previous feature image $J_{f(k)-1,1,0}$ and the future feature image $J_{f(k)+1,1,0}$ from the first video with the rest of the videos within a search window w. The search window w is created in the rest of the video centred at the location of the synchronization frame from the first video given by $f(k)$.

$$
\begin{aligned}
&\text{for } k \leftarrow 1 \text{ to No of Synchronization Frames} \\
&\quad \text{for } i \leftarrow 2 \text{ to No of Videos Per Person} \\
&\quad\quad \text{for } w \leftarrow f(k)-5 \text{ to } f(k)+5 \\
&SF_i = \arg\min_{J} \frac{\sum\sum((J_{f(k)-1,1,0})^2 - (J_{f(k)-1,i,w})^2) + \sum\sum((J_{f(k),1,0})^2 - (J_{f(k),i,w})^2) + \sum\sum((J_{f(k)+1,1,0})^2 - (J_{f(k)+1,i,w})^2)}{(M * N)}
\end{aligned}
$$

Fig. 11. Syncronization frame matching algorithim

Where SFi is the final matrix that contains the synchronization frames for all the videos Vi for one person.

3.3 Person recognition

Classification was carried out using the eigenface technique (Turk & Pentland, 1991). The pre-processing step consists of histogram equalisation and image vectorisation (image pixels are arranged in long vectors).

We apply a linear transformation from the high dimensional image space, to a lower dimensional space (called the face space). More precisely, each vectorised image $\mathbf{s}_n$ is approximated with its projection in the face space $\mathbf{v}_n \in \Re^D$ by the following linear transformation, equation 5.

$$\mathrm{v}_n = \mathrm{W}^T(\mathrm{s}_n - \mu) \tag{5}$$

where $\mathbf{W}$ is a projection matrix with orthonormal columns, and $\boldsymbol{\mu} \in \Re^D$ is the mean image vector of the whole training set, equation 6.

$$\boldsymbol{\mu} = \frac{1}{JN}\sum_{j=1}^{J}\sum_{n=1}^{N}\mathbf{s}_{j,n} \tag{6}$$

in which J is the total number of sequences in the training set, and $\mathbf{s}_{j,n}$ is the n-th vectorised image belonging to video Φ_j. The optimal projection matrix W is computed using the principal component analysis (PCA).
After the image data set is projected into the face space, the classification is carried out using a nearest neighbour classifier which compares unknown feature vectors with client models in feature space. The similarity measure adopted S, equation 7, is inversely proportional to the cosine distance.

$$S(y_i, y_j) = 1 - \frac{{y_i}^T y_j}{||y_i|| \, ||y_j||} \tag{7}$$

and has the property to be bounded into the interval [0, 1].

3.4 Experiments and results

Tests were carried out on Valid Database (Fox et al., 2005) which consists of five recording sessions of 106 subjects using the third utterance. The videos contain head and shoulder region of the subjects and the subjects are present in front of the camera from the beginning till the end.
The first video V_1 was selected for the synchronization frame selection module and the rest of the 4 videos were then matched with the first video using the synchronization frame matching module. To estimate the improvement due to our synchronization process we have compared the synchronization frames SF_i and randomly selected frames using the person recognition module. The first video was excluded from training and testing due to its unrealistic recording conditions, 2nd and 3rd videos were used for training and 4th and 5th were used for testing both synchronization and random frames.
We apply PCA to the enrolment subset to compute a reduced face space of 243 dimensions. Then, the client models are registered into the system using their centroid vectors, which are calculated by taking the average of the feature vectors in the enrolment subset; in the end, recognition is achieved using a nearest neighbour classifier with cosine distances.
We have created 8 datasets from our database by varying the parameters such as selection method, the type of feature image and the number of synchronization frames. The results are summarized in Table 3, the first column gives dataset number, the second column the method for selecting frames, the first 4 datasets use the proposed synchronization frame selection method and the last 4 datasets were created by selecting random frames from the

videos. The third column signifies which lip features were used in the synchronization frame matching module. The fourth column is the number of synchronization frames *K* that were used for each video, in this study we have limited *K* to only 7 and 10 frames as most of the video in our database ranged from 60 to 110 frames. In case of last 4 datasets the number of synchronization frames simply signifies the number of random frames selected. The last column gives the identification rates.

Dataset	Method	Lip Feature	Number of Synchronization Frames	Identification Rates
1	Synchronization	MI	7	71.80 %
2	Synchronization	MI	10	74.18 %
3	Synchronization	LSA	7	72.28 %
4	Synchronization	LSA	10	74.02 %
5	Random	-	7	69.01 %
6	Random	-	10	69.92 %
7	Random	-	7	69.64 %
8	Random	-	10	68.85 %

Table 3. Person Recognition Results

The main result of this study is the overall improvement of identification results from synchronization frames as compared to random frames, which is evident from the Table 3. If we compare the identification results from the first 4 and last 4 datasets, it is obvious that there is an average improvement of around 4% between the 2 group of datasets. The second result that can be deduced is the improvement of recognition rates when more synchronization frames are used. The number of synchronization frames in the case of random frames simply signifies how many random frames were used and as it can be seen from the Table 3, using more random frames has no impact on the identification results. The third is insignificant change with regards to using *MI* or *LSA* as features. Here we would like to emphasize that the amount of testing for the second and third results is rather limited but this was not the main focus of this study.

4. Normalization

This section of the chapter consists of a temporal normalization algorithm that takes the synchronization frames from the previous module and normalizes the length of the video by lip morphing. Firstly the videos are divided into segments defined by the location of the synchronization frames. Next the normalization is carried out independently for each segment of the video by first selecting an optimal number of frames for each segment and then adding and removing frames to normalize the length of the video. The evaluation is carried out by comparing normalized videos with the original videos in a person recognition scenario.

4.1 Optimal number of frames

Given the video V_i , it is first divided into segments S_q , where q is the number of segments and is equal to the number of synchronization frames plus one. Next the optimal number of frames O_q for each corresponding segment S_q is calculated by averaging the number of frames $F_{i,q}$ in the corresponding segment of the videos V_i.

$$
\begin{aligned}
&\textit{for } q \leftarrow 1 \textit{ to } Q \\
&\quad \textit{for } i \leftarrow 1 \textit{ to } I \\
&\qquad O_q = \frac{\sum_{i=1}^{I} F_{i,q}}{I}
\end{aligned}
$$

Fig. 12. Optimal number algorithim

4.2 Transcoding

The next step is to add/remove frames (commonly known as transcoding) from each segment of the video so as to make them equal to the optimal number of frames. The simplest techniques for transcoding like up/down-sampling and interpolation results in jerky and blurred videos respectively. Advanced technique such as motion compensated frame rate conversion (Ugiyama et al., 2005), use block matching to estimate and compensate for motion but are imperfect as they lack information about the type of motion and thus frequently consider a uniform linear model of motion. As for this study we already have an estimation of lip motion from previous modules, we decided to use image morphing instead of block matching/compensation which results in visually superior results.

Morphing is the process of creating intermediate or missing frames from existing frames. Mesh morphing (Wolberg, 1996), one of the well studied techniques consists of creating a morphed frame I_m from source frame I_s and target frame I_t by selecting corresponding feature points in I_s and I_t , creating a mesh based on these feature points, warping I_s and I_t and finally interpolating warped frames to obtain the morphed frame I_m. In our study morphing was carried out only on the lip ROI as this region exhibits the most significant motion in the video. Lip ROI was first isolated and outer lip contour detected as in the previous section. These Lip ROI formed the I_s and I_t frames, feature points consisted of the 4 extremas of the outer lip contour (top, bottom, left, right). Mesh morphing was then carried out. Finally the morphed Lip ROI was superimposed on the original image to obtain the morphed frame (cf. Figure 13).

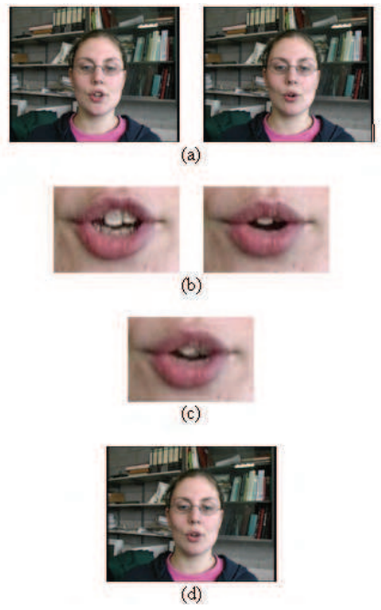


Fig. 13. (a) Existing Frames (b) Lip ROI (c) Morphed Lip ROI (d) Morphed Frame

Decision regarding the number of frames to be added/removed is taken by comparing the number of frames in each segment S_q to the optimal number of frames; the frames are then added/removed at regularly spaced intervals of the segment. Addition of a frame consists of creating a morphed frame I_i from previously existing frames, I_{i-1} and I_{i+1}. Similarly frame I_i is removed by morphing frames I_{i-1} and I_i and replacing I_{i-1} with the morphed frame, and replacing frame I_{i+1} with the morphed frame from I_i and I_{i+1}. Finally deleting the frame I_i. Thus

$$\begin{aligned} &\textit{Frame Addition} \\ &\quad I_i \leftarrow \textit{Morph}(I_{i-1}, I_{i+1}) \end{aligned}$$

$$\begin{aligned} &\textit{Frame Removal} \\ &\quad I_{i-1} \leftarrow \textit{Morph}(I_{i-1}, I_i) \\ &\quad I_{i+1} \leftarrow \textit{Morph}(I_{i+1}, I_i) \\ &\quad \textit{Delete}(I_i) \end{aligned}$$

Fig. 14. Frame addition/deletion algorithim

4.3 Person recognition

For testing our normalization algorithm we used a spatio-temporal method proposed by (Matta & Dugelay, 2008). It consists of two modules: Feature Extraction, which transforms input videos into "X-ray images" and extracts low dimensional feature vectors, and Person Recognition, which generates user models for the client database (enrolment phase) and matches unknown feature vectors with stored models (recognition phase).

4.3.1 Feature extraction

Inspired by the application of discrete video tomography (Akutsu & Tonomura, 1994) for camera motion estimation, we compute the temporal X-ray transformation of a video sequence, to summarize the facial motion information of a person into a single X-ray image. It is important to notice that we restrict our framework to a fixed camera; hence, the video X-ray images represent the motion of the facial features and some appearance information, which is the information that we use to discriminate identities.

Given an input video of length T_i, $V_i \equiv \{I_i,1, \ldots, I_i,T_i\}$, the Feature Extractor module first calculates the edge image sequence E_i, obtained by applying the Canny edge-finding method (Canny, 1986) frame by frame, equation 8.

$$E \equiv \{J_{i,1}, ..., J_{i,T,i}\} = f_{EF}(V_i) \tag{8}$$

Then, the resulting binary frames, $J_{i,t}$, are temporally added up to generate the X-ray image of the sequence, equation 9.

$$X_i = C\sum_{t=1}^{T_i} J_{i,t} \tag{9}$$

where C is a scaling factor to adjust the upper range value of the X-ray image.

Fig. 15. Original Frames and Temporal X-ray Image.

After that, the Feature Extractor reduces the X-ray image space to a low dimensional feature space, by applying the principal component analysis (PCA) (also called the Karhunen-Loeve transform (KLT)): PCA computes a set of orthonormal vectors, which optimally represent the distribution of the training data in the root mean squares sense. In the end, the optimal projection matrix, P, is obtained by retaining the eigenvectors corresponding to the M largest eigenvalues, and the X-ray image is approximated by its feature vector, $y_i \in \Re^M$ calculated using the linear projectionin equation 10.

$$y_i = P^T(x_i - \mu) \tag{10}$$

where x_i is the X-ray image in a vectorial form and μ is the mean value.

4.3.2 Person recognition

During the enrolment phase, the Person Recognition module generates the client models and stores them into the system. These representative models of the users are the cluster centres in feature space that are obtained using the enrolment data set.

For the recognition phase, the system implements a nearest neighbour classifier which compares unknown feature vectors with client models in feature space. The similarity measure adopted S, equation 11, is inversely proportional to the cosine distance.

$$S(y_i, y_j) = 1 - \frac{y_i^T y_j}{||y_i|| \, ||y_j||} \tag{11}$$

and has the property to be bounded into the interval [0, 1].

4.4 Experiments and results

Tests were carried out on Valid Database (Fox et al., 2005) which consists of five recording sessions of 106 subjects using the third utterance. The first video was selected for the synchronization frame selection module and the rest of the 4 videos were then synchronized with the first video using the synchronization frame matching module. Finally all videos were temporally normalized.

To estimate the improvement due to our normalization process we have compared the normalized videos generated by our algorithm to original non-normalized videos using the person recognition module described above. First 3 videos were used for training and the rest 2 were used for testing. The number of synchronization frames in this study have been set to 7, as the average number of frames per video in our database was approximately 70. The recognition system has been tested using a feature space of size 190, constructed with the enrolment data set. The video frames are also pre-processed using histogram equalization, in order to reduce the illumination variations between different sequences.

Method	CIR % (1st)	CIR % (5th)	CIR % (10th)	EER %
Normalized Video	69.02 %	82.60 %	89.13 %	10.1 %
Original Video	65.21 %	81.52 %	85.86 %	11.9 %

Table 4. Person Recognition Results

The identification and verification results are summarized in Table 4; its columns report the correct identification rates (CIR), computed using the best, 5-best and 10-best matches, and the equal error rates (EER) for the verification mode. We notice that the recognition system using normalized videos performs better than the analogous one working with non-normalized videos. Detailed Identification and EER Rates are given in figure 16.

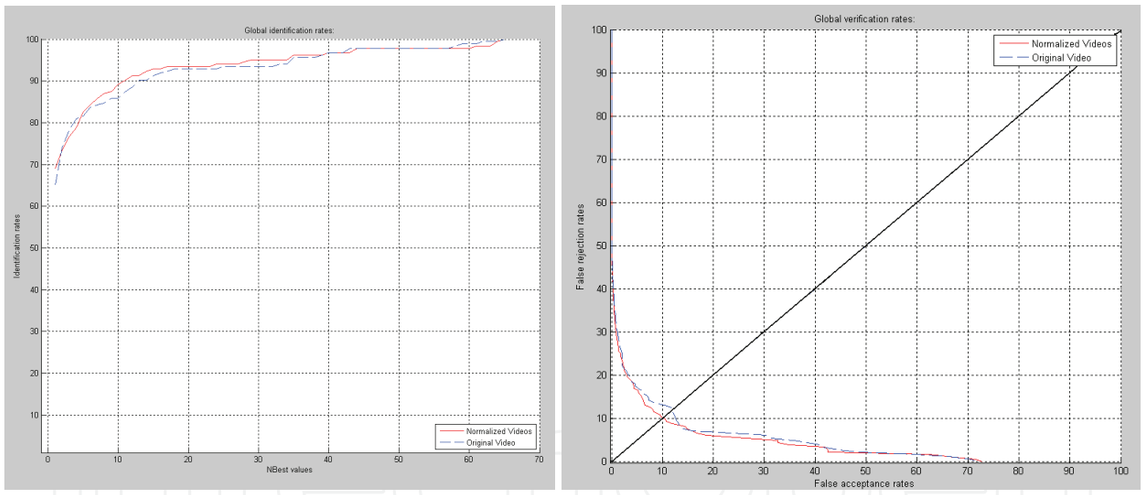


Fig. 16. Correct Identification Rates (CIR) and Verification Rates (EER)

5. Conclusions

In this chapter at first, we have presented a novel lip detection method based on the fusion of edge based and segmentation based methods, along with empirical results on a dataset of considerable size with illumination and speech variation. We observed that the edge based technique is comparatively more accurate, but is not so robust and fails if lighting conditions are not favourable, thus it ends up selecting some other facial feature. On the other hand the segmentation based method is robust to lighting but is not as accurate as the edge based method. Thus by fusing the results from the two techniques we achieve comparatively better results which can be achieved by using only one method. The proposed methods

were tested on a real world database of considerable size and illumination/speech variation with adequate results.

Then we have presented a temporal synchronization algorithm based on mouth motion for compensating variation caused by visual speech. From a group of videos we studied the lip motion in one of the videos and selected synchronization frames based on a criterion of significance. Next we compared the motion of these synchronization frames with the rest of the videos and selects frames with similar motion as synchronization frames. For evaluation of our proposed method we use the classical eigenface algorithm to compare synchronization frames and random frames extracted from the videos and observed an improvement of 4%.

Lastly we have presented a temporal normalization algorithm based on mouth motion for compensating variation caused by visual speech. Using the synchronization frames from the previous module we normalized the length of the video. Firstly the videos were divided into segments defined by the location of the synchronization frames. Next normalization was carried out independently for each segment of the video by first selecting an optimal number of frames and then adding/removing frames to normalize the length of the video. The evaluation was carried out by using a spatio-temporal person recognition algorithm to compare our normalized videos with non-normalized original videos, an improvement of around 4% was observed.

6. References

Blanz, V. and Vetter, T. (2003). Face recognition based on fitting a 3D morphable model. *PAMI*, Vol. 9, (2003), pp. 1063-1074

Matta, F. Dugelay, J-L. (2008). Tomofaces: eigenfaces extended to videos of speakers, *In Proc. of International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, USA, March 2008

Lee, K. and Kriegman, D. (2005). Online learning of probabilistic appearance manifolds for video-based recognition and tracking, *In Proc of CVPR*, San Diago, USA, June 2005

Georghiades, A. S. Kriegman, D. J. and Belhumeur, P. N. (1998). Illumination cones for recognition under variable lighting: Faces, *In Proc of CVPR*, Santa Barbara, USA, June 1998

Tsai, P. Jan, T. Hintz, T. (2007). Kernel-based Subspace Analysis for Face Recognition, *In Proc of International Joint Conference on Neural Networks*, Orlando, USA, August 2007

Ramachandran, M. Zhou, S.K. Jhalani, D. Chellappa, R. (2005). A method for converting a smiling face to a neutral face with applications to face recognition, *In Proc of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, USA, March 2005.

Liew, A.W.-C. Shu Hung, L. Wing Hong, L. (2003). Segmentation of color lip images by spatial fuzzy clustering, *IEEE Transactions on Fuzzy Systems*, Vol.11, No.4, (2003), pp. 542-549

Guan, Y.-P. (2008). Automatic extraction of lips based on multi-scale wavelet edge detection, *IET Computer Vision*, Vol.2, No.1, March 2008, pp.23-33

Canzler, U. and Dziurzyk, T. (2002). Extraction of Non Manual Features for Videobased Sign Language Recognition, *In Proceedings of the IAPR Workshop on Machine Vision Application*, Nara, Japan, June 2002

Michael, K. Andrew, W. and Demetri, T. (1987). Snakes: active Contour models, *International Journal of Computer Vision*, Demetri, Vol. 1, (1987), pp. 259-268

Thejaswi N. S and Sengupta, S. (2008). Lip Localization and Viseme Recognition from Video Sequences, *In Proc of Fourteenth National Conference on Communications*, Bombay, India, 2008

Chan, T.F. Vese, L.A. (2001). Active contours without edges, *IEEE Transactions on Image Processing* , Vol.10, No.2, (2001) pp.266-277

Bourel, F. Chibelushi, C. C. and Low, A. (2000). Robust Facial Feature Tracking, *In Proceedings of the 11th British Machine Vision Conference,* Bristol, UK, September 2000

Fox, N. O'Mullane, A. B. and Reilly, R.B. (2005). The realistic multi-modal VALID database and visual speaker identification comparison experiments, *In Proc of 5th International Conference on Audio- and Video-Based Biometric Person Authentication*, New York, USA, July 2005

Turk, M. and Pentland, A. (1991). Eigenfaces for recognition, *J. Cog. Neurosci.* Vol. 3, (1991) pp. 71–86

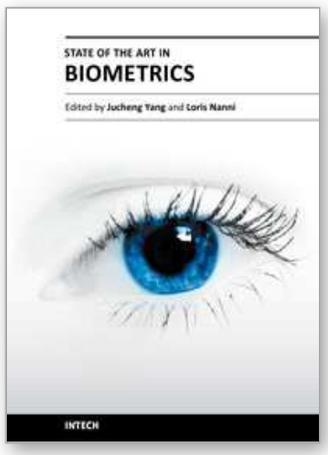
Ugiyama, K. Aoki, T. Hangai, S. (2005). Motion compensated frame rate conversion using normalized motion estimation, *In Proc. IEEE Workshop on Signal Processing Systems Design and Implementation*, Athens, Greece, November 2005.

Wolberg, G. (1996). Recent Advances in Image Morphing, In Proceedings of the International Conference on Computer Graphics, USA, 1996

Huang, C.L. and Huang, Y.M. (1997). Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification, *Journal of Visual Communication and Image Representation*, Vol. 8, (1997), pp. 278-290

Akutsu, A. and Tonomura, Y. (1994). Video tomography: an efficient method for camerawork extraction and motion analysis, *In Proceedings of the Second ACM international Conference on Multimedia*, USA, 1994

Canny, J. (1986). A computational approach to edge detection, *IEEE Trans. Pattern Anal. Mach. Intell*. Vol. 8, (1986), pp. 679-698



State of the art in Biometrics

Edited by Dr. Jucheng Yang

ISBN 978-953-307-489-4
Hard cover, 314 pages
Publisher InTech
Published online 27, July, 2011
Published in print edition July, 2011

Biometric recognition is one of the most widely studied problems in computer science. The use of biometrics techniques, such as face, fingerprints, iris and ears is a solution for obtaining a secure personal identification. However, the â€œoldâ€ biometrics identification techniques are out of date. This goal of this book is to provide the reader with the most up to date research performed in biometric recognition and descript some novel methods of biometrics, emphasis on the state of the art skills. The book consists of 15 chapters, each focusing on a most up to date issue. The chapters are divided into five sections- fingerprint recognition, face recognition, iris recognition, other biometrics and biometrics security. The book was reviewed by editors Dr. Jucheng Yang and Dr. Loris Nanni. We deeply appreciate the efforts of our guest editors: Dr. Girija Chetty, Dr. Norman Poh, Dr. Jianjiang Feng, Dr. Dongsun Park and Dr. Sook Yoon, as well as a number of anonymous reviewers

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