A New Visual Speech Recognition Approach for RGB-D Cameras

Ahmed Rekik[1(✉)], Achraf Ben-Hamadou[2], and Walid Mahdi[1]

[1] Multimedia Information Systems and Advanced Computing Laboratory (MIRACL), Sfax University Pôle technologique de Sfax, route de Tunis Km 10, BP 242, 3021 Sfax, Tunisia
[2] Valeo Driving Assistance Research Center, 34 rue St-André Z.I. des Vignes, 93012 Bobigny, France
rekikamed@gmail.com, achraf.ben-hamadou@valeo.com, walid.mahdi@isimsf.rnu.tn

Abstract. Visual speech recognition remains a challenging topic due to various speaking characteristics. This paper proposes a new approach for lipreading to recognize isolated speech segments (words, digits, phrases, *etc*.) using both of 2D image and depth data. The process of the proposed system is divided into three consecutive steps, namely, mouth region tracking and extraction, motion and appearance descriptors (HOG and MBH) computing, and classification using the Support Vector Machine (SVM) method. To evaluate the proposed approach, three public databases (MIRALC, Ouluvs, and CUAVE) were used. Speaker dependent and speaker independent settings were considered in the evaluation experiments. The obtained recognition results demonstrate that lipreading can be performed effectively, and the proposed approach outperforms recent works in the literature for the speaker dependent setting while being competitive for the speaker independent setting.

Keywords: Visual speech recognition · Lip-reading · Visual communication · Face tracking · Human-computer-interaction

1 Introduction

Visual Lip-Reading (LR) systems play an important role for Human-Machine-Interation applications in noisy environments where audio speech recognition still very challenging (*i.e.,* overcoming signals alternate the recognition). However, visual LR systems face their own challenges. Indeed, there is an important variation in terms of lip shapes, skin colors, speaking speeds and intensities *etc.* which make difficult to develop generative LR systems. Consequently, most of works in this research field are restricted to limited number of classes (*i.e.,* words, phrases, commands, *etc.*) and speakers [14,15].

We can divide LR systems into two categories. The first category includes methods where a non-rigid tracking of the mouth region is needed. After performing the lips boundaries tracking, different features are extracted from the tracking results and used in the recognition process. For example, [12] applied an AAM (Active Appearance

A. Campilho and M. Kamel (Eds.): ICIAR 2014, Part II, LNCS 8815, pp. 21–28, 2014.
DOI: 10.1007/978-3-319-11755-3_3

Model) for deformable face tracking (including lips) and then used directly the animation units of the AAM as a descriptor of the lips deformations. Features like angles and distances computed between reference points located on the lips boundaries are also used with a KNN classifier in [14]. Recently, [9] used a combination of descriptors like HOG (Histogram of Oriented Gradients) [5] and LBP (Local binary patterns) [6] computed on small patches centred on reference points of the tracked lip boundaries. The advantage of such approaches is that lips deformation trajectory is directly modelized from the lips non-rigid tracking. However, lips deformation tracking is a complex task and very sensitive to illumination and texture variations. In opposition to these methods, the second category of LR systems gathers methods that do not rely on non-rigid tracking of the speaker's lips. In these methods, a rectangular mouth region is cropped from input data, then, features are computed on the cropped data to describe the mouth shape. For example, Zhao *et al.* [15] proposed a spatio-temporal version of LBP descriptor for lipreading. This descriptor is computed on the extracted mouth regions and used as input for an SVM classifier. Shaikh and Kumar [11], used optical flow for extracting the whole word features from the mouth region. These methods are characterized by their simplicity where no deformable lip tracking is needed. However, usually they are sensitive to the orientation of the speaker's face and assume a limited face motion.

In this paper, we propose a new LR system using both of 2D images and depth maps. We describe a rectification process to handle the motion of the speaker's face for a robust mouth region extraction. Also, we introduce an effective combination of appearance and motion descriptors to be used for the classification process.

The remaining of this paper is organized as follows. First, a general system overview is presented in section 2.1. A 3D face tracking and mouth region extraction process will be described in section 2.2. The appearance and motion descriptors used in the recognition step are presented in section 2.3. Experimental results will be shown in section 3. Finally, a conclusion and some perspectives and future work are given in section 4.

2 Lip-Reading System

2.1 System General Overview

As illustrated in Figure 1, the proposed LR system takes a multi-modal video clip (*i.e.,* 2D images and depth maps) representing a speech portion to recognize (*e.g.,* a word, a command, a phrase, *etc.*). The system is then divided in three main steps, namely, mouth region extraction, descriptors computing, and classification. In the first step, the user's face is detected and its 3D pose is tracked over the input sequence. This step allows to robustly extract the mouth region yielding a sequence of rectified patches centred on the user's mouth. Then, the second step takes the rectified patches as input to compute both of appearance and motion descriptors which serve for the final SVM classification step. All of these steps will be detailed in the following sections.

2.2 Mouth Region Extraction

Since an LR system user can obviously move his head while speaking, it is mandatory to track the position and the orientation of the user's face over the input sequence to

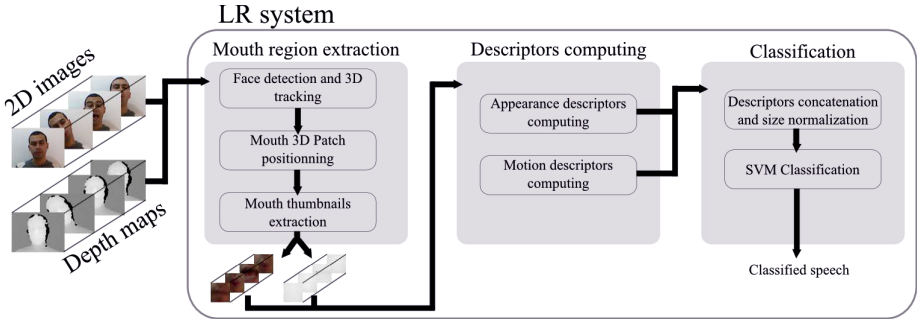

Fig. 1. General overview of our LR system

robustly extract the mouth region. To this end, we used a 3D rigid face tracking method [10]. This tracking method handles illumination changes and offers a very low jitter. The face tracking is modelized as an optimization problem to find the optimal 3D rigid motion between two consecutive video frames. We denote the optimal face pose of a given frame from the input sequence, by $\widehat{x} \in \mathbb{R}^6$ involving the 6 degree of freedom (*i.e.,* 3 translations and 3 rotation angles) of a 3D rigid motion. We presented the mouth region by a 3D rectangular patch (see figure 2(a)). This patch rigidly follows the face motion by means of the estimated face poses (*i.e.,* the 3D patch is rigidly fixed to the 3D face model). The size of the mouth patch is automatically scaled to adapt the user's face size using the tracking initialization parameters [10]. The patch is densely sampled to $n_h \times n_w$ 3D points $\{P_{i,j} | i = 1, \ldots, n_h, j = 1, \ldots, n_w\}$. For a given instant of the input sequence, the new position $\widehat{P}_{i,j}$ of each point $P_{i,j}$ is computed by applying the face pose as follows:

$$\widehat{P}_{i,j} = \mathbf{R}\, P_{i,j} + \mathbf{t} \tag{1}$$

where $\mathbf{R}$ and $\mathbf{t}$ are respectively the 3×3 rotation matrix and the 3D translation vector generated in a standard way from $\widehat{x}$. The definition of the mouth patch in 3D simplifies the mouth thumbnails rectification in terms of position, orientation, and size. Indeed, for a given face pose, the patch points $\widehat{P}_{i,j}$ are simply projected into the 2D image and the depth map yielding two rectified thumbnails (see figures 2(b), 2(c), and 2(d) for an example). Let denote by $\boldsymbol{p}^c_{i,j}$ and $\boldsymbol{p}^d_{i,j}$ the projection of $\widehat{P}_{i,j}$ in the 2D image and the depth map respectively. These projection points are computed as follows:

$$p^c_{i,j} = \mathcal{H}\left(\mathcal{K}_c\, \widehat{P}_{i,j}\right) \tag{2}$$

$$p^d_{i,j} = \mathcal{H}\left(\mathcal{K}_d\, \mathbf{T} \begin{bmatrix} \widehat{P}_{i,j} \\ 1 \end{bmatrix}\right) \tag{3}$$

In equation (2) and (3), $\mathcal{K}_c$ and $\mathcal{K}_d$ stand for the intrinsic matrix of the color and the depth cameras, respectively. Also, $\mathcal{H} : \mathbb{R}^3 \longmapsto \mathbb{R}^2$ is the homogeneous coordinate function where $\mathcal{H}\left([X\ Y\ Z]^T\right) \stackrel{\Delta}{=} [X/Z\ Y/Z]^T$, and $\mathbf{T}$ is a 3×4 matrix corresponding to the global 3D transformation between the color and depth cameras. $\mathbf{T}$, $\mathcal{K}_c$, and $\mathcal{K}_d$ are known when calibrating the acquisition system [2,3]. Since $p^c_{i,j}$ and $p^d_{i,j}$ are not

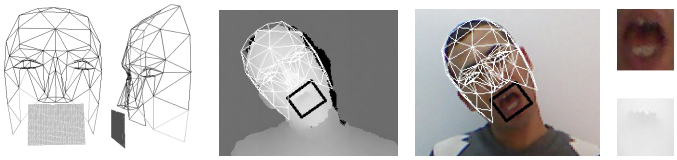

Fig. 2. Mouth region extraction. (a): 3D model used for 3D face tracking (viewed from two points of view). The rectangular mesh corresponds to the mouth patch rigidly fixed to the face model. (b) and (c): Projection of the 3D face model and the 3D mouth patch in the depth map and the 2D image, respectively. (d): Extracted 2D and depth rectified mouth thumbnails.

necessarily round coordinates, we perform a bilinear interpolation to fill pixel values of the 2D image and the depth thumbnails.

2.3 Motion and Appearance Descriptors Computing

To describe the mouth shape and deformation during a speech segment, we need to compute descriptors on the extracted mouth thumbnails. We combined both of appearance and motion descriptors to modelize temporal deformations of the mouth.

Appearance Descriptor: We consider the well known Histogram of Oriented Gradient (HOG) descriptor [4]. It is worth to notice that we tested other descriptors like Local Binary Patten (LBP) [7], but the best performance were provided using HOG descriptor. The HOG method tiles the mouth thumbnails with a dense grid of cells, with each cell containing a local histogram over orientation bins. At each pixel of the thumbnail, the image gradient vector is calculated and converted to an angle, voting into the corresponding orientation bin with a vote weighted by the gradient magnitude. Votes are accumulated over the pixels of each cell. The cells are grouped into blocks and a robust normalization process is run on each block to provide strong illumination invariance. The normalized histograms of all of the blocks are concatenated to give the window-level visual descriptor vector for learning. In the remaining sections, we denote by HOG_c and HOG_d the HOG descriptor computed on the 2D and depth thumbnail respectively.

Motion Descriptor: The motion descriptor allows somehow to modelize the lips movement. In [5], motion descriptors like Internal Motion Histograms (IMH) and Motion Boundary Histograms (MBH) were used for pedestrian detection in video sequences. A IMH variant gave the best recognition rates. However, in our case, the lips motion is much less complex than pedestrians motions. We tested both of the IMHcd (variant of IMH) and the MBH descriptors for our LR system. Briefly, the MBH computes the optical flow between two consecutive mouth thumbnails. Then, each flow component is treated as an independent thumbnail, takes their local gradients separately, find the corresponding gradient magnitudes and orientations, and use these as weighted votes into local orientation histograms in the same way as for the standard gray scale HOG.

2.4 Classification

Given an input sequence of N frames, the computed appearance and motion descriptors for each frame are concatenated to generate a single descriptor vector (denoted by $\mathcal{D}_i|_{i=1,\ldots,N}$). Then, following [12], all $\mathcal{D}_i$ are interpolated to squeeze them in a fixed length vector. To give an example, in our experiments on words recognition we fixed the size of the final descriptor vector to 20 frames $\times$ size of $\mathcal{D}_i$.

This paper investigates the use of support vector machines (SVMs) [13]. SVMs are selected due to their ability to find a globally optimum decision function to separate the different classes of data. In this work, we compared the linear, the polynomial, and the RBF kernels for our LR system. The best performance was given by linear kernel.

3 Experiments and Results

3.1 Evaluation Datasets

To evaluate our LR system we used several datasets with different kind of input data (2D images, RGB-D sequences) and different configurations (distance of face from acquisition vision system) mouth resolution in the images. One of the contributions of this paper is the construction of a new dataset MIRACL-VC.

MIRACL-VC: It consists of 1500 word data (15 persons$\times$10 words$\times$10 times/word) and 1500 phrases (15 persons$\times$10 phrases$\times$10 times/phrase). The dataset covers words like *navigation*, *connection*, *etc.*. We took also phrases like *Nice to meet you*, *I love this game*, *etc.*. We used the Kinect sensor to acquire 2D images and depth maps with a resolution of 640×480 and at an acquisition rate of 15 fps. The distance between the speaker and the Kinect is about $1m$.

OuluVS: [15] It consists of ten different everyday phrases. Each phrase is uttered by 20 subjects up to five times. The frame rate was set to 25 fps at a resolution of 720×576. Only color images are provided for each phrase sequence.

CUAVE: [8] It consists of a speaker-independent corpus of over 7,000 utterances of both connected and isolated digits spoken by 37 subjects. The database was recorded at a resolution of 720×480 with a frame rate of 29.97 fps where only color images are provided for each sequence.

3.2 Visual Speech Recognition Results

We adopt two test settings for visual speech recognition: speaker independent and speaker dependent.

Speaker Independent (SI): In this experiment, the training and the query data are from different speakers. We employ the leave-one-out strategy where data from a single speaker are used as the validation data, and the remaining speakers are used as the training data. This is repeated for each speaker in the dataset.

Speaker Dependent(SD): In this experiment, the training and the testing data are from the same speaker.

MIRACL-VC Results. To evaluate our system on the MIRACL-VC dataset, we have applied our method to compute different descriptors on the extracted mouth thumbnails. Then, we have compared different combination of descriptors HOG_c, HOG_d and MBH. Table 1 illustrates the obtained recognition rates. For **SD** settings, the best performance is obtained for the combination HOG_c+HOG_d (96.0% for phrases and 95.4% for words). That is, when we added MBH (*i.e.,* HOG_c+HOG_d+MBH), we ameliorate the **SI** setting results (*i.e.,* gain of 12.5% and 3.3% of recognition rate) while keeping the same recognition rates for **SD** setting. We can see also in table 1 that the performance of our LR system is better for phrases than words. This is can be explained by the fact that a phrases contains a longer series of visemes than words. Thus, it is easier to discriminate between phrases than between words.

Table 1. Lipreading performances of speaker dependent (SD) and speaker independent (SI) experiments on MIRACL-VC dataset

	Phrases		Words	
Descriptors	**SD**	**SI**	**SD**	**SI**
HOG_c	95.4%	54.4%	92.8%	48.1%
HOG_d	94.2%	64.7%	94.2%	54.5%
HOG_c+HOG_d	96.0%	66.7%	95.4%	59.8%
MBH_c	87.4%	49.4%	86.2%	45.1%
HOG_c+HOG_d+MBH_c	96.0%	79.2%	95.3%	63.1%

OuluVS Results. We have compared our method with the recent lipreading works [1],[15],[16] on OuluVS dataset, where we combined only HOG_c and MBH_c descriptors since depth maps are not provided in this dataset. Table 2 shows the recognition rates of our method (first row) and recent works in the literature. Our method outperforms all the other works for **SD** configuration with 93.2% against 64.2%, 73.5%, 85.1% for [15], [1], and [16] respectively. However, we all obtained roughly a comparable results for **SI** configuration (*i.e.,* between 58.6% and 70.6%). Additionally, we can see that our system perform better in MIRACL-VC (see results for phrases in table 1). This difference of performance can be explained by the absence of the depth data in the OuluVS dataset.

Table 2. Lipreading performances of speaker dependent (SD) and speaker independent (SI) experiments on OuluVS dataset and comparison with other recent works

	SD	**SI**
Our method	93.2%	68.3%
[15]	64.2%	58.6%
[1]	73.5%	62.3%
[16]	85.1%	70.6%

CUAVE Results. In [9], only two sample on CUAVE where used for evaluation. They obtained 100% of recognition rate for the **SD** setting. For the same two sample and the same setting, we obtained 97.0%. We decided to perform more representative evaluation with CUAVE dataset and we used all the available samples (36 samples). We obtained 90% for **SD** setting and 70.1% for **SI** setting.

4 Conclusion

We proposed a new visual speech recognition system using both of 2D images and depth maps acquired typically with RGB-D cameras. Our approach is based on 3D rigid tracking of the speaker's face to robustly extract image and depth thumbnails. These are then used to compute motion and appearance descriptors which are used as input for an SVM classifier. For the evaluation of our system, we construct a new dataset (MIRACL-VC) and we used also two other public datasets (CUAVE and OuluVS). The obtained evaluation results on these three datasets show that our system is very competitive with other recent works in the literature. Additionally, we showed also that depth data ameliorate sensibly the LR system performance and the combination of the appearance and motion descriptors improve the **SI** performance.

In our future work, we will try to ameliorate the **SI** performance by considering other descriptors. Also, we would like to investigate the speech spotting in continuous speech flow.

References

1. Bakry, A., Elgammal, A.: Mkpls: Manifold kernel partial least squares for lipreading and speaker identification. In: CVPR, pp. 684–691. IEEE (2013)
2. Ben-Hamadou, A., Soussen, C., Daul, C., Blondel, W., Wolf, D.: Flexible projector calibration for active stereoscopic systems. In: 2010 IEEE International Conference on Image Processing, pp. 4241–4244 (September 2010)
3. Ben-Hamadou, A., Soussen, C., Daul, C., Blondel, W., Wolf, D.: Flexible calibration of structured-light systems projecting point patterns. Computer Vision and Image Understanding **117**(10), 1468–1481 (2013)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
5. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006)
6. Huang, D., Shan, C., Ardabilian, M., Wang, Y., Chen, L.: Local binary patterns and its application to facial image analysis: a survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews **41**(6), 765–781 (2011)
7. Nanni, L., Lumini, A., Brahnam, S.: Survey on lbp based texture descriptors for image classification. Expert Syst. Appl. **39**(3), 3634–3641 (2012)
8. Patterson, E.K., Gurbuz, S., Tufekci, Z., Gowdy, J.: Cuave: A new audio-visual database for multimodal human-computer interface research. In: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 2, pp. II-2017-II-2020. IEEE (2002)

9. Pei, Y., Kim, T.K., Zha, H.: Unsupervised random forest manifold alignment for lipreading. In: ICCV, pp. 129–136 (2013)
10. Rekik, A., Ben-Hamadou, A., Mahdi, W.: Face pose tracking under arbitrary illumination changes. In: VISAPP (2014)
11. Shaikh, A.A., Kumar, D.K., Yau, W.C., Che Azemin, M., Gubbi, J.: Lip reading using optical flow and support vector machines. In: 2010 3rd International Congress on Image and Signal Processing (CISP), vol. 1, pp. 327–330. IEEE (2010)
12. Shin, J., Lee, J., Kim, D.: Real-time lip reading system for isolated korean word recognition. Pattern Recognition **44**(3), 559–571 (2011)
13. Vapnik, V.: The nature of statistical learning theory. Springer (2000)
14. Yargic, A., Dogan, M.: A lip reading application on ms kinect camera. In: 2013 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA), pp. 1–5. IEEE (2013)
15. Zhao, G., Barnard, M., Pietikainen, M.: Lipreading with local spatiotemporal descriptors. IEEE Transactions on Multimedia **11**(7), 1254–1265 (2009)
16. Zhou, Z., Zhao, G., Pietikainen, M.: Towards a practical lipreading system. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 137–144. IEEE (2011)