Sparse Coding Based Lip Texture Representation For Visual Speaker Identification

Jun-Yao Lai, Shi-Lin Wang, Xing-Jian Shi
School of EIEE
Shanghai Jiao Tong University
Shanghai, China
tambdc,wsl@sjtu.edu.cn

Alan Wee-Chung Liew
School of Info. and Comm. Technology
Griffith University
Brisbane, Australia
a.liew@griffith.edu.au

Abstract—**Recent research has shown that the speaker's lip shape and movement contain rich identity-related information and can be adopted for speaker identification and authentication. Among all the static lip features, the lip texture (intensity variation inside the outer lip contour) is of high discriminative power to differentiate various speakers. However, the existing lip texture feature representations cannot describe the texture information adequately and provide unsatisfactory identification results. In this paper, a sparse representation of the lip texture is proposed and a corresponding visual speaker identification scheme is presented. In the training stage, a sparse dictionary is built based on the texture samples for each speaker. In the testing stage, for any lip image investigated, the lip texture information is extracted and the reconstruction errors using all the dictionaries for every speaker are calculated. The lip image is identified to the speaker with the minimum reconstruction error. The experimental results show that the proposed sparse coding based scheme can achieve much better identification accuracy (91.37% for isolate image and 98.21% for image sequence) compared with several state-of-the-art methods when considering the lip texture information only.**

Keywords—Lip texture; visual speaker identificaiton; sparse coding; lip biometrics.

I. INTRODUCTION

In the past several years, biometric features, such as fingerprint, iris and human face, have been widely used for human identity identification and authentication. Recent study [1-9] have shown that visual information about the lip region and its movement contain abundant speaker identity related information and can be regarded as a new biometric feature in many multi-modal person verification systems.

Different from some widely used biometric features such as the fingerprint, human face, etc., the lip biometrics contain both physiological and behavioral information about the speaker's identity. Different people have different lips. The unique lip shape and texture (intensity variation of the lip region) can be regarded as the physiological lip features. On the other hand, the lip movements during utterances which reflect the distinctive talking styles can be regarded as the behavioral lip features.

Many researches have proposed various lip feature representations for speaker authentication and identification [3-7]. For the physiological part, Luettin et al. [3] employ the Active Shape Model (ASM) to describe the outer lip contour and both the lip shape and intensity profile along the contour points have been adopted to describe the static human lip. Broun et al. [4] incorporated the inner mouth information (visibility of the teeth and tongue during utterance) into the geometric lip shape descriptors to better describe the lip region. Matthews et al. [5] have proposed the Active Appearance Model for lip modeling and both the lip shape and the intensity variation inside the outer lip contour (referred to as the lip texture) have been adopted to represent the lip region. In our previous work [6], we have demonstrated that the ICA-based presentation of the lip features (both shape and texture) can be a better choice compared with the traditional PCA-based representations in [5]. Recently, Goswami et al. proposed a new local texture descriptor (LOCP) to describe the lip texture and achieved a satisfactory performance for speaker identification [7]. For the behavioral part, the first order derivatives of the static features have been widely used to infer the dynamic information [3-6].

Based on the conclusions in [8,9], the static lip textures contain abundant information related to the speaker identity. However, the identification results solely based on the lip texture information from an isolate lip image are usually very low compared with the dynamic features [9]. In this paper, we have employed the sparse coding technique to provide a representative lip texture descriptor for each speaker and designed an efficient and accurate visual speaker identification scheme accordingly. The identification results have demonstrated that the proposed scheme solely based on the lip texture information can achieve a highly reliable performance for visual speaker identification.

The paper is organized as follows. Section II briefly presents the methods to obtain the lip texture information based on our previous work. In section III, the proposed lip texture representation based on the sparse coding technique has been elaborated. The new visual speaker identification scheme based on the proposed feature is also presented in this section. The optimal parameter selection of the proposed feature has been discussed in Section IV and the identification results by the proposed method compared with several state-of-the-art techniques are also given in this section. Finally, Section V draws the conclusion.

978-1-4799-4612-9/14/$31.00 © 2014 IEEE

II. LIP TEXTURE EXTRACTION

A. Lip Outer Contour Extraction

In order to extract the lip texture information, the lip region should be accurately located first. In our previous work, a robust lip region segmentation [10] and a lip contour extraction algorithm [11] have been proposed to extract the lip outer contour. For a lip image, the MS-FCM algorithm [10] is first applied to segment the entire image into two regions, i.e. the lip region and non-lip region. Based on the segmentation result, a point-driven lip contour extraction scheme [11] is adopted to obtain the lip contour in the image. A 5-2-7 lip model (5 points representing the lower lip contour, 2 points for the lip corners and 7 points for the upper lip contour) is adopted to describe the outer lip contour. Fig. 1 illustrates some lip outer contour extraction results using [10] and [11].

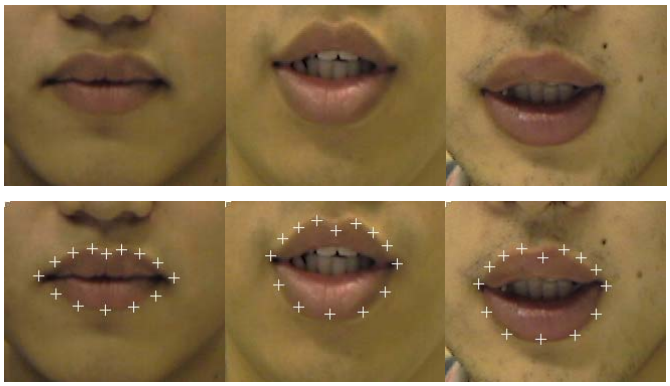

Figure 1. Lip outer contour extraction results. The first row: the original images; the second row: the extraction results.

B. The Normalized Lip Texture Extraction

Since the main scope of this paper is focused on the lip texture information, variations caused by different lip shapes, lighting conditions, speaker's poses during utterances, distances between the speaker and the camera, etc. should be excluded from investigation. Hence, the following normalization procedure has been carried out to extract the lip texture information.

First, a linear projection has been performed to align the extracted lip outer contour with the reference shape and the entire region inside the outer lip contour is then projected onto the reference shape. The original lip texture feature (denoted by $\boldsymbol{I}$) is constructed by concatenating all the intensity values inside the reference shape in a specific order. The detailed lip shape alignment method can be found in [11].

Then, an intensity normalized scheme in [5] is applied to obtain the normalized lip texture feature (denoted by $\boldsymbol{I}_{normalized}$), as given in Eqn.1 & 2.

$$\boldsymbol{I}_{normalized} = (\boldsymbol{I} - mean\_i \cdot \boldsymbol{1})/amp \quad (1)$$

$$mean\_i = \boldsymbol{I} \cdot \boldsymbol{1}/m,\ amp = \boldsymbol{I} \cdot \boldsymbol{I}_{ref} \quad (2)$$

where is $\boldsymbol{I}_{ref}$ is the reference intensity vector, *mean_i* is the average intensity value of $\boldsymbol{I}$, m is the number of elements in the vector $\boldsymbol{I}$. Fig.2 illustrates some normalized lip texture extraction results. Due to there are negative values in the features, the normalized features are linearly projected onto [0,1] when illustrating using grayscale images.

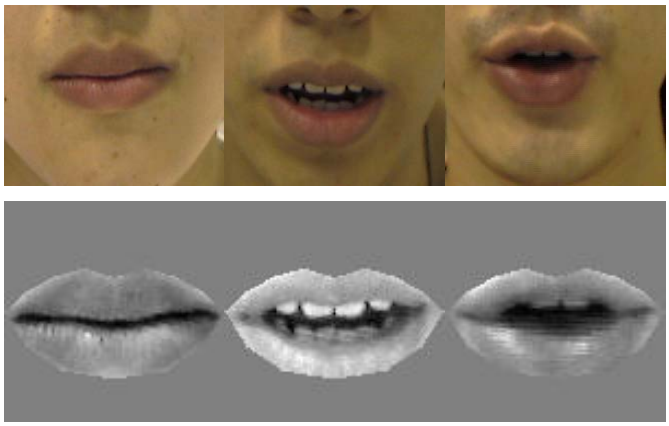

Figure 2. Lip texture extraction results. The first row: the original images; the second row: the normalized texture.

III. LIP TEXTURE REPRESENTATION BY SPARSE CODING

A. Sparse Representation of Lip Textures

Sparse coding [12] is a generative model for various kinds of signals. The model describes a specific signal by a combination of atomic elements in a certain dictionary. The sparsity of the model limits the number of atomic elements used in the description of a single signal.

The normalized lip texture feature obtained in Section *II-B* is of high dimension, which leads to low efficiency in information representation and calculation. Hence, the sparse coding technique is adopted to provide an efficient representation for lip textures, which help reduce the redundancy and keep the key information as well. Given a lip texture feature $\boldsymbol{I}$, the sparse code w can be obtained by Eqn. 3 as follows:

$$w = \text{argmin}_c \|\boldsymbol{I} - \boldsymbol{D}c\|_{l^2} \ subject\ to\ \|c\|_{l^0} \leq s \quad (3)$$

where the original feature $\boldsymbol{I}$ is represented by a weighted summation of several elements in the dictionary $\boldsymbol{D}$ with the sparsity s and $\|\cdot\|_{l^2}$ and $\|\cdot\|_{l^0}$ denote the *l-2* and *l-0* norm distance, respectively. The final coding result w is obtained by minimizing the reconstruction error.

The dictionary used in sparse coding usually reflects the underlying characteristics of the input signal. In our approach, a series of texture features for a specific speaker are selected to build the speaker's texture feature dictionary. Considering both accuracy and efficiency, the K-SVD [12] is employed to generate the dictionary from the training texture features. Given N speakers investigated, a set of N dictionaries, i.e. $\{D_1, D_2, \ldots, D_N\}$ can be obtained.

B. Speaker Identification Scheme Based on Sparse Coding

Based on the analysis in *Section III-A*, the sparse coding technique provides an efficient way to represent the high-dimensional lip texture feature with a low-dimensional sparse code. The original lip texture can be reconstructed directly with

the knowledge of the sparse code and the dictionary. Using an appropriate dictionary which well matches the lip texture, the reconstruction error will be small. On the other hand, if the dictionary adopted is irrelevant to the lip texture, the reconstruction error will be large. Fig.3 provides an illustration for such phenomenon. From Fig.3, it is observed that the reconstructed lip texture by the speaker's dictionary better matches the original lip texture.

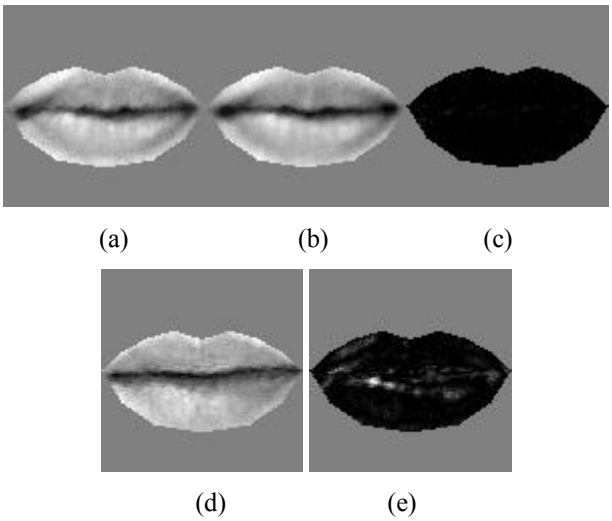

Figure 3. (a) The original lip texture; (b) & (d): The reconstructed lip textures using (b) the speaker's dictionary and (d) the other speaker's dictionary, respectively; (c) & (e): The reconstruction error using (c) the speaker's dictionary and (e) the other speaker's dictionary, respectively.

Based on the fact mentioned above, a new speaker identification scheme based on sparse coding has been proposed, which runs as follows:

i) For a lip image investigated, extract the normalized lip texture feature (denoted by $\boldsymbol{I}$) using the methods introduced in *Section II*.

ii) For the dictionary of *i-th* speaker (denoted by D_i, i=1,2…,N), calculate the sparse code w_i based on Eqn.3.

iii) Reconstruct the lip texture (denoted by $\boldsymbol{I_{rec,i}}$) with w_i and D_i. Calculate the reconstruction error by $\boldsymbol{E_{rec,i}} = \left\|\boldsymbol{I} - \boldsymbol{I_{rec,i}}\right\|_{l^2}$

iv) Finally, the lip image is identified to the *k-th* speaker with the minimum reconstruction error, i.e., $k = \text{argmin}_i\, \boldsymbol{E_{rec,i}}$

IV. Experimental Results

A. Database and Experimental setup

To comprehensively evaluate the performance of the proposed lip identification approach, the database in [9] has been used for investigation. The database contains 40 speakers (29 male and 11 female) and each speaker is asked to repeat the phrase "3725" for ten times. This phrase is selected to cover various movement and different shape of human lip and provides rich information for human identification. Each sequence contains 90 lip images of size 220×180 lasting for three seconds.

Three sequences (including 3×90=270 lip images) for each speaker are randomly selected to train the speaker's dictionary. The rest seven sequences are used for testing.

B. Discriminative Ability of the Reconstruction Error

In order to investigate the discriminative ability of the reconstruction error used in our identification scheme, a speaker is randomly selected from the database. All the reconstruction errors of the speaker's testing samples using the speaker's dictionary and the other 39 speaker's dictionaries are analyzed. Fig. 4 shows the distributions of the reconstruction errors, where the blue line shows the distribution of the reconstruction error using the speaker's dictionary and the red line shows the distribution using the other speaker's dictionaries. From the figure, it is observed that the two distributions are well separated to some extent, which demonstrates that the reconstruction error can be a good measure to differentiate the speaker's identity.

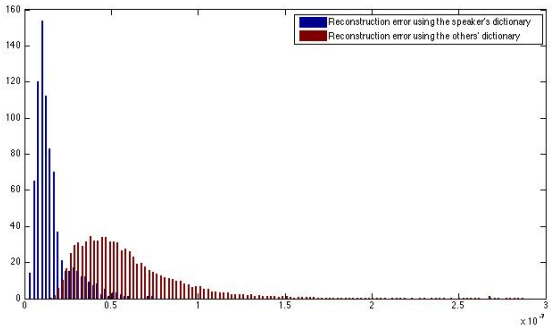

Figure 4. Distribution of reconstruction error

C. Selection of the Size of the Dictionary

The size of dictionary is an important parameter in sparse coding, which influence the efficiency and accuracy in reconstruction and classification. In order to evaluate the influence caused by this parameter and thus select an optimal value, the following experiments have been carried out. Four kinds of sizes, i.e. 8, 16, 32, 64, are investigated and the identification accuracies are given in Table I. From the table, it is observed that the size of 32 provides the best identification results. In addition, the identification accuracy does not change much (less than 2%) with different selections of the size, which demonstrates the proposed method is not sensitive to the parameter selection.

TABLE I. Speaker Identification Accuracy For Isolate Frames In % Using Various Sizes of the Dictionary

Size of the Dictionary	8	16	32	64
Identification Accuracy	89.76	91.35	91.37	90.33

D. Identificaion Performance Comparison with Existing Techniques

In order to evaluate the performance of the proposed speaker identification scheme, three widely-used lip features, i.e. Mathews et al.'s texture feature [5] (Mathews's in short),

Wang and Liew's ICA texture feature [6] (Wang's in short), and Chan et al.'s LOCP feature [7] (Chan's in short), are adopted for comparison.

Note that in our experiments for all the features, three sequences are adopted for training and the remaining seven sequences for testing. Two kinds of identification accuracies are analyzed to evaluate the performance of the identification schemes: i) the identification accuracy for an isolated lip image; ii) the identification accuracy for a lip image sequence (containing 90 images). Since only the static lip texture feature is analyzed, the identification result for a lip sequence can be obtained by majority voting over the identification results for all the lip images in the sequence.

TABLE II. SPEAKER IDENTIFICATION ACCURACY IN % BY VARIOUS KINDS OF SPEAKER IDENTIFICATION SCHEMES

Identification scheme	***Isolated Images***	***Image Sequences***
Matthews's	60.22	86.86
Wang's	61.34	87.27
Chan's	54.58	85.71
The proposed scheme	91.37	98.21

Table II shows the identification results for all the schemes investigated. It is observed from the table that the proposed sparse coding based scheme achieves much better identification performance compared with the other three approaches investigated. The high identification accuracy demonstrates that sparse coding can provide a fine representation of various lip texture in an efficient manner. Compared with the existing lip texture representations (PCA-based for Matthews's, ICA-based for Wang's and local texture based for Chan's), the proposed sparse representation shows advantages in both accuracy and feature dimension (the sparse code is only 32-dimensional).

V. CONCLUSION

This paper proposes a sparse coding based lip texture representation. Based on the representation, a new visual speaker identification scheme solely based on the static lip texture has also been designed. For each speaker, the sparse coding technique is adopted to learn a representative dictionary from the training samples of the speaker's lip texture. Then the reconstruction error is adopted as a discriminative measure to differentiate various speakers based on the fact that when using the appropriate dictionary (the speaker's own dictionary), the reconstructed lip texture can well match the original one while large reconstruction error will occur when using the inappropriate dictionary (the other speaker's dictionary). From the experimental results, it is observed that the proposed identification scheme outperforms the three approach investigated.

ACKNOWLEDGMENT

The work described in this paper is supported by NSFC Fund (61271319), Key Laboratory of Infrared System Detection and Imaging Technology, Shanghai Institute of Technical Physics, Chinese Academy of Sciences, and Key Lab of Information Network Security, Ministry of Public Security.

REFERENCES

[1] A. Kanak, E. Erzin, Y. Yemez, A.M. Tekalp, "Joint audio-video processing for biometric speaker identification", Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 3, pp. 561-564, Hong Kong SAR, China, July 2003.

[2] H. E. Cetingul, Y. Yemez, E. Erzin and A. M. Teklap, "Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading", IEEE Transactions on Image Processing, vol. 15, issue 10, pp. 2879-2891, Oct. 2006.

[3] J. Luettin, N. A. Thacker and S. W. Beet, "Learning to recognise talking faces", Proceedings of 13th International Conference on Pattern Recognition, vol. 4, pp. 55-59,Vienna, Austria, Aug. 1996.

[4] C. C. Broun, X. Zhang, R. M. Mersereau, and M. Clements, "Automatic speechreading with application to speaker verification," Proceedings of International Conference on Acoustics, Speech and Signal Processing, vol.1, pp. 685–688, 2002.

[5] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, R. Harvey, "Extraction of visual features for lipreading", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.24, issue 2, pp. 198-213, Feb. 2002.

[6] S. L. Wang and A. W. C. Liew, "ICA-Based Lip Feature Representation for Speaker Authentication", Proceedins of Third International IEEE Conference on Signal-Image Technologies and Internet-Based System, pp. 763-767, 2007.

[7] C. H. Chan, B. Goswami, J. Kittler and W. Christmas, "Local Ordinal Contrast Pattern Histograms for Spatiotemporal, Lip-Based Speaker Authentication", IEEE Transactions on Information Forensics and Security, vol. 7, issue 2, pp.602-612, 2012.

[8] J. S. Mason and J. D. Brand, "The Role of Dynamics in Visual Speech Biometrics", Proceedings of International Conference on Acoustics, Speech and Signal Processing, vol.4, pp. 4076-4079, May 2002.

[9] S. L. Wang and A.W. C. Liew, "Physiological and behavioral lip biometrics: A comprehensive study of their discriminative power", Pattern Recognition, vol. 45, issue 9, pp. 3328-3335, Sept. 2012.

[10] S. L. Wang and A.W.C. Liew, "Robust Lip Region Segmentation for Lip Images with Complex Background", Pattern Recognition, vol.40, no. 12, pp. 3481-3491, Dec. 2007.

[11] K.L. Sum, W.H. Lau, S.H. Leung, A.W.C Liew and K.W. Tse, "A new optimization procedure for extracting the point-based lip contour using active shape model", Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, vol.3, pp.1485-1488, Salt Lake City, USA, May 2001.

[12] M. Aharon, M. Elad and A. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation", IEEE Transactions on Signal Processing, vol. 54, issue 11, pp. 4311-4322, 2006.