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

Chi Ho Chan, Budhaditya Goswami, Josef Kittler, *Member, IEEE,* and William Christmas,

***Abstract*—Lip region deformation during speech contains biometric information and is termed *visual speech*. This biometric information can be interpreted as being genetic or behavioural depending on whether static or dynamic features are extracted. In this paper, we use a texture descriptor called Local Ordinal Contrast Pattern (LOCP) with a dynamic texture representation called Three Orthogonal Planes (TOP) to represent both the appearance and dynamics features observed in visual speech. This feature representation, when used in standard speaker verification engines, is shown to improve the performance of the lip-biometric trait compared to the state-of-the-art. The best baseline state-of-the-art performance was Half Total Error Rate (HTER) of 13.35% for the XM2VTS database. We obtained HTER of less than 1%. The resilience of the LOCP texture descriptor to random image noise is also investigated. Finally, the effect of the amount of video information on speaker verification performance suggests that with the proposed approach, speaker identity can be verified with a much shorter biometric trait record than the length normally required for voice-based biometrics. In summary, the performance obtained is remarkable and suggests that there is enough discriminative information in the mouth-region to enable its use as a primary biometric trait.**

***Index Terms*—biometrics, lip, spatiotemporal, speaker verification, texture descriptor, ordinal contrast, dynamic texture**

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. The latter two arise naturally in the process of human speech production. The video of a talking face contains lip deformation during speech and can be termed *visual speech*. The McGurk effect [1] demonstrates the fact that visual speech contains different information to the audio speech signal. Both these sources of information are used during human speech perception. This effect may be experienced when a video of one phoneme's production is dubbed with a sound-recording of a different phoneme being spoken. Often, the perceived phoneme is a third, intermediate phoneme. The McGurk effect was used to improve the accuracy of speech recognition systems by using visual speech to reinforce the quality of input auditory information especially in noisy and crowded scenarios [2], [3]. Visual speech was subsequently used to perform speaker recognition [4]–[6]. In this paper, we investigate the usefulness of a novel representation of this lip deformation in visual speech as a biometric.

The lip is a twin biometric containing both genetic and behavioural information. This information is contained within individual mouth-region appearance as well as lip dynamics during speech. The individual mouth-region appearance is influenced by DNA and thus encapsulates genetic biometric information. Lip dynamics results from muscular movement and mandibular deformation and therefore also contains genetic information. Similarly, the mouth-region appearance is, to some extent, affected by behavioural factors such as the presence of facial hair or lipstick. Behavioural factors linked to language, emotional condition and socio-economic background also affect lip dynamics. The resulting measurement is akin to the idea of "mouth-signature".

Besides being a twin biometric, there are a variety of other factors that make lip features a compelling biometric for industrial deployment. The advent of increasingly cheaper cameras facilitates the non-intrusive capture of visual speech signals, making it easier than ever before to obtain lip-region features. The use of these features also naturally increases system robustness to attempts at faking *"liveness"*. Lip biometrics can be described as being passive since they do not require active user participation. These factors together suggest that such a biometric has potential for industrial deployment and user acceptance. 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 features from the lip region in order to maintain a sufficient inter-person to intra-person variation ratio for accurate verification?

In [7], we proposed a novel method of texture representation called Local Ordinal Contrast Patterns (LOCP) based on the concept of ordinal contrast. This texture representation was combined with a configuration called Three Orthogonal Planes (TOP). The combination of LOCP and TOP enabled the quantisation of spatiotemporal appearance observed within visual speech. This feature representation was demonstrated to have excellent performance in speaker verification.

In this paper, we make three contributions. First, the LOCP operator presented in [7] is extended to include pattern history. This minimises transitional binary decisions to ensure that the texture pattern does not inset any false texture related to edges when there are none in the real image. Second, the robustness of the LOCP texture descriptor is evaluated with respect to image noise intensity. The purpose of this investigation is to ensure that the proposed texture descriptor is suitable for charactering the texture properties of the lip region. Finally, the effect of the amount of video information on speaker verification performance is also investigated to ensure that it is usable in real-world situations.

The remainder of this paper is structured as follows. A review of the state-of-the art in lip-based speaker verification is

The authors are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, Surrey, GU2 7XH UK. E-mail: {c.chan, w.christmas, j.kittler}@surrey.ac.uk, budgoswami@gmail.com

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 novel feature descriptor. The detailed treatment of the computation of LOCP-TOP features is presented in Section III. An overview of the visual speaker verification systems used to evaluate this novel descriptor is provided in Section IV. The paper concludes with an experimental evaluation in Section V and concluding remarks in Section VI.

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 by investigators such as Fischer and Locard [8]. Lip prints contained information about the individual grooves and eccentricities of the lip surface. The application of lip prints specifically as a biometric trait was first introduced in [9]. Relevant work can be segregated 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.

Static Methods use shape, geometric or appearance features extracted from the lip-region. Additionally, most of these methods either operate on static images or on a visual speech video per-frame [4], [5], [10]–[13].

Dynamic Methods use features related to the changes observed in the mouth-region during speech production. Within these systems, there are two categories. Most deployed biometric systems are based on scenarios with cooperative users speaking fixed string passwords from a small vocabulary. These generally employ what is known as ***text-dependent***(TD) systems [14]. 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 [7], [15]–[17].

Hybrid Methods use a combination of static and dynamic information [6], [18]–[22].

A. Summary of Relevant Work

TABLE I
PERFORMANCE OF LIP BIOMETRIC SYSTEMS FOR SPEAKER VERIFICATION SHOWING SYSTEMS USING **ONLY** LIP-FEATURES

SYSTEM	LIP FEATURE	DB	CLIENTS	PERF.(%)	
FARAJ [15], [16]	DYNAMIC TI	**XM2VTS**	295	EER	22
SANCHEZ [14]	DYNAMIC TD	**XM2VTS**	295	HTER	**13.35**
AUCKENTHALER [6]	STATIC	DAVID	7	% ERROR	2.2
CETINGUL [19]	STATIC (INTENSITY)	MVGL-AVD	50	EER	5.6
CETINGUL [19]	DYNAMIC TD	MVGL-AVD	50	EER	5.2
CETINGUL [18]	STATIC(TEXTURE)	MVGL-AVD	50	EER	1.7
GOMEZ [11]	STATIC(GEOMETRIC)	CUSTOM	50	EER	0.015
JOURLIN [5]	STATIC(SHAPE)	M2VTS	37	HTER	15.4
SAMAD [17]	DYNAMIC TI	AMP CMU	10	HTER	0.0
WARK [22]	DYNAMIC TI	TULIPS1	12	EER	0.0

TABLE II
PERFORMANCE OF MULTI-MODAL BIOMETRIC SYSTEMS FOR SPEAKER VERIFICATION THAT USE LIP FEATURES AS A FUSED MODALITY

SYSTEM	FEATURE FUSION	DB	CLIENTS	PERF.(%)	
BROUN [12]	STATIC(GEOMETRIC) + AUDIO	**XM2VTS**	261	HTER	6.3
FARAJ [15], [16]	DYNAMIC TI + AUDIO	**XM2VTS**	295	EER	2
SANCHEZ [23]	DYNAMIC TD + FACE	**XM2VTS**	295	HTER	2.62
SANCHEZ [23]	DYNAMIC TD + AUDIO	**XM2VTS**	295	HTER	0.70
SANCHEZ [23]	DYNAMIC TD + FACE + AUDIO	**XM2VTS**	295	HTER	0.66
SANCHEZ [23]	DYNAMIC TD + 2FACE + 2AUDIO	**XM2VTS**	295	HTER	0.15
ABDULLA [21]	HYBRID(SHAPE AND INTENSITY)	CUSTOM	35	EER	18.0
CETINGUL [19]	HYBRID(TEXTURE AND MOTION)	MVGL-AVD	50	EER	3.6
CETINGUL [18]	STATIC(TEXTURE)+DYNAMIC+AUDIO	MVGL-AVD	50	EER	0.4
JOURLIN [5]	STATIC(SHAPE) + AUDIO	M2VTS	37	HTER	1.65

Table I provides an overview of the performance of various lip-biometric systems that perform speaker verification using only lip features. Table II provides an overview of the performance of various lip-biometric systems that perform speaker verification using lip features fused with other biometric traits such as audio and face. The performance figures 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 [24].

In order for various visual 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 [25] with established verification protocols that enable a fair comparison of systems. However, because most of these databases are not free, a number of publications in the area of lip-based biometric systems use custom-built datasets and evaluation protocols.

The disadvantage of using custom-built datasets for this task is that, in addition to reducing the comparability of the systems, often the classification task is made easier. This can be seen from Tables I and II where some methods achieve perfect performance (e.g. [17], [22]) on databases with few subjects. This is because success in speaker verification depends on a the ratio of feature dimensions to the number of clients being sufficiently high so as to capture inter-person variation whilst at the same time minimising intra-person variation. 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 are XM2VTS [25] (used by 3 authors) and Lausanne Protocols [26] respectively. The best performance obtained using lip features *only* on this database is by [14](HTER of 13.35%). Multi-modal fusion with two face detectors and two audio systems [23] yields HTER of 0.15%. In our experiments, we use the XM2VTS database to ensure the comparability of our results.

The next sections describe the proposed texture and dynamic texture representations. These together enable the spatiotemporal information within visual speech to be quantised for use in visual speaker verification experiments.

III. Ordinal Contrast Patterns

Ordinal contrast patterns represent image texture by conveying information about the relative differences in the local neighbourhood of a pixel. In computer vision, the absolute information contained within a pixel including intensity, colour and texture can vary dramatically with illumination. 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. Research has shown that it is these mutual ordinal relationships rather than precise image properties that are used by the primary visual cortex [27].

Ordinal contrast encodings are a computationally efficient representation of such neighbourhood relationships and have consequently become popular in information representation. 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 binary decision. The code is efficient to compute and the information entropy of the measure is maximised because the code has nearly equal probability of being 0 or 1 for arbitrary patterns. The authors in [28] explain that the ordinal measure is invariant to any monotonic transformation of the gray scale. As long as the order of pixel values stays the same, the output of any ordinal contrast measurement does not change. This makes them an attractive feature descriptor where the gray scale is subject to changes due to, for instance, varying illumination conditions. Additionally, their computational efficiency presents an advantage in time-critical applications.

Local Binary Pattern (LBP) [29] is an example of an ordinal measure. It offers a powerful texture descriptor showing excellent results in terms of representation accuracy and computational complexity in many empirical studies. The LBP operator encodes the ordinal contrast pairs between a local neighbour value and the centre pixel value as a binary result. The pattern is obtained by concatenating these binary results into a bit string. The pattern value in the resulting LBP image consists of the above bit string. Given its advantages as a texture descriptor, we use LBP as a benchmark texture descriptor for visual speech.

The application of any texture descriptor to represent the lip surface for use in biometrics requires the consideration of two factors. Firstly, biometric applications involve visual speech signals that are particularly affected by varying illumination [30], [31]. As a result of this variation, amplifier noise from the imaging system may affect the signal quality of observed visual speech. Amplifier noise can be modelled as being additive, Gaussian, independent and identically distributed. The primary effect of such noise is to degrade the quality of texture description particularly in uniform or near-uniform intensity regions e.g. the cheeks and skin area around the lip. Random noise could result in a situation where the reference value, i.e. the centre pixel of an LBP, changes by a single unit, thereby altering all 8 neighbourhood ordinal contrast measurements. This results in LBP misrepresenting local structure. Various methods have been proposed to enhance the robustness of the LBP operator to such an effect. [31] have proposed Local Ternary Patterns (LTP), which extend LBP by increasing the feature dimensionality depending on the sign of the centre bit. However, LTP is sensitive to monotonic transformation. The authors in [30] proposed Improved LBP which performs ordinal contrast measurement with respect to the average of the pixel neighbourhood instead of the centre pixel to reduce the effect of a single noisy reference.

The second consideration in using a texture descriptor to parametrise the lip as a biometric is the lip surface itself. The first published use of lip texture in the field of person identification can be found in [9]. They used *lip prints* as a means of identification. According to the authors in [9], *lip prints* are *"the normal lines and fissures in the form of wrinkles and grooves present in the zone of transition of the human lip, between the inner labial mucosa and the outer skin"*. These wrinkles and grooves manifest themselves in an image as lines with varying orientations. Consequently, a desirable property for a lip texture descriptor is sensitivity to line orientation variation across the lip surface.

A. Local Ordinal Contrast Patterns

In [7], we propose a novel texture descriptor called Local Ordinal Contrast Pattern (LOCP) which attempts to fulfil both requirements. LOCP does this by diversifying the source of reference values. Instead of computing the contrast with respect to a fixed reference, LOCP uses pairwise ordinal contrast measurement of pixels from a circular neighbourhood starting at the centre pixel. In terms of information representation, LBP suggests that the ordinal relationship between a single reference pixel and its neighbourhood contains texture information. LOCP on the contrary, suggests a new paradigm where texture is represented by pairwise ordinal relationships of the entire neighbourhood. LOCP thus improves on LBP since a change in the value of a single pixel affects at most 2 ordinal contrast encodings.

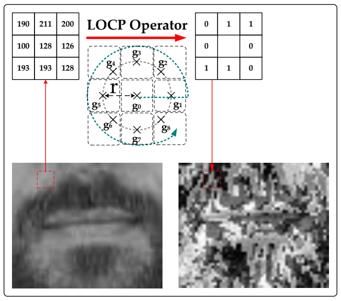


Fig. 1. LOCP Feature Computation: Compute pairwise ordinal contrast measure along the direction of the overlaid green arrow. r is the radius of the operator.

Additionally, linearly interpolating the pixel values allows the choice of any radius, r and any number of pixels in the circular neighbourhood, p, to form LOCP. Varying r therefore enables the modelling of arbitrarily large scale structures. During the operation of LOCP, we choose p pixel pairs for ordinal contrast encoding defined in Equation 1. The pixel indices are shown in Figure 1. At any location $\boldsymbol{x}$ in the LOCP image $\mathcal{I}$, the pattern is calculated as:

$$\mathcal{I}^{p,r}[\boldsymbol{x}] = \sum_{i=0}^{p-1} s(g_{i+1} - g_i)\, 2^i$$

$$\text{where } s(v_i) = \begin{cases} 1 & v_i > 0 \\ 0 & v_i < 0 \\ 0 & v_i = 0 \text{ and } i = 0 \\ s(v_{i-1}) & v_i = 0 \text{ and } i > 0, \end{cases} \qquad (1)$$

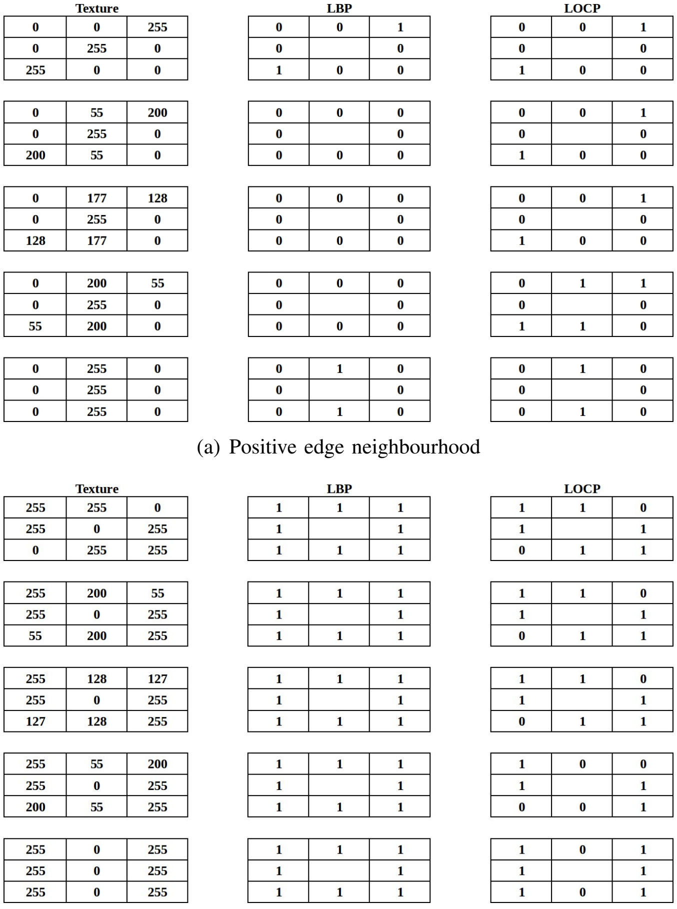
and g_i is the intensity value of the i^{th} pixel at a distance r away from the current location $\boldsymbol{x}$ and $0 \leq i < p$. The pattern is obtained by concatenating the binary numbers from the encoding into a p-bit code.

The above LOCP operator also includes some notion of pattern history to deal with instances where the pair of pixels have the same value. In this case, the operator uses the immediately previous encoding as the current ordinal contrast measurement. This spatial history enables the transitional binary decisions to be minimised in case of no ordinal contrast measurement and is an extension to the operator initially presented in [7][1]. The effect of incorporating this spatial history is to ensure that the texture pattern does not insert any false texture related edges when there are none in the real image. Note that from hereon, the operator originally presented in [7] will be referred to as LOCP$_o$.

Figure 2 illustrates the line orientation sensitivity of LOCP by comparing the ordinal contrast encoding of a gradual shift in orientation of diagonal line in a pixel neighbourhood with LBP. Unlike LBP, LOCP still manages to conserve the local neighbourhood pixel structure for every intermediate change of orientation. This suggests that it may be more appropriate for use as a lip-texture descriptor.

It is also illustrative to observe the distribution of patterns from all possible permutations of a given pixel neighbourhood. Consider a 3×3 pixel neighbourhood containing gray level information at each pixel. Such a pixel neighbourhood results from setting $p = 8$ and $r = 1$ for instance. There are $2^9 = 512$ possible permutations of binary relationships within this neighbourhood. LOCP and LBP differ in how they distribute the $2^p = 2^8 = 256$ possible pixel labels (p-bit codes) amongst these permutations. This distribution is shown by Figure 3 which demonstrates that the proposed LOCP encoding naturally lends itself to a more uniform distribution of ordinal contrast patterns. Unlike LOCP, LBP groups half of the 512 patterns into a single bin. Given this distribution of ordinal contrast encodings, the key question is whether LOCP encodes sufficient discriminative information about an individual's lip texture to be useful as a primary biometric trait.

The LOCP texture descriptor represents local, pairwise neighbourhood derivatives. The use of LOCP enables the encapsulation of compact, local structure within this descriptor. Additionally, LOCP has the same computational complexity as LBP since the number of ordinal contrast comparisons per pixel is unchanged. While LOCP is useful as a texture descriptor, its application to the parametrisation of spatiotemporal information requires it to be combined with a method of dynamic texture representation.



Texture			LBP			LOCP		
0	0	255	0	0	1	0	0	1
0	255	0	0		0	0		0
255	0	0	1	0	0	1	0	0
0	55	200	0	0	0	0	0	1
0	255	0	0		0	0		0
200	55	0	0	0	0	1	0	0
0	177	128	0	0	0	0	0	1
0	255	0	0		0	0		0
128	177	0	0	0	0	1	0	0
0	200	55	0	0	0	0	1	1
0	255	0	0		0	0		0
55	200	0	0	0	0	1	1	0
0	255	0	0	1	0	0	1	0
0	255	0	0		0	0		0
0	255	0	0	1	0	0	1	0

(a) Positive edge neighbourhood

Texture			LBP			LOCP		
255	255	0	1	1	1	1	1	0
255	0	255	1		1	1		1
0	255	255	1	1	1	0	1	1
255	200	55	1	1	1	1	1	0
255	0	255	1		1	1		1
55	200	255	1	1	1	0	1	1
255	128	127	1	1	1	1	1	0
255	0	255	1		1	1		1
127	128	255	1	1	1	0	1	1
255	55	200	1	1	1	1	0	0
255	0	255	1		1	1		1
200	55	255	1	1	1	0	0	1
255	0	255	1	1	1	1	0	1
255	0	255	1		1	1		1
255	0	255	1	1	1	1	0	1

(b) Negative edge neighbourhood

Fig. 2. LBP Vs LOCP Feature Descriptor, showing insensitivity of LBP to line orientations in both a positive and negative edge neighbourhood. Notice how the LBP does not encode intermediate orientations.

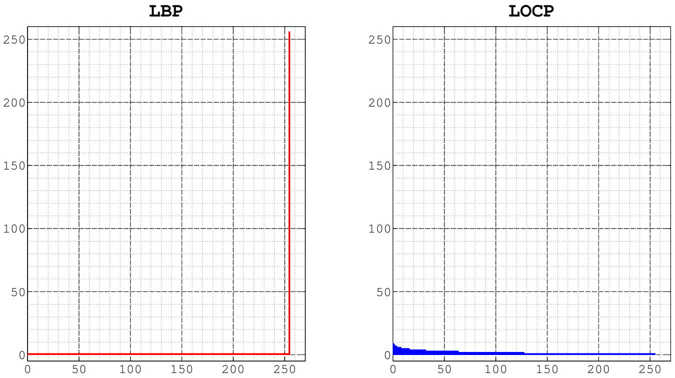


Fig. 3. LBP Vs LOCP Feature Descriptor, showing pattern distribution in a 3×3 pixel neighbourhood

B. Three Orthogonal Planes

Recently, the use of local binary patterns on three orthogonal planes (LBP-TOP) [32] has been proposed to extend the LBP to a spatiotemporal representation for dynamic texture analysis. LBP-TOP extracts the LBP in three orthonormal planes within a spatiotemporal volume. Motivated by [32], we extended [7] our new operator for dynamic texture analysis by extracting the LOCP in three orthogonal planes (i.e. XY, XT and YT) within a volume. Figure 4 demonstrates the lip images from three planes.

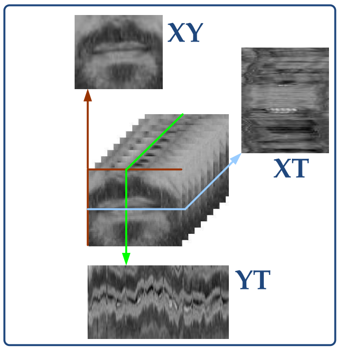


Fig. 4. Extraction of images using TOP. (a) XY Image (b) YT Image (c) XT Image

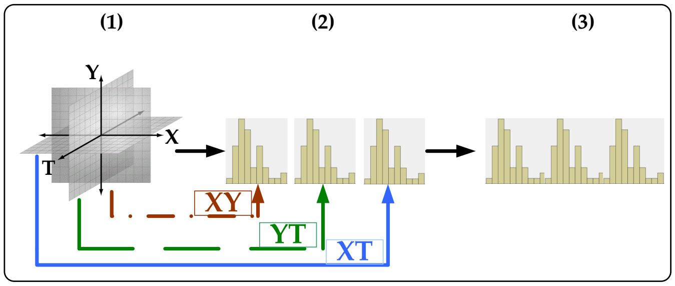


Fig. 5. LOCP-TOP Feature Description:(a) Represents feature parametrisation along TOP planes using LOCP operators (b) Represents the histogram of the LOCP features from each TOP plane (c) Represents the concatenation of these histograms for use in dynamic texture analysis

In each plane, for each speaker video (with the mouth region detected), the LOCP image, $\mathcal{I}^{p,r,\beta}$, is extracted and the corresponding plane-pattern histogram, $\boldsymbol{h}^{\beta}_{p,r} \in \mathbb{R}^{2^p}$ is computed using the function $h^{\beta}(i), i \in [0, 2^p)$ where $\beta \in \{XY, XT, YT\}$ represents a plane:

$$h^{\beta}_{p,r}(i) = \sum_{(x',y') \in \boldsymbol{M_\beta}} Bool(\mathcal{I}^{p,r,\beta}(x',y') = i) \tag{2}$$

where $i \in [0, 2^p)$ is the value of the LOCP, $\boldsymbol{M_\beta}$ is the region in the plane for which we are computing the histogram and the function $Bool()$ represents a boolean operator:

$$Bool(\gamma) = \begin{cases} 1 & \text{when } \gamma \text{ is true} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Then the histogram of each plane is concatenated into one single histogram, $\boldsymbol{b}^{\alpha}$ shown in Figure 5 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 4.

$$\boldsymbol{b}^{XYXTYT} = [\boldsymbol{h}^{XY}_{p,r}, \boldsymbol{h}^{XT}_{p,r}, \boldsymbol{h}^{YT}_{p,r}] \tag{4}$$

An 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. Visual Speaker Verification

For this system, the method in [23] 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 in each frame. These extracted regions were, in turn, parameterised using LOCP-TOP. Each extracted region can be visualised as a cuboid containing spatiotemporal information. This cuboid is first subdivided into overlapping sub-cuboids. For each sub-cuboid, we use LOCP-TOP to extract histograms $\boldsymbol{h}^{\beta,j}_{p,r}$ where j represents the sub-cuboid index. These are then further concatenated to form $\boldsymbol{b}^{\alpha,j}$. These combined histograms conceptually represent the intra-modal feature-level fusion of extracted LOCPs in the different planes. The concatenated histograms are then input into one of two classification engines as described below. Each classifier use the nearest neighbour principle with a different distance metric.

Chi-squared Histogram Distance (X2): In order to measure the similarity between two input LOCP-TOP histograms resulting from a probe and an enrolled gallery video, we use a similarity measure $Sim_{\chi^2}(\boldsymbol{G}, \boldsymbol{I})$ based on Chi-squared distance between the histograms (with bin index i) of two input videos $\boldsymbol{G}$ and $\boldsymbol{I}$.

$$Sim_{\chi^2}(\boldsymbol{G}, \boldsymbol{I}) = -\sum_j \sum_i \frac{(\boldsymbol{b}^{\alpha,j}_G(i) - \boldsymbol{b}^{\alpha,j}_I(i))^2}{\boldsymbol{b}^{\alpha,j}_G(i) + \boldsymbol{b}^{\alpha,j}_I(i)} \tag{5}$$

Normalised Correlation (NC): In order to extract the discriminative features we project the sub-cuboid histograms, $\boldsymbol{b}^{\alpha,j}$, into LDA space as: $\boldsymbol{d}^{\alpha,j} = (\boldsymbol{W}^{\alpha,j}_{lda})^T \boldsymbol{b}^{\alpha,j}$. After projection, we perform normalized cross-correlation across all sub-cuboids using two videos $\boldsymbol{G}$ and $\boldsymbol{I}$ as specified in Equation 6.

$$Sim_{NC}(\boldsymbol{G}, \boldsymbol{I}) = \sum_j \frac{(\boldsymbol{d}^{\alpha,j}_G)^T \boldsymbol{d}^{\alpha,j}_I}{\|\boldsymbol{d}^{\alpha,j}_G\| \|\boldsymbol{d}^{\alpha,j}_I\|} \tag{6}$$

V. Evaluation and Discussion

A. Experimental Setup

The experimental evaluation uses the XM2VTS database [25] as mentioned in Section II. The XM2VTS 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 videos of 295 subjects divided into training, development and test sets. For these experiments, we followed the Configuration 1 (C1) and Configuration 2 (C2) of the Lausanne protocol [26] that accompanies this database for speaker verification. This leads to the following test statistics:

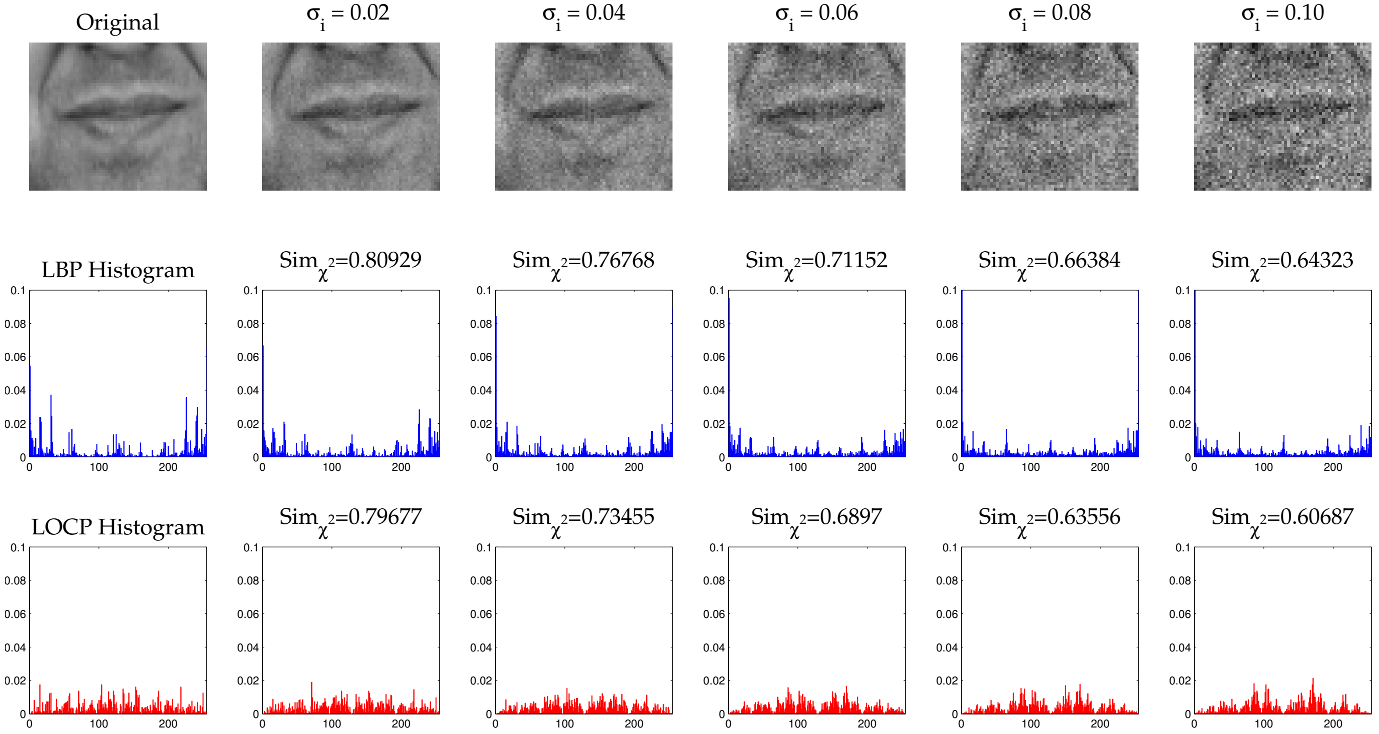


Fig. 6. Texture histogram similarity variation with increasing noise intensity. σ_i denotes the noise intensity and Sim_{χ^2} denotes Chi-squared histogram similarity.

- Client training examples: 3 per client in Configuration I, 4 per client in Configuration II
- Development samples (clients): 600 in Configuration I and 400 in Configuration II
- Development samples (impostors): 40000 $(25 \times 4 \times 2 \times 200)$
- Test client accesses: $400 (200 \times 2)$
- Test impostor accesses: 1120000 $(70 \times 4 \times 2 \times 200)$

The mouth-region localisation for the XM2VTS database was set to be 61×51 pixels. LOCP feature parameters p and r were set to 8 and 3 respectively [32]. Additionally, they were set to be the same for all planar configurations. Each spatiotemporal video cuboid was subdivided into 5 sub-cuboids along the XY direction and 3 sub-cuboids along the T axis. Each of these sub-cuboids overlapped each other by 70%. The reason for this overlap was to enhance the robustness on misalignment in spatial and temporal domains.

Three sets of experiments were performed to evaluate:

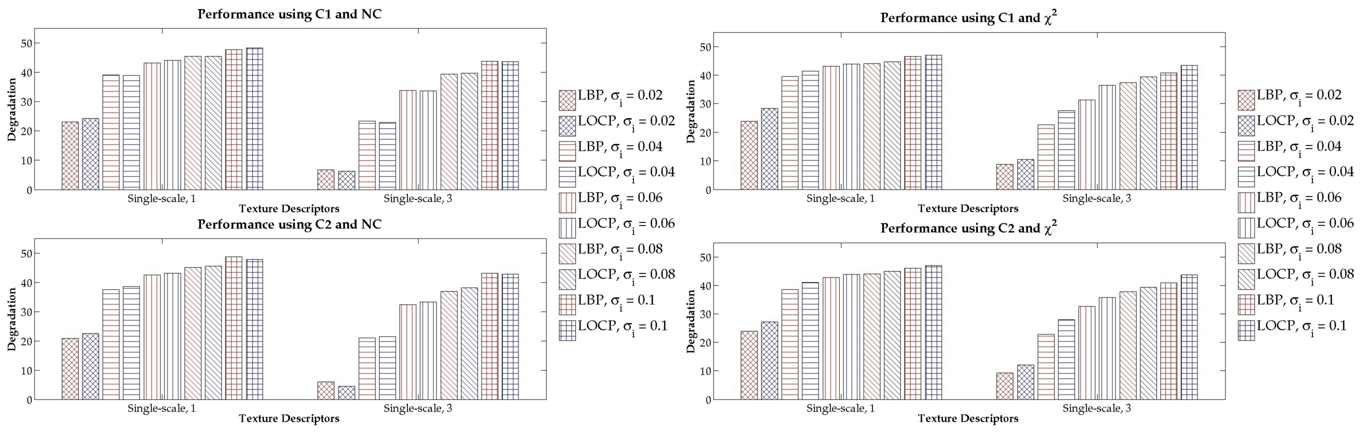
1) The performance of LOCP vs LBP as a texture descriptor by measuring resilience to camera noise (Section V-B)
2) The effect of visual speech video length on verification performance (Section V-C)
3) The performance of the LOCP-TOP lip features in a speaker verification experiment on the XM2VTS database using the Lausanne protocol (Section V-D)

The experiments in Sections V-B and V-C aim to provide a comparative evaluation of the optimum results for each of the proposed feature parameterisations. Equal Error Rate (EER) can be obtained after a full authentication experiment has been performed and represents the performance at which the number of False Accepts is equal to the number of False Rejects. The results of Sections V-B and V-C are therefore reported using EER.

Section V-D presents a comparative evaluation of the proposed feature parameterisation within a verification experiment. In order to facilitate a fair comparison with the state-of-the-art systems reported in Table I, we follow the Lausanne protocol and report the Half Total Error Rate (HTER). Additionally, [25] also reports that EER does not capture a real authentication scenario and might not predict the expected system performance well, given an unseen test set. To accommodate this shortcoming, the HTER is reported using the similarity threshold value corresponding to the EER obtained on the evaluation set. In addition to comparing against the state-of-the-art, the results of Section V-D are used to compare the LBP texture descriptor and the LOCP descriptors proposed in this paper and originally proposed in [7]. To validate the contribution of the various texture descriptors , we perform the McNemar's test as described on pages 12–15 in [33]. McNemar's test computes $M(S_1, S_2)$ where M is the McNemar's test statistic (incidentally, distributed as Chi-squared with one degree of freedom) and S_1 and S_2 are the two systems in being compared. As suggested in [33], if $M > 3.841$, this indicates a level of significance 0.05 and we accept that systems S_1 and S_2 have significantly different HTER.

B. Evaluation of the robustness of LOCP to intensity noise

Resilience to illumination variation is an important consideration when choosing a texture descriptor to parameterise the



(a) NC (b) X2

Fig. 7. LOCP Vs LBP Feature Descriptor, showing degradation in EER(%) of the test set when increasing Gaussian noise, σ_i

lip region. The effect of illumination variation is that uniform or near uniform regions in an image may be affected by noise from the digital imaging systems. This noise can be modelled as an additive, Gaussian, independent and identically distributed (i.i.d.) random process.

This experiment was designed to compare the performance of the LOCP to LBP in a simulated environment where the variance of the Gaussian i.i.d. process, σ_i was systematically increased: $\sigma_i = \{0.02, 0.04, 0.06, 0.08, 0.1\}$ to model the effect of worsening illumination conditions. In order to compare the performance, a still lip image verification experiment was designed using only the first frame of each visual speech video in the XM2VTS database. In other words, only the first XY image from the visual speech signal was used to compute texture features for speaker authentication. This data was chosen because accurate lip localisation information was available. The NC and X2 classifiers were used to compare the performance degradation of the texture descriptors. The original images were used to form the training and evaluation sets required by the Lausanne protocol. The degradation measurement used the test set EER and was computed as:

$$\frac{EER_{noisy,\sigma_w} - EER_{clean}}{EER_{clean}}.$$

Two configurations of the texture descriptors were used in this experiment:

- LBP / LOCP with full image (denoted by Single-scale 1 in the X-labels in Figure 6)
- LBP / LOCP with 3 Y sub-cuboid images overlapping each other by 70% (denoted by Single-scale 3 in the X-labels in Figure 6). The sub-cuboid partitioning along the Y axis enables greater image resolution.

The values of the test-set EER (%) obtained with the original, clean images are shown in Table III.

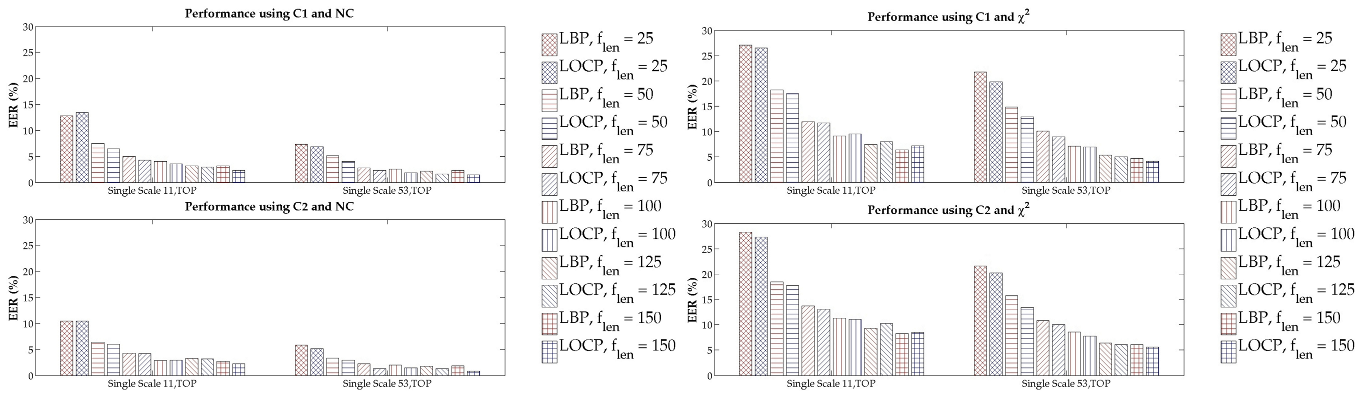
The bar charts in Figure 7(a) and 7(b) show the degradation in EER(%) of the test set when increasing Gaussian noise, σ_i. The LBP and LOCP bars are placed beside each other for easy comparison. In these figures, a higher value of degradation indicates worse resilience to increasing image noise. Figure 6 shows the variation of histogram similarity (Chi-squared histogram distance) with respect to increasing Gaussian noise using both the LBP and LOCP feature descriptors. Figure 7 illustrates the variation in accuracy of the systems (EER(%) on the test set). In this figure, a lower accuracy line is indicative of a better performing texture descriptor.

The results of this experiment demonstrate that in terms of texture representation in the original histogram space, measured using the X2 classifier, LOCP performs worse than LBP since both the degradation as well as system accuracy is worse. Consequently, this suggests that it is less robust to image noise than LBP. However, its performance in LDA space demonstrates that it is a more discriminative texture descriptor despite its poorer resilience to image noise variation. This is a seemingly contradictory phenomenon; we would expect discriminative performance to be proportional to intensity noise robustness. It is illustrative to consider the example histogram representations presented in Figure 6 to explain the above contradiction.

Ordinal contrast measures represent texture by attributing labels to observed image micro-structure. For a 3×3 pixel neighbourhood, we expect to encounter 512 different ordered combinations of binary micro-structure. LBP and LOCP are different in how they distribute the 256 available labels amongst the total number of observable micro-structure patterns. This is illustrated by Figure 6. As can be seen in Figure 6, LOCP distributes the labels more evenly than LBP. This is because the effect of any pixel change in the image is absorbed by the entire pixel neighbourhood in LOCP. It can also be seen that an increase in intensity noise causes a disproportionate increase in the bin counts of a select number of labels in the LBP histogram. This effect is reduced in the LOCP histogram.

TABLE III
TEXTURE DESCRIPTOR PERFORMANCE ($EER_{clean}(\%)$) IN RECOGNITION EXPERIMENT

Protocol	System	LBP,1	LBP,3	LOCP,1	LOCP,3
C1	NC	7.82	2.01	6.78	2.25
C2	NC	7.26	1.99	7.02	1.26
C1	X2	8.33	4.02	8.03	4.02
C2	X2	9.08	4.75	9.21	3.46



(a) NC

(b) X2

Fig. 8. Test-set EER(%) variation against increasing frame length f_{len}

The X2 classifier measures histogram distance by computing per-label (bin) differences. Since LOCP makes more even use of the available bin-space, there are more discrepancies between histograms resulting in a poorer degradation compared to LBP. However, this same effect proves useful when the NC classifier is employed. NC operates in a dimension-reduced LDA space. Dimensionality reduction methods select a subset of the bins from the histogram. Since LOCP provides a more uniform distribution of labels in the histogram, the effect of any major stimulus to the image content (in this case due to image noise) is tempered by the histogram since all bins contribute in absorbing that stimulus. This results in better discriminative performance in lower dimensional space. Consequently, LOCP can be considered a more effective ordinal descriptor.

C. Evaluation of the effect of the volume of video information on speaker verification accuracy

In this experiment, we quantify the effect of reducing the amount of video information on the accuracy of the proposed lip biometric systems. This analysis is particularly relevant from the point of view of the commercial application potential of the lip modality. The XM2VTS data was used for this experiment. The highest, lowest and average frame lengths of the XM2VTS videos were $673, 167$ and 319 frames respectively. Given a frame-rate of 25 frames-per-second, the lowest frame length represented just over 6 seconds of video. This experiment was run by cropping the amount of video information supplied to the lip-based speaker verification systems in steps of 25 frames starting from the first frame of video. Note that both the gallery and probe videos were cropped to the same length per experiment. The performance improvement was then measured by increasing the amount of video information by 25 frames up to 150 frames i.e. 6 seconds of video. The results of the system accuracy measurements (test set EER (%)) are shown in Figure 8. For both LBP and LOCP texture descriptors, the following systems were used:

- S11, TOP : single-scale texture descriptor with no sub-cuboid partitioning using TOP
- S53, TOP : single-scale texture descriptor with $5 \times Y$ and $3 \times T$ sub-cuboid partitioning (with 70% overlap) using TOP

The results demonstrate that increasing the amount of video information improves system accuracy as expected. They also demonstrate that LOCP outperforms LBP in terms of resilience to shorter talking face records. Additionally, increasing video resolution by sub-cuboid partitioning improves performance. The NC system reaches a steady state more quickly implying that it extracts discriminative information quicker. This also suggests that the discriminative information contained in the lip biometric can be extracted with only a few samples of visual speech. With about 3 seconds of video information in this case, HTER values of around 1% were obtained with the NC system. This finding corroborates the claim that the lip biometric can be used as a primary biometric trait not only in terms of performance but also in terms of the amount of information required for accurate verification.

D. Evaluation of the LOCP-TOP descriptor for speaker authentication

Tables IV and V show the HTER of the test-set and the EER of the evaluation-set of the various LOCP-TOP histograms with the chi-squared and LDA verification systems respectively. The best performances (highlighted in bold) with LOCP-TOP were obtained using XYYT histograms with the chi-squared system for C1 and the XYXTYT histograms with the LDA system for C2. Note that in each experiment the best performance was chosen based on the figures of the lowest EER in the evaluation test. The comparative results using LBP-TOP and LOCP$_o$-TOP[2]are also shown.

Tables VI and VII show the results of computing McNemar's test. These tables are colour-coded to facilitate comprehension. Each cell in the table displays the result of $M(S_1,S_2)$, where S_1 and S_2 are the two systems being compared. If the result of the McNemar's test is not significant, i.e. $M(S_1,S_2) \leq 3.841$, then the corresponding text is in black. If the HTER of S_2 is lower than that of S_1 and $M(S_1,S_2) > 3.841$, then the text colour is in bold blue. This indicates that S_2 is statistically significantly more

[2] Note that LOCP$_o$ refers to the LOCP implementation originally proposed in [7].

TABLE IV
FEATURE PERFORMANCE OF THE CHI-SQUARED SYSTEM ON XM2VTS SHOWING EER (EVAL) AND HTER (TEST)

	Configuration 1						Configuration 2					
TOP Input	**LBP**		**LOCP$_o$**		**LOCP**		**LBP**		**LOCP$_o$**		**LOCP**	
	Eval	**Test**	**Eval**	**Test**	**Eval**	**Test**	**Eval**	**Test**	**Eval**	**Test**	**Eval**	**Test**
XY	5.30	4.09	3.70	3.70	4.19	3.75	5.45	4.56	4.25	4.27	5.49	4.14
XT	16.83	17.44	18.33	19.85	16.85	18.25	17.94	18.61	19.73	19.75	17.78	18.86
YT	8.31	8.98	9.05	10.41	9.07	10.63	10.19	9.29	11.55	10.73	11.59	10.76
XYXT	3.68	3.61	3.46	3.75	3.28	3.77	4.80	4.25	4.72	4.38	4.48	4.35
XYYT	**2.97**	**3.02**	**2.71**	**2.79**	**2.74**	**3.02**	**4.01**	**3.47**	**2.98**	**3.31**	**4.17**	**3.68**
XTYT	9.77	10.47	11.70	13.14	10.09	11.35	10.66	9.85	13.17	13.62	12.19	12.00
XYXTYT	3.18	3.52	3.17	3.90	2.99	3.86	4.46	3.61	4.26	4.43	4.27	3.97

TABLE V
FEATURE PERFORMANCE OF THE NORMALISED CORRELATION SYSTEM ON XM2VTS SHOWING EER (EVAL) AND HTER (TEST)

	Configuration 1						Configuration 2					
TOP Input	**LBP**		**LOCP$_o$**		**LOCP**		**LBP**		**LOCP$_o$**		**LOCP**	
	Eval	**Test**	**Eval**	**Test**	**Eval**	**Test**	**Eval**	**Test**	**Eval**	**Test**	**Eval**	**Test**
XY	2.50	1.84	1.16	1.04	1.16	0.94	2.49	2.08	1.28	1.29	1.50	1.24
XT	7.32	10.10	7.97	8.59	7.36	9.10	8.68	10.76	9.06	10.19	8.58	9.89
YT	3.36	4.67	2.80	5.03	2.84	4.64	4.00	5.38	4.13	5.38	4.21	4.90
XYXT	1.50	1.26	0.50	0.84	0.54	0.83	2.13	1.91	1.29	1.22	1.46	0.87
XYYT	1.14	1.90	0.51	0.82	0.53	0.36	1.51	1.70	0.98	0.99	**0.77**	**0.49**
XTYT	2.67	3.82	2.01	3.56	1.96	3.51	3.01	3.57	2.52	4.22	2.75	4.11
XYXTYT	**0.87**	**1.29**	**0.33**	**0.65**	**0.25**	**0.36**	**1.50**	**1.67**	**0.76**	**0.95**	0.99	0.49

TABLE VI
MCNEMAR'S TEST ON APPEARANCE FEATURE PERFORMANCE USING THE CHI-SQUARED SYSTEMS ON XM2VTS TEST SET

	Configuration 1			**Configuration 2**		
TOP Input	M(LBP , **LOCP$_o$**)	M(LBP , **LOCP**)	M(LOCP$_o$, **LOCP**)	M(LBP , **LOCP$_o$**)	M(LBP , **LOCP**)	M(LOCP$_o$, **LOCP**)
XY	**244.58**	**113.40**	49.61	**141.68**	**36.56**	**503.92**
XT	80.33	3.09	**83.05**	224.14	13.00	**208.60**
YT	309.35	208.70	41.92	542.49	767.47	0.00
XYXT	19.74	18.51	0.34	0.39	0.42	2.56
XYYT	2.2	0.00	3.29	**232.66**	11.60	500.33
XTYT	679.55	24.37	**661.94**	993.61	537.22	**209.29**
XYXTYT	0.44	3.24	8.63	1.75	0.06	1.84

TABLE VII
MCNEMAR'S TEST ON APPEARANCE FEATURE PERFORMANCE USING THE NORMALISED CORRELATION (WITH LDA) SYSTEMS ON XM2VTS TEST SET

	Configuration 1			**Configuration 2**		
TOP Input	M(LBP , **LOCP$_o$**)	M(LBP , **LOCP**)	M(LOCP$_o$, **LOCP**)	M(LBP , **LOCP$_o$**)	M(LBP , **LOCP**)	M(LOCP$_o$, **LOCP**)
XY	**404.7**	**685.62**	46.05	**374.76**	**553.1**	11.12
XT	**43.17**	0.00	60.95	77.18	**49.96**	**145.11**
YT	19.95	**39.00**	1.15	**150.52**	**222.32**	**91.64**
XYXT	**428.6**	**462.84**	0.47	**117.38**	**323.79**	**61.64**
XYYT	**32.24**	**193.14**	**69.02**	**4.09**	**561.57**	**635.15**
XTYT	**31.64**	**216.19**	109.12	2.19	19.37	**44.63**
XYXTYT	**140.87**	**286.56**	**44.41**	**210.67**	**462.46**	**53.4**

accurate than S_1. If the HTER of S_1 is lower than that of S_2 and $M(S_1,S_2) > 3.841$, then the text colour is in red. This indicates that S_1 is statistically significantly more accurate than S_2. Please note that the HTER can be found in Tables IV and V.

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, since LDA performs subspace projection of the histograms into a discriminative space. Chi-squared distance is applied to the LOCP-TOP histograms directly in an unsupervised manner.

It is also interesting to compare the performance of the various TOP planes. XY outperforms XT and YT, indicating that spatial appearance is a more discriminative feature than temporal mouth-region variation. Another interpretation of this result is that the genetic aspect of the lip biometric is a stronger trait than the behavioural aspect. YT outperforms XT in all systems. YT represents mouth opening during speech while XT represents mouth widening. The fact that YT outperforms XT indicates that mouth opening is a more discriminative feature than mouth-widening. The results also show that the feature-level fusion of spatial and temporal information contained in the XYXTYT feature consistently leads to the best performance when used with template matching. This ties in well with the intuition behind the use of lip as a biometric i.e. it being a twin biometric containing both genetic (in this case characterised by the spatial appearance) as well as behavioural (characterised by XT and YT) information.

The results also enable the comparison between LOCP, LOCP$_o$ and LBP which belong to the same family of ordinal contrast measures. The results of Table VI demonstrate that the average performance in the X2 system, over TOP inputs, of LBP is slightly better than both LOCP and LOCP$_o$. However, no conclusive comparison can be drawn between LOCP and LOCP$_o$. Additionally, no conclusive performance improvement can be drawn by considering *only* the best performing TOP plane in the X2 system i.e. XYYT as well as the performance of the XYXTYT plane. The results of using the NC system in Table VII demonstrate that both LOCP and LOCP$_o$ perform statistically significantly more accurately than LBP over most TOP inputs in both configurations. Additionally, of the 14 feature parameterisations used to compare LOCP and LOCP$_o$, LOCP is statistically significantly more accurate over 8 planes as compared to 4 for LOCP$_o$. Thus we suggest that in LDA space, the enhancement of LOCP$_o$ with spatial history results in a performance improvement.

A final point to note is that this experiment investigates how our system compares with the state-of-the-art benchmarks. The best baseline performance obtained using lip features *only* on this database was by [14](HTER of 13.35%) as shown in Table I. Multi-modal fusion with two face detectors and two audio systems [23] yielded HTER of 0.15% as shown in Table II. Our experiments result in HTER of 0.36% using C1 and 0.99& using C2. In both these cases, an improvement of two orders of magnitude can be observed compared to the performance of state of the art systems using lip-features alone. Furthermore, the performance is almost comparable to that obtained using multi-modal fusion. The obtained results strongly challenge the commonly held perception of the lip being a soft biometric. They also suggest that the proposed feature representation method i.e. LOCP-TOP is well suited to extracting discriminative genetic and behavioural information from visual speech.

VI. Conclusions and Future Work

We have proposed a novel ordinal contrast measure called LOCP. This has been used in a TOP configuration to represent video of the mouth region of a talking face as input into a speaker verification system in the form of LOCP histogram. The verification was accomplished using chi-squared histogram distance or LDA classifiers. The resulting biometric systems have been used to evaluate the discriminatory content of the mouth-region biometrics on the XM2VTS database using the standard Lausanne protocols. The application of this novel feature representation has been demonstrated comprehensively to outperform previous feature descriptors encountered in the state-of-the-art review presented in the paper. The findings confirm that there is sufficient discriminative information within the spatiotemporal evolution of the mouth-region appearance during speech production for use as a primary biometric trait. This can be especially useful in circumstances where auditory information may not be available for fusion. The proposed LOCP histograms are also computationally simpler compared to the more exotic feature parametrisations encountered in the literature.

Several interesting research directions arise as possible avenues for future research as part of this work. The presented LOCP feature descriptor is a novel ordinal contrast encoding. Consequently, the utility of its application purely as a texture descriptor warrants investigation. The combination of LOCP-TOP is a method for spatiotemporal feature quantisation that could also be used in other application areas such as talking face generation, automatic speech recognition and tele-presence rendering to name a few. Feature fusion with alternative lip-based features would also be of interest especially if they result in increased robustness to degraded video capture conditions. Since image based information can encode more information about identity than simply the vocal-tract based models used in audio-based speaker recognition, it is a very real possibility for lip-based biometric systems to be a complementary modality to audio. Additionally, it is of interest to investigate the performance gains that the combination of lip, speech and face would bring in increasing the robustness of a video-based speaker authentication.

Acknowledgment

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

References

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[2] E. D. Petajan, N. M. Brooke, B. J. Bischoff, and D. A. Bodoff, "Experiments in automatic visual speech recognition," in *Proc. 7th FASE Symp.*, 1988, pp. 1163 – 1170.

[3] M. E. Hennecke, K. V. Prasad, and D. G. Stork, "Using deformable templates to infer visual speech dynamics," in *Proceedings of 28th Asilomar Conference on Signals Systems and Computers*, 1994, pp. 578 – 582.

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

[5] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, "Acoustic labial speaker verification," in *AVBPA*, 1997, pp. 319–334.

[6] R. Auckenthaler, J. Brand, J. Mason, C. Chibelushi, and F. Deravi, "Lip signatures for automatic person recognition," in *MMSP*, 1999, pp. 457 – 462.

[7] B. Goswami, C. Chan, J. Kittler, and W. Christmas, "Local ordinal contrast patterns for spatiotemporal, lip-based speaker authentication," in *BTAS*, sept. 2010, pp. 1 –6.

[8] J. Kasprazak, "Possibilities of cheiloscopy," *Forensic Science International*, vol. 46, pp. 145–151, 1990.

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

[10] H. Çetingül, Y. Yemez, E. Erzin, and A. Tekalp, "The use of lip motion for biometric speaker identification," in *SIU*, 2004, pp. 148 – 151.

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

[12] C. Broun, X. Zhang, R. Mersereau, and M. Clements, "Automatic speechreading with application to speaker verification," in *ICASSP*, vol. 1, 2002, pp. 685 – 688.

[13] M. Choraś, "Human lips as emerging biometrics modality," in *ICIAR*, 2008, pp. 993 – 1002.

[14] M. Sánchez and J. Kittler, "Fusion of talking face biometric modalities for personal identity verification," in *ICASSP*, vol. 5, 2006, pp. 1073 – 1076.

[15] M. I. Faraj and J. Bigün, "Person verification by lip-motion," in *CWPRW*, 2006, pp. 37–44.

[16] M. I. Faraj and J. Bigün, "Motion features from lip movement for person authentication," in *ICPR*, 2006, pp. 1059–1062.

[17] S. Samad, D. A. Ramli, and A. Hussain, "Lower face verification centered on lips using correlation filters," *Information Technology Journal*, vol. 6, no. 8, pp. 1146–1151, 2007.

[18] H. Çetingül, Y. Yemez, E. Erzin, and A. Tekalp, "Multimodal speaker/speech recognition using lip motion, lip texture and audio," *Signal Process.*, vol. 86, no. 12, pp. 3549–3558, 2006.

[19] H. Çetingül, E. Erzin, Y. Yemez, and A. Tekalp, "Discriminative analysis of lip motion features for speaker identification and speech-reading," *Image Processing, IEEE Trans.*, vol. 15, no. 10, pp. 2879–2891, 2006.

[20] S. Tamura, K. Iwano, and S. Furui, "Multi-modal speech recognition using optical-flow analysis for lip images," *J. VLSI Signal Process. Syst.*, vol. 36, no. 2/3, pp. 117–124, 2004.

[21] W. Abdulla, P. Yu, and P. Calverly, "Lips tracking biometrics for speaker recognition," *International Journal of Biometrics*, vol. 1, no. 3, pp. 288–306, 2009.

[22] T. Wark, D. Thambiratnam, and S. Sridharan, "Person authentication using lip information," in *IEEE TENCON*, 1997, pp. 153–156.

[23] M. Sánchez, "Aspects of facial biometrics for verification of personal identity," Ph.D. dissertation, University of Surrey, 2000.

[24] S. Bengio, J. Mariethoz, and S. Marcel, "Evaluation of biometric technology on XM2VTS," IDIAP, Tech. Rep., 2001.

[25] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *AVBPA*, 1999, pp. 72–77.

[26] J. Luettin and G. Maître, "Evaluation protocol for the extended M2VTS database (XM2VTSDB)," IDIAP, Tech. Rep. Idiap-Com-05-1998, 1998.

[27] C. Chan, "Multi-scale local binary pattern histogram for face recognition," Ph.D. dissertation, University of Surrey, 2008.

[28] R. Zabih and J. Woodfill, “Non-parametric local transforms for computing visual correspondence,” in *ECCV (2)*, 1994, pp. 151–158.

[29] 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*, vol. 2354, no. 1, pp. 197–204, 1994.

[30] H. Jin, Q. Liu, H. Lu, and X. Tong, “Face detection using improved lbp under bayesian framework,” in *ICIG*, 2004, pp. 306–309.

[31] X. Tan and B. Triggs, “Enhanced local texture feature sets for face recognition under difficult lighting conditions.” in *AMFG*, 2007, pp. 168–182.

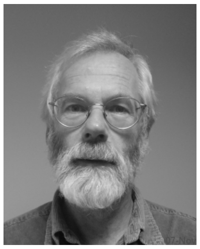
[32] G. Zhao and M. Pietikäinen, “Local binary pattern descriptors for dynamic texture recognition,” in *ICPR (2)*, 2006, pp. 211–214.

[33] L. I. Kuncheva, *Combining Pattern Classifiers Methods and Algorithms*. Wiley, 2004.

Chi Ho Chan received his Ph.D. degree from the University of Surrey, U.K. in 2008. He is currently a research fellow at Centre for Vision, Speech and Signal Procssing, University of Surrey. From 2002 to 2004, he served as a researcher at ATR International (Japan). His research interests include Image Processing, Pattern Recognition, Biometrics, and Vision-Based Human-Computer Interaction.

Budhaditya Goswami received a B.Eng in Electronic Engineering in 2005 from the University of Surrey. Upon completion of this degree, he then moved on to work on his PhD which he completed in 2011. The PhD was titled 'Lip Behavioural Biometrics' and supervised by Prof J Kittler at the Centre for Vision, Speech and Signal Processing. A major finding of this research was that the appropriate quantisation of lip deformation during speech could lead to a stand-alone biometric trait. His research interests lie in the areas of computer vision and machine learning.

Josef Kittler received the B.A., Ph.D., and D.Sc. degrees from the University of Cambridge in 1971, 1974, and 1991, respectively. He heads the Centre for Vision, Speech and Signal Processing at the School of Electronics and Physical Sciences, University of Surrey, U.K. He teaches and conducts research in the subject area of machine intelligence, with a focus on biometrics, video and image database retrieval, automatic inspection, medical data analysis, and cognitive vision. He published a Prentice-Hall textbook on Pattern Recognition: A Statistical Approach and several edited volumes, as well as more than 600 scientic papers, including in excess of 170 journal papers. He serves on the Editorial Board of several scientific journals in pattern recognition and computer vision.



William Christmas received the bachelors degree from the University of Oxford and the PhD degree from the University of Surrey. His PhD research work was focused on the use of probabilistic methods for matching geometric features. After graduating from the University of Oxford, he was with the British Broadcasting Corp. as a research engineer, working on a wide range of projects related to broadcast engineering. Then, he was with the BP Research International as a senior research engineer, working on research topics that included both hardware and software aspects of parallel processing, real-time image processing, and computer vision. He is currently a university fellow in technology transfer in the Centre for Vision, Speech and Signal Processing, University of Surrey. His research interests include the integration of machine vision algorithms to create complete applications. He is currently working on a range of projects concerned with the automatic annotation and indexing of multimedia material. A list of his publications can be found at http://www.ee.surrey.ac.uk/Personal/W.Christmas/list.pdf.