Kernel-based Speaker Verification Using Spatiotemporal Lip Information

Chi Ho Chan, Budhaditya Goswami, Josef Kittler and William Christmas
Centre for Vision, Speech and Signal Processing,
University of Surrey, Guildford, Surrey, U.K., GU2 7XH.
{c.chan, b.goswami, j.kittler, w.christmas}@surrey.ac.uk

Abstract

The lip-region can be interpreted as either a genetic or behavioural biometric trait. Despite this breadth of biometric content, lip-based biometric systems are scarcely developed in the literature. A recent trend in lip biometrics is to use a spatiotemporal texture representation of visual speech to generate biometric features. In this paper we make two contributions related to the above biometric traits. We investigate whether the application of non-linear discriminant analysis on spatiotemporal texture improves its biometric performance. Spatiotemporal texture representation of visual speech results in performance that suggests that the lip can be used as a hard *biometric. We investigate the effect of the amount of video information on speaker verification performance. The results show that using non-linear discriminant analysis improves speaker verification performance. Additionally, we also demonstrate that using over 3 seconds of video is sufficient to achieve satisfactory accuracy.*

1 Introduction

Numerous measurements and signals have been investigated for use in biometric systems. Among the most popular measurements are fingerprint, face and voice. The latter two arise naturally in the process of human communication. The process of human speech production results in two information sources in the form of video and audio signals. These signals can be measured independently. The video information contains the deformation of the lip during speech and can be termed visual speech. This deformation can be parameterised for use as a complement to or a replacement of audio-biometric information in noisy environments. Lip-region features straddle the area between the face and voice biometric. The lip-region can be interpreted as either a genetic (individual lip appearance) or behavioural (lip dynamics) biometric trait.

The task of lip-based speaker verification involves verifying the claimed identity of an input *probe* visual speech signal. The various methods of parameterising the biometric information contained within visual speech can be segregated into static, dynamic or hybrid methods depending on the amount of temporal information used. The authors in [11] have recently introduced a novel lip-biometric system. The proposed system was the first biometric system to utilise histogram features which combined a method of dynamic texture representation called Three Orthogonal Planes (TOP) with a novel texture descriptor called Local Ordinal Contrast Pattern (LOCP).

The system in [11] uses linear classifiers in conjunction with the extracted histogram features to perform speaker verification. In this paper we investigate the effect of non-linear discriminative analysis of these features on speaker verification performance. Additionally, we also quantify the effect of reducing the amount of video information available to the speaker verification system. Such an analysis is relevant to the application of lip biometrics in an industrial context.

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

SYSTEM	LIP FEATURE	DATABASE	CLIENTS	PERF.(%)	
Faraj[9, 8]	Dynamic TI	**XM2VTS**	295	EER	22
Goswami[11]	Dynamic TI	**XM2VTS**	295	HTER	**0.65**
Sanchez[18]	Dynamic TD	**XM2VTS**	295	HTER	13.35
Auckenthaler[2]	Static	DAVID	7	% Error	2.2
Cetingul[6]	Static (intensity)	MVGL-AVD	50	EER	5.6
Cetingul[6]	Dynamic TD	MVGL-AVD	50	EER	5.2
Cetingul[7]	Static(texture)	MVGL-AVD	50	EER	1.7
Gomez[10]	Static(geometric)	Custom	50	EER	0.015
Jourlin[12]	Static(shape)	M2VTS	37	HTER	15.4
Samad[16]	Dynamic TI	AMP CMU	10	HTER	0.0
Wark[20]	Dynamic TI	TULIPS1	12	EER	0.0
SYSTEM	**FEATURE FUSION**	**DATABASE**	**CLIENTS**	**PERF.(%)**	
Broun[4]	Static(geometric),Audio	**XM2VTS**	261	HTER	6.3
Faraj[9, 8]	Dynamic TI,Audio	**XM2VTS**	295	EER	2
Sanchez[17]	Dynamic TD,Face	**XM2VTS**	295	HTER	2.62
Sanchez[17]	Dynamic TD,Audio	**XM2VTS**	295	HTER	0.70
Sanchez[17]	Dynamic TD,Face,Audio	**XM2VTS**	295	HTER	0.66
Sanchez[17]	Dynamic TD,2Face,2Audio	**XM2VTS**	295	HTER	0.15
Abdulla[1]	Hybrid(shape,intensity)	Custom	35	EER	18.0
Cetingul[6]	Hybrid(texture,motion)	MVGL-AVD	50	EER	3.6
Cetingul[7]	Static(texture),Dynamic,Audio	MVGL-AVD	50	EER	0.4
Jourlin[12]	Static(shape),Audio	M2VTS	37	HTER	1.65

Section 2 presents a survey of relevant work. A summary of the feature parameterisation of visual speech is presented in Section 3. The non-linear discriminant analysis projection is described in Section 4. The developed system is compared to the appropriate benchmarks in Section 5. Finally, Section 6 provides some concluding remarks.

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 [13]. Lip prints contained information about the lip texture. The application of lip prints specifically as a biometric was first introduced in [19]. 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 [12, 10, 4].
- **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 [18] and *Text-independent*(TI) speaker recognition [9, 16].
- **Hybrid Methods**: use both static and dynamic information by performing either score-level or feature-level fusion [7, 6, 1, 20, 11]

Performance Review: Commonly, lip-based features are evaluated in terms of the performance improvement they provide through fusion with more established 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 reduce the comparability of the systems. In these systems, the classification task is often made easier by using a relatively small ratio of trait feature dimensions to the number of clients.

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 [3]. As shown in Table 1, the most commonly used database and protocol are XM2VTS [14] and Lausanne Protocols respectively. This database has 295 subjects. The best performance obtained using lip features *only* on this database is 0.65% [11] Half Total Error Rate (HTER). Multimodal fusion with audio features [18] yields HTER of 0.15%. In this paper, we use the XM2VTS database to ensure the comparability of our results with these benchmarks.

3 Feature Parameterisation

In this section, we describe the feature parameterisation of the visual speech signal as originally proposed in [11]. The features use a novel LOCP texture descriptor in a TOP configuration.

Local Ordinal Contrast Pattern: LOCP is an example of an ordinal contrast pattern. 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 signal in a neighbourhood is then turned into a binary code. The ordinal measure is invariant to any monotonic transformation such as image gain, bias or gamma correction [21]. A popular example of an ordinal contrast based texture descriptor is the Local Binary Pattern (LBP) [15][1].

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 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 [11] to incorporate ordinal polarity cases where there is no contrast between pixel pairs. When computing the LOCP at location $\boldsymbol{x} = (x, y)$, we choose P pixel pairs for ordinal contrast encoding in Eqn. 1.

$$LOCP_{P,R}(\mathbf{x}) = \sum_{p=0}^{P-1} s(g_{p+1} - g_p)2^p, \text{ where}$$
$$s(\gamma_p) = \begin{cases} 1 & \gamma > 0 \\ 0 & \gamma < 0 \\ 0 & \gamma = 0 \quad \text{and} \quad p = 0 \\ s(\gamma_{p-1}) & \gamma = 0 \quad \text{and} \quad p > 0 \end{cases} \tag{1}$$

The pattern is obtained by concatenating the binary numbers from the encoding into a P-bit sequence. LOCP represents local, pixel intensity derivatives.

LBP suggests that the ordinal relationship between a single reference pixel and its neighbourhood contains texture information. With LOCP, texture is represented by the contents of the entire neighbourhood, not by the relationship of the neighbourhood with a single reference value. LOCP therefore 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.

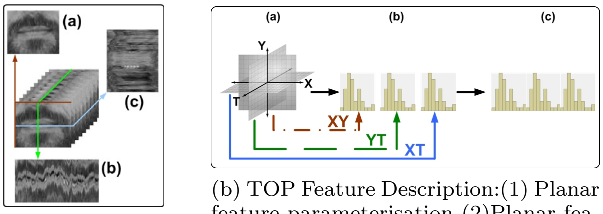
Three Orthogonal Planes: While LOCP is useful as a texture descriptor, its application to the parameterisation of spatiotemporal information requires it to be twinned with a method of dynamic texture representation. TOP is an example of a Dynamic Texture (DT) descriptor[23]. The authors [11] combined the LOCP texture representation with TOP to quantise visual speech information. TOP is computationally simple as it extracts the texture feature in each of the three orthonormal planes(i.e. XY, XT and YT) within a spatiotemporal volume. Figure 1a demonstrates the lip images from three planes. In each plane, the LOCP is extracted and the i^{th} element of the plane-pattern histogram, $\boldsymbol{h}^{\beta}_{P,R} \in \mathbb{R}^{1\times 2^P}$ is computed where $\beta \in \{XY, XT, YT\}$ represents a plane.

$$h^{\beta,i}_{P,R} = \sum_{\boldsymbol{x}\in\boldsymbol{M}} B(LOCP^{\beta}_{P,R}(\boldsymbol{x}) = i), \;\; i \in [0, 2^P - 1] \tag{2}$$

where the function $B()$ represents a boolean indicator and $\boldsymbol{M}$ is the region within which we are computing the histogram.

Then the histogram of each plane is concatenated into a single histogram, $\boldsymbol{f}$ shown in Figure 1b to provide the dynamic texture information. The best performing TOP configuration is given by Equation 3.

$$\boldsymbol{f} = [\boldsymbol{h}^{XY}_{P,R}, \boldsymbol{h}^{XT}_{P,R}, \boldsymbol{h}^{YT}_{P,R}]^{\intercal} \tag{3}$$



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

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

Figure 1: TOP configuration

4 Kernel Discriminant Analysis (KDA)

The system proposed in [11] used a linear method of classification to perform speaker verification. The primary contribution of this paper is to investigate the performance of non-linear discriminant analysis. This is an example of a supervised speaker verification system, i.e. it requires a set of training examples.

Consider a data matrix, $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_m]$ of m training samples. It can be easily shown that the Linear Discriminant Analysis (LDA) projection which maximises the ratio of between class to total class scat-

[1] LBP is used as the benchmark texture descriptor in these experiments.

ters can be found by optimising:

$$\max_{\mathbf{w}} \quad J(\mathbf{w}) = \frac{\mathbf{w}^\intercal\, \mathbf{K}\,\mathbf{A}\,\mathbf{K}\,\mathbf{w}}{\mathbf{w}^\intercal\, \mathbf{K}\,\mathbf{K}\,\mathbf{w}} \tag{4}$$

where $\mathbf{K}$ is a kernel matrix defined as $\mathbf{K} = \mathbf{F}^\intercal\mathbf{F}$ and $\mathbf{A}$ is a block diagonal matrix of constants, reflecting the number of training samples in each class. Note that each element $k_{u,v}$ of the kernel matrix, $\mathbf{K}$, is given as a scalar product of two training vectors, i.e.

$$k_{u,v} = \mathbf{f}_u^\intercal\, \mathbf{f}_v \tag{5}$$

which, where suitably normalised, measures their correlation i.e. similarity. The formulation in Eqn. (4) renders LDA to be extendable to its non-linear form by replacing the definition of similarity in Eqn. (5) by a non linear function, $\varphi(\mathbf{f}_u, \mathbf{f}_v)$, of samples $\mathbf{f}_u$ and $\mathbf{f}_v$. Such a function maps the original vectors into a higher dimensional space which potentially can be infinite and in which the class separation is enhanced. In this paper, we use the Radial-basis function (RBF) as described by:

$$k_{u,v} = \varphi(\mathbf{f}_u, \mathbf{f}_v) = e^{-\frac{1}{\sigma}\, dist(\mathbf{f}_u,\mathbf{f}_v)} \tag{6}$$

where $dist(\mathbf{f}_u, \mathbf{f}_v)$ is the Euclidean distance,$e^{(\cdot)}$ is the exponential function and σ is a scalar which normalises the distance. Following [22], σ is set to the average Euclidean distance between all elements of the kernel matrix on the training data.

4.1 KDA using Spectral Regression (SR-KDA)

It is shown in [5] that instead of solving the eigenproblem in Eqn. (4), the KDA projections can be obtained from the following two linear equations:

$$\begin{aligned} \mathbf{A}\phi &= \lambda\phi \\ (\mathbf{K} + \delta\mathbf{I})\mathbf{w} &= \phi \end{aligned} \tag{7}$$

where ϕ is an eigenvector of $\mathbf{A}$, $\mathbf{I}$ is the identity matrix and $\delta > 0$ is a regularisation parameter. Eigenvectors ϕ are obtained directly from the Gram-Schmidt method. Since $(\mathbf{K} + \delta\mathbf{I})$ is positive definite, the Cholesky decomposition: $(\mathbf{K} + \delta\mathbf{I}) = \mathbf{R}^\intercal\mathbf{R}$ is used to solve the linear equations in Eqn. (7). The obtained result, $\mathbf{R}$ is a upper triangular matrix. Thus, the solution of the linear system becomes:

$$(\mathbf{K} + \delta\mathbf{I})\mathbf{w} = \phi \Leftrightarrow \left\{ \begin{array}{l} \mathbf{R}^\intercal\theta = \phi \\ \mathbf{R}\mathbf{w} = \theta \end{array} \right. \tag{8}$$

i.e., first solve the system to find vector θ and then vector $\mathbf{w}$. In summary, the C-class SR-KDA projection matrix, $\mathbf{W}^{kda} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_{C-1}]$, only needs to solve a set of regularised regression problems and there is no eigenvector computation involved. This results in great improvement of computational cost compared to LDA computations using eigen-decomposition and allows the system to handle large kernel matrices.

Speaker Verification Systems: Mouth-region features were extracted using per-frame localisation. These extracted regions were then used as input information for parameterisation using LOCP-TOP. Each extracted region can be visualised as a cuboid containing spatiotemporal information. This cuboid is first subdivided into overlapping sub-cuboids to increase the resolution of available video information. For the j^{th} sub-cuboid, we use LOCP-TOP to extract histograms $\boldsymbol{f}^j$. These combined histograms conceptually represent the intra-modal, feature-level fusion of extracted LOCPs in the different planes. The histograms are then compared using a similarity function.
Normalised Correlation & KDA (NC+KDA): In order to extract the discriminative features we project the sub-cuboid histograms, $\boldsymbol{f}^j$, into KDA space as: $\boldsymbol{d}^j = (\boldsymbol{W}^{kda,j})^\intercal \boldsymbol{f}^j$. The similarity of the two videos is measured using normalised cross-correlation:

$$Sim_{NC}(\boldsymbol{G}, \boldsymbol{I}) = \sum_j \frac{(\boldsymbol{d}_G^j)^\intercal \boldsymbol{d}_I^j}{\|\boldsymbol{d}_G^j\| \|\boldsymbol{d}_I^j\|} \tag{9}$$

where $\boldsymbol{G}$ and $\boldsymbol{I}$ are the input videos and i, the bin index.

5 Experimental Set-up and Results

Visual speech videos consisted of 61 × 51 pixel mouth-region windows localised in the XM2VTS database video frames. LOCP and LBP feature parameters P and R were set to 8 and 3 respectively 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. These sub-cuboids overlapped each other by 70% to ensure quantisation of temporally continuous information. The experimental evaluation used the XM2VTS and Configuration 1 (C1) and Configuration 2 (C2) Lausanne protocol. The default value of regularisation parameter for SR-KDA, δ, was set to 0.01. Two experiments were performed.

The first experiment aimed to quantify the effect of reducing the amount of video information available to both systems. The highest, lowest and average frame lengths of the XM2VTS videos were 673, 167 and 319 frames respectively. Given a frame-rate of 25 fps, 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 TOP systems in steps of 25 frames starting from the first frame of video. Note that both gallery and probe videos are 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 improvement in performance with increasing video information are plotted in Figures 2 and 3. They show that by increasing the amount of video information we gain improvement in system accuracy as expected. The results also demonstrate that LOCP is consistently better than LBP as a texture descriptor. As a benchmark, this experiment was also performed using the Chi-squared (X2) system [11]. The accuracy of NC+KDA is better than X2. The system also reaches a

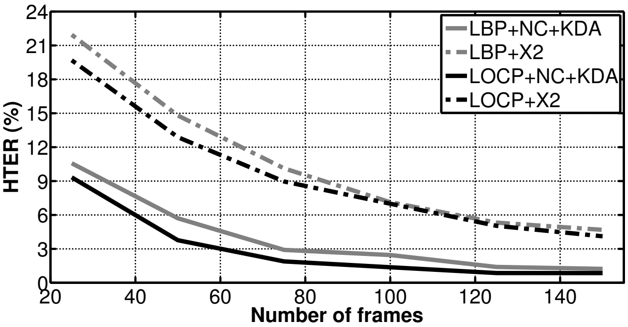


Figure 2: HTER(%) of TOP systems Versus Visual Speech Frame Length for C1

steady-state more quickly implying that it extracts discriminative information quicker. The results demonstrate that around 3 seconds of video information provides good HTER performance. This suggests that lip biometrics can play an important role in applications with short probe signal videos. The second experiment compared the performance of using all available video information. The aim of this experiment was to quantify the performance of using the NC+KDA verification system with the state of the art benchmarks. Table 2 shows the Equal Error Rate (EER) and HTER performance in % for the LBP/LOCP-TOP features when input into the NC+KDA system. The results obtained by [11] are also included for reference which used X2 and normalised correlation with LDA (NC+LDA). The NC-LDA system incidentally is also the best performing lip biometric system encountered in the literature.

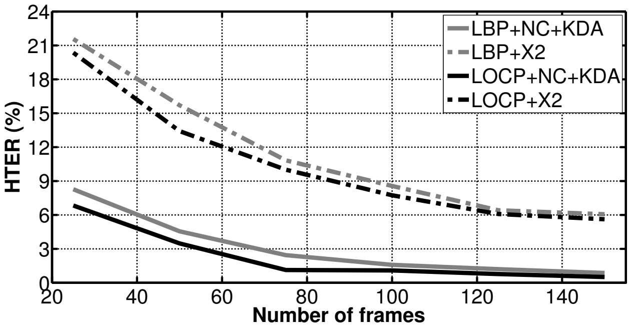


Figure 3: HTER(%) of TOP systems Versus Visual Speech Frame Length for C2

Table 2: TOP systems EER and HTER (in %)

System	Configuration I				Configuration II			
	LBP		LOCP		LBP		LOCP	
	Eval	Test	Eval	Test	Eval	Test	Eval	Test
NC+KDA	0.48	0.56	0.50	0.24	0.51	0.79	0.29	0.40
X2 [11]	-	-	2.99	3.86	-	-	4.27	3.97
NC+LDA [11]	-	-	0.33	0.65	-	-	0.76	0.95

Two inferences can be drawn from the results. The first is that the NC+KDA system outperforms the state of the art. This system compares the probe and gallery features in KDA projected space. The performance improvement suggests that the use of a nonlinear similarity function like RBF within the KDA formulation is better than using simple vector correlation used in LDA. It can again be seen that LOCP outperforms LBP in the NC+KDA system in this database.

6 Conclusions

We have demonstrated that measuring normalised correlation in KDA space results in a better speaker verification system than the state of the art using dynamic texture representation of visual speech. We have also measured the effect of considering the amount of video information available to a lip-based speaker verification system. The results using XM2VTS suggest that over 3 seconds of video provides HTER performance of less than 1%. This implies that the lip biometric is useful in situations with short probe videos.

References

[1] W. Abdulla, P.W.T. Yu, and P. Calverly. Lips tracking biometrics for speaker recognition. *International Journal of Biometrics*, 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] D. Cai, X. He, and J. Han. Efficient kernel discriminat analysis via spectral regression. In *ICDM*, 2007.

[6] 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.

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

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

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

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

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

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

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

[15] M. 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.

[16] 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.

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

[18] 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.

[19] 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.

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

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

[22] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *IJCV*, 73(2):213–238, jun 2007.

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