Local Ordinal Contrast Pattern Histograms for Spatiotemporal, Lip-based Speaker Authentication

Budhaditya Goswami, Chi Ho Chan, Josef Kittler and Bill Christmas

Abstract— **The lip-region can be interpreted as either a genetic or behavioral biometric trait depending on whether static or dynamic information is used. Despite this breadth of possible application as a biometric, lip-based biometric systems are scarcely developed in scientific literature compared to other more popular traits such as face or voice. This is because of the generalized view of the research community about the lack of discriminative power in the lip region. In this paper, we propose a new method of texture representation called Local Ordinal Contrast Pattern (LOCP) for use in the representation of both appearance and dynamics features observed within a given lip-region during speech production. The use of this new feature representation, in conjunction with some standard speaker verification engines based on Linear Discriminant Analysis and Histogram-distance based methods, is shown to drastically improve the performance of the lip-biometric trait compared to the existing state-of-the-art methods. The best, reported state-of-the-art performance was an HTER of 13.35% for the XM2VTS database. We obtained HTER of less than 1%. The improvement obtained is remarkable and suggests that there is enough discriminative information in the mouth-region to enable its use as a primary biometric modality as opposed to a "soft" biometric trait as has been done in previous research.**

I. INTRODUCTION

Numerous measurements and signals have been proposed and investigated for use in biometric recognition systems. Among the most popular measurements are fingerprint, face and voice. Each of these biometric traits have their pros and cons with respect to accuracy and deployment. The use of lip-region features as a biometric straddles the area between the face and voice biometric.

There are various factors that make the use of lip features a compelling biometric. Since speech is a natural, non-invasive signal to produce, the associated lip-motion can also be captured in a non-intrusive manner. With the advent of cheap camera sensors for imaging, it is easier than ever before to isolate the lip-region features and use them in combination with other biometric traits to enhance the robustness of multi-modal biometric systems. The use of talking face features also naturally increases the robustness of the system with respect to any attempts at faking "liveness". Since the lip data can be captured at a distance, it represents a passive biometric as it requires no active user participation. The challenges of using the lip as a biometric lie in the areas of uniqueness and circumvention. The research question is therefore: how do we extract accurate and person-specific information from the lip region at a distance and still maintain a sufficient inter-person variation to intra-person variation ratio for accurate verification?

All authors are with the Centre for Vision, Speech and Signal Processing, Faculty of Electronics and Physical Sciences, University of Surrey, Guildford, GU2 7XH, United Kingdom {b.goswami, c.chan, j.kittler, w.christmas}@surrey.ac.uk

The physical attributes of the lip region are affected by the craniomaxillofacial structure of an individual. Human lip movement actually occurs through the use of the flexible mandible and consequently, the shape, appearance and movement of an individual's lip are a direct physical manifestation of their DNA resulting in its usability as a genetic biometric. Additionally, the lip is used by humans to control speech production. The means and forms of its use depend upon the language being spoken and an individual's pronunciation which is affected by numerous socio-economic factors. The manifestation of individual behaviour leads to behavioural dynamics of the lip region which in turn can also be used as a biometric somewhat akin to the idea of a "mouth-signature".

In this paper, we propose a new method of texture representation called Local Ordinal Contrast Patterns (LOCP). We use this texture representation within a configuration called Three Orthogonal Planes (TOP). The TOP configuration is increasingly being used within the field of speech or action recognition and segmentation. To the best of our knowledge, this is the first application of TOP to speaker authentication (verification or identification). TOP specifies planar directions along which LOCP features can be computed. This effective enables LOCP-TOP to quantise the dynamic texture and appearance information in the mouth-region. This sort of feature description is demonstrated to have excellent performance in extracting identity specific information from within a visual speech signal when used with some simple text-independent speaker verification systems based on Linear Discriminant Analysis (LDA) and chi-squared histogram matching based methods.

A taxonomy of the state-of-the art in lip-based speaker verification is presented in Section II. A summary of the current performance characteristics of the field is presented in Table I. A discussion of the approaches and their merits and failings leads to the motivation behind the development of the current, novel feature descriptor. The detailed treatment of the use of LOCP features for dynamic texture description is provided in Section III. An overview of the speaker verification systems used to evaluate the usefulness of this novel descriptor is provided in Section IV. The paper concludes with the experimental evaluation in Section. V and some concluding remarks in Section VI.

978-1-4244-7580-3/10/$26.00 ©2010 Crown

II. Literature Review

The use of the lip region as a means of human identification was first proposed through the concept of "lip-prints" in the field of forensic anthropology as early as the 20th century by forensic investigators such as Fischer and Locard [13]. Lip prints contained information about the individual grooves and eccentricities of the lip surface. The application of lip prints specifically as a biometric trait for security applications was introduced in [20].

The state-of-the-art approaches to lip biometrics can be segregated into approaches that either make use of genetic or behavioural lip characteristics. From a systems point of view, an alternative taxonomy can also be based on whether the approach uses static or dynamic information from the lip-region. This also enables the incorporation of a hybrid class of methods which attempt to capture both types of information.

A. Static Methods

When the lip is used as a genetic biometric, the features extracted from it either corresponds to a shape representation of its contour, geometric properties or appearance. Additionally, most of these methods either operate on static images using only single-frame information, or they operate on a sequence of speech and the genetic biometric features can conceptually be represented as a 3 dimensional volume of the temporal shape evolution.

[7] used hand-labelling to segment the lip contour for recognition. The extracted, geometeric lip features were compared to the performance of long-established acoustic features with largely similar performance. However, an obvious drawback to this system was the use of hand-labelling which restricted the scope of the experimental validation to only a small data-set. This idea was extended in [17] where an automatic lip-segmentation system based on Active Shape Models(ASM) was used to extract the shape and intensity information from the mouth-region during speech. Principal Components Analysis (PCA) was then used to perform shape-independent, intensity based feature extraction. These features were then used in conjunction with acoustic features to perform speaker recognition. The authors in [11] achieved some very promising results using geometrical features. They parameterised the shapes of observed lips on a frame-by-frame basis using cartesian and polar co-ordinates. Recognition was then performed using a multiparameter Hidden Markov Model (HMM) with the polar co-ordinates and a multilayer neural network is applied to the Cartesian coordinates. In [4] automatic lip segmentation using pixel-based thresholding in the HSV colour space was used to extract the lip region. A geometric feature vector was computed from this region using information like lip width and height. A score level fusion was then performed with acoustic features for speaker verification. In [8], a comparative evaluation of the representation of a segmented lip region in terms of a variety of geometric features is presented. The authors suggest the use of 3rd order Zernike moments as the optimal geometric representation of a lip region for use as a shape based biometric.

B. Dynamic Methods

Dynamic methods make use of features related to the changes observed in the mouth-region during speech production. Within these systems, there are two catergories. Most deployed biometric systems are based on scenarios with cooperative users speaking fixed string passwords or repeating prompted phrases from a small vocabulary. These generally employ what is known as *text-dependent*(TD) systems. Such constraints are quite reasonable and can greatly improve the system accuracy. However, there are cases when such constraints can be impossible to enforce. In situations requiring greater flexibility, systems are required that are able to operate without explicit speaker cooperation and independent of the spoken utterance. This mode of operation is referred to as *text-independent*(TI) speaker recognition.

a) Text-dependent Methods: In the work presented in [19] lip tracking is performed using a simple Bayes filter, an apriori, generative, eigen-model of lip-shape and a simple first order temporal evolution model. The lip contour is then segmented using probabilistic boundary searching using colour models. The motion parameters are computed by considering all the eigenlips that form the sequence of some speech. This sequence is then compared against the claimed identity of a sequence using the Dynamic Time Warping (DTW) algorithm.

b) Text-independent Methods: In the work presented in [10] and [9], a conceptually similar technique to optical flow estimation, called the 3-D structure-tensor method is used to estimate the spatiotemporal motion flow vectors of the lip contour. These features are then fused with acoustic features to perform speaker recognition with Support Vector Machines (SVM). The difference between each method lies in the method used for quantisation of the 3-d structure tensors and the use of Gaussian Mixture Models (GMMs) as opposed to HMMs for speaker verification.

In the work of [18], Minimum Average Correlation Energy (MACE) filters are used on frames containing only the mouth-region to perform lower face based person verification. The aim is to decorrelate all the variant information present in this region and use the resulting features to build a discriminative model for an individual.

C. Hybrid Methods

Hybrid methods use information in both a static and dynamic manner. The authors in [2] improved the quality of automatic lip feature extraction by using the Discrete Cosine Transform(DCT) to orthogonalise the lip region data into static and dynamic features. These features were then individually added to acoustic data for use as a biometric.

The authors in [6] use a combination of audio, lip texture and lip motion features. The lip texture features are represented using 2d-DCT coefficients. Discriminative analysis of the dense motion vectors contained in a bounding box around the mouth region is used to obtain the lip

motion information. The feature level comparison is then performed using the reliability weighted summation (RWS) decision rule. Additionally, the authors have extended their experimentation on the explicit usefulness and type of lip motion information using dense motion features to perform a comparative evaluation in [5].

In [21], motion estimation is performed using optical flow. The optical flow information is used to generate two kinds of visual feature sets in each frame. The first feature set consists of variances of vertical and horizontal components of optical-flow vectors. These are useful for estimating silence/pause periods in noisy conditions since they represent movement of the speakers mouth. The second feature set consists of maximum and minimum values of integral of the optical flow. Each of the feature sets is combined with an acoustic feature set in the framework of HMM-based speaker recognition. In [1] particle filters are used to track the shape of the lip during speech production. GMMs are then used to build speaker models based on the extracted shape and intensity features. These models are then used within a speaker verification engine. In [23], ASMs are used to perform lip tracking. LDA is then performed on the extracted temporal sequences connected with user speech. This serves to identify the most discriminative features. These features can then be fused with intensity features as a biometric.

D. State-of-the-art Performance Review

In order for various speaker verification systems to be compared, a variety of factors need to be considered. Commonly, lip-based features are evaluated in terms of the performance improvement they provide through feature-level fusion with more established biometric traits such as audio and face. For the testing of speaker verification systems, there exist only a few databases such as [14] with established verification protocols that enable a fair comparison of systems. However, because most of these databases are not free, most publications in the area of lip-based biometric systems use custom-built datasets and evaluation protocols. Table I provides an overview of the performance of various lip-biometric systems. The performance values are for the respective metric used to evaluate the verification performance. For a more thorough description of the various metrics related to speaker verification, the reader is referred to [3]. Please note that some methods did only speaker identification, in which case their results are not included. Additionally, some methods presented in the section above were referred to for ideas about lip-region feature parameterisation. The applications of this were also sometimes in the field of visual speech recognition. The disadvantage of using custom-built datasets for this endeavor is that in addition to reducing the comparability of the systems, often the classification task is made easier. In speaker verification, success depends on the ratio of trait feature dimensions to the number of clients. In real-world scenarios, this ratio is heavily skewed towards the number of clients and consequently, creates an unfavourable environment for successful classification. As shown in Table I, the most commonly used database and protocol for speaker verification using lip-features is XM2VTS (used by 3 authors) and Lausanne Protocols respectively. The performance obtained using lip features *only* on this database are by [10](Equal Error Rate (EER) of 22.0%) and [19](Half Total Error Rate (HTER) of 13.35%). Given the definition of HTER and EER respectively, whilst we cannot infer the HTER of [10], we can however say that it is going to be at best equal to the EER value. Consequently, the performance measure to beat using only lip information (i.e. without multi-modal feature fusion) is an HTER of 13.35%. For the purposes of this experiment, we will compare our findings with those systems that made use of the XM2VTS database and the Lausanne protocols for experimental evaluation.

III. Spatiotemporal Descriptors Using Local Ordinal Contrast Patterns

Ordinal contrast is a measure from the same family as Local Binary Patterns (LBP) [16]. It represents the relative difference in the immediate local neighbourhood of a given pixel. In computer vision, the absolute information contained within a pixel, including intensity, color and texture can vary dramatically under various illumination conditions. However, the mutual ordinal relationships between neighbours at the pixel level or region level continue to reflect the intrinsic nature of the object and provide a degree of response stability in the presence of such changes. An ordinal contrast encoding is used to measure the contrast polarity of values between a pixel pair (or average intensities between a region pair) as either brighter than or darker than some reference. This polarity is then turned into a result value in a binary decision. [24] have explained that the ordinal measure is invariant to any monotonic transformation, such as image gain, bias or gamma correction.

LBP is an example of such an ordinal measure. It offers a prowerful and attractive texture descriptor showing excellent results in terms of accuracy and computational complexity in many empirical studies. The LBP operator meaures the ordinal contrast pairs between a local neighbour value and the centre pixel value. The LBP is obtained by concatenating these binary results and then converting the sequence into the decimal number. Recently however, [12] and [22] have pointed out that LBP misses the local structure if the centre pixel is affected by noise. This is because LBP captures the mutual information between a neighbourhood and its centre value. In order to tackle this problem, [12] proposed, Improved LBP which performs odinal contrast measurement with respect to the average of the pixel neighborhood instead of the centre pixel. [22] have proposed Local Ternary Patterns (LTP), which extends LBP to 3-valued codes. The LTP is split into two LBPs: positive and negative. In other words, LTP increases the feature dimension.

In this paper, we propose a novel approach to ordinal contrast measurement called Local Ordinal Contrast Patterns(LOCP). LOCP uses the circular neighbourhoods for ordinal contrast measurement. Instead of computing the ordinal contrast with respect to any fixed value such as that at the centre pixel or the average intensity value, it computes the

TABLE I

PERFORMANCE OF LIP BIOMETRIC SYSTEMS FOR SPEAKER VERIFICATION

System	Feature	Database	Subjects	Metric	Performance
Abdulla[1]	Hybrid(lip shape and lip intensity)	Custom	35	EER	18.0
Broun[4]	Static(lip geometric) + Audio	XM2VTS(Modified Lausanne Protocol C2)	261	HTER	6.3
Cetingul[5]	Hybrid(lip texture and lip motion)	MVGL-AVD	50	EER	5.2
Cetingul[6]	Static(lip texture)	MVGL-AVD	50	EER	1.7
Faraj[10]	Lip Dynamic TI	XM2VTS (Lausanne Protocol)	295	EER	22
Faraj[10]	Lip Dynamic TI + Audio	XM2VTS (Lausanne Protocol)	295	EER	2
Gomez[11]	Static(lip geometric)	Custom	50	EER	0.015
Jourlin[17]	Static(lip shape)	M2VTS	37	HTER	6.85
Jourlin[17]	Static(lip shape) + Audio	M2VTS	37	HTER	0.3
Sanchez[19]	Lip Dynamic TD)	XM2VTS(Lausanne Protocol)	295	HTER	13.35
Sanchez[19]	Lip Dynamic TD + Face	XM2VTS(Lausanne Protocol)	295	HTER	4.72
Sanchez[19]	Lip Dynamic TD + Audio	XM2VTS(Lausanne Protocol)	295	HTER	0.74
Sanchez[19]	Lip Dynamic TD + Face + Audio	XM2VTS(Lausanne Protocol)	295	HTER	7.06
Samad[18]	Lip Dynamic TI	Custom from AMP CMU	10	HTER	0.0
Wark[23]	Lip Dynamic TI	TULIPS1	96	EER	0.0
Wark[23]	Static(lip shape)	TULIPS1	96	EER	6.3
Wark[23]	Static(lip intensity)	TULIPS1	96	EER	0.0

pairwise ordinal contrasts for the chain of pixels representing the circular neighbourhoods starting from the centre pixel. Additionally, linearly interpolating the pixel values allows the choice of any radius, R and the number of pixels in the cicular neighbourhood, P, to form an operator. This enables the modelling of arbitrarily large scale structure by varying R. During the operation of LOCP, we choose P pixel pairs for ordinal contrast encoding presented in Equation 1. The pixel indices are shown in Figure 1. The pattern is obtained by concatenating the binary numbers coming from the encoding and then converting the sequence into the decimal number. LOCP is a 2-dimensional texture descriptor. It represents local, pairwise neighbourhood derivatives and the non-dependence on a fixed point of reference implies that it is implicitly conditioned to be more robust to noise. The use of LOCP enables the encapsulation of compact, local structure within this descriptor.

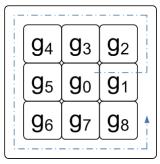

Fig. 1. LOCP Feature Computation: Compute pairwise ordinal contrast measure along the direction of the dotted arrow

$$LOCP_{P,R}(\mathbf{x}) = \sum_{p=0}^{P} s(g_{p+1} - g_p)2^p \mid s(v) = \begin{cases} 1 & v \geq 0 \\ 0 & v < 0 \end{cases} \quad (1)$$

Recently, local binary patterns on three orthogonal planes (LBP-TOP) [25] have been proposed to extend the LBP to a spatiotemporal representation for dynamic texture analysis. LBP-TOP is computationally simple as it extracts the LBP in three orthonormal planes within a spatiotemporal volume. Motivated by [25], we extend our new operator for dynamic texture analysis by extracting the LOCP in three orthonormal planes (i.e. XY, XT and YT) within a volume. Figure 2 demonstrates the lip images from three planes. In each plane, the LOCP is extracted and the plane-pattern histogram, $\boldsymbol{h}^{\beta}_{P,R}(i)$ is computed where $\beta \in \{XY, XT, YT\}$ represents a plane.

$$\boldsymbol{h}^{\beta}_{P,R}(i) = \sum_{(x',y') \in \boldsymbol{M}} B(LOCP^{\beta}_{P,R}(x', y') = i) \quad (2)$$

where the function $B()$ represents a boolean indicator, i is the value of the LOCP, $\boldsymbol{M}$ is the region for which we are computing the histogram.

Then the histogram of each plane is concatenated into one single histogram, $\boldsymbol{f}^{\alpha}$ shown in Figure 3 to provide the dynamic texture information. Here, α represents a member from the set of possible TOP configuration combinations $\alpha \in \{XY, XT, YT, XYXT, XYYT, XTYT, XYXTYT\}$. Consequently, for a concatenation of all features i.e. $\alpha = XYXTYT$, we would obtain the histogram shown by Equation 3.

$$\boldsymbol{f}^{XYXTYT} = [\boldsymbol{h}^{XY}_{P,R}, \boldsymbol{h}^{XT}_{P,R}, \boldsymbol{h}^{YT}_{P,R}] \quad (3)$$

One important consideration in the application of the TOP configuration is the parameter value of P and R for the LOCP feature descriptor along each place. These values relate to the sampling rate in the XY, XT or YT planes. Since the sampling rates in each plane are used to capture sufficient dynamic evolution, the input parameter values for P and R need to be tailored to each plane.

IV. SPEAKER VERIFICATION SYSTEMS

For this system, the method in [15] was first used to generate estimates of tracked outer lip contours for all videos. The estimated lip contours were then used to localise the mouth-region on a per-frame basis. These extracted regions were then used as input information for parameterisation using LOCP-TOP. Each extracted region can be visualised as a cube containing spatiotemporal information. This cube is first subdivided into overlapping sub-cubics. For each subcubic region, we use LOCP-TOP to extract histograms

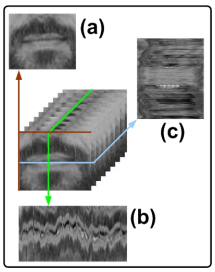

Fig. 2. Extraction of images using TOP. (a) XY Image (b) YT Image (c) XT Image

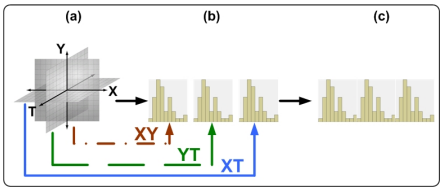

Fig. 3. LOCP-TOP Feature Description:(a) Represents feature parameterisation along TOP planes using LOCP operators.(b) Represents the histogram of the LOCP features from each TOP plane.(c) Represents the concatenation of these histograms for use in dynamic texture analysis

$\boldsymbol{h}_{P,R}^{\beta,j}$ where j represents the subcubic index. These are then further concatenated to form $\boldsymbol{f}^{\alpha,j}$. These combined histograms conceptually represent the intra-model feature-level fusion of extracted LOCPs in the different planes. The concatenated histograms are then input into one of two classification engines are described below.

Chi-squared Histogram Matching: In order to measure the similarity between two input LOCP-TOP histograms resulting from a probe and an enrolled gallery video, we use a simple, direct measure $Sim_{chi}(\boldsymbol{G}, \boldsymbol{I})$ based on Chi-squared distance between the histograms (with bin index i) of two input videos $\boldsymbol{G}$ and $\boldsymbol{I}$.

$$Sim_{\chi}(\boldsymbol{G}, \boldsymbol{I}) = -\sum_{j}\sum_{i} \frac{(\boldsymbol{f}_G^{\alpha,j}(i) - \boldsymbol{f}_I^{\alpha,j}(i))^2}{\boldsymbol{f}_G^{\alpha,j}(i) + \boldsymbol{f}_I^{\alpha,j}(i)} \quad (4)$$

Linear Discriminant Analysis: In order to extract the discriminative features we project the subcubic histograms, $\boldsymbol{f}^{\alpha,j}$, into LDA space as: $\boldsymbol{d}^{\alpha,j} = (\boldsymbol{W}_{lda}^{\alpha,j})^T \boldsymbol{f}^{\alpha,j}$. After projection, we perform normalized cross-correlation across all subcubics using two videos $\boldsymbol{G}$ and $\boldsymbol{I}$ as specified in Equation 5.

$$Sim_{LDA}(\boldsymbol{G}, \boldsymbol{I}) = \sum_{j} \frac{(\boldsymbol{d}_G^{\alpha,j})^T \boldsymbol{d}_I^{\alpha,j}}{\|\boldsymbol{d}_G^{\alpha,j}\| \|\boldsymbol{d}_I^{\alpha,j}\|} \quad (5)$$

V. Results and Evaluation

A. Experimental Set-up

For our experiments, the following set-up was used. The mouth-region localisation for the XM2VTS database was set to be 61 by 51 pixels. LOCP feature parameters P and R were set to 8 and 3 respectively. Additionally, they were set to be the same for all planar configurations. Each spatiotemporal video cube was subdivided into 5 subcubics along the XY direction and 3 subcubics along the T axis. Each of these subcubics overlapped each other by 70%. The reason for this overlap was to ensure quantisation of temporally continuous information. The XM2VTSDB [14] database, is a large multi-modal database intended for training and testing multi-modal verification systems. It contains synchronised video and speech data along with image sequences that allow multiple views of the face. The database consists of digital video of 295 subjects. For these experiments, we followed the Configuration 1 (C1) and Configuration 2 (C2) of the Lausanne protocol that accompanies this database for speaker verification.

B. Discussion

TABLE II
HTER and EER Performance for LOCP Histograms with Chi-Squared Histogram Matching in %

LOCP-TOP Histogram	Configuration I		Configuration II	
	Evaluation EER	Test	Evaluation	Test
XYXTYT	3.17	3.9	4.26	4.43
XY	3.7	3.7	4.25	4.27
XT	18.33	19.85	19.73	19.75
YT	9.05	10.41	11.55	10.73
XYXT	3.46	3.75	4.72	4.38
XYYT	**2.71**	**2.79**	**2.98**	**3.31**
XTYT	11.7	13.14	13.17	13.62

TABLE III
HTER and EER Performance for LOCP Histograms with LDA in %

LOCP-TOP Histogram	Configuration I		Configuration II	
	Evaluation EER	Test	Evaluation	Test
XYXTYT	**0.33**	**0.65**	**0.76**	**0.95**
XY	1.16	1.04	1.28	1.29
XT	7.97	8.59	9.06	10.19
YT	2.8	5.03	4.13	5.38
XYXT	0.5	0.84	1.29	1.22
XYYT	0.51	0.82	0.98	0.991
XTYT	2.01	3.56	2.52	4.22

Tables II and III show the HTER of the test-set and the EER of the evaluation-set of the various LOCPTOP histograms with the chi-squared and LDA verification systems respectively. The ROC curves are shown in Figs 4,5. The best performances (highlighted in bold) were obtained using XYYT histograms with the chi-squared system for C1 and the XYXTYT histograms with the LDA system for C2. The first notable observation is that the performance of the speaker verification engine using LDA is significantly better (2 times better in the worst case) than using chi-squared. This is unsurprising, since LDA performs subspace projection of the histograms onto a discriminative space. Chi-squared distance is applied to the LOCP-TOP histograms directly in an unsupervised manner. Another interesting observation is that the performance along the XT plane in any configuration degrades the system performance except in LDA space. This is because mandibular deformation during speech production primarily manifests itself in the YT direction.

The results obtained demonstrated a marked and remarkable improvement on the best performance observed on this database in the literature (HTER=13.35) and indeed the result of the XYXTYT LOCP-TOP histograms using LDA is comparable to the state-of-the-art system performance using multi-modal fusion with audio and face features. This is due to the encapsulation of discriminative, dynamic and appearance textures using LOCP-TOP and the implicit intra-model fusion of both genetic and behavioural properties of the observed subject lip-regions. A final point to note was the comparison between LOCP and LBP which belong to the same family of ordinal contrast measures. LOCP consistently outperformed LBP suggesting that it is a viable alternative for texture description. For the XY plane, this improvement was to the tune of 40%.

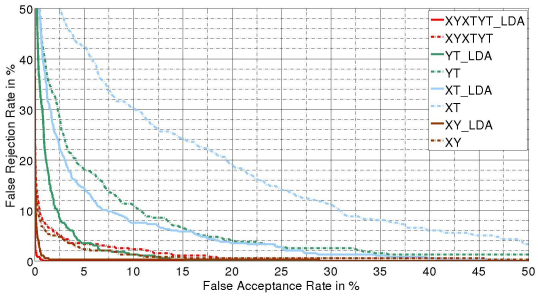

Fig. 4. ROC Curves for C1 using LOCP-TOP. Dashed lines are for the Chi-squared system, solid lines are LDA system

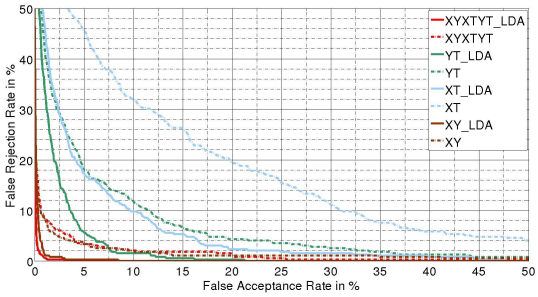

Fig. 5. ROC Curves for C2 using LOCP-TOP. Dashed lines are for the Chi-squared system, solid lines are LDA system

VI. Conclusions and Future Work

We first presented a thorough review of the current state-of-the-art lip biometric systems. In this paper, we have proposed a novel ordinal contrast measure called LOCP. This has been used in a TOP configuration as input into speaker verification systems using chi-squared histogram distance and LDA respectively. The resulting biometric systems have been used to evaluate the performance of mouth-region biometrics in the XM2VTS database using the standard Lausanne protocols. The application of this novel feature representation has been demonstrated to comprehensively outperform previous feature descriptors encountered in the state-of-the-art. The findings also suggest that there is sufficient discriminative information within the spatiotemporal evolution of the mouth-region during speech production for its use as a primary biometric trait. This can be especially useful in circumstances where auditory information may not be available for fusion. Finally, LOCP histograms are computationally simple compared to the more exotic feature parameterisations encountered in the literature.

Acknowledgments: This work is supported by the EU-funded MOBIO project grant IST-214324.

References

[1] W. Abdulla, P.W.T. Yu, and P. Calverly. Lips tracking biometrics for speaker recognition. 1(3):288–306, 2009.

[2] R. Auckenthaler, J. Brand, J. Mason, C. Chibelushi, and F. Deravi. Lip signatures for automatic person recognition. In *MMSP*, pages 457 – 462, 1999.

[3] S. Bengio, J. Mariethoz, and S. Marcel. Evaluation of biometric technology on XM2VTS, 2001.

[4] C.C.Broun, X.Zhang, R.M.Mersereau, and M.Clements. Automatic speechreading with application to speaker verification. In *ICASSP*, volume 1, pages 685 – 688, 2002.

[5] H.E. Çetingül, E. Erzin, Y. Yemez, and A.M. Tekalp. Discriminative analysis of lip motion features for speaker identification and speech-reading. *Image Processing, IEEE Trans.*, 15(10):2879–2891, 2006.

[6] H.E. Çetingül, Y. Yemez, E. Erzin, and A.M. Tekalp. Multimodal speaker/speech recognition using lip motion, lip texture and audio. *Signal Process.*, 86(12):3549–3558, 2006.

[7] C.C. Chibelushi, S. Gandon, J.S.D. Mason, F. Deravi, and R.D. Johnston. Design issues for a digital integrated audio-visual database. *Integrated Audio-Visual Processing for Recognition, Synthesis and Communication, IEE Colloquium on*, pages 711–717, 1996.

[8] M. Chorasś. Human lips as emerging biometrics modality. In *ICIAR*, pages 993 – 1002, 2008.

[9] M.I. Faraj and J. Bigün. Motion features from lip movement for person authentication. In *ICPR*, pages 1059–1062, 2006.

[10] M.I. Faraj and J. Bigün. Person verification by lip-motion. In *CWPRW*, pages 37–44, 2006.

[11] E. Gomez, C.M. Travieso, J.C. Briceno, and M.A. Ferrer. Biometric identification system by lip shape. In *ICCST*, pages 39 – 42, 2002.

[12] H. Jin, Q. Liu, H. Lu, and X. Tong. Face detection using improved lbp under bayesian framework. In *ICIG*, pages 306–309, 2004.

[13] J. Kasprazak. Possibilities of cheiloscopy. *Forensic Science International*, 46:145–151, 1990.

[14] K.Messer, J.Matas, J.Kittler, J.Luettin, and G.Maitre. XM2VTSDB: The extended M2VTS database. In *AVBPA*, 1999.

[15] M.U.R.Sánchez. *Aspects of facial biometrics for verification of personal identity*. PhD thesis, University of Surrey, UK, 2000.

[16] M. Pietikäinen, T. Ojala, J. Nisula, and J. Heikkinen. Experiments with two industrial problems using texture classification based on feature distributions. *Intelligent Robots and Computer Vision XIII: 3D Vision, Product Inspection, and Active Vision*, 2354(1):197–204, 1994.

[17] P.Jourlin, J.Luettin, D.Genoud, and H.Wassner. Acoustic labial speaker verification. In *AVBPA*, pages 319–334, 1997.

[18] S.A. Samad, D. A. Ramli, and Aini Hussain. Lower face verification centered on lips using correlation filters. *Information Technology Journal*, 6(8):1146–1151, 2007.

[19] M.U.R. Sánchez and J. Kittler. Fusion of talking face biometric modalities for personal identity verification. In *ICASSP*, volume 5, pages 1073 – 1076, 2006.

[20] K. Suzuki, Y. Tsuchihashi, and H. Suzuki. A trail of personal identification by means of lip print. *I. Jap. J. Leg. Med.*, 22:392, 1968.

[21] S. Tamura, K. Iwano, and S. Furui. Multi-modal speech recognition using optical-flow analysis for lip images. *J. VLSI Signal Process. Syst.*, 36(2/3):117–124, 2004.

[22] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. In *AMFG*, pages 168–182, 2007.

[23] T.Wark, D. Thambiratnam, and S.Sridharan. Person authentication using lip information. In *IEEE TENCON*, pages 153–156, 1997.

[24] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In *ECCV (2)*, pages 151–158, 1994.

[25] G. Zhao and M. Pietikäinen. Local binary pattern descriptors for dynamic texture recognition. In *ICPR (2)*, pages 211–214, 2006.