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

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

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

I. INTRODUCTION

Numerous measurements and signals have been proposed and investigated for use in biometric recognition systems. Among the most popular measurements are fingerprint, face and voice. Each of these biometric traits have their pros and cons with respect to accuracy and deployment. The use of lip-region features as a biometric straddles the area between the face and voice biometric. There a two main factors that make the use of lip features a compelling biometric. Firstly, since speech is a natural, non-invasive signal to produce, the associated lip-motion can also be captured in a non-intrusive manner. Additionally, with the advent of cheap camera sensors for imaging, it easier than ever before to isolate the lip-region features and use them in combination with other biometric traits to enhance the robustness of multi-modal biometric systems. The use of talking face features also naturally increases the robustness of the system with respect to any attempts at faking "liveness".

The use of the lip region as a means of human identification was first proposed in the field of forensic anthropology as early as the 20th century when forensic investigators such as Fischer and Locard [14] used it for identification in homicidal investigations. In lip prints, the individual grooves and eccentricities of the lip surface are used to identify individuals. The application of lip prints specifically as a biometric trait for security applications was introduced in [25],[24]. A review of the use and evolution of lip prints for use as a human identifier is presented in [17]. Lips can be envisioned as the combination of a genetic and a behavioral biometric.

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

A taxonomy of the state-of-the art in lip-based speaker verification is presented in Section I-A. A summary of the current performance characteristics of the field is presented in (INCLUDE REFERENCE TO TABLE CONTAINING THE EVALUATION SUMMARY). A discussion of the approaches and their merits and failings leads to the motivation behind the development of the current, novel feature descriptor. This motivation is presented in Section I-B and the detailed treatment of the use of OCP features for locally spatiotemporal description is provided in Section II. An overview of the speaker verification systems used to evaluate the usefulness of this novel descriptor is provided in Section III. The paper concludes with the experimental evaluation in Section. IV and some concluding remarks in Section V.

A. Literature Review

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

1) Static Methods: The physical and geometric structure of the lip can be used to extract appropriate shape and appearance related features. These features are used as

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

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

[7] used hand-labelling to segment the lip contour. This sort of segmentation approach enabled the recognition experiments to approach the upper limits of accuracy governed by the use of facial characteristics themselves without any segmentation errors. The extracted, geomteric lip features were compared to the performance of long-established acoustic features with largely similar performance. However, an obvious drawback to this system was the use of hand-labelling which restricted the scope of the experimental validation to only a small data-set. This idea was extended in [20] where an automatic means of lip-region segmentation based on Active Shape Models(ASM) [27] was used to extract the shape and intensity information related to an individual producing speech. Principal Components Analysis (PCA) [21] was then used to perform shape-independent, intensity based feature extraction. These features were then used in conjunction with acoustic features to perform speaker recognition. The accuracy of this system was affected by the quality of the obtained segmentation and consequently, the visual lip features were weighted during feature-level-fusion with the acoustic data. In [5] 2d-DCT coefficients were used as lip-texture features and applied to a multi-modal speaker/speech recognition system. The authors in [13] achieved some very promising results using geometrical parameters, HMM and PCA methods. They parameterised the shapes of observed lips on a frame-by-frame basis using cartesian and polar co-ordinates. Recognition was then performed using a multiparameter HMM with the polar co-ordinates and a multilayer neural network is applied to the Cartesian coordinates. In [3] automatic lip segmentation using pixel based thresholding in the HSV colour space was used to extract the lip region. A geometric feature vector was computed from this region using information like lip width and height. A score level fusion was then performed with acoustic features and the resulting biometric system tested.

2) Dynamic Methods: Most deployed speaker applications of biometric systems are based on scenarios with co-operative users speaking fixed string passwords or repeating prompted phrases from a small vocabulary. These generally employ what is known as text-dependent systems. Such constraints are quite reasonable and can greatly improve the system accuracy. However, there are cases when such constraints can be cumbersome and 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 speaker recognition.

a) Text-dependent systems: In the work presented in [18], [23] and [15] lip tracking is performed using a simple bayes filter and an apriori trained lip shape eigen model. An instance of a lip is represented as a combination of affine parameters representing global deformation and eigenvalues representing local refinement. Lip tracking is then performed by synthesizing an example of an eigenlip using a simple first order temporal evolution model. The lip contour is then segmented using a means of probabilistic boundary searching using colour models. The motion parameters are computed by considering all the eigenlips that form the sequence of some speech. This sequence is then compared against the claimed identity of a sequence using the Dynamic Time Warping (DTW) algorithm presented in [19].

b) Text-independent systems: In [28], ASMs are used to perform lip tracking. Linear discriminant analysis is then performed on the extracted temporal sequences connected with user speech. This serves to identify the most discriminative features and consequently, these features can be used with acoustic features as a biometric. In [8], a comparative evaluation of the representation of a segmented lip region in terms of a variety of geometric features such as central moments, Hu moments, Malinowska ratio, Lp1 and Lp2 ratio and Feret ratio is presented. The authors suggest the use of 3rd order Zernike moments as the optimal geometric representation of a lip region for use as a shape based biometric.

In the work presented in [12], [9], [11] and [10], a modified, conceptually similar technique to optical flow estimation, called the 3-D structure-tensor method is used to estimate the motion flow vectors of the lip contour in spatiotemporal space. This method has an advantage in that it does not require the successful tracking of the entire lip contour since it extracts motion features from the entire mouth region. These features are then fused, at the feature level with acoustic features to perform speaker recognition and also uses Support Vector Machines (SVM) [2], [29] for the classiciation process. The difference between each method lies in the method used for quantisation of the 3-d structure tensors and the use of Gaussian Mixture Models (GMMs) as opposed to HMMs for speaker verification.

In the work of [22], Minimum Average Correlation Energy (MACE) filters are used on frames containing only the mouth-region to perform lower face based person verification. The aim is to decorrelate all the variant information present in this region and use the resulting features to build a discriminative model for an individual.

3) Hybrid Methods: Hybrid methods use information in both a static and dynamic manner. The authors in [1] attempted to improve the quality of automatic lip feature extraction by using the Discrete Cosine Transform(DCT) to orthogonalise the lip region data into static and dynamic features. The respective features were then individually added to acoustic data for use as a biometric.

The authors in [6] use a combination of audio, lip texture and lip motion features to prove the usefulness of the lip biometric. The lip texture features are obtained as explained in the previous section [5]. Discriminative analysis of the dense motion vectors contained in a bounding box around the mouth region is used to obtain the lip motion information. The feature level comparison is then performed using the reliability weighted summation (RWS) decision rule. Additionally, the authors have extended their experimentation on the explicit usefulness and type of lip motion information using dense motion features to perform a comparative evaluation in [4].

In [26], motion estimation is performed using optical flow. The optical flow information is used to generaet two kinds of visual feature sets in each frame. The first feature set consists of variances of vertical and horizontal components of optical-flow vectors. These are useful for estimating silence/pause periods in noisy conditions since they represent movement of the speakers mouth. The second feature set consists of maximum and minimum values of integral of the optical flow. These are expected to be more effective than the first set since this feature set has not only silence/pause information but also open/close status of the speakers mouth. Each of the feature sets is combined with an acoustic feature set in the framework of HMM-based recognition. Triphone HMMs are trained using the combined parameter sets extracted from clean speech data.

B. *Motivation*

II. Local Spatiotemporal Descriptors Using Ordinal Contrast Patterns

- include brief reference to LBP-TOP paper for speaker verification - Better than LBP because: not dependent on only central pixel. It represents local, pairwise neighbourhood derivatives and therefore is implicitly conditioned to be more robust to noise and indeed provide a richer texture representation than say LBP. -

III. Speaker Verification

- text independent speaker verification using gmm or histogram matching

A. *Histogram Matching*

B. *Linear Discriminant Analysis*

IV. Experimental Evaluation

The XM2VTSDB [16] database, available since 1999, 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 recorded at one month intervals over a period of five months. The entire database was captured using a Sony VX1000E digital cam-corder and DHR1000UX digital VCR under a controlled environment with uniform illumination conditions and a blue background to facilitate face segmentation. Because of the recording conditions, the database is primarily intended for developing 2D personal identity verification systems where it is reasonable to assume that the client will be cooperative.

- describe comparison - describe configuration - evaluation metrics - we are using EER

χ^2				

A. *Performance as a genetic trait*

B. *Performance as a behavioral trait*

C. *Performance using feature-fusion*

V. Conclusions and Future Work

Why is XT worse than YT? Because mandibular deformation in speech production primarily manifests itself in the YT direction. Anything combined with XT degrades performance

References

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

[2] C.J. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[3] 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.

[4] 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 Transactions on*, 15(10):2879–2891, 2006.

[5] H.E. Çetingül, Y. Yemez, E. Erzin, and A.M. Tekalp. The use of lip motion for biometric speaker identification. In *SIU*, pages 148 – 151, 2004.

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

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

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

[9] M. I. Faraj and J. Bigun. Lip biometrics for digit recognition. In *CAIP*, pages 360 – 365, 2007.

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

[11] M.I. Faraj and J. Bign. Audio-visual person authentication using lip-motion from orientation maps. *Pattern Recognition Letters*, 28(11):1368–1382, 2007.

[12] M.I. Faraj and J. Bign. Synergy of lip-motion and acoustic features in biometric speech and speaker recognition. *IEEE Trans. Computers*, 56(9):1169–1175, 2007.

[13] E. Gomez, C.M. Travieso, J.C. Briceno, and M.A. Ferrer. Biometric identification system by lip shape. In *Security Technology, International Carnahan Conference on*, pages 39 – 42, 2002.

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

[15] J. Kittler, Y. Li, J. Matas, and M.U.R. Sánchez. Lip-shape dependent face verification. In *AVBPA*, pages 61–68, 1997.

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

[17] T.N.U. Maheswari. *Lip Prints*. PhD thesis, Saveetha Dental College and Hospitals, 1995.

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

[19] M. Pandit and J. Kittler. Feature selection for a dtw-based speaker verification system. In *ICASSP*, volume 2, page 769772, 1998.

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

[21] R.O.Duda, P.E.Hart, and D.G.Stork. *Pattern Classification*. Wiley, 2001.

[22] 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.

[23] 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.

[24] K. Suzuki, Y. Tsuchihashi, and H.Suzuki. A trail of personal identification by means of lips print. *I. Jap. J. Leg. Med.*, 23:324–325, 1969.

[25] 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.

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

[27] Cootes T.F., Taylor C.J., Cooper D.H., and Graham J. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(8):36–59, 1995.

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

[29] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.