A Survey of Visual Lip Reading and Lip-Password Verification

Seksan Mathulaprangsan[1], Chien-Yao Wang[1], Aufaclav Zatu Kusum[1], Tzu-Chiang Tai[2], and Jia-Ching Wang[1]
[1]Department of Computer Science and Information Engineering,
National Central University, Taoyuan, Taiwan, R.O.C.
[2]Department of Computer Science and Information Engineering,
Providence University, Taichung, Taiwan, R.O.C.
mr.sekk@gmail.com, x102432003@yahoo.com.tw, 102522607@cc.ncu.edu.tw, tctai@pu.edu.tw, and jcw@csie.ncu.edu.tw

***Abstract*— In this paper, we have reviewed the main process of visual lip reading and lip motion password (simply called lip-password hereinafter) verification that is the useful and flexible method to apply in many applications, especially in security field since it can do double checks to verify both the speaker and his/her password. The reviewed content includes selectively significant preprocessing, visual feature extractions, and classification schemes that are all important parts in this tasks' processes.**

***Keywords*— Lip-password, lip biometric, visual lip reading, visual speech recognition, visual speaker identification**

I. Introduction

In the last decade, biometric features have been very popular to use in the research area about human recognition and verification because they provide rich information to use in classification step. There are many types of biometric features e.g. face, fingerprint, iris, lip, speech, etc. Visual lip reading is the technique to analyze speech by transforming the movement and/or the concrete appearance of the any physical biometric parts that making speech to be visual features. Normally, we use this technique when the normal sound is not available or the speech signal is incomplete e.g. in noisy environment. This method can be applied in various applications of security e.g. password entry in surveillance system [5], user authentication system [15], mobile phone apps authentication [16], word recognition [12], and hearing impaired persons [8], [11].

Recently, lip motion password (simply called lip-password hereinafter) verification has introduced [1], [6] as the quite new biometric scheme to verify human which composes of a password embedded in speaker's utterances and the underlying characteristic of speaker's lip motion. This technique uses not only audio modality from the speaker's speech, but also the video modality of lip motion extracted at the same time. Therefore, we can use this technique to improve the security level because the system can do double checks to verify the speaker: the right target speaker and his/her right password. From this prominent point, lip-password verification can be applied in various applications in security protection, human-computer interfaces, authentication, and so forth.

In the rest of this paper, we have reviewed about the main steps of visual lip reading and lip-password verification that starts by localization and segmentation. Then the visual feature extractions have been discussed. After that, the classification schemes are illustrated. Finally the conclusion of the paper has presented.

II. Localization and Segmentation

In visual lip reading as well as lip-password verification, similar to other image processing tasks, the preprocessing is the vital step that mainly affects to the verification and the recognition performances. In this section, we have discussed about two important tasks in preprocessing step including localization and segmentation.

A. Localization

The preprocessing to seek for the location of lip region is the key issue in the automatic lip-reading process that is the important part of lip-password task [9]. The challenges of localization are the variety of many factors that affect the quality of lip images such as changing illumination, various face and mouth poses, image with shadow. Therefore, to reach high recognition rate, the lip-reading system performance is depended on the preciseness of the lip localization. In [23], they have used multi-class, shape-guided FCM (MS-FCM) algorithm to localize the input images. This algorithm can divide the whole image into two parts, lip and non-lip regions, and uses the image's spatial information as well as the color to make description. Moreover, this algorithm can deal with the difficult image, including beards, images noise, and ambiguity. In [9], they solved the problem of illumination and shadow of face when do lip localization that is one of the most challenging in localization. They have used the close contrast values to represent the gray level of both mouth area and shadow. From their proposed method, the gray level image is divided into two parts i.e. lip and surrounding skin region. They then applied contrast stretching adjustment on both sub-image parts. Whereas Cheung *et al.* [13] used the framework of localized color active contour model (LCACM) that is used to find the contour line and localize the foreground (mouth part) and the background regions in different color spaces.

B. Segmentation

Lip motion segmentation is the process to determine the proper starting and ending frames of the distinguishable subunit of utterance. This preprocessing is the key for visual lip reading and lip-password verification. In [1], they have employed the forward-backward filtering technique to analyze the input signal in both directions i.e. forward and backward. Then they can attain the positions of peak points and valley

978-1-4673-8237-3/15/$31.00 ©2015 IEEE

points that refer to the status of mouth opening and mouth closing. After that, they focused on the valley points that represent to the connecting position of two adjacent subunits and used them to create subunits of distinguishable utterances that will be transformed to the feature.

III. Visual Feature Extractions

According to automatic lip reading and lip-password verification, in this paper, we have focused on visual feature extraction of lip and lip motion. There are several ways to categorize visual feature extraction methods in lip recognition. In general visual feature extractions can be roughly divided into three branches [1]: contour-based feature, appearance-based feature, and motion-based features. The first type of lip visual features is extracted from the geometric shape information e.g. mouth's height and width, mouth area, and perimeter [22]. The second type feature considers on the physical aspect of the speaker such as teeth and tongue that appear between utterances. The motion-based features are extracted from the sequential images of lip movement. Moreover, another interesting way to categorize visual features is based on physiological and behavioral properties [4]. The physiological lip feature is derived from appearance characteristics of the speaker's lip. In general, it is extracted from an isolated lip image and normally static. On the other hand, when the speaker makes lip movement during utterance, the behavioral lip feature is extracted from this temporal information. So that this feature is derived from a sequence image of lip that contain temporal information and normally dynamic.

However, if we focus on problem-oriented perspective, visual features extractions can be categorized into three branches [3]:

A. *Speaker Dependency* [3]

This kind of feature consider on the different appearance of the speakers. Generally, it uses the information upon the speakers' visual variability, as illustrated in Fig. 1, to do extraction. Almost speaker dependency features use the linear discriminant analysis to do feature extraction because this technique has good discriminative power for classification. There are some previous research that also used the linear discriminant methods in feature extraction step such as discrete wavelet transform (DWT) [15], discrete cosine transform (DCT) [1], [15], two dimensional discrete cosine transform (2D-DCT) [2], and principle component analysis (PCA) [1] to extract features.

In [20], they have proposed the alternative way to extract the feature by using the image transformation on ROI images. Then they removed the mean from feature vectors' output by the transformation over each utterance. After that they reduced the dimensionality of these features by LDA. Finally, they extended their method through applying the inter-frame LDA on the concatenation of consecutive feature vectors output by the previous LDA which referred to the intra-frame.

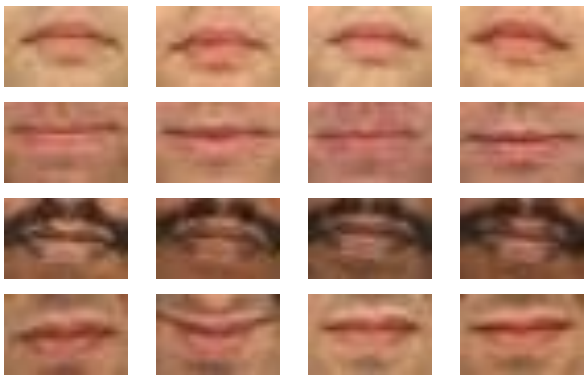

Fig. 1. Sample mouth images from the XM2VTS database [24].

B. *Pose Variation* [3]

These kind of features are the extraction methods that concern the pose variation among the speaker relative to the camera view that can affect to the appearance of the speaking mouth. Rather than focusing on the frontal view (FV) image, these extraction method try to compensate the angle variation i.e. the non-frontal view (NFV) image. The pose-dependent features (PDFs) from NFV image is one of important feature in this category. Moreover, the pose-independent features (PIFs) that transformed from the PDFs is another method that more flexible to apply with the FV and NFV images [3].

For PDFs extracting method, the advantages are lossless information and no noise addition during the transformation process. In another way, the disadvantage is that this system needs more train for every particular system according to PDFs. These methods may suffer from the lack of representative training data for a particular view. Hence, we may have to collect extra NFV data for training. On the contrary, the latter methods attempt to transform PDFs into a common PIF space such that they are comparable. Therefore, only one system needs to be trained based on the available training data. However, they often suffer some substantial performance drop due to the information loss or added noise caused by the feature transformation.

C. *Temporal Information* [3]

Since the human speaking is a dynamic process that happen in a period of time, we can get the temporal information for dynamic mouth shape deformation while uttering. In [19], linear discriminant analysis (LDA) was proposed to extract the feature by concatenate sequential feature vectors and employ LDA to obtain the final compact features encoding temporal information. Because the traditional linear approach may not be sufficient to capture dynamic linear information, in [7], they use optical flow that is computed by using two consecutive frames to capture the motion information.

A video can be viewed as a sequence or a stack of images that composes a particular row/column of each frame along the temporal axis. From this concept, Zhao and Pietikäinen [18] exploited the texture information within temporal patterns (TPs) to characterize video dynamics. They have applied the

local binary pattern (LBP) descriptors on both TPs to capture temporal information and video frames to extract spatial information. To represent the dynamic feature in visual speech, they proposed a spatiotemporal LBP histograms, named LBP-TOP [18], which derived from a dynamic texture representation method named three orthogonal planes. Then they used this technique for dynamic texture classification. Moreover, Chan *et al*. [14] proposed the new texture descriptors, called local ordinal contrast pattern (LOCP). Similar to [18], they applied LOCP with the TOP dynamic texture representation, called LOCP-TOP, which got a high score in speaker authentication task on XM2VTS dataset.

IV. Classification Schemes

A visual lip reading is complicated classification problem that can be applied relevant classifiers depend on the extracted feature. Normally, there are two classifiers that are used in many researches i.e. hidden Markov models (HMMs) and Gaussian mixture models (GMMs). Moreover, some modern researches try to increase the classification performance by boosting them [1], [10].

A. Hidden Markov Models

The HMM classifier is widely used in temporal pattern recognition [19]. In the previous works, HMMs were successfully used in acoustic signal classification tasks [25] and audio-visual classification/verification [7], [22]. However, nowadays, they are also used in visual lip reading and lip-password verification [1], [4], [6]. The HMM is a finite set of states that consists of a first-order Markov chain. Each of its states are considered to be hidden from the observer. Normally, these hidden states are the random process that can be used to generate the observation sequence. From this ability, the temporal structure of the data, which have the consecutive characteristic, can be captured by these hidden states [17].

To model these relations mathematically, let $x_1, \ldots, x_T$ be a succession of observable data vectors, we can apply HMM to assume that the appearance of a hidden Markov chain is to generate these observable data vectors. Then the HMM parameters can be symbolized by using K to represent the state number, $\pi_i = 1, \ldots, K$ to represent the initial state probabilities using in the hidden Markov chain, and a_{ij}, $i = 1, \ldots, K$ and $j = 1, \ldots, K$ stand for the state transition probability from state i to state j. In general, we can use the ML principle to estimate the HMM parameters. Assigning $s_1, \ldots, s_T$ are the true state sequence, we can compute the likelihood of the observable data by [17]

$$p(x_1 s_1, \ldots, x_T s_T) = \pi_{s_1} b_{s_1}(x_1) a_{s_1,s_2} b_{s_2}(x_2), \ldots, a_{s_{T-1},s_T} b_{s_T}(x_T)$$
$$= \pi_{s_1} b_{s_1}(x_1) \prod_{t=2}^{T} a_{s_{t-1},s_t} b_{s_t}(x_t) \quad (1)$$

where the ith state's observation density can be derived by $b_i(x_t) \equiv P(x|s_t = i)$. One of important characteristics, this observation density can be consider to be both discrete and continuous. When it is applied to discrete HMM, it is discrete. On the other hand, when it is applied to continuous HMM, it is a mixture of Gaussian densities. In general we do not know the HMMs' true state sequence, therefore we use another way to find the likelihood of a given data sequence by do summation over all sequences of possible states that can be written as [17]

$$p(x_1, \ldots, x_T) = \sum_{s_1,\ldots,s_T} \left(\pi_{s_1} b_{s_1}(x_1) \prod_{t=2}^{T} a_{s_{t-1},s_t} b_{s_t}(x_t) \right) \quad (2)$$

Moreover, in some modern researches, they try to improve the performance of HMMs classifier. In [1], they have proposed multi-boosted HMMs learning approach that is used with their proposed segmentation. From this method, the discrimination power is better than a single HMM classifier significantly.

B. Gaussian Mixture Models

GMMs are the classification approach that normally used to related visual feature such as audio-visual speech recognition [26] and speech spectral features [27]. GMMs are statistical approach that is normally applied for implementing maximum likelihood estimation in classification step. When contain only one state, they become a special case of continuous HMM. Consequently, their requirements for training and testing are much less than the requirements of a general continuous HMM [17].

From their properties, they are suitable to be applied for physiological visual lip features [4] that their input data are static. However, to classify the temporal structure's features, GMMs are not relevant to model temporal information of the training data since the assumption that all vectors are independent is applied to all the training and testing equations. Moreover, GMMs can be used with Baum-Welch re-estimation in training process to improve the recognition accuracy in audio-visual person authentication [7].

V. Conclusion

In this paper, we have reviewed the whole process of visual lip reading and lip-password verification, including some significant preprocessing, feature extraction methods, and classification schemes. In the preprocessing, the essential works are localization and segmentation. The prior task is about how to locate the lip region and separate it from the background. The latter task focuses on selecting the appropriate starting and ending frames that contain the desired information, distinguishable utterance.

For the feature extraction methods, there are many types of visual speech feature extractions as well as the way to categorize them. However, we have classified them based on the problem-oriented perspective that divides them into three categories: speaker dependency, pose variation, and temporal information. In the speaker dependency, the visual variability of the speaker is the vital information to extract the features. The next category, pose variation, is the method that focuses on how to deal with the variation of the mouth pose. The last group, temporal information, is the processing of temporal information of a speaker's utterance normally derived from mouth shape deformation.

The classification schemes in visual lip reading and lip password verification depend on what extracted features. Many researches in this field use one from two classifiers: Hidden Markov models (HMMs) and Gaussian mixture models (GMMs). HMMs are relevant to use with the feature in temporal information type while GMMs are more appropriate to apply to physiological features that deal with the static data input.

REFERENCES

[1] X. Liu and C. Yiu-ming "Learning multi-boosted HMMs for lip-password based speaker verification," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 2, pp. 233-246, Feb. 2014.

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

[3] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, "A review of recent advances in visual speech decoding," *Image and Vision Computing Trans.*, vol. 32, no.9, pp. 590-605, Sep. 2014.

[4] S. L. Wang, A. Wee, and C. Liew, "Physiological and behavioral lip biometrics: a comprehensive study of their discriminative power," *Pattern Recognition Trans.*, vol. 45, no. 9, pp. 3328-3335, Sep. 2012.

[5] H. A. Mahmoud, F.B. Muhaya, and A. Hafez, "Lip reading based surveillance system," in *Procs. IEEE Int. Future Infor. Technol.*, May 2010, pp. 1-4.

[6] X. Liu and C. Yiu-ming "A multi-boosted HMM approach to lip password based speaker verification," in *Procs. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 2197-2200, Mar. 2012.

[7] M. I. Faraj and J. Bigun, "Audio–visual person authentication using lip-motion from orientation maps," *Pattern Recognition Letters*, vol. 28, no.11, pp. 1368-1382, Aug. 2007.

[8] Y. L. Lay, C.H. Tsai, and H.J. Yang, "The application of extension neuro-network on computer-assisted lip-reading recognition for hearing impaired," *Expert Systems with Applications Trans.*, vol. 34, no.2, pp. 1465-1473, Feb. 2008.

[9] M. Li and Y.M. Cheung, "Automatic lip localization under face illumination with shadow consideration," *Signal Processing Trans.*, vol. 89, no. 12, pp. 2425-2434, Dec. 2009.

[10] M. Deypir, S. Alizadeh, T. Zoughi, and R. Boostani, "Boosting a multi-linear classifier with application to visual lip reading," *Expert Systems with Applications Trans.*, vol. 38, no. 1, pp. 941-948, Jan. 2011.

[11] N. Puviarasan and S. Palanivel, "Lip reading of hearing impaired persons using HMM," *Expert Systems with Applications Trans.*, vol. 38, no. 4, pp. 4477-4481, Apr. 2011.

[12] S. Jongju, L. Jin, and K. Daijin, "Real-time lip reading system for isolated Korean word recognition," *Pattern Recognition Trans.*, vol. 44, no. 3, pp. 559-571, Mar. 2011.

[13] Y. M. Cheunga, X. Liua, and X. Youb, "A local region based approach to lip tracking," *Pattern Recognition Trans.*, vol. 45, no. 9, pp. 3336-3347, Sep. 2012.

[14] C. H. Chan, B. Goswami, and J. Kittler, "Local ordinal contrast pattern histograms for spatiotemporal, lip-based speaker authentication," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 2, pp. 602-612, Apr. 2012.

[15] V. Gupta and S. Sengupta, "Automatic speech reading by oral motion tracking for user authentication system," in *Procs. IEEE Int. Advanced Computing and Communication Technol.*, Apr. 2013, pp. 50-54.

[16] F. S. Lesani, F.F. Ghazvini, and R. Dianat, "Mobile phone security using automatic lip reading," in *Procs. IEEE Int. Conf. e-Commerce, e-Business*, Apr. 2015, pp. 1-5.

[17] M. E. Ayadi, M.S. Kamel, and F. Karray, "Survey on speech emotion recognition: features, classification schemes, and databases," *Pattern Recognition Trans.*, vol. 44, no.3, pp. 572-587, Mar. 2011.

[18] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915-928, Jun. 2007.

[19] G. Potamianos, C. Neti, and G. Gravier, "Recent advances in the automatic recognition of audiovisual speech," in *Procs. IEEE Int. Conf.* Mar. 2003, vol. 91, no. 9, pp. 1306-1326,.

[20] G. Potamianos, C. Neti, G. Iyengar, A. Senior, A. Verma, "A cascade visual front end for speaker independent automatic speechreading," *Int. J. Speech Technol.*, vol. 4, pp. 193-208, 2001.

[21] X. Liu and C. Yiu-Ming, "An exemplar-based hidden Markov model with discriminative visual features for lipreading," in *Proc. Tenth int. Conf. Computational Inteligence and Security (CIS)*, Nov. 2014, pp. 90-93.

[22] T. J. Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," *IEEE Trans., Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 1082-1089, May 2006.

[23] W. Shi-Lin, L. Wing-Hong, L. A. Wee-Chung, and L. Shu-Hung, "Robust lip region segmentation for lip images with complex background," *Pattern Recognition Trans.*, vol. 40, no. 12, pp. 3481-3491, Mar. 2007.

[24] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. "Xm2vtsdb: the extended m2vts database." *in Proc. Second Int. Conf., Audio and Video-based Biometric Person Authentication*, Washington, USA, Mar. 1999, pp. 72-77.

[25] A. Jitendra, M. Iain, and B. Herve, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework." *Speech Communication Trans.*, vol. 40, no. 3, pp. 351-363, May 2003.

[26] C. Tsuhan, "Audiovisual speech processing," *IEEE Trans., Signal Processing*, vol. 18, no. 1, pp. 9-21, Jan. 2001.

[27] M. Zuheng, B. Denis, and F. Gang, "GMM mapping of visual features of cued speech from speech spectral features," *12th International Conference on Auditory-Visual Speech Processing (AVSP 2013)*, Aug 2013, St Jorioz, France, pp. 191-196, 2013.