TEMPORALLY CONSISTENT KEY FRAME SELECTION FROM VIDEO FOR FACE RECOGNITION

Usman Saeed, Jean-Luc Dugelay

Eurecom Sophia Antipolis
2229 Route des Cretes, Sophia Antipolis, France
phone: + 33 (0)4 93 00 82 48, email: {Saeed, Dugelay}@eurecom.fr
web:www.eurecom.fr

ABSTRACT

Even after almost three decade of research on automatic face recognition, identification results cannot be considered comparable to superior biometrics. Reasons have been attributed to various modes of variations such as pose, illumination and expression. With the advent of video based face recognition a decade ago we were presented with some new opportunities, algorithms were developed to take advantage of the abundance of data and behavioral aspect of recognition. But this modality introduced some new challenges also, one of them was the variation introduced by speech. In this paper we present a novel method of handling this variation by selecting keyframes from videos based on the temporal analysis of lip motion. Evaluation was carried out by comparing face recognition results obtained by using keyframes selected by the proposed method and frames randomly selected from the videos.

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 [1], 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 AFR also has several advantages over image based techniques, the two main being, more data for pixel-based techniques, and availability of temporal information. Techniques that do not take advantage of temporal information are mostly extensions of image based algorithms adapted for video such as statistical models [2], kernel based [3] or GMM based [4]. Technique that use temporal information can be further divided as Holistic, Feature based and Hybrid. In Holistic approaches, [5] computes a discrete video tomography to summarize the head and facial dynamics of a sequence into a single image. In [6] Aggarwal et al. have modeled the moving face as a linear dynamical system using an autoregressive and moving average (ARMA) model. The second group exploits individual facial features, like the eyes. In [7], they propose to use the optical flow extracted from the motion of the face for creating a feature vector used for identification. The Hybrid approach combines holistic and feature based methods, Colmenarez et al. in [8] have proposed a Bayesian framework which combines face recognition and facial expression recognition to improve results.

Degraded performance in face recognition has mostly been attributed to three main sources of variation in the human face, these being pose, illumination and expression. Of 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 [9] and then use the parameters of the model for pose compensation, second are subspace based such as eigenspace [5]. And the third type are those which build separate subspaces for each pose of the face such as view-based eigenspace [10].

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 [11] and 3D morphable models [9]. Lastly expression invariant face recognition technique can be divided in two categories, first are based on subspace methods that model the facial deformations, such as by Tsai et al. [12]. Next are techniques that use morphing techniques, like Ramachandran et al. [13], who morph a smiling into a neutral face.

In this paper we have focused on another mode of variation that has been conveniently neglected by the research community 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. We propose a key frame selection 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 key frames based on a criterion of significance (optical flow). Next we search these key frames from the first video with the rest of the videos of the same person, within a predefined window created around the location where the key frames were located in the first video. For evaluation of our proposed method we use the classical eigenface algorithm to compare key frames selected by the

proposed method and random frames to observe the improvement in a face recognition scenario.
The rest of the paper is divided as follows. In Section 2 we elaborate the proposed key frame selection method. In Section 3 we give a face recognition method, after that we report and comment our results in section 4 and finally in section 5 we give the concluding remarks and future works.

2. PROPOSED METHOD

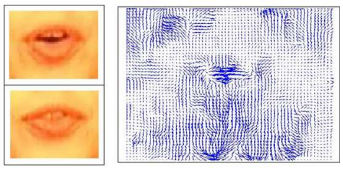
The proposed method consists of two modules, in the first module we propose a key frame selection method that, given a group of videos for a person repeating the same phrase in all videos, studies the lip motion in one of the videos and selects key frames based on a criterion of significance (optical flow). The next module then compares the motion of these key frames with the rest of the videos and selects frames with similar motion as key frames. These frames will be later compared with random frames using the classical eigenface algorithm to observe the improvement in a face recognition scenario.

2.1 Key Frame Selection

The aim of this module is to select key 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 key frames SF_1 , that are considered useful for matching 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 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. Fig. 1.) 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 motion in the frame we then calculate the absolute mean of the motion vectors Of_t for each frame t.

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

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



(a) (b)
Fig. 1 (a) Lip ROI. (b) LK optical flow.

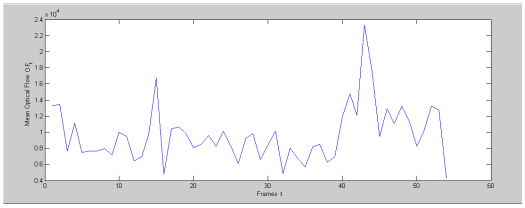


Fig. 2. Mean Optical flow Of_t for Video

The next step is to select key frames SF_1 based on the mean optical flow Of_t, if we select frames that exhibit maximum 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 and then select the frame with local maxima as key frames.

$$\begin{aligned} &for\ t \leftarrow 1\ to\ (N-D)\ with\ increments\ of\ D \\ &\quad SF_1 = Frame\ with\ value\ (\max(Of_t\ to\ Of_{t+D})) \\ &end \\ &where\ D = \frac{N}{k} \end{aligned} \quad (2)$$

Where N is the total number of frames in the video. k is the number of key 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.

2.2 Key Frame Matching

In the previous module we have selected some key 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 key frame matrix SF_i for each person.

2.2.1 Feature Extraction

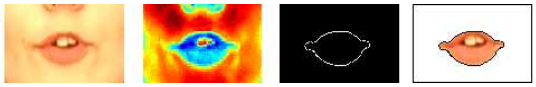
For the matching algorithm we have studied the suitability of two lip features, the first one is quite simply the mouth ROI (MI_t) as used in the previous module, the second is based on lip shape and appearance (LSA) and its extraction is described below:

Color Transform: The first step is to transform the color space so as to enhance the difference between the skin and lip. From several color transform proposed in the literature

we have selected the one proposed by [14], It is defined in eq. 3.

$$I = \frac{(2G - R - 0.5B)}{4} \quad (3)$$

Lip Contour Detection: The next step is the extraction of the outer lip contour, for this we have used active contours. The contour was initialized as an oval, half the size of the ROI with node separation of four pixels.



(a) (b) (c) (d)

Fig. 3. (a) Lip ROI. (b) Colour transform. (c) Snake edge. (d) Lip SA.

Feature Definition and Extraction: Finally the background is removed based on the outer lip contour. The final feature is depicted in Fig. 3. 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.

2.2.2. 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.

2.2.3. Key Frame Matching

The last module consists of a search algorithm, which tries to find frames having similar lip motion as key frames selected from the first video in the remaining 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 key frame index, *f(k)* is the location of the key frame in the video, *i* describes the video number and *w* the search window, which is fixed to +/-5 frames. Thus the search algorithm (Eq. 4) tries to find key frames SF_i by matching the current feature image $J_{f(k),1}$ previous feature image $J_{f(k)-1,1}$ and the future feature image $J_{f(k)+1,1}$ from the first video with the remaining videos within a search window *w*. The search window *w* is created in the remaining video centered at the location of the key frame from the first video given by *f(k)*.

Where SF_i is the final matrix that contains the key frames for all the videos V_i for one person.

3. PERSON RECOGNITION

Classification was carried out using the classical eigenface technique [15]. 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 S_n is approximated with its projection in the face space v_n by the following linear transformation:

$$\mathbf{v}_n = \mathbf{W}^T(\mathbf{s}_n - \boldsymbol{\mu}) \quad (5)$$

where W is a projection matrix with orthonormal columns, and $\boldsymbol{\mu}$ is the mean image vector of the whole training set:

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

in which *J* is the total number of sequences in the training set, and $s_{j,n}$ 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, is inversely proportional to the cosine distance:

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

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

$$\begin{aligned} &for\ \ k \leftarrow 1\ to\ No\ of\ \text{Synchronization}\ Frames \\ &\quad for\ i \leftarrow 2\ to\ No\ of\ Videos\ \ Per\ Person \\ &\quad\quad for\ w \leftarrow f(k) - 5\ to\ f(k) + 5 \\ &SF_i = \text{argmin}\frac{\sum\sum((J_{f(k)-1,1})^2 - (J_{f(k)-1,i,w})^2) + \sum\sum((J_{f(k),1})^2 - (J_{f(k),i,w})^2) + \sum\sum((J_{f(k)+1,1})^2 - (J_{f(k)+1,i,w})^2)}{(M * N)} \end{aligned} \quad (4)$$

4. EXPERIMENTS AND RESULTS

In this section we elaborate the experimental setup and discuss the results obtained. Tests were carried out on a subset of the Valid database [17], which consists of 106 subjects. The database contains five sessions for each subject, where one session has been recorded in studio conditions while the others in uncontrolled environments such as the office or corridors. In each session the subjects repeat the same sentence, “Joe took father's green shoe bench out”. The videos contain head and shoulder region of the subjects and the subjects are present in front of the camera form the beginning till the end.

Fig. 4. Image example from Valid Database

The first video V_1 was selected for the key frame selection module and the rest of the 4 videos were then matched with the first video using the key frame matching module.

To estimate the improvement due to our selection process we have compared the key frames SF_i generated by our algorithm to randomly selected frames from the videos using the person recognition module described above. 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 key and random frames. In our experiments the eigenspace had a dimensionality of 240.

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 key frames. The results are summarized in the Table 1. , the first column gives dataset number, the second column the method for selecting frames, the first 4 datasets use the proposed key 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 key frame matching module. The fourth column is the number of key 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 keyframes simply signifies the number of random frames selected. The last column gives the identification rates.

Table 1. Person Recognition Results

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

The main result of this study is the overall improvement of identification results from key frames as compared to random frames, which is evident from the Table 1. 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 key frames are used. The number of key frames in the case of random frames simply signifies how many random frames were used and as it can be seen from the table 1, 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.

5. CONCLUSIONS

In this paper we have presented a key frame selection algorithm based on mouth motion for compensating variation caused by visual speech. The proposed algorithms were tested in a face recognition scenario using eigenface algorithm and results compared keyframes selected by the proposed method with randomly selected frames; an improvement of 4% was observed.

Further improvements to the proposed work could be in the form studying variation in number of key frames. Another interesting improvement could be testing the method with other databases and person classifiers.

6. REFERNCES

[1] W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld, “Face recognition: A literature survey,” In *ACM Comp. Surv*., vol. 35, no. 4, pp. 399–458, 2003.

[2] G. Shakhnarovich, J. W. Fisher, and T. Darrel, “Face recognition from long-term observations,” In *Proc. of ECCV*, pp.851-868, 2002.

[3] L. Wolf and A. Shashua, “Learning over sets using kernel principal angles,” In *JMLR*, vol.4, pp.913-931, 2003.

[4] T. Kim, O. Arandjelovic, and R. Cipolla, “Learning over sets using boosted manifold principal angles,” In *Proc. of BMVC*, pp. 779–788, 2005.
[5] F. Matta, J-L. Dugelay, “Tomofaces: eigenfaces extended to videos of speakers,” In *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, pp.1793-1796, 2008.
[6] G. Aggarwal, A.K.R. Chowdhury, and R. Chellappa, “A System Identification approach for Video-based Face Recognition,” In *Proc. of the International Conference on Pattern Recognition*, vol. 4, pp. 175-178, 2004.
[7] L. Chen, H. Liao, and J. Lin, “Person identification using facial motion,” In P*roc. of International Conference on Image Processing*, pp. 677-680, 2001.
[8] A. Colmenarez, B. Frey, and T.S. Huang, “A Probabilistic Framework for Embedded Face and Facial Expression Recognition,” In *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 592-597, 2003.
[9] V. Blanz and T. Vetter, “Face recognition based on fitting a 3D morphable model,” In *PAMI*, vol. 9, pp. 1063-1074, 2003.
[11] A. S. Georghiades, D. J. Kriegman, and P. N Belhumeur, “Illumination cones for recognition under variable lighting: Faces.” In *Proc of CVPR*, pp. 52-59, 1998.
[10] K. Lee and D. Kriegman, “Online learning of probabilistic appearance manifolds for video-based recognition and tracking,” In Proc of CVPR, pp. 852-859, 2005.
[12] P. Tsai, T. Jan, T. Hintz, “Kernel-based Subspace Analysis for Face Recognition," In *Proc of International Joint Conference on Neural Networks*, pp.1127-1132, 2007.
[13] M. Ramachandran, S.K. Zhou, D. Jhalani, R. Chellappa, “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*, vol.2, pp. 977-980, 2005.
[14] U. Canzler, and T. Dziurzyk, “Extraction of Non Manual Features for Video based Sign Language Recognition,” In *Proc. of the IAPR Workshop on Machine Vision Application*, pp. 318–321, 2000.
[15] M. Turk and A. Pentland, “Eigenfaces for recognition,” In *J. Cog. Neurosci*. Vol. 3, pp. 71–86, , 1991.
[16] R.O. Duda, P.E. Hart and D.G. Stork, “Pattern classification,” John Wiley & Sons, *Interscience*, 2nd edition, 2000.
[17] N.A. Fox, B.A. O'Mullane, and R.B. Reilly, “The Realistic Multi-modal VALID database and Visual Speaker Identification Comparison Experiments,” In *Lecture Notes in Computer Science*, vol. 3546, pp. 777, 2005.