Improvements in Lip-Based Video Biometric Performance

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 the nearest neighbour classifier, 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 investigated for use in biometric recognition systems. Among the most popular measurements are fingerprint, face and voice. Lip-region features straddle the area between the face and voice biometric.

There are various factors that make the use of lip features a compelling biometric. Lip-motion can be capture non-intrusively since it is a result of speech. 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?

The physical attributes of the lip region are affected by the craniomaxillofacial structure of an individual. Since this structure is a physical manifestation of their DNA, the lip can be considered a genetic biometric. Human lip movement during speech production occurs due to the flexible mandible. The means and forms of lip movement depend upon the language and an individual's pronunciation which is affected by numerous socio-economic factors. This manifestation of individual behaviour leads to behavioural dynamics of the lip region which in turn can also be used as a biometric. Notionally, this is akin to the idea of a "mouth-signature".

In this paper, we demonstrate some remarkable improvements in the performance of lip-based biometric systems. These improvements are realised through the use of a novel texture descriptor called Local Ordinal Contrast Patterns (LOCP). We use this texture representation within a novel configuration called Windowed Three Orthogonal Planes (WTOP), an extension to an approach called Three Orthogonal Planes (TOP). TOP was first proposed in the field of speech or action recognition and segmentation. It specifies planar directions along which texture features can be computed enabling the quantisation of dynamic texture and appearance information in the mouth-region. The combination of LOCP and WTOP 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 Normalised Correlation(NC) and Chi-squared (X2) histogram matching 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 relevant approaches 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. Section IV describes the proposed WTOP configuration. The speaker verification systems used to evaluate this novel descriptor are described in Section V. The paper concludes with the experimental evaluation in Section VI and some concluding remarks in Section VII.

II. Relevant Work

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 investigators such as Fischer and Locard [14]. Lip prints contained information about the individual eccentricities of the lip surface. The application of lip prints specifically as a biometric trait was introduced in [21]. A taxonomy of contemporary relevant work can be based on

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

System	Lip Feature	Database	Clients	Performance	
Faraj[11]	Lip Dynamic TI	**XM2VTS**	295	EER	22
Sanchez[20]	Lip Dynamic TD	**XM2VTS**	295	HTER	**13.35**
Gomez[12]	Static(lip geometric)	Custom	50	EER	0.015
Jourlin[18]	Static(lip shape)	M2VTS	37	HTER	6.85
Samad[19]	Lip Dynamic TI	AMP CMU	10	HTER	0.0
Wark[24]	Lip Dynamic TI	TULIPS1	96	EER	0.0
Wark[24]	Static(lip shape)	TULIPS1	96	EER	6.3
Wark[24]	Static(lip intensity)	TULIPS1	96	EER	0.0
System	**Feature Fusion**	**Database**	**Clients**	**Performance**	
Broun[4]	Static(lip geometric) + Audio	**XM2VTS**	261	HTER	6.3
Faraj[11]	Lip Dynamic TI + Audio	**XM2VTS**	295	EER	2
Sanchez[20]	Lip Dynamic TD + Face	**XM2VTS**	295	HTER	4.72
Sanchez[20]	Lip Dynamic TD + Audio	**XM2VTS**	295	HTER	**0.74**
Sanchez[20]	Lip Dynamic TD + Face + Audio	**XM2VTS**	295	HTER	7.06
Abdulla[1]	Hybrid(lip shape and lip intensity)	Custom	35	EER	18.0
Cetingul[5]	Hybrid(lip texture and lip motion)	MVGL-AVD	50	EER	5.2
Cetingul[7]	Static(lip texture)	MVGL-AVD	50	EER	1.7
Jourlin[18]	Static(lip shape) + Audio	M2VTS	37	HTER	0.3

TABLE I

PERFORMANCE OF LIP BIOMETRIC SYSTEMS FOR SPEAKER VERIFICATION SHOWING LIP PERFORMANCE AND FUSION PERFORMANCE

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.

Static Methods: use features extracted from the lip-region to describe its shape, geometric properties or appearance. Additionally, most of these methods either operate on static images using only single-frame information or on a sequence of speech video on a per-frame basis. Examples of such methods are in [8], [18], [6], [12], [4], [9]

Dynamic Methods: use features related to the changes observed in the mouth-region during speech production. These systems can be further segregated into two categories. Most deployed biometric systems are based on scenarios with co-operative users speaking fixed string passwords from a small vocabulary. These generally employ ***text-dependent***(TD) systems [20]. 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 [11], [10], [19].

Hybrid Methods: exploit both static and dynamic information by performing either score-level or feature-level fusion. [2], [7], [5], [22], [1], [24].

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 multi-modal fusion with more established biometric traits (e.g. audio and face). For the testing of speaker verification systems, there exist only a few databases such as [15] with established verification protocols that enable a fair comparison of systems. However, some publications use custom-built datasets and evaluation protocols. The disadvantage of using such datasets is that in addition to reducing the comparability of the systems, often the classification task is made easier by skewing the ratio of trait feature dimensions to the number of clients in favor of the systems. Table I provides an overview of the performance of various lip-biometric systems. For a more thorough description of the various metrics related to speaker verification, the reader is referred to [3].

As shown in Table I, the most commonly used database and protocol are XM2VTS (used by 3 authors) and Lausanne Protocols respectively. The performance obtained using lip features *only* on this database are by [20](Half Total Error Rate (HTER) of 13.35%). Multi-modal fusion with audio features [20] yields HTER of 0.74%. In this paper, we use the XM2VTS to ensure the comparability of our results with these reference benchmarks.

III. SPATIOTEMPORAL DESCRIPTORS USING LOCAL ORDINAL CONTRAST PATTERNS

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. The ordinal measure is invariant to any monotonic transformation such as image gain, bias or gamma correction [25]. The pattern 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.

Local Binary Patterns (LBP) [17] are an example of ordinal contrast patterns. LBP offers a powerful and attractive texture descriptor showing excellent results in terms of accuracy and computational complexity in many empirical studies. The LBP operator measures the ordinal contrast pairs between a local neighbour value and the reference (centre pixel) value. The LBP is obtained by concatenating these binary results and then converting the sequence into the decimal number. Recently [13] and [23] have pointed out that LBP misses the local structure if the reference value is affected by noise. In an LBP, the effect of measurement noise could result in a situation where the reference value changes by a single unit. This change affects all of the 8 ordinal contrast encodings. [23] have proposed Local Ternary Patterns (LTP), which extend LBP by increasing the feature dimensionality depending on the sign of the centre bit, at the expense of not being invariant to monotonic transformation. [13] proposed, Improved LBP which performs ordinal contrast measurement with respect to the average of the pixel neighborhood instead of the centre pixel to reduce the effect of a single noisy reference.

In this paper, we propose a novel approach to ordinal contrast measurement called Local Ordinal Contrast Patterns(LOCP) by diversifying the source of reference values. LOCP uses 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 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 represents local, pairwise neighbourhood derivatives.

$$LOCP_{P,R}(\mathbf{x}) = \sum_{p=0}^{P} s(g_{p+1} - g_p)2^p \text{ where}$$

$$s(v_p) = \begin{cases} 1 & v > 0 \\ 0 & v < 0 \\ 0 & v = 0 \quad \text{and} \quad p = 0 \\ s(v_{p-1}) & v = 0 \quad \text{and} \quad p > 0 \end{cases} \tag{1}$$

In terms of information representation, LBP suggests that the ordinal relationship between a single reference pixel and its neighbourhood contains texture information. LOCP suggests a new paradigm where texture is represented by the contents of the entire neighbourhood, not by the relationship of the neighbourhood with a single reference value. LOCP thus improves on LBP since a change in the value of a single pixel only affects at most 2 ordinal contrast encodings. Put another way, LOCP increases the robustness of the texture representation since a change in all 8 ordinal contrast encodings would require 4 alternate pixel values to change as opposed to just the single reference for LBP.

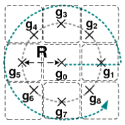

Fig. 1. LOCP Feature Computation: Compute pairwise ordinal contrast measure along the direction of the dotted arrow

Recently, local binary patterns on three orthogonal planes (LBP-TOP) [26] have been proposed to extend the LBP to a spatiotemporal representation for dynamic texture analysis. Motivated by [26], we extend our new operator for dynamic texture analysis by extracting the LOCP using the TOP configuration.

IV. Windowed Three Orthogonal Planes

WTOP is an example of a Dynamic Texture (DT) and a generalisation of the TOP configuration[26]. TOP allows the quantisation of spatial appearance information from the XY plane. Temporal quantisation in terms of co-occurence statistics of horizontal and vertical motion are in turn obtained from the XT and YT planes. TOP aims to describe these co-occurence statistics for each subject using volumetric, single-pixel-width image profiles in the X and Y directions to create XT and YT images. These profiles are extacted from the central column and central row of the frames in the visual speech volume.

TOP assumes temporal alignment of the texture object being quantised. Temporal alignment in this context ensures that feature correspondence exists in between frames of the same speaker. Unfortunately, in real-world scenarios, this may not be the case because tracking systems could get lost irrecoverably or alternatively display a large amount of temporal-"jitter". As a result, there is a need to extend this configuration in case of misalignment.

Since TOP uses only single-pixel width image profile layers, it represents a sampling of the volume of visual speech in both the spatial and temporal directions. It is clearly obvious that if the sampled planar cross-section does not display adequate temporal feature alignment i.e. in the XT and YT direction, the information in these planes will be degraded. TOP assumes that the object is centrally located within the spatiotemporal volume. Its use of central columns and rows for the XT and YT directions is an attempt to encapsulate maximal temporal variation. However, for deformable shapes such as the human lip during speech production, this assumption does not necessarily hold true if the lip is mislocalized and not central.

A simple approach to rectify this is to use a larger number of sample profiles to describe the XT and YT directions. However, this approach exponentially increases the feature dimensionality making the process computationally more expensive as well more redundant. Additionally it may negatively impact the signal-to-noise ratio, degrading performance.

A novel method to increase robustness to this effect is to use a windowing function that better samples the planar information content. As a special case of the above argument, TOP uses a Kronecker delta windowing function (or impulse response function) to sample the information contained in each planar direction.

In order to balance the choice of windowing function from every pixel layer to a single pixel, whilst at the same time, controlling the sample rate in planar directions, we chose to use the Gaussian windowing function. Potentially, various windowing functions can be applied to this process depending on the computational resources available e.g. rectangular, Hamming, triangular etc.

This novelty is clearer upon mathematical formulation of these ideas. A spatiotemporal cube of observed video of T frames can be considered as a set, $C : XY_{\{0 \dots T-1\}}$ where XY_i is each image frame of height N pixels and width M pixels so that $XY \in \mathcal{R}^{N \times M}$. The extraction of the $XT \in \mathcal{R}^{T \times M}$ images involves the application of a windowing mask, $\mathbf{\Omega}_Y \in \mathcal{R}^{N \times 1}$ to each element in the set C. The extraction of the $YT \in \mathcal{R}^{N \times T}$ images involves the application of a windowing mask, $\mathbf{\Omega}_X \in \mathcal{R}^{M \times 1}$ to each element in the set C. The two orthogonal temporal images, XT and YT can then be written as:

$$XT(i) = \mathbf{\Omega}_Y^T \times XY_i, \quad (0 \leq i \leq T) \tag{2}$$

and

$$YT(i) = XY_i \times \mathbf{\Omega}_X, \quad (0 \leq i \leq T) \tag{3}$$

In the case of the TOP representation, the mask $\mathbf{\Omega}$ is obtained as a constant vector computed using a Kronecker delta function. As mentioned above, we use a Gaussian mask to perform this task:

$$\mathcal{N}(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{(\boldsymbol{x} - \mu)^2}{2\sigma^2} \tag{4}$$

where μ here is the seeding point in the spatial axis whilst σ is the width of the windowing mask. $\boldsymbol{x}$ represents the axis index vector e.g. for $\mathbf{\Omega}_Y$, $\boldsymbol{x} = [0 \quad 1 \dots N-1]^T$. This idea is shown in Figure 2

Another method to increase robustness to spatial misalignment is to use the histogram of obtained mid-level features as input to the speaker verification system. This serves to remove the location information of each individual ordinal contrast pattern ensuring greater robustness to rotation and translation effects on the object in the visual speech segment. The histogram therefore ensures that it is the overall structural content of the visual speech signal that contributes to the spatiotemporal representation.

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 WTOP plane.

$$\boldsymbol{h}^{\beta}_{P,R}(i) = \sum_{(x',y') \in \boldsymbol{M}} B(LOCP^{\beta}_{P,R}(x', y') = i) \tag{5}$$

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 WTOP 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 6.

$$\boldsymbol{f}^{XYXTYT} = [\boldsymbol{h}^{XY}_{P,R}, \boldsymbol{h}^{XT}_{P,R}, \boldsymbol{h}^{YT}_{P,R}] \tag{6}$$

One important consideration in the application of the WTOP 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.

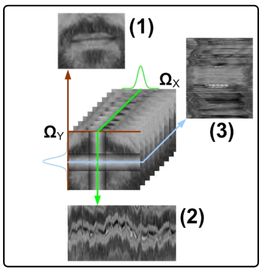

Fig. 2. Extraction of images using WTOP. (a) XY Image (b) YT Image (c) XT Image

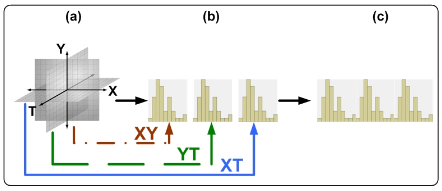

Fig. 3. LOCP-TOP Feature Description:(a) Represents feature parameterisation along TOP planes using LOCP operators.(b) Represents the histogram of the LOCP features from each TOP plane.(c) Represents the concatenation of these histograms for use in dynamic texture analysis

V. Speaker verification systems

For this system, the method in [16] 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-WTOP. Each extracted region can be visualised as a cube containing spatiotemporal information. The spatiotemporal video cube is divided into 3 or 6 sub-cubes along the T axis and 5 sub-cubes along the Y axis. These sub-cubes overlapped each other by 70%. The reason for this overlap was to ensure quantisation of temporally continuous information. Additionally, the number of cube partitions values in the T and Y axes respectively enabled us to evaluate the relative performance of the spatial and temporal information in greater detail. For each subcube, we use LOCP-WTOP to extract histograms $\boldsymbol{h}_{P,R}^{\beta,j}$ where j represents the subcube index. These are then further concatenated to form $\boldsymbol{f}^{\alpha,j}$. The combined histograms conceptually represent the feature-level fusion of extracted LOCPs in the different planes. These 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-WTOP 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 $\boldsymbol{I}$.

$$Sim_{\chi}(\boldsymbol{G}, \boldsymbol{I}) = -\sum_{j}\sum_{i} \frac{(\boldsymbol{f}_G^{\alpha,j}(i) - \boldsymbol{f}_I^{\alpha,j}(i))^2}{\boldsymbol{f}_G^{\alpha,j}(i) + \boldsymbol{f}_I^{\alpha,j}(i)} \tag{7}$$

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 8.

$$Sim_{LDA}(\boldsymbol{G}, \boldsymbol{I}) = \sum_{j} \frac{(\boldsymbol{d}_G^{\alpha,j})^T \boldsymbol{d}_I^{\alpha,j}}{\|\boldsymbol{d}_G^{\alpha,j}\| \|\boldsymbol{d}_I^{\alpha,j}\|} \tag{8}$$

VI. Results and Evaluation

A. Experimental Set-up

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. The XM2VTSDB [15] 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 I (C1) and Configuration II (C2) of the Lausanne protocol that accompanies this database for speaker verification.

B. Results

TOP Planar Configuration Evaluation: Tables II and III show the HTER of the test-set and the EER of the evaluation-set of the various LOCP-TOP histograms with the X2 and NC verification systems respectively. 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 NC is significantly better (2 times better in the worst case) than using X2. This is unsurprising, NC is performed in the discriminative space(LDA projection) while Chi-squared distance is applied to the raw feature histograms. The XY plane outperforms both the temporal planes implying that the appearance differences between clients are more discriminative than their lip dynamics. 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.

TOP Plane	Configuration I		Configuration II	
	Eval	Test	Eval	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 II

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

TOP Plane	Configuration I		Configuration II	
	Eval	Test	Eval	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

TABLE III

EER and HTER Performance for LOCP Histograms with Normalised Correlation(with LDA) Histogram Matching in %

TOP Input	Configuration I				Configuration II			
	LBP		LOCP		LBP		LOCP	
	Eval	Test	Eval	Test	Eval	Test	Eval	Test
TOP	3.18	3.52	2.99	3.86	4.46	3.60	4.27	3.97
WTOP, $\sigma = 0.1$	3.32	3.77	3.21	3.69	4.48	3.68	4.23	3.92
WTOP, $\sigma = 0.5$	3.41	4.03	3.22	3.69	4.71	3.83	4.41	3.97
WTOP, $\sigma = 1$	3.36	3.95	3.17	3.59	4.26	4.07	3.99	3.74
WTOP, $\sigma = 10$	3.00	3.19	2.51	3.04	4.43	4.11	4.01	3.49
WTOP, $\sigma = 100$	3.92	4.28	3.82	3.82	6.51	5.63	6.26	5.8

TABLE IV

LOCP/LBP-WTOP performance using the Chi-squared system

Performance of WTOP: Tables IV and V show the performance figures for LOCP-WTOP histograms. We have also included the results of the LBP-WTOP for reference. Note that in these tables, only the best performing TOP configuration i.e. XYXTYT was used. Various width (σ)

TOP Input	Configuration I				Configuration II			
	LBP		LOCP		LBP		LOCP	
	Eval	Test	Eval	Test	Eval	Test	Eval	Test
TOP	0.87	1.29	0.25	0.36	1.5	1.67	0.99	0.49
WTOP, $\sigma = 0.1$	0.86	1.64	0.33	0.38	1.3	1.57	0.76	0.5
WTOP, $\sigma = 0.5$	0.87	1.43	0.33	0.38	1.51	1.53	0.75	0.62
WTOP, $\sigma = 1$	0.83	1.75	0.19	0.44	1.53	1.53	0.75	0.62
WTOP, $\sigma = 10$	0.84	1.01	0.33	0.09	**1.25**	**0.96**	**0.61**	**0.27**
WTOP, $\sigma = 100$	0.86	1.18	0.33	0.35	1.52	1.23	0.96	0.55

TABLE V
LOCP/LBP-WTOP PERFORMANCE USING THE NORMALISED CORRELATION SYSTEM

values for the Gaussian windowing function were used. A final point to note was the comparison between LOCP and LBP which belong to the same family of ordinal contrast measures. The performance of LOCP and LBP in the X2 system were comparable. However, LOCP outperformed the LBP representation when combined with the NC system.

Comparison to literature: 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.

VII. 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. The use of lip motion for biometric speaker identification. In *SIU*, pages 148 – 151, 2004.

[7] 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.

[8] 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.

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

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

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

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

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

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

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

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

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

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

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

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