SPEAKER AUTHENTICATION USING VIDEO-BASED LIP INFORMATION

B Goswami[⋆] *C Chan*[⋆] *J Kittler*[⋆] *W Christmas*[⋆]

[⋆] CVSSP, FEPS, University of Surrey
Guildford, GU2 7XH, Surrey, United Kingdom

ABSTRACT

The lip-region can be interpreted as either a genetic or behavioural biometric trait depending on whether static or dynamic information is used. In this paper, we use a texture descriptor called Local Ordinal Contrast Pattern (LOCP) in conjunction with a novel spatiotemporal sampling method called Windowed Three Orthogonal Planes (WTOP) to represent both appearance and dynamics features observed in visual speech. This representation, with standard speaker verification engines, is shown to improve the performance of the lip-biometric trait compared to the state-of-the-art. The improvement obtained suggests that there is enough discriminative information in the mouth-region to enable its use as a primary biometric as opposed to a "soft" biometric trait.

***Index Terms*—** Biometrics, lip, spatiotemporal

1. 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. The lip-region can be interpreted as either a genetic or behavioural biometric trait. 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 general view of the research community about the lack of discriminative power in the lip region.

In this paper, we propose a method that results in improvements in the performance of lip-based biometrics. These improvements are realised through the use of a texture descriptor called Local Ordinal Contrast Patterns (LOCP). The primary contribution of this paper is a novel dynamic texture representation 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 text-independent speaker verification systems based on Normalised Correlation(NC) and Chi-squared (χ^2) histogram matching methods.

A review of the state-of-the art is presented in Section 2. The description of LOCP is provided in Section 3. Section 4 describes the proposed WTOP configuration. The speaker verification systems used to evaluate this novel descriptor are described in Section 5. The paper concludes with the experimental evaluation in Section 6 and some concluding remarks in Section 7.

Table 1: Performance of Lip Biometric Systems for Speaker Verification Showing Lip Performance And Fusion Performance

SYS.	LIP FEATURE	DATABASE	CLIENTS	PERF.(%)	
[7]	Dynamic TI	**XM2VTS**	295	EER	22
[6]	Dynamic TD	**XM2VTS**	295	HTER	13.35
[9]	Dynamic TI	**XM2VTS**	295	HTER	**0.65**
[4]	Static(geometric)	Custom	50	EER	0.015
[3]	Static(shape)	M2VTS	37	HTER	6.85
[8]	Dynamic TI	AMP CMU	10	HTER	0.0
[10]	Dynamic TI	TULIPS1	96	EER	0.0
[10]	Static(intensity)	TULIPS1	96	EER	0.0
SYS.	**FEATURE FUSION**	**DATABASE**	**CLIENTS**	**PERF.(%)**	
[5]	Static(geometric) + Audio	**XM2VTS**	261	HTER	6.3
[7]	Dynamic TI + Audio	**XM2VTS**	295	EER	2
[6]	Dynamic TD + Audio	**XM2VTS**	295	HTER	**0.74**
[6]	Dynamic TD + Face + Audio	**XM2VTS**	295	HTER	7.06
[11]	Hybrid(shape and intensity)	Custom	35	EER	18.0
[12]	Hybrid(texture and motion)	MVGL-AVD	50	EER	5.2
[13]	Static(texture)	MVGL-AVD	50	EER	1.7
[3]	Static(shape) + Audio	M2VTS	37	HTER	0.3

2. RELEVANT WORK

The use of the lip features for human identification was first proposed through the concept of "lip-prints" by forensic anthropologists Fischer and Locard [1]. Lip prints contained information about the lip texture. The application of lip prints specifically as a biometric was first introduced in [2]. A taxonomy of contemporary relevant work can be based on whether the approach uses static or dynamic information from the lip-region. This also allows for 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 [3, 4, 5].

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: *Text-dependent*(TD) systems [6] and *Text-independent*(TI) speaker recognition [7, 8].

Hybrid Methods: use both static and dynamic information by performing either score-level or feature-level fusion [13, 12, 11, 10, 9]

State-of-the-art Performance Review: Commonly, lip-based features are evaluated in terms of the performance improvement they provide through multi-modal fusion with more established biometric traits. For the testing of speaker verification systems, only a few databases such as [14] provide established verification protocols that enable a fair comparison of systems. However, some publications use custom-built datasets and evaluation protocols which reduces the comparability of the systems. In these systems, the classifica-

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

tion task is often made easier by skewing the ratio of trait feature dimensions to the number of clients to be small.

Table 1 provides an overview of the performance of reviewed lip-biometric systems. For a more thorough description of the various speaker verification metrics, the reader is referred to [15].

As shown in Table 1, the most commonly used database and protocol are XM2VTS (used by 4 authors) and Lausanne Protocols respectively. The best performance obtained using lip features *only* on this database are by [9](Half Total Error Rate (HTER) of 0.65%. Multi-modal fusion with audio features [6] yields HTER of 0.74%. In this paper, we use the XM2VTS to ensure the comparability of our results with these reference benchmarks.

3. 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 [16]. Local Binary Patterns (LBP) [17] are an example of ordinal contrast patterns. The LBP operator measures the ordinal contrast pairs between a local neighbour value and the reference (centre pixel) value.

In this paper, we use an alternative 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 circular neighbourhood, P, to form an operator. This enables the modelling of arbitrarily large scale structure by varying R. In this paper, we improve the LOCP operator originally presented by [9] to incorporate ordinal polarity cases where there was no contrast between pixel pairs. When applying LOCP, we choose P pixel pairs for ordinal contrast encoding in Eqn. 1.

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

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

where g_p is the intensity measurement at the p^{th} pixel,$p \in [0, P)$, a distance R away from the location $\boldsymbol{x}$. The pattern is obtained by concatenating the binary contrast encodings into a P-bit sequence. LOCP represents local, pairwise neighbourhood derivatives.

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.

Recently, three orthogonal planes (TOP) [18, 9] have been proposed to provide a spatiotemporal representation for dynamic texture analysis. The extension of the TOP idea is discussed in Section 4 and its evaluation results using XM2VTS is shown to outperform the state-of-the-art.

4. WINDOWED THREE ORTHOGONAL PLANES

WTOP is an example of a Dynamic Texture (DT) and a generalisation of the TOP configuration[18]. TOP allows the quantisation of spatial appearance information from the XY plane. Temporal quantisation in terms of co-occurrence statistics of horizontal and vertical motion are in turn obtained from the XT and YT planes. TOP aims to describe these co-occurrence 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 extracted from the central column and 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.

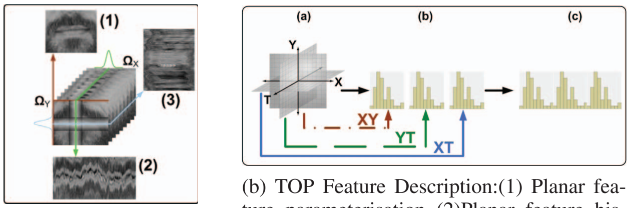
Since TOP uses only single-pixel width image profile layers, it represents a sampling of the volume of visual speech in both the spatial and temporal directions. It is clearly obvious that if the sampled planar cross-section does not display adequate temporal feature alignment i.e. in the XT and YT direction, 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 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 clear from the mathematical formulation of these ideas. A spatiotemporal cube of observed video of T frames can be considered as a set, $C : XY_{\{0...T-1\}}$ where XY_i is each image frame of height N pixels and width M pixels so that $XY \in \mathbb{R}^{N\times M}$. The extraction of the $XT \in \mathbb{R}^{T\times M}$ images involves the application of a windowing mask, $\mathbf{\Omega}_Y \in \mathbb{R}^N$ to each element in the set C. The extraction of the $YT \in \mathbb{R}^{N\times T}$ images involves the application of a windowing mask, $\mathbf{\Omega}_X \in \mathbb{R}^M$ to each element in the set C. The i^{th} column of XT and i^{th} row of YT can then be written as:

$$\begin{aligned} XT(i) &= \mathbf{\Omega}_Y^{\mathsf{T}} \cdot XY_i, \quad (0 \le i < T) \\ YT(i) &= XY_i \cdot \mathbf{\Omega}_X, \quad (0 \le i < T) \end{aligned} \tag{2}$$



(a) Extraction of images using TOP. (1) XY (2) YT (3) XT Images

(b) TOP Feature Description:(1) Planar feature parameterisation (2)Planar feature histograms (3) Concatenated histograms for dynamic texture

Fig. 1: TOP configuration

In the case of the TOP representation, the mask $\boldsymbol{\Omega}$ is obtained as a constant vector computed using a Kronecker delta function. As mentioned above, we use a Gaussian mask to perform this task $\mathcal{N}(\mu,\sigma)$. Here μ 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 $\boldsymbol{\Omega}_Y$, $\boldsymbol{x} = [0 \quad 1 \ldots N-1]^T$. This idea is shown in Fig. 1a

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.

$$\boldsymbol{h}^{\beta}_{P,R}(i) = \sum_{(x',y')\in \boldsymbol{M}} B(LOCP^{\beta}_{P,R}(x',y') = i) \qquad (3)$$

where $B()$ 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 Fig. 1b to provide the dynamic texture information. Here, $\alpha = XYXTYT$ represents the best performing planar TOP configuration [9]. Eqn. 4 shows the obtained histogram:

$$\boldsymbol{f}^{\alpha} = [\boldsymbol{h}^{XY}_{P,R}, \boldsymbol{h}^{XT}_{P,R}, \boldsymbol{h}^{YT}_{P,R}] \qquad (4)$$

An important consideration in the WTOP configuration is the parameters P and R of the LOCP descriptor along each place. These values relate to the sampling rate in the XY, XT or YT planes. Since the planar sampling rates are used to capture sufficient dynamic evolution, the parameters P and R need to be tailored to each plane.

5. SPEAKER VERIFICATION SYSTEMS

For this system, the method in [19] 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 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 sub-cube, we use LOCP-WTOP to extract histograms $\boldsymbol{h}^{\beta,j}_{P,R}$ where j represents the sub-cube 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 to the classification engines described below.

Chi-squared Histogram Matching (χ^2): 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 $\boldsymbol{I}$.

$$Sim_{\chi}(\boldsymbol{G},\boldsymbol{I}) = -\sum_{j}\sum_{i} \frac{(\boldsymbol{f}^{\alpha,j}_{G}(i) - \boldsymbol{f}^{\alpha,j}_{I}(i))^2}{\boldsymbol{f}^{\alpha,j}_{G}(i) + \boldsymbol{f}^{\alpha,j}_{I}(i)} \qquad (5)$$

Normalised Correlation (NC): In order to extract the discriminative features we project the sub-cubic histograms, $\boldsymbol{f}^{\alpha,j}$, into Linear Discriminant Analysis (LDA) space as: $\boldsymbol{d}^{\alpha,j} = (\boldsymbol{W}^{\alpha,j}_{lda})^T \boldsymbol{f}^{\alpha,j}$. After projection, we measure NC across all sub-cubes using two videos $\boldsymbol{G}$ and $\boldsymbol{I}$ as specified in Eqn. 6.

$$Sim_{LDA}(\boldsymbol{G},\boldsymbol{I}) = \sum_{j} \frac{(\boldsymbol{d}^{\alpha,j}_{G})^{\top} \boldsymbol{d}^{\alpha,j}_{I}}{\|\boldsymbol{d}^{\alpha,j}_{G}\| \|\boldsymbol{d}^{\alpha,j}_{I}\|} \qquad (6)$$

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 [9]. The XM2VTSDB [14] 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.

6. RESULTS AND EVALUATION

Tables 2 and 3 show the performance figures for LOCP-WTOP histograms with the best performances shown in bold. Please note that the best HTER performance in the test sets are chosen based on the lowest EER scores in the evaluation set. 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. $\alpha = XYXTYT$ was used [9]. Various widths (σ) for the Gaussian window were used.

Performance of WTOP: The results clearly show that a Gaussian windowing function outperforms TOP. As theorised, TOP uses a delta window function which under-samples spatiotemporal data. Increasing the width of the windowing function using a Gaussian provides a more detailed and discriminative representation. 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 χ^2 system were comparable. LOCP outperformed LBP when combined with the NC system.

Comparison to literature: The results obtained demonstrated a marked and remarkable improvement on the best performance using lip features alone observed on this database in the literature (HTER=0.65 [9]). The best performance using the χ^2 system for C1

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 2: LOCP/LBP-WTOP HTER (in %) using the χ^2 system

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 3: LOCP/LBP-WTOP HTER (in %) using the NC system

was obtained using LOCP-WTOP ($\sigma = 10$) with HTER of 3.04% and for C2, LOCP-WTOP ($\sigma = 1$) with HTER 3.74%. The best performance using the NC system for C1 was obtained using LOCP-WTOP ($\sigma = 1$) with HTER of 0.44% and for C2, LOCP-WTOP ($\sigma = 10$) with HTER 0.27%.

As expected, the NC system outperformed the χ^2 system. The overall improvement is due to the encapsulation of discriminative, dynamic and appearance textures using LOCP-WTOP and the implicit intra-modal fusion of genetic and behavioural properties of the observed lip-regions. The obtained results were better than those provided in [9] which used LOCP-TOP to obtain HTER 0.65%(C1) and 0.95%(C2) using the NC system. The obtained results even outperformed the best multi-modal fusion results (lip with audio) [6] which resulted in HTER of 0.74%. Note that the systems in [6] were text-dependent, while our representation is text-independent.

7. CONCLUSIONS AND FUTURE WORK

We first presented a state-of-the-art review of lip biometric systems. In this paper, we have used an ordinal contrast measure called LOCP. This has been applied in a novel WTOP configuration as input into speaker verification systems. WTOP is a generalisation of TOP. The resulting biometric systems have been used to evaluate the performance of mouth-region biometrics in the XM2VTS database. A Gaussian mask in WTOP is shown to outperform TOP. LOCP was also evaluated against LBP and found to be a better texture descriptor in LDA space. The application of this novel spatiotemporal feature representation has been demonstrated to outperform the state-of-the-art. The findings in this paper suggest that there is sufficient discriminative information in the spatiotemporal evolution of the mouth-region during speech production for its use as a hard biometric.

8. REFERENCES

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

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

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

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

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

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

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

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

[9] B. Goswami, C.H. Chan, J. Kittler, and W. Christmas, "Local ordinal contrast patterns for spatiotemporal, lip-based speaker authentication," in *BTAS*, 2010.

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

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

[12] 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.*, vol. 15, no. 10, pp. 2879–2891, 2006.

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

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

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

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

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

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

[19] M.U.R. Sánchez, *Aspects of facial biometrics for verification of personal identity*, Ph.D. thesis, University of Surrey, 2000.