Temporal Normalization of Videos Using Visual Speech

Usman Saeed
EURECOM Sophia Antipolis
2229 Route Des Cretes
Sophia Antipolis, France
+33(0)493008248
Usman.Saeed@eurecom.fr

Jean-Luc Dugelay
EURECOM Sophia Antipolis
2229 Route Des Cretes
Sophia Antipolis, France
+33(0)493008141
Jean-Luc.Dugelay@eurecom.fr

ABSTRACT

Pose and illumination variation has been considered the major cause of poor recognition results in automatic face recognition as compared to other biometrics. 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 for handling this variation by using temporal normalization based on lip motion. Evaluation was carried out by comparing face recognition results from original non-normalized videos and normalized videos.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding – *motion, video analysis;* I.4.6 [**Image Processing and Computer Vision**]: Segmentation – *edge and feature detection;* I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis – *color, motion, object recognition, tracking.*

General Terms

Algorithms, Security.

Keywords

Biometrics, Face Recognition, Image and Video Analysis.

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 the most 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 temporal normalization method that, given a group of videos for a person studies the lip motion in one of the videos and selects synchronization frames based on a criterion of significance

MiFor'09, October 23, 2009, Beijing, China.
Copyright 2009 ACM 978-1-60558-755-4/09/10...$10.00.

(optical flow). The next module compares these synchronization 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 synchronization frames were located in the first video. Finally videos are normalized temporally using lip morphing. 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.

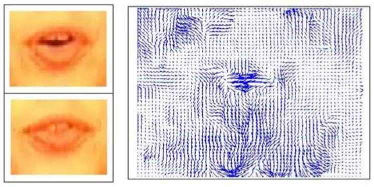
The rest of the paper is divided as follows. In Section 2 we elaborate the proposed method. In Section 3 we give a face recognition method, after that we report and comment our results in section 4 and finally we conclude this paper with remarks and future works in section 5.

2. VIDEO SYNCHRONIZATION

The proposed normalization method can be divided into three parts; first is an algorithm which selects frames in one of the video that are considered significant, second is synchronization frame matching in which the synchronization frames selected in the first video are synchronized with the remaining videos. Third is the temporal normalization, which uses the synchronization frames from the previous module to normalize videos temporally by lip morphing.

2.1 Synchronization Frame Selection

This module takes the first video from the group of videos that have to be synchronized as input and selects frames that are considered useful for synchronization. The criteria for frame selection 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 the mouth region is isolated based on tracking points provided along the database. Then frame by frame optical flow is calculated using the Lucas Kanade method for the entire video (see figure 1).



(a) (b)

Figure 1: (a) Lip ROI (b) LK optical flow

As we are interested in a general description of the amount of motion in the frame we calculate the absolute mean of each frame as

$$\sum_{n=1}^{M}\sum_{m=1}^{N}(abs(u_{m,n})+abs(v_{m,n}))$$

Where *m*, *n* are image row and columns respectively. *v*, *u* are the horizontal and vertical components of the motion vectors, this results in a signal as follows for the entire video.

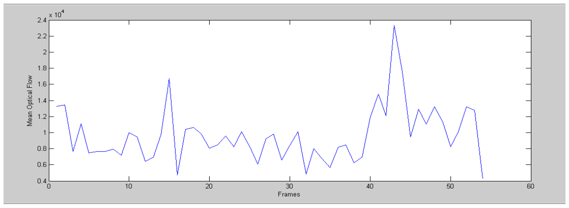


Figure 2: Motion signal for video

The next step is to select frames based on the above signal, 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 local maxima as synchronization frames.

2.2 Synchronization Frame Matching

This module synchronizes the synchronization frames selected from the first video with the rest of the videos. Lip Shape and Appearance (Lip *SA*) features are first extracted from all videos, aligned and then matched using an adapted mean-square error algorithm.

2.2.1 Feature Extraction

Synchronization frame matching is carried out using Lip *SA* features that are based on the shape and appearance of the lip and their extraction is described below:

2.2.1.1 Color Transform

The first step is to transform the colour space so as to enhance the difference between the skin and lip. From several colour transform proposed in the literature we have selected the one proposed by [14], It is defined as.

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

2.2.1.2 Lip Contour Detection

The next step is the extraction of the lip contours, for this we have used active contours [15]. The contour was initialized as an oval half the size of the ROI with node separation of four pixels.

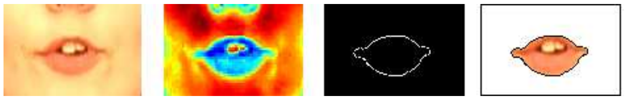


Figure 3: (a) Lip ROI (b) Color transform (c) Snake edge (d) Lip *SA*

2.2.1.3 Feature Definition and Extraction

Finally the background is removed based on the outer lip contour. The final feature is depicted in figure 3(d). It contains the shape information in the form of lip contour and the appearance as pixel values inside the outer lip contour.

2.2.2 Alignment

Before the actual matching step, it is imperative that the feature images 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 *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.

```
for $k \leftarrow 1$ to No of Synchronization Frames
  for $i \leftarrow 2$ to No of Vids Per Person
    for $w \leftarrow f(k)-5$ to $f(k)+5$
$\arg\min \frac{\sum\sum((I_{f(k)-1,1})^2-(I_{f(k)-1,i,w})^2)+\sum\sum((I_{f(k),1})^2-(I_{f(k),i,w})^2)+\sum\sum((I_{f(k)+1,1})^2-(I_{f(k)+1,i,w})^2)}{(M*N)}$
```

Figure 4: Synchronization Frame Matching Algorithm

Let $I_{f(k),i,w}$ be the feature image extracted from the synchronization frame to be matched, 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 $I_{f(k),1}$ by matching the current frame $I_{f(k),1}$ previous frame $I_{f(k)-1,1}$ and the future frame $I_{f(k)+1,1}$ from the first video with the rest of the videos within a search window w. The search window is created in the rest of the video centered at the location of the synchronization frame from the first video given by $f(k)$.

2.3 Temporal Normalization

Once the synchronization frames have been obtained for all the videos of a person, the next step is to normalize the length of each segment of the videos. Normalization is carried out independently for each person by first selecting an optimal number of frames for each segment of the video based on the synchronization frames and then adding and removing frames to normalize the length of the video.

2.3.1 *Optimal Number of Frames.*

Optimal number of frames $O_{S,P}$ for each corresponding segment S of the video is calculated by averaging the number of frames F in the corresponding segment of the video for person P.

```
for $S \leftarrow 1$ to $k$
  {
      $O_{S,P} = \frac{\sum_{n=1}^{N} F_{n,S,P}}{N}$
  }
```

where n represents the videos for person P .

2.3.2 *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 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 [16], 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 rectilinear 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 [17], one of the well studied techniques consists of morphing a 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. 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 explained in the previous modules. 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 as explained above. Finally the morphed Lip ROI was superimposed on the original image to obtain the morphed frame.

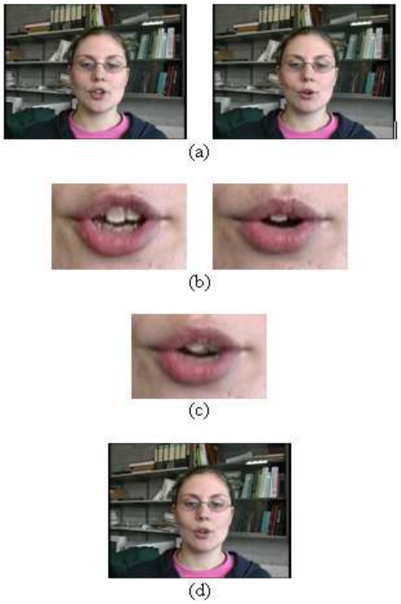


Figure 5:(a) Existing Frames (b) Lip ROI (c) Morphed Lip ROI (d) Morphed Frame

Decision regarding the number of frames to be added/removed is taken by comparing the number of frames in each segment 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.

3. PERSON RECOGNITION

Our person recognition system [5] is composed of two modules: a Feature Extractor, which transforms input videos into “X-ray images” and extracts low dimensional feature vectors, and a Person Recognizer, which generates user models for the client database (enrolment phase) and matches unknown feature vectors with stored models (recognition phase).

3.1 Feature Extractor

Inspired by the application of discrete video tomography [18] for camera motion estimation, we compute the temporal X-ray transformation of a video sequence, to summaries 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.

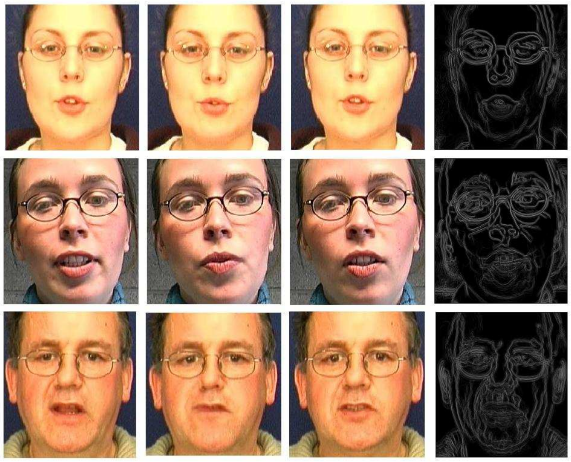


Figure 6: Original Frames and Temporal X-ray Image.

Given an input video of length T_i, $V_i \equiv \{I_{i,1}, \ldots, I_{i,Ti}\}$, the Feature Extractor module first calculates the edge image sequence E_i, obtained by applying the Canny edge-finding method [19] frame by frame:

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

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

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

where C is a scaling factor to adjust the upper range value of the 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 following linear projection:

$$y_i = P^T(x_i - \mu)$$

where $\mathbf{x}_i$ is the X-ray image in a vectorial form and μ is the mean value.

3.2 Person Recognizer

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

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

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

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

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 Valid Database [20], 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 are 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 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.

Table 1: Person Recognition Results

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

The identification and verification results are summarized in Table 1; 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.

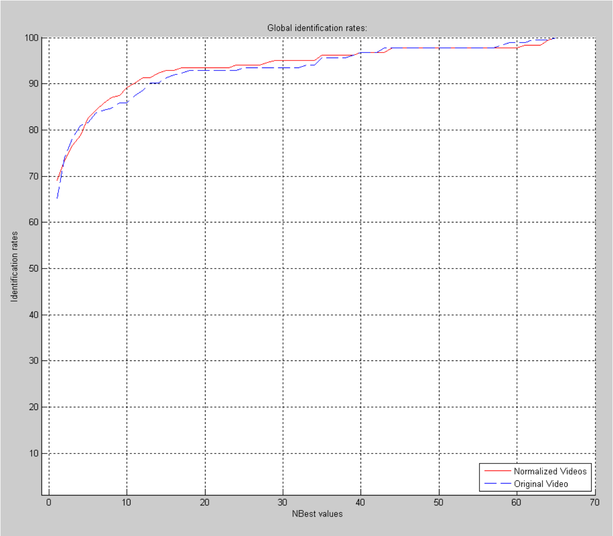


Figure 7 : Correct Identification Rates (CIR)

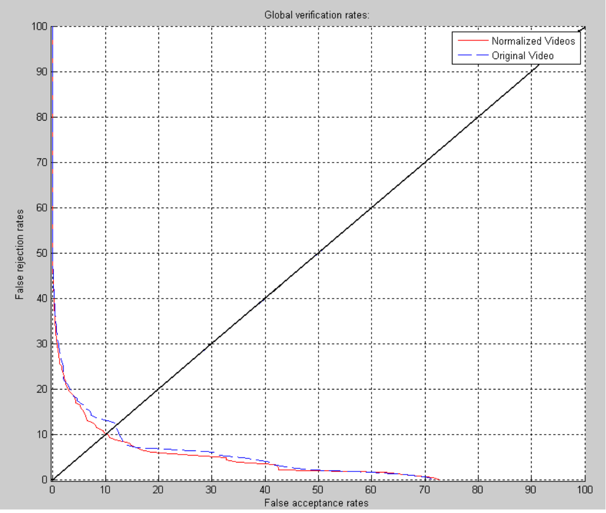


Figure 8 : Verification Rates (EER)

5. CONCLUSIONS

In this paper we have presented a temporal normalization algorithm based on mouth motion for compensating variation caused by visual speech. The proposed algorithm was tested in a face recognition scenario using a spatio-temporal person recognition algorithm and results compared with original non-normalized videos, with an improvement of 4%.

The database used in these experiments consisted of short sentences; it would be interesting to see results of normalization on other databases. Another specificity of the database used was that although it did not contain any pose variation, strong illumination variation was present, which has effected recognition results. Further improvements to the proposed work could be inclusion of other forms of normalization, such as spatial and illumination.

6. REFERENCES

[1] Zhao, W., Chellappa, R., Phillips, P. J., and Rosenfeld, A. 2003. Face recognition: A literature survey. ACM Comput. Surv. 35, 4 (Dec. 2003), 399-458. DOI= http://doi.acm.org/10.1145/954339.954342

[2] Shakhnarovich, G., Fisher, J. W., and Darrell, T. 2002. Face Recognition from Long-Term Observations. In Proceedings of the 7th European Conference on Computer Vision-Part III (May 28 - 31, 2002). A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, Eds. Lecture Notes In Computer Science, vol. 2352. Springer-Verlag, London, 851-868.

[3] Wolf, L. and Shashua, A. 2003. Learning over sets using kernel principal angles. J. Mach. Learn. Res. 4 (Dec. 2003), 913-931.

[4] Kim, T.-K., Arandjelović, O. and Cipolla, R. 2005. Learning over sets using boosted manifold principal angles (BoMPA). In Proceedings of the British Machine Vision Conference. 779-788.

[5] Matta, F., Dugelay, J.-L.2008. Tomofaces: Eigenfaces extended to videos of speakers. In Proceedings of the Acoustics, Speech and Signal Processing, IEEE International Conference on (March 31 -April 4 2008).1793-1796.

[6] Aggarwal, G., Chowdhury, A. K., and Chellappa, R. 2004. A System Identification Approach for Video-based Face Recognition. In Proceedings of the Pattern Recognition, 17th international Conference on (Icpr'04) Volume 4 - Volume 04 (August 23 - 26, 2004). ICPR. IEEE Computer Society, Washington, DC, 175-178. DOI= http://dx.doi.org/10.1109/ICPR.2004.107

[7] Chen, L., Liao, H., and Lin, J. 2001. Person identification using facial motion. In Proc. of International Conference on Image Processing (7-10 Oct), 2. 677-680.

[8] Colmenarez, A., Frey, B., and Huang, T.S. 2003. A Probabilistic Framework for Embedded Face and Facial Expression Recognition. In Proc. of IEEE International Conference on Computer Vision and Pattern Recognition. 592-597.

[9] Blanz, V. and Vetter, T. 2003. Face Recognition Based on Fitting a 3D Morphable Model. IEEE Trans. Pattern Anal. Mach. Intell. 25, 9 (Sep. 2003), 1063-1074. DOI= http://dx.doi.org/10.1109/TPAMI.2003.1227983

[10] Lee, K. and Kriegman, D. 2005. Online Learning of Probabilistic Appearance Manifolds for Video-Based Recognition and Tracking. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cvpr'05) - Volume 1 - Volume 01 (June 20 - 26, 2005). CVPR. IEEE Computer Society, Washington, DC, 852-859. DOI= http://dx.doi.org/10.1109/CVPR.2005.260

[11] Georghiades, A. S., Kriegman, D. J., and Belhumeur, P. N. 1998. Illumination Cones for Recognition under Variable Lighting: Faces. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (June 23 - 25, 1998). CVPR. IEEE Computer Society, Washington, DC, 52.

[12] Tsai, P., Jan, T., Hintz, T. 2007. Kernel-based Subspace Analysis for Face Recognition. In Proc of International Joint Conference on Neural Networks.1127-1132.

[13] 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. 2, 977-980.

[14] Canzler, U., and Dziurzyk, T. 2000. Extraction of Non Manual Features for Video based Sign Language Recognition. In Proc. of the IAPR Workshop on Machine Vision Application. 318–321.

[15] Michael, K., Andrew, W., and Demetri ,T. 1987. Snakes: active Contour models. In Proc. International Journal of Computer Vision.1, 259-268.

[16] Sugiyama, 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. 663-668.

[17] Wolberg, G. 1996. Recent Advances in Image Morphing. In Proceedings of the Conference on Computer Graphics international (June 24 - 28, 1996). Computer Graphics International. IEEE Computer Society, Washington, DC, 64.

[18] 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 (San Francisco, California, United States, October 15 - 20, 1994). MULTIMEDIA '94. ACM, New York, NY, 349-356. DOI= http://doi.acm.org/10.1145/192593.192697

[19] Canny, J. 1986. A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. 8, 6 (Nov. 1986), 679-698. DOI= http://dx.doi.org/10.1109/TPAMI.1986.4767851

[20] Fox, N., O'Mullane, B., Reilly, R.B. 2005. The Realistic Multi-modal VALID database and Visual Speaker Identification Comparison Experiments. Kanade, T., Jain, A.K., Ratha, N.K. Eds. Lecture Notes in Computer Science, vol. 3546. Netherlands: Springer-Verlag.