REGRESSION BASED LANDMARK ESTIMATION AND MULTI-FEATURE FUSION FOR VISUAL SPEECH RECOGNITION

Hong Liu, Xuewu Zhang, Pingping Wu

Key Laboratory of Machine Perception (Ministry of Education)
Engineering Lab on Intelligent Perception for Internet of Things (ELIP)
Shenzhen Graduate School, Peking University, China
{hongliu,xuewuzhang,pingpingwu}@pku.edu.cn

ABSTRACT

Visual speech recognition also known as lipreading can improve robustness of automatic acoustic speech recognition especially under noisy environments. However, it remains a challenging topic considering the variety of speaking characteristics and confusion between visual speech features. In this paper, we propose an automatic lipreading method by using a new lip tracking method and multiple visual information fusion to tackle the problem. First, a method of face landmark estimation based on regression is employed for lip detection, based on which a geometric-based shape invariant feature (SIF) is put forward. Moreover, it can also be applied to the removal of the non-speaking utterance. Then the motion interchange patterns and spatial-temporal descriptors are also adopted to describe the lip information, where the Bayes combination strategy is applied. The proposed method is explored on three benchmark data sets: Avletters2, OuluVS and PKUVS. Experimental results demonstrate promising results and show effectiveness of the proposed approach.

Index Terms— Visual Speech Recognition, Shape Invariant Features, Motion Interchange Patterns, Bayes Combination

1. INTRODUCTION

Human speech perception is a multimodal process which involves information not only from what we hear but also from what we see. Therefore, audio visual speech recognition (AVSR) [1–3] have drawn much attention in recent years, in which the visual signals are regarded as a supplement to improve the performance of speech recognition especially when audio is corrupted or even inaccessible. Distinguished from acoustic signals, visual signals are not affected by acoustic noise and can benefit human speech perception, in cases such as human robot interaction (HRI) with no media or audio transmission.

However, visual speech recognition (VSR), also called lipreading, is still a challenging problem due to the confusion between visemes [4]. Specifically, the variations of lip shapes, style related to speaking speeds and intensities, skin texture and different accents could significantly affect spatiotemporal appearances of a speaking mouth. For lipreading, a comprehensive survey of previous studies can be found in [5]. Specifically, there have been various models for lipreading, such as PCA [6], DCT [1], AAM [7] and HMM [8] etc. From a different perspective, Zhao et al. [9] employed the local binary pattern from three orthogonal planes to extract features and Zhou et al. [10] combined these features with graph embedding techniques for lipreading. However, most of them do not pay much attention to the accurate lip detection and multi-feature representation.

We distinguish two challenges for lipreading, which are lip detection and feature extraction. The first aims at finding and tracking a specific facial part (mouth, lip contours etc), which is of crucial importance for the feature extraction. The main methods are AAM [11], ASM [12] and Haar Cascaded AdaBoost [13], which are not accurate enough for lip detection as reported in [10]. The second challenge comprises the extraction and representation of visual information. Despite of the many years of research, we have not seen any visual feature set universally accepted for representing visual speech, in contrast to the well-established features (e.g. MFCC [14]) for acoustic speech. Hence, we make improvements in both aspects.

Accordingly, contributions of this paper are: 1) Usage of regression based accurate landmark estimation for lip detection; 2) Proposal of shape invariant features and combination with other two motion based and spatial-temporal based descriptors for multi-feature fusion; 3) Establishment of a new mandarin visual speech dataset named PKUVS, which is designed to benchmark lipreading algorithms in mandarin, including similar utterances. The dataset[1] is available online.

This work is supported by National Natural Science Foundation of China (NSFC, No.61340046, 60875050, 60675025), Science and Technology Innovation Commission of Shenzhen Municipality (No.JCYJ20120614152234873, No.JCYJ20130331144631730, No.JCYJ20130331144716089).

[1] http://robotics.szpku.edu.cn/datasets/pkuvs.html

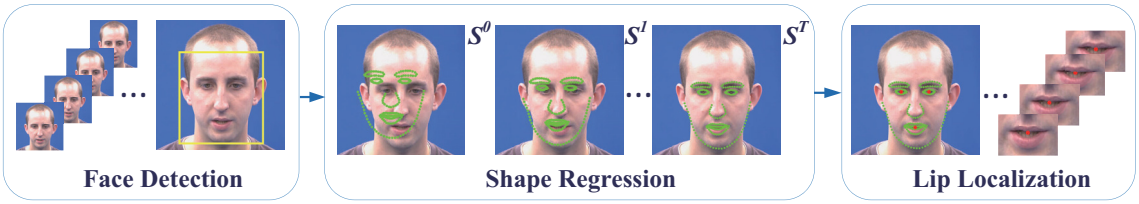


Fig. 1. Flowchart of lip detection. Left: input utterance video, then perform the coarse face detection. Middle: fast shape regression in a coarse to fine manner. Right: face alignment and lip localization.

2. LIP DETECTION BY SHAPE REGRESSION

Recently, there has been much focus on face alignment or facial feature localization [15–17]. Inspired by these works, a face shape or semantic facial landmarks are adopted for lip detection. To introduce our lip detection framework, we firstly introduce the shape regression model. Then the lip region can be cropped by our landmark detection procedure.

2.1. Shape Regression Model

A face shape $S = [x_1, y_1, ..., x_N, y_N]^T$ consists of N facial landmarks. Given a facial image I , the goal of face landmark detection is to estimate a shape S that is as close as possible to the true shape $\hat{S}$, i.e., minimizing $\left\|S - \hat{S}\right\|$. Then the boosted regression [18] is used to combine T weak regressor $(R^1, ..., R^t, ..., R^T)$ in an additive fashion.

Given an initial face shape S^0, each regressor computes a shape increment δS from image features and then updates the face shape. Accordingly, given N training examples $\left\{(I_i, \hat{S}_i)\right\}_{i=1}^{N}$, the regressors are sequentially learnt until the training error no longer decreases:

$$R^t = \arg\min_{R} \sum_{i=1}^{N} \left\| \hat{S}_i - (S_i^{t-1} + R(I_i, S_i^{t-1})) \right\| \quad (1)$$

where S_i^{t-1} is the estimated shape in previous stage.

2.2. Lip Detection

Inspired by the shape regression model, Cao et al. [16] proposed a Explicit Shape Regression (ESR) method for face alignment. Later Burgos et al. [17] proposed the Robust Cascaded Pose Regression (RCPR) method which focused on robustness to occlusion and large shape variations. Compared with [16], a smart restarts approach [17] was added for predicting failure cases early on.

However, considering that most faces in the databases for lipreading are frontal, we do not apply the occlusion mechanism and interpolated shape-indexed features [17] because of the special characteristic of databases and the need for fast computation. Similar to [16, 17], a two-level cascaded regression and correlation-based feature selection are adopted. Thus, the lip detection procedures are as follows: First, a rough face box is detected, then the landmark is estimated in a coarse-to-fine way. Next the geometric center of eyes and month can be calculated. As a result, the mouth region is cropped depending on the mouth center after normalizing the face using pre-defined ratio parameters. The flowchart of our lip detection is shown in Fig. 1.

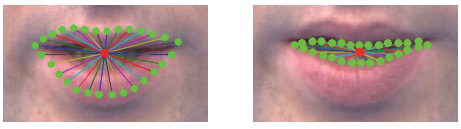
3. MULTI-FEATURES AND FUSION METHOD

Based on the accurate lip detection, we proposed a geometric-based Shape Invariant Feature (SIF), which describes the shape change information during uttering. However, features designed to describe the motion information are equally important. Therefore, The Motion Interchange Patterns (MIP) [19] is firstly adopted to describe the lip motion information. Further, the local spatial-temporal descriptor [10] is also applied. Finally, the Bayes combination strategy is employed to make full use of the multi-feature information to improve performance.

3.1. Shape Invariant Feature

Feature Extraction. As the efficiency of our shape regression model, lip shape can be accurately estimated. The problem is how to precisely represent the shape or geometric information, such as lip width, height, contour and area. With this goal in mind, the shape invariant feature is proposed to maximally use these information.

Given the lip shape $S_{lip} = [x_i, y_i, ..., x_M, y_M]^T$, M is the number of lip landmarks including the mouth center, then the Euclidean distance between two points is calculated. Denoting D_t^r is the distance matrix of the tth frame and rth type. Apparently, D_t^r is a symmetric matrix, only the upper triangular value is focused. Moreover, D_t^r is the uniform representation with different type r for the shape as shown in Fig. 2. Considering the variation from individual mouth appearance, the difference matrix ΔD_t^r is defined between two adjacent frames as Eq. (2) shows.



(a) (b)

Fig. 2. Matrix D_t^r represents different shapes or geometric information with different point size M. (a) illustrates the lip outer contour with the distance between landmark and mouth center. (b) shows the inner lip contour.

$$\Delta D_t^r = |D_{t+1}^r - D_t^r|, t \in \{1, 2, ..., T\} \quad (2)$$

where r is the kind of shape type and T is the total frames of an utterance. Then, the final representation of shape invariant features (SIF) can be obtained by transforming ΔD_t^r into a vector by concatenating its elements of upper triangular matrixes column by column. Here, both the outer and inner contour shape types with mouth center are adopted, since they contain the main shape change when one is uttering.

Removal Approach. Since the shape invariant features are extremely facile to compute, they can also be used to remove the non-speaking mouth. Given a video for normalization, the non-speaking mouth frames have to be firstly removed to maximize performance. In this work, the removal is done by using the shape invariant features, which is different from training an SVM classifier in [10]. The first non-speaking frame is regarded as the reference frame, and the shape invariant features for each frame are extracted. Then the kth frame's difference of intensity is defined as:

$$I_k = ||D_k - D_{ref}||_2 \tag{3}$$

where D_k and D_{ref} are the feature matrix of kth frame and the reference frame respectively. As a result, the threshold for different subjects to get utterance frames is adaptively chosen as shown in Fig. 3.

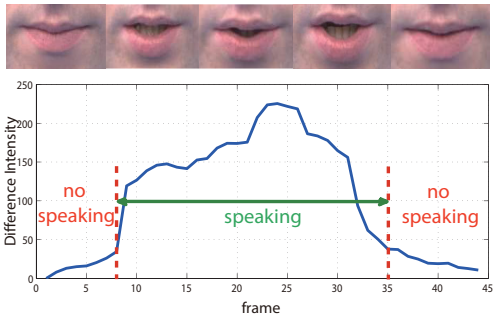


Fig. 3. Removal approach corresponding to a non-speaking mouth

3.2. Bayes Combination Fusion

Several classifiers can be trained with different descriptors and parameters. Thus, we combine the multiple classifiers based on Bayes rule. Many approaches [20] such as the voting methods that combine the results of individual classifiers are only based on the label output by each classifier. It is equally treated as one vote without considering the error of each classifier itself. Therefore, we take the error of each classifier itself into consideration. Consequently, the Bayes combination is adopted

Bayes Combination. Assume that the classifiers are mutually independent. Given L individual classifiers $D_i, i = 1, ..., L$, the error of each classifier D_i is described by its confusion matrix [20] that is given by:

$$CM^i = \begin{pmatrix} n_{11}^i & n_{12}^i & \cdots & n_{1c}^i \\ n_{21}^i & n_{22}^i & \cdots & n_{2c}^i \\ \vdots & \vdots & \ddots & \vdots \\ n_{c1}^i & n_{c2}^i & \cdots & n_{cc}^i \end{pmatrix} \tag{4}$$

where c is the total classes. The (k, s)th entry of this matrix, $cm_{k,s}^i$ is the number of elements of the data set whose true class label is ω_k, and is assigned by D_i to class ω_s.

Denote by $\mathbf{s} = [s_1, ..., s_L]^T$ the vector with the label output of the ensemble, where $s_i \in \Omega$ is the label suggested for $\boldsymbol{x}$ by classifier D_i. The conditional independence allows for the following representation:

$$P(\mathbf{s}|\omega_k) = P(s_1, s_2, ..., s_L|\omega_k) = \prod_{i=1}^{L} P(s_i|\omega_k) \tag{5}$$

where ω_k is the true class label with $k = 1, ..., c$. Then the posterior probability needed to label $\boldsymbol{x}$ is:

$$P(\omega_k|\mathbf{s}) = \frac{P(\omega_k)P(\mathbf{s}|\omega_k)}{P(\mathbf{s})} = \frac{P(\omega_k)\prod_{i=1}^{L} P(s_i|\omega_k)}{P(\mathbf{s})} \tag{6}$$

The denominator does not depend on ω_k and can be ignored, so the support for class ω_k is calculated as:

$$\mu_k(\mathbf{x}) \propto \frac{1}{N_k^{L-1}} \prod_{i=1}^{L} cm_{k,s_i}^i \tag{7}$$

where $cm_{k,s_i}^i/N_k$ is an estimate of the probability $P(s_i|\omega_k)$ and N_k is the number of elements of total data N from class ω_k. Finally, the predicted label $\hat{\omega}_k$ that a test sample $\boldsymbol{x}$ belongs to can be obtained according to the following equation.

$$\hat{\omega}_k = \arg\max_k \mu_k(\mathbf{x}) \tag{8}$$

4. EXPERIMENTS AND ANALYSIS

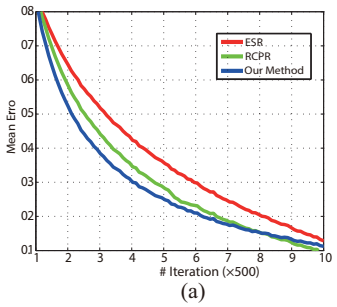
4.1. Preprocessing

The lip regression model is trained on the Helen [15] dataset. Then three datasets Avletters2 [21], OuluVS [9] and PKU-VS are all preprocessed by our lip regression model, where PKUVS is a phrase dataset in Chinese including 30 subjects uttering 10 phrases five times each with high resolution of 1920×1072. We recorded this new dataset to explore the performance in a different language. Further, a 100×80 mouth region is cropped off from each of video frames. Due to the different utterance speed of different subjects, the same path graph based video normalization scheme in [10] is employed and all the utterances are normalized to be 30-frame long.

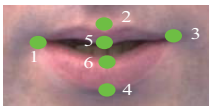
4.2. Experiment A: Evaluation of landmark

Experiments are conducted in the subject independent (SID) and subject dependent (SD) respectively. In the SID experiments, the training and test data are from different subjects and the leave-one-subject-out is explored. In the SD experiments, the training and the test data are from the same subject and the leave-one-utterance-out is performed.

To evaluate the performance of our landmark detection, we report the average error and speed which is measured in frames per second (fps). Errors are measured as the average



(a)



Method	Mean Error	FPS
ESR [15]	7.1	2
RCPR [16]	6.5	6
Ours	6.9	8

(b)

Fig. 4. (a) Fast convergence and accurate estimation with no occlusion mechanism and interpolated shape-index feature compares with [17], but has smart restarts compared with [16]. (b) Average error of selected six landmarks in mouth region and speed .

landmark distance to ground-truth, normalized as percentages with respect to interocular distance [17]. Here, only the six landmarks are selected to measure the mean error. As Fig. 4 shows, our method has a rapid convergence speed without significantly reducing the accuracy, which is much better for the following lipreading. Next, to demonstrate our accurate landmark performance for lipreading, experiment on OuluVS is conducted with two kinds of preprocessed video. One is processed by our lip detection method, the other is detected with Haar-Cascade in Opencv as described in [9]. Then we compare with [10] under the same feature and normalization as shown in Table. I.

Table I. Result of lip detection with landmark and Haar-cascade in the SD experiment on OuluVS.

Method	Using Landmark	Haar-Cascade
$\text{LBP}^{u2}_{(8,3)}$	87.8%	81.5%
$\text{LBP}^{u2}_{(16,4)}$	89.6%	83.3%
$\text{LBP}^{u2}_{(8,3)} + \text{LBP}^{u2}_{(16,4)}$	91.2%	85.1%

As expected, the accurate lip detection boots the performance for lipreading. In particular, the result shows that the power lies in the accurate landmark estimation. From another perspective, the accurate lip detection can distinguish the noise from clean data, where the noise means the lip detection error while the clean data means real lip motion.

4.3. Experiment B: Evaluation of Feature Fusion

To compare our results directly to others, the same phrase indexes in [9] are adopted. As shown in Fig. 5, we demonstrate the relative accuracy improvement on per phrase using accuracy of our feature fusion subtracted by single SIF (in blue), MIP (in green) and LBP-TOP [9] (in red) described in [19] on OuluVS. Accordingly, almost every phrase has an improvement when comparing the feature fusion to the single feature only. Further, it can be found that the improvements on both MIP and LBP-TOP are small while on SIF is great. In brief, the contribution should be attributed to the MIP's motion encoding mechanism and spatial-temporal representation of LBP-TOP. As a result, our fusion method make full use of these information with different patterns and outperforms the single descriptors.

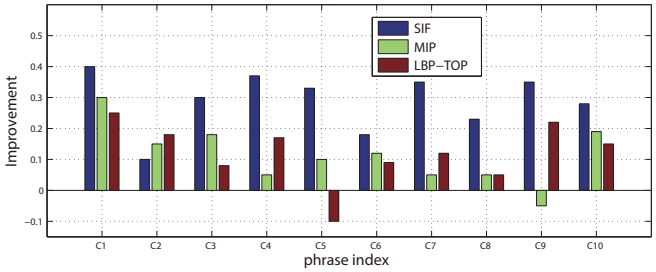


Fig. 5. The relative improvement on per phrase

4.4. Experiment C: Comparison with previous work

We benchmark our method to recent works on three datasets, then the overall results are calculated both in SID and SD experiments in Table II. In particular, experiments on PKUVS is conducted with specific descriptors and feature fusion. As can be seen , our fusion method boosts the performance and the recognition rate of SD is much higher than that of SID experiments. This is reasonable because the uttering characteristics from the same subject are quite similar. Nevertheless, the results shows that our method can match or slightly outperform recent works on lipreading, which we attribute it to our accurate landmark and fusion method.

Table II. Lipreading performances of SID and SD experiments on three uttering databases.

Data sets	Methods	Accuracy (%)	
		SID	SD
Avletters2	Cox *et al.* [21]	-	62.56
	Feature Fusion	**64.38**	**81.54**
OuluVS	Zhao *et al.* [9]	58.85	64.20
	Zhou *et al.* [10]	84.70	85.10
	Pei *et al.* [22]	89.70	97.30
	Feature Fusion	**86.38**	**93.52**
PKUVS	Shape Invariant Feature	40.06	52.65
	$\text{MIP}_{2\times2,\alpha=0,3}$	61.76	69.23
	$\text{LBP}^{u2}_{(16,4)}$	71.76	72.23
	Feature Fusion	**81.54**	**89.25**

5. CONCLUSIONS

We have presented a novel lipreading method by using regression based lip detection and multi-feature fusion. Specifically, a geometric-based feature is proposed and combined with motion interchange patterns and spatial-temporal descriptors to describe the lip information. Our framework takes advantage of an accurate landmark estimation and multi-feature fusion, which makes full use of the uttering information. Experiments results show that our approach achieve better performance on three benchmarks with frontal faces. In the future, we will focus on the large variation of lip and employ the more robust representation for multi-features.

6. REFERENCES

[1] J. Gowdy, A. Subramanya, C. Bartels, and J. Bilmes, "DBN based multi-stream models for audio-visual speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 993–996, 2004.

[2] E. Benhaim, H. Sahbi, and G. Vitte, "Continuous visual speech recognition for multimodal fusion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4618–4622, 2014.

[3] H. Liu, T. Fan, and P. Wu, "Audio-visual keyword spotting based on adaptive decision fusion under noisy conditions for Human-Robot Interaction," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6644–6651, 2014.

[4] T. Ezzat and T. Poggio, "Miketalk: A talking facial display based on morphing visemes," in *Computer Animation 98. Proceedings*, pp. 96–102, 1998.

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

[6] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[7] I. Matthews, T. Cootes, J. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 24, no. 2, pp. 198–213, 2002.

[8] J. Lee and C. Park, "Hybrid simulated annealing and its application to optimization of hidden Markov models for visual speech recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 4, pp. 1188–1196, 2010.

[9] G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1254–1265, 2009.

[10] Z. Zhou, G. Zhao, and M. Pietikainen, "Towards a practical lipreading system," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 137–144, 2011.

[11] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 23, no. 6, pp. 681–685, 2001.

[12] S. Alizadeh, R. Boostani, and V. Asadpour, "Lip feature extraction and reduction for HMM-based visual speech recognition systems," in *International Conference on Signal Processing (ICSP)*, pp. 561–564, 2008.

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 511–518, 2001.

[14] M. Hasan, M. Jamil, M. Rabbani, and M. S. Rahman, "Speaker identification using mel frequency cepstral coefficients," in *International Conference on Electrical & Computer Engineering ICECE*, vol. 2004, 2004.

[15] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. Huang, "Interactive facial feature localization," in *European Conference on Computer Vision (ECCV)*, pp. 679–692, 2012.

[16] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *International Journal of Computer Vision (IJCV)*, vol. 107, no. 2, pp. 177–190, 2014.

[17] X. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 1513–1520, 2013.

[18] N. Duffy and D. Helmbold, "Boosting methods for regression," *Machine Learning*, vol. 47, no. 2-3, pp. 153–200, 2002.

[19] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf, "Motion interchange patterns for action recognition in unconstrained videos," in *European Conference on Computer Vision (ECCV)*, pp. 256–269, 2012.

[20] L. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons, 2004.

[21] S. Cox, R. Harvey, Y. Lan, J. Newman, and B. Theobald, "The challenge of multispeaker lip-reading.," in *International Conference on Auditory-Visual Speech Processing (AVSP)*, pp. 179–184, 2008.

[22] Y. Pei, T. Kim, and H. Zha, "Unsupervised random forest manifold alignment for lipreading," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 129–136, 2013.