Multimodal feature fusion for video forgery detection

Girija Chetty
Faculty of Information Sciences and Engineering
University of Canberra
girija.chetty@canberra.edu.au

Matthew Lipton
AdSoftech R & D Pty. Ltd.
Melbourne, Australia
mLipton@adsoftech.net

Abstract - In this paper we propose a novel local feature analysis and feature level fusion technique for detecting tampering or forgery for facial-biometric based on-line access control scenarios. The local features are extracted by analysing facial image data in the chrominance colour space and hue-saturation colour space. A feature level fusion of local features consisting of hue and saturation gradients with global features obtained from principal component analysis showed that a significant improvement in performance can be achieved in detecting tampered or forged images from genuine images in low bandwidth online streaming video access control contexts. The performance evaluation of the proposed fusion technique for a multimodal facial video corpus showed that an equal error rate of less than 1% could be achieved with feature level fusion of local features and global features.

Keywords: feature fusion, local feature analysis, forgery , tamper detection

1 Introduction

In this paper we propose that multimodal feature fusion, based on combining local image features and global image features offers an opportunity to discriminate genuine image from a tampered or forged image. We propose that in video footage depicting certain human communication and interaction tasks such as speaking, talking, acting or expressing emotions, different regions in faces such as lips, eyes, and eyebrows undergo different levels of motion, and by exploiting the spatio-temporal dynamics from different regions in face images, it is possible to discriminate a genuine video from tampered or forged video. This is an emerging problem in on-line access control, and security and forensic investigation contexts [1, 2, 3, and 4].

The detection and tracking of local features from images in video sequences (such as lips, eyes or eyebrows in a facial video for example), is a fundamental and challenging problem in signal and image processing. This research area has many applications in emerging security, surveillance and forensic scenarios, in addition to human-computer interaction and human robot interaction areas. Most of the techniques proposed so far work very well for off-line application scenarios, but for real time application scenarios, there is a need to establish the authenticity of the images in the video footage- thanks to advancements in computer graphics and animation technologies. It has become very easy to create a photographic fake of the person's face using image cloning and animation tools. In this paper we consider the context of facial biometric access control, and propose a simple technique based on extracting the local feature in alternative colour spaces. With subsequent feature level fusion with global features based on principal component analysis, we try to discriminate the genuine facial image from tampered or forged images.

The paper is organized as follows. Next Section describes the region of interest (ROI) segmentation technique (face and lip region detection here) from talking face video sequences. The local feature extraction technique (lip feature extraction) is described in Section 3. Section 4 describes the feature level fusion technique and details of the experimental results are given in Section 5. The paper concludes with conclusion and plans for further work in Section 6. For simplicity and considering the limits on the paper length, only details of lip region feature extraction is described in detail here. A similar technique was used for extracting other prominent regions of the face such as eye region and eyebrow region from face video sequences.

2 Face ROI Segmentation

The region of interest (ROI) segmentation for detecting faces and regions within faces (lips, eyes, eyebrows, nose) is done in the first frame of the video sequence. The tracking of the face and lip region in subsequent frames is done by projecting the markers from the first frame. This is followed by measurements on the lip region boundaries based on pseudo-hue edge detection and tracking. The advantage of this segmentation technique based on exploiting the alternative colour spaces is that it is simpler and more powerful in comparison with methods based on deformable templates, snakes and pyramid images [7]. Moreover, the method can be easily extended to the detection and tracking of local features in image sequences with multiple faces or degrading operating

environments with illumination and affine transformation artifacts.

The ROI segmentation scheme for detecting faces consists of three image processing stages. The first stage is to classify each pixel in the given image as a skin or non-skin pixel. The second stage is to identify different skin regions in the image through connectivity analysis. The last stage is to determine for each of the identified skin regions- whether it represents a face. This is done using two parameters. They are the aspect ratio (height to width) of the skin-coloured blob, and template matching with an average face image at different scales and orientations. A statistical skin colour model was generated by means of supervised training, using a set of skin colour regions, obtained from a coloured-face database. A total of 10,000 skin samples from 100 colour images were used to determine the colour distribution of human skin in chromatic colour space. The extracted skin samples were filtered using a low pass filter to reduce the effect of noise. Red-blue chrominance colour vectors were concatenated from the rows of the red and blue pixel matrices, *Cr* and *Cb* as shown in Equation (1).

$$x = (Cr_{11},\ldots,Cr_{1m},Cr_{21},\ldots Cr_{nm}, \\ Cb_{11},\ldots,Cb_{1m},Cb_{21}\ldots,Cb_{nm}) \quad (1)$$

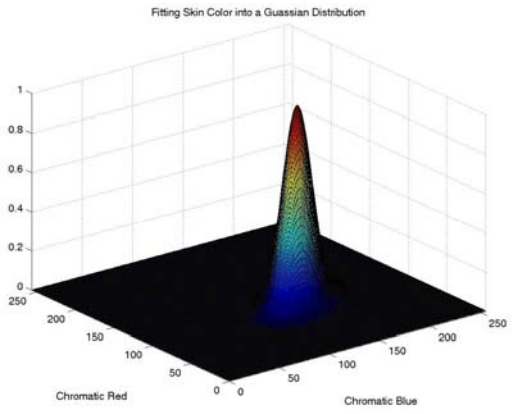

Figure 1: Gaussian distribution of skin-colour sample in red-blue chromatic space

The colour histogram reveals that the distribution of skin colour for different people is clustered in this chromatic colour space and can be represented by a Gaussian model *N(μ, C)* shown in Equation (2):

$$\text{mean vector } \hat{\imath} = E[x], \text{ and}$$
$$\text{covariance matrix } C = E[(x - \mu)(x - \mu)T]. \quad (2)$$

One such skin patch with the Gaussian distribution N(μ, C) fitted to our data in red-blue chromatic space are shown in Figure 1. With this Gaussian skin colour model, we can now obtain the "skin likelihood" for any pixel of an image.

Figure 2: Face detection by skin colour in red-blue chromatic space

The skin probability image thus obtained is thresholded to obtain a binary image. The morphological segmentation involves the image processing steps of erosion, dilation and connected component analysis to isolate the skin-coloured blob. By applying an aspect ratio rule and template matching with an average face, it is ascertained the skin-coloured blob is indeed a face. Figure 2 shows an original image from a facial video database (VidTIMIT), and the corresponding skin-likelihood and skin-segmented images, obtained by the statistical skin colour analysis and morphological segmentation.

3 Local Feature extraction

Once the ROI is segmented and face region is localised as shown in Figure 2, the local features from different sections of the face (lip region for example) are detected using another colour space. For example, to detect lip regions in the faces a hue-saturation colour space is used.. It is easier to detect the lips in hue/saturation colour space because hue/saturation colour for the lip region is fairly uniform for a wide range of lip colours and fairly constant under varying illumination conditions and different human skin colours [8].

We derive a binary image B from the hue and saturation images H and S using thresholds as shown in Equation (3) below:

$$H0 = 0.4 \text{ and}$$
$$S0 = 0.125$$
$$\text{such that}$$
$$Bij = 1 \text{ for } Hij > H0 \text{ and}$$
$$Sij > S0, \text{ and } Bij = 0 \text{ otherwise.} \quad (3)$$

Figure 3 shows ROI region extraction (lip region) based on hue-saturation thresholding. To derive the local features from lip region, such as lip dimensions and key points within the lip region of interest, we detect edges based on combining pseudo-hue colour and intensity information.

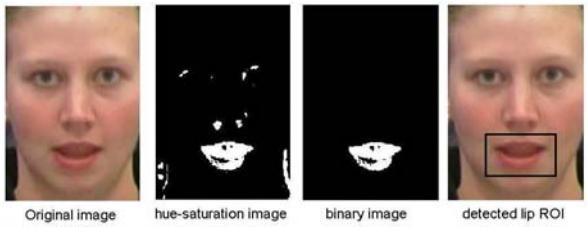

Figure 3: Lip region localization using hue-saturation thresholding

The pseudo-hue matrix P is computed from the red and green image components as in Equation (4) and then normalised to the range [0,1].

$$Pij = Rij \,/\, (Gij + Rij) \tag{4}$$

Pseudo-hue is higher for lips than for skin. In addition, luminance is a good cue for lip key point extraction and we compute a luminance matrix L, which is also normalised to the range of [0,1]. We then compute the following gradient matrices from P and L (Equation (5)):

$$R^{mid} = \nabla^{y}(P\nabla^{x}L),$$
$$\textit{where the matrices } P \textit{ and } \nabla^{x}L \textit{ are multiplied element by element}$$
$$R^{top} = \nabla^{y}(P - L);$$
$$\textit{and}$$
$$R^{low} = \nabla^{y}(P + L); \tag{5}$$

R_{mid} has high values for the middle edge of the lips, R_{top} has high values for the top edge of the lips, and R_{low} has large negative values for the lower edge of the lip (shown in Figure 4). From these observations it is possible to do a robust estimation of the key points in the lip region. The horizontal mid-line is determined by finding the maximum value over several central columns of R_{mid}. The row-index of that maximum determines the horizontal mid-line y_m, as shown in Figure 5. To detect the left and right corners of the lips, we consider the pseudo-hue on the horizontal mid line. Pseudo-hue is quite high around the lip corners and the lip corners appear clearly when we compute the gradient of the pseudo-hue image $\nabla^{x}P$ along the horizontal mid-line. The maxima and minima of this gradient over the cross-section, with appropriate geometric constraints, determine the outer left lip corner (*olc-left*), inner left lip corner (*ilc-left*), outer right lip corner (*olc-right*) and inner right lip corner (*ilc-right*), which are shown in Figure 5. The two edges of the upper lip are determined by two zero-crossings, again with appropriate constraints, of the gradient R_{top} taken along the vertical mid-line of the lips and the two edges of the lower lip are determined by two zero-crossings of the gradient R_{low} taken along the vertical mid-line of the lips. Consequently, the upper lip width *(ULW)* and the lower lip width (*LLW*) can be determined.

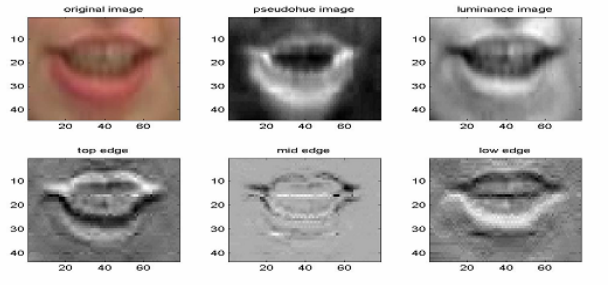

Figure 4: Detecting lip boundaries using pseudo hue/luminance edge images

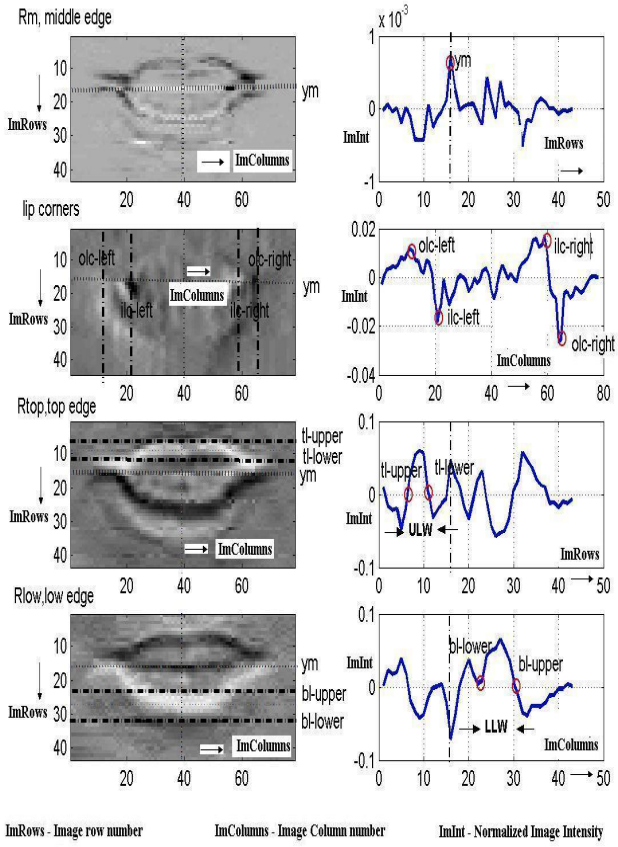

Figure 5: Key lip point extraction from lip-edge images

An illustration of the geometry of the extracted features and measured key points is shown in Figure 6. The major dimensions extracted are the inner lip width (w_1), outer lip width (w_2), inner lip height (h_1), outer lip height (h_2), upper and lower lip widths (h_3 *and* h_4) and the distance $h_5 = h_2/2$ between the mid-horizontal line and the upper lip. The extracted local features from the lip region are used as visual feature vector. This method was applied to the video sequences showing a speaking face for all subjects in the database. In all sequences, the lip region of the face was identified correctly with more than 95% accuracy.

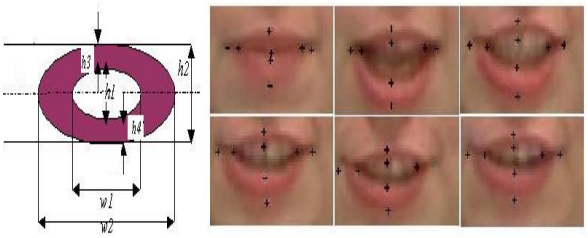

Figure 6: Lip-ROI key points for different lip openings of a speaking face

4 Global Feature Extraction

For extracting global features, principal component analysis (PCA) or Eigen analysis was performed on face and segmented facial regions corresponding to lips, eyes, eyebrows, forehead and nose. Each image was histogram-normalised and principal component analysis (PCA) was undertaken to obtain Eigen image vectors as described below.

The ROI from each image region in the video frame was projected onto n-dimensional subspace generated by the first n eigenvectors with n = 1…40, which represent the directions of maximum variance of the data. For example, each lip ROI yielded a visual feature vector with components between 1 and 40. Figure 7 shows the normalisation of a video frame and extraction of the lip ROI for a male subject, and six original lip images and their projections onto *n* dimensional subspace generated by the first n eigenvectors with *n = 1, 2, 4, 8, and 10* are shown.

The local and global features from other salient regions of the face such as eyes, eyebrows. forehead and nose region were extracted using a similar approach described in Section 3 and 4. The feature fusion involved concatenation of local and global features corresponding to each facial region. Next Section describes the experimental details for evaluating the proposed approach.

5 Experimental Details

For evaluating the performance of the proposed local feature extraction and feature fusion technique experiments with a facial video sequence corpus, VidTIMIT database developed by Sanderson in [9] was used. The VidTIMIT database consists of facial video and corresponding audio recordings of 43 people (19 female and 24 male), reciting short sentences selected from the NTIMIT corpus. Each video clip contains only one talking person making it a well controlled environment for conducting the experiments. The data was recorded in 3 sessions, with a mean delay of 7 days between Session 1 and 2, and of 6 days between Session 2 and 3. The mean duration of each sentence is around 4 seconds, or approximately 100 video frames. A broadcast-quality digital video camera in a noisy office environment was used to record the data. Figure 8 shows some sample images from this video corpus.

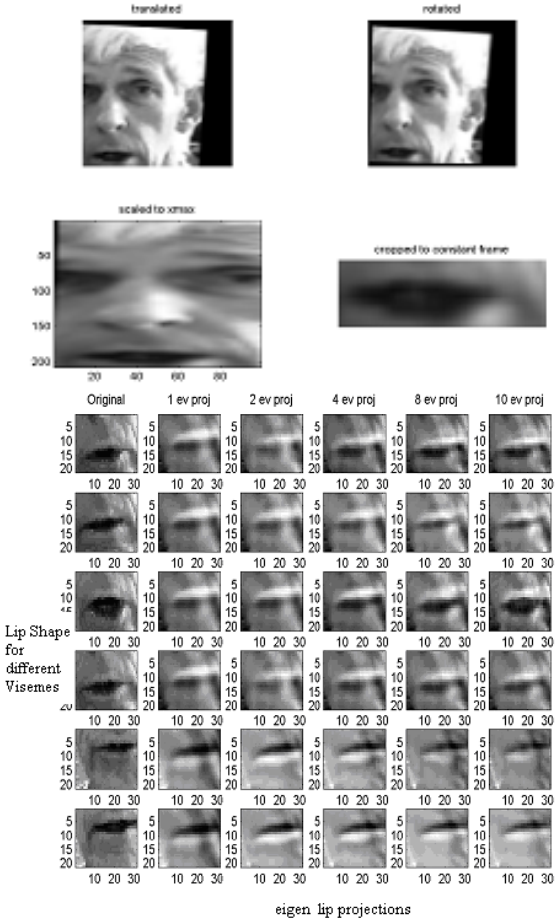

Figure 7: Face normalization and global feature extraction

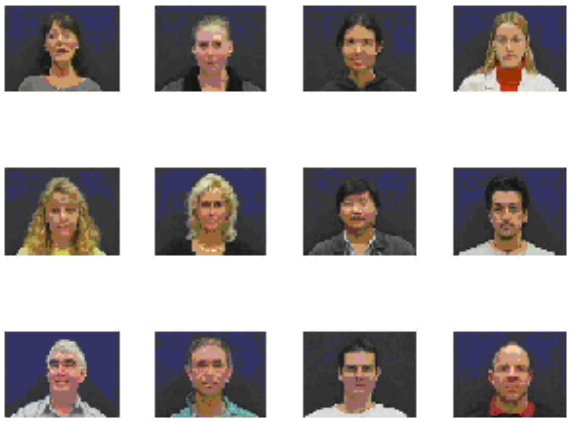

Figure 8: Sample Images from the Facial Video Corpus (VidTIMIT)

In order to examine the performance achieved, a feature level fusion of local features with global features is used. The fusion of local and global features was evaluated for three different experimental scenarios. In the first experiment, the local feature vector for each facial part (lips, eyes, eyebrows, forehead and nose) was obtained. The local feature vector for lip region comprised 6 lip parameters h1, h2, h3, h4, w1, and w2, described in Section 3. In the second experiment, global features in terms of 10 eigen projections of each facial part, based on principal component analysis as proposed in [11] and [12], was used. Before performing principal component analysis of the lip region images, each image was normalised by affine transformations involving translation, rotation and scaling operations (shown in Figure 7). For the third experiment, the local features from the each part of the face were concatenated with global features.

A statistical modeling technique based on Gaussian Mixture Models (GMM) was used for modeling each part from the training data [11]. The 10-mixture GMM was constructed was built with local feature vector, global feature vector and feature fusion vector during training phase for each person(client). First 2 sessions out of 3 sessions for each person in the data corpus was used for training phase.

The third session is used for test phase. The test phase comprised two scenarios - the genuine test scenario and tamper or forgery test scenario. For genuine scenario original Session 3 images were used. For tamper or forgery scenario, an artificially synthesized fake image database was created as it is very difficult to get a image databases depicting forgery. From the video of each person from session 3, the image frames were first extracted using Windows Movie Maker® tool and audio was discarded. For examining the discrimination ability of the proposed approach to detect forged or tampered images from genuine images, a set of synthetic fake images for each subject in the database was created by tampering the image frames with a set of affine transform operations and subsequent contamination with noise and compression artifacts emulating tampered images.

Discriminating a forged/tampered image from genuine image was done in Bayesian framework and was approached as a two class classification task based on hypothesis testing, i.e., a decision between the following two hypotheses H_0 and H_1:

- H_0: That the test image sequence O belongs to a genuine person model λ_C.
- H_1: That the test image sequence O belong to the tampered/forgery model λ_B

To test the two hypotheses, a ratio of the probability that the utterance belonged to the genuine person model, $P(O|\lambda_C)$, and the probability that the utterance belonged to the tampered/forged model, $P(O|\lambda_B)$ is obtained. By using Bayes' rule, the ratio is written as:

$$\frac{P(\lambda_c|O)}{P(\lambda_B|O)} = \frac{\frac{P(O|\lambda_c)P(\lambda_c)}{P(O)}}{\frac{P(O|\lambda_B)P(\lambda_B)}{P(O)}} \tag{6}$$

Assuming equal priors and cancelling the term $P(O)$ *in* Equation (6), we get the decision scores or log-likelihood ratio (LLR) as:

$$LLR(O) = \log\{p(O|\lambda_C)\} - \log\{p(O|\lambda_B)\} \tag{7}$$

A decision threshold, θ is used to decide the choice of H_0 or H_1.

$$\begin{aligned} LLR(O) \geq \theta \quad & accept(choose\ H_o) \\ LLR(O) < \theta \quad & reject(choose\ H_1) \end{aligned} \tag{8}$$

Setting the decision threshold θ is an important task and can be adjusted to implement a system with the desired efficiency.

In access control scenarios the decision threshold θ is set for equal error rate (EER), where the decision threshold is set such that FAR is equal to FRR. The details are explained below:

The performance of the proposed approach was evaluated in terms of measures used in security and access control contexts, i.e DET curves (detector error tradeoff) and

EERs (equal error rates). DET curves depict the variation of FARs (false accept rates) w.r.t. FRRs (False Reject Rates) in normal deviate scale. The FARs represent the likelihood that forged/tampered test images gets accepted as genuine images, and FRR represent the likelihood that genuine images gets rejected as tampered/forged images. The EER represents Equal Error Rates, a measure where the detection threshold θ is set such that FAR is equal to FRR. Ideally an EER score of 0 % is excellent and desirable. Practically a system with EER greater than 10% is useless. Any automatic approach that results in less than 5% is considered satisfactory in medium security access control scenarios.

It should be noted that decision scores are computed from for each facial part separately and combined using a late fusion approach with equal weights. Further, instead of making decision from one image, the decision about genuine and tampered/forged image sequence is done using at-least 100 correlated image frames (as they come from an audio-visual talking face sequences).

The DET curves showing EERs achieved for the proposed feature extraction and feature fusion of local and global features from the images are shown in Table 1, and Figure 8, 9 and 10. The DET curve for experiment I in Figure 8 represents the performance with a global feature vector consisting of Eigen image projections from each facial region of each frame. The three curves show the error rates for the best case frame level performance out of all frames, average performance across all frames, and the worst case frame level performance for each person, with tampering (involving affine transformation, additive noise and compression artefacts). The average case EER is about 4.8%. The DET curve for experiment II in the Figure 9 represents the performance with a local feature vector obtained from each facial part in each frame. Here an average EER of 2.2% is achieved, with the best case EER being 1.8% and the worst case EER of 3.6%. The DET curve for experiment III, shown in Figure 10 is obtained with a feature vector consisting of a feature fusion of local and global features from each facial part. The local features for lip region for example are the 6 lip region parameters (as in Experiment I), and global features are the 10 eigenlip projections (as in Experiment II), and for this combined feature vector, the average EER is 0.8%, with best-case EER being 0.5% and worst-case EER being 1.9 %.

Table 1: EER performance for three experiments

Facial Features	Best case Tamper	Average case Tamper	Worst case Tamper
Expt 1	4.8	5.4	6.5
Expt 2	1.8	2.2	3.6
Expt 3	**0.5**	0.8	1.9

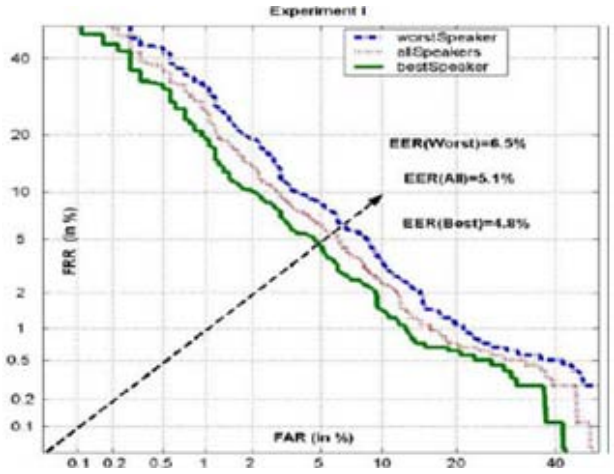

Figure 8: DET curve and EERs for Experiment I

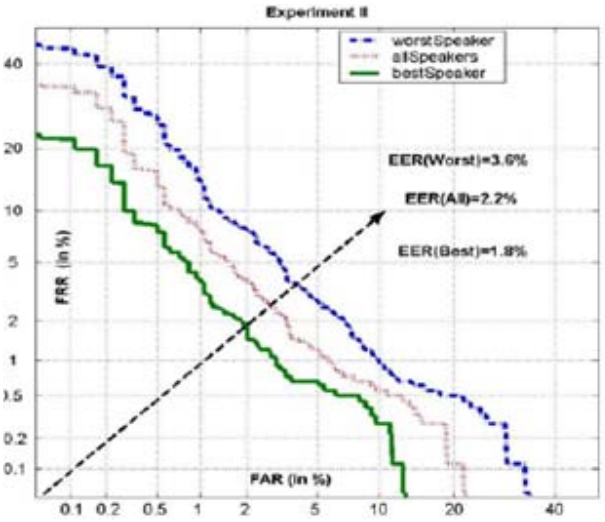

Figure 9: DET curve and EERs for Experiment II

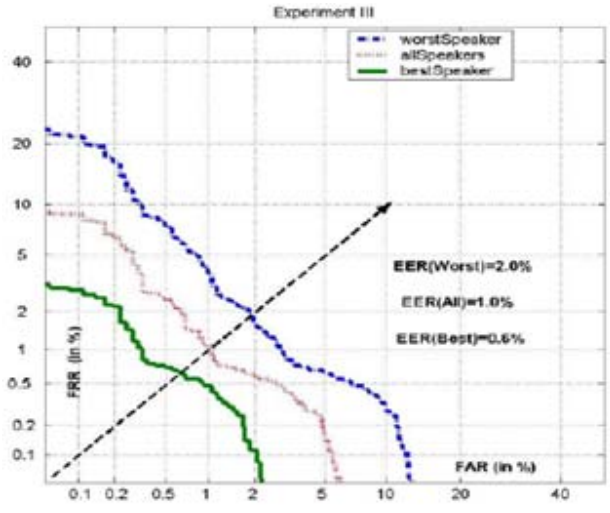

Figure 10: DET curve and EERs for Experiment III

6 Conclusion

The results of the tamper/forgery detection experiments have shown that the feature fusion of local features and global features from different facial regions undergoing motion (such as a human communication activity) contains inherently rich spatio-temporal information which can be exploited to detect tampering or forgery in on-line facial biometric access control scenarios. The proposed feature fusion of local features extracted from alternative colour spaces such as chrominance colour space and pseudo-hue colour space coupled with global features obtained with principal component analysis shows a significant improvement in performance in detecting tamper/forgery as compared to single global or local features. The technique demonstrates a simple and powerful method of verifying authenticity of images. Further investigations include extending the proposed modelling techniques for extracting and estimating blindly the internal camera processing techniques targeted for complex video surveillance, secure access control, and forensic investigation scenarios.

References

[1] Stork, D. and M. Hennecke "Speechreading by man and machine: data, models and systems", Springer-Verlag, Berlin, 1997

[2] Frischholz, R. and A. Werner, "Avoiding replay attacks in a face-recognition system using head pose estimation", Proceedings IEEE Int Workshop on Analysis and Modeling of Faces and Gestures, AMFG'03, 2003.

[3] Matsui, T and S. Furui, "Concatenated phoneme models for text-variable speaker recognition", Proceedings International Conference on Acoustics, Speech and Signal Processing, ICSLP-1993, 391-394, 1993.

[4] Choudhury, T., B. Clarkson, T. Jebara and A. Pentland, "Multimodal person recognition using unconstrained audio and video", in AVBPA-1999, Audio- and Video-based biometric person authentication, International Conference on Audio and Video-Based Biometric Person Authentication, 1999.

[5] Chetty, G. and M. Wagner, "Liveness" verification in audiovideo authentication", Proceedings International Conference on Spoken Language Processing, ICSLP-2004, 2004.

[6] Chetty, G. and M. Wagner, "Liveness verification in Audio Video Speaker Authentication", accepted for Australian International Conference on Speech Science and Technology SST-04, 2004.

[7] Eveno, N., A. Caplier and P.Y. Coulon, "Jumping Snakes and Parametric Model for Lip Segmentation", International Conference on Image Processing, Barcelona, Spain, 2003.

[8] Matthews, I., J. Cootes, J. Bangham, S. Cox and R. Harvey, "Extraction of visual features for lipreading" , IEEE Trans PAMI, vol.24, no.2,pp, 198-213, 2002.

[9] Sanderson, C. and K.K. Paliwal, "Fast features for face authentication under illumination direction changes", Pattern Recognition Letters 24, 2409-2419, 2003.

[10] Atal, B.. "Effectiveness of linear prediction characteristics of the speech waveform for automatic speaker identification and verification", Journal of Acoustical Society of America 55, 1304-1312. 1974

[11] Turk, M and A. Pentland, "Eigenfaces for recognition", Journal of Cognitive Neuroscience 3, 71-86, 1991

[12] Bregler, C. and Y. Konig, " "Eigenlips" for robust speech recognition", ICASSP-1994, Proc. Int. Conf. On Acoustics, Speech and Signal Processing, 1994.