Computer Vision and Image Understanding 148 (2016) 153–163

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Lip event detection using oriented histograms of regional optical flow and low rank affinity pursuit

Xin Liu [a], Yiu-ming Cheung [b,*], Yuan Yan Tang [c]

[a] College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China
[b] Department of Computer Science and Institute of Research and Continuing Education, Hong Kong Baptist University, Hong Kong SAR, China
[c] Department of Computer and Information Science, University of Macau, Macau SAR, China

ARTICLE INFO

Article history:
Received 15 April 2015
Accepted 28 November 2015

Keywords:
Lip event detection
Oriented histograms
Regional optical flow
Low rank affinity

ABSTRACT

Lip event detection is of crucial importance to the better understanding of visual speech perceptually between humans and computers. In this paper, we address an efficient lip event detection approach using oriented histograms of regional optical flow (OH-ROF) and low rank affinity pursuit. First, we align the extracted lip region sequences to reduce the impact of irrelevant motion caused by the moving cameras. Then, an optical flow field is calculated from these sequentially stabilized images and an efficient descriptor, namely OH-ROF, is presented to discriminatively code the visual appearance of each motion frame, whereby each lip motion clip can be represented by a sequence of OH-ROF vectors as its signature. Subsequently, we detect the visual silence event based on the small flow magnitude, and further propose a low rank affinity pursuit method to determine the visual speech event that incorporates the lip-dynamic states of mouth opening and closing. As a result, various kind of lip motion events can be appropriately estimated. The proposed approach neither requires any training set on the labeled videos nor learns the lip motion priors of each visual event in an unconstrained video. Experiments show a promising result in comparison with the state-of-the-art counterparts.

1. Introduction

In general, the visual information of lip motion, which is completely independent of background noise and tightly correlated with the acoustical signals, is much helpful for speech recognition, particulary in the noisy environment. In recent years, lip event analysis in videos has been extensively studied because of its attractable applications including lipreading [1], visual speaker identification [2,3], audio–visual speech recognition (AVSR) [4], human computer/robot interaction, facial expression analysis [5,6] and so forth. Among these applications, one of the key issues is to precisely detect the lip motion events so that the corresponding lip-dynamic states can be well obtained for further speaking analysis and lip behavior investigation. For instance, the sequential variations of the lip motion appearances can be regarded as the visual counterpart of voice activities, and lip event detection can therefore be utilized to overcome the poor performance of voice activity detection when the background noise is noticeable. In addition, the detection of the visual speech often plays a key role in the speech recognition, while the detection of lip-dynamic states about the mouth opening and closing is of crucial importance to the facial appearance analysis. Nevertheless, to the best of our knowledge, it is still a non-trivial task to perform a reliable lip motion event detection due to its elastic shape, non-rigid motion, and large variations caused by the intra-personal lip appearance changes, surrounding clutters, uncontrollable lighting condition, and so forth.

In the past years, a few specific techniques have been developed to realize lip event detection, which can be roughly grouped into three categories: shape-based approaches, motion-based approaches and model-based approaches.

The shape-based approaches generally assume that the variations of lip shape are mainly found within the speaking interval and the stationary lips are overwhelmingly found in the non-speaking interval. Along this line, Sodoyer et al. [7] first conducted a comprehensive analysis of lip shape parameters about spontaneous speech corpus, and then smoothed the visual information in terms of the interolabial width and height of the lip regions. Accordingly, the visual differences between the natural silence and non-silence sections of a given speaker can be well characterized. Later, paper [8] extends this work to adapt the difficult case of convolutive mixtures even if the recording sources are highly non-stationary. In particular, the

* Corresponding author.
E-mail addresses: starxliu@gmail.com (X. Liu), ymc@comp.hkbu.edu.hk (Y.-m. Cheung), yytang@umac.mo (Y.Y. Tang).

http://dx.doi.org/10.1016/j.cviu.2015.11.015
1077-3142/

experiments conducted in these two approaches are especially designed on the make-up lip video databases, through which the shape parameters can be well obtained. Furthermore, Aoki et al. [9] first extracted the lip shapes of the target speaker by an elastic bunch graph matching method, and subsequently measured the lip aspect ratio to prevent the wrong voice activity detection. However, this approach is specially utilized to handle the infrared image sequence. Therefore, the aforementioned three approaches are unsuitable for lip event detection in real conditions, e.g., lips without make-up or image sequences captured under natural environment. Recently, Talea et al. [10] first made a series of mouth area subtractions and then employed a smoothing filtering to detect the syllable event. Nevertheless, it is very difficult to extract the lip shape parameters with great reliability when the mouth image incorporates very low resolution and the poor contrast between the lip and surrounding skin pixels. In addition, these shape-based approaches are somewhat sensitive to the poor lighting conditions.

The motion-based approaches suppose that the appearances of the consecutive mouth regions are different when the human lip moves in speaking. From this viewpoint, Yau et al. [11] computed the motion history images (MHIs) of lip motions and utilized the Zernike moment features to detect the starting and ending frames of isolated utterances, in which the magnitude of Zernike moments corresponding to the uttering frames is much greater than the one of the frames within the period of pause or silence. However, it is found that this approach is quite sensitive to the illumination changes. To resist this attack, Libal et al. [12] calculated the accumulated intensity difference in a bar mask and compared it to a running average histogram, through which the motion states of lip opening/closing can be determined by investigating the significant changes of such comparisons. In this approach, they defined a speaking period provided that the states of mouth are opened and closed semi-regularly during speech. Nevertheless, this condition is obviously too strong because it does not consider any uncertainty in observations. Later, Siatras et al. [13] found that the increased average value and standard deviation of the mouth region pixels with low intensities can be well utilized as the visually distinctive cues to depict visual speech from those that depict visual silence. Such an approach does not require a complex feature extraction procedure, e.g., the geometric features within the lip shapes. However, their performances would be instable when there exist poor lighting conditions or insufficient mouth information. Furthermore, Karlsson et al. [14] have utilized the recently developed optical flow differential invariants to exploit the divergence of the flow field at a coarse scale, whereby the lip-dynamic states corresponding to the mouth opening and closing can be determined. This lip event detection approach has an advantage of fast computation and has shown to perform well on the XM2VTS database. However, this type of approach might be prone to suffer from the tiny movements of the muscles around the lips. Recently, Shaikh et al. [15] have utilized the pair-wise pixel comparison of consecutive images to segment the isolated utterances temporally. Nevertheless, this approach incorporating the pair-wise pixel comparing is very sensitive to the irrelevant motion caused by unstable camera. Until most recently, Taeyup et al. [16] first calculated a phase space plot over the joint histogram of a Gaussian blurred image pair (closed lip vs. open lip) and then extracted the chaos inspired similarity measure for visual speech/silence detection. This approach has found to be adaptive to the illumination changes, but which often degrades its performance when the located lip sequences are unstable.

The model-based approaches empirically learn a reference model to characterize the lip activities such that the corresponding event states can be identified. Following this idea, Luthon et al. [17] utilized a spatiotemporal neighborhood of each pixel associated with the Markov Random Field (MRF) to label the motion states of mouth opening and closing. Under natural lighting conditions, this pioneer work is able to detect the mouth states without any particular make-up. Nevertheless, such an approach exploiting the horizontal and vertical spatial gradients, is somewhat sensitive to the image noise and the changes of lighting conditions. To handle this problem, the active shape model (ASM) [18] and active appearance model (AAM) [19] employ a set of landmark points to describe the lip movements, and these points are controlled within a few previously derived modes in the training set. Nevertheless, inevitably, such kind of systems is generally required to label a group of landmark points and to perform a training process to determine the corresponding model parameters. Moreover, it is very difficult to apply these two models on very low-resolution image sequences. Differently, Liu et al. [20] first applied principal component analysis (PCA) to extract the visual features on the detected mouth region, and then modeled the distribution of speech and non-speech events using two different Gaussian mixture models (GMMs). Accordingly, the corresponding voice activities can be well detected. In general, the desired parameters of these two models are estimated from the feature vectors derived from the training data. The decision of the speech/non-speech event is taken by evaluating the likelihood of each frame conditioned on both model distributions. Even though the mouth appearances during the speaking and non-speaking intervals exhibit the different distributions, there always exist the overlap between two models and the reliable decision boundaries may not be well determined for robust event detection. Later, Aubrey et al. [21] computed the optical flows within the successive mouth regions in a training dataset and modeled the temporal variation of these motion vectors via a hidden Markov model (HMM). Accordingly, each frame of the new motion data can be classified as either speech or non-speech periods by comparing the probability generated by this model to a threshold value. That is, the frames below the threshold are assigned as non-speech event and the frames above the threshold are designated as speech event. Furthermore, Navarathna et al. [22] first divided the incoming speech utterance into a number of fixed-length frames and then embedded the extracted lip region features into the GMM visual speech classifier, through which the corresponding score list of each frame state can be obtained. Recently, Tiawongsombat et al. [23] have employed the mouth image energy as a visual cue and proposed a bi-level HMM embracing both the lip moving states and speaking states to assist voice activity detection in human robot interaction. Among these model-based approaches, it is found that the related model parameters and the threshold value should be sufficiently learned from the training dataset, which, from the practical viewpoint, limits their application domains.

In general, the successful achievement of reliable lip motion event detection lies in a closer investigation of the physical process within the corresponding lip motion activities. Meanwhile, the robust lip event detection algorithms should be capable of adapting to various illumination conditions. In this paper, we present an efficient lip event detection approach by using oriented histograms of regional optical flow and low rank affinity pursuit. Without learning priors, the proposed approach aims not only to distinguish frames depicting visual speech from those depicting visual silence, but also to investigate the lip-dynamic states of mouth opening and closing. Experiments have shown that the proposed approach performs favorably compared to the state-of-the-art methods.

The remaining part of this paper is structured as follows: Section 2 briefly introduces the optical flow framework. Section 3 describes the pipeline and procedures of the proposed framework, and Section 4 shows the experimental results, together with the discussions. Finally, we draw a conclusion in Section 5.

2. Overview of optical flow

Lip event analysis is a challenging research topic due to its complexity and variation of mouth appearances. As a visual descriptor, optical flow is able to describe the distribution of the apparent

velocities of brightness patterns in a sequence, and there has been significant interest in exploiting the motion vector that derives from the optical flow to characterize the lip movements. The main merits are three-fold: 1) This motion descriptor is able to well describe the lip activity even if the extracted mouth regions are of low-resolutions; 2) The visual features exploited in optical flow need not extract the lip shapes or require any prior knowledge about the lip structure; 3) The optical flow vector has found to be robust against the complex lighting conditions because the lighting changes are smooth between the neighboring frames.

In the past, different kinds of optical flow methods have been exploited in the literature [24–26]. In this paper, we utilize the Lucas–Kanade technique to compute the optical flow vectors [26], which has found to be more robust under noisy environment. Let us consider a group of consecutive image sequence $f(x, y, t)$, where (x, y) denotes the pixel position within an $M \times N$ rectangular image region, and t represents the time notation. Many different optical flow estimation methods are based on the assumption that the intensity values of an image object in subsequent frame do not change over time, i.e. $f(x+u, y+v, t+1) = f(x, y, t)$, where the displacement field u and v are the horizontal and vertical components of the optical flow field to be estimated from the image pair at time t and $t+1$, respectively. For small displacements, the linearized version of the intensity value constancy assumption yields the famous first-order optical flow constraint: $f_x u + f_y v + f_t = 0$, where subscripts denote the partial derivatives.

In general, if the spatial gradient of the image is sufficiently large and its direction varies sufficiently within the neighborhood, the flow constraint is well-conditioned and the reliable flow value can be estimated. Nevertheless, if the spatial gradient is close to zero or its direction is nearly a constant, the flow constraint is not sufficient to uniquely compute the two unknowns: u and v. To tackle this issue, it is effective to assume that the unknown optic flow vector is a constant within its neighborhood of size ρ, and it is possible to determine the two constants u and v from a weighted least square fit:

$$E_{LK}(u, v) = K_\rho * \left((f_x u + f_y v + f_t)^2\right). \tag{1}$$

Specifically, the standard deviation ρ of the Gaussian function serves as an integration scale over the main contribution of the computed least square fit. Accordingly, the minimum (u, v) of E_{LK} satisfies the conditions $\partial_u E_{LK} = 0$ and $\partial_v E_{LK} = 0$, which gives the following linear system:

$$\begin{pmatrix} K_\rho * (f_x^2) & K_\rho * (f_x f_y) \\ K_\rho * (f_x f_y) & K_\rho * (f_y^2) \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} -K_\rho * (f_x f_t) \\ -K_\rho * (f_y f_t) \end{pmatrix}. \tag{2}$$

According to this framework, the lip motion during a speaking interval often produces a more radical increase in the velocity of pixels than it does during a non-speaking interval.

3. The proposed approach

In this section, we present the proposed lip event detection approach in detail. First, we introduce a sequence stabilization scheme to reduce the impact of irrelevant motions, and then present an oriented histograms of regional optical flow to characterize the visual appearance of each lip motion frame. Finally, we show the details of the implemented algorithm.

3.1. Image sequence stabilization

In general, the optical flow field is calculated from the spatiotemporal gradients of the stabilized image sequence, whereby the motion vectors derived from these raw flows can be utilized to characterize the target motions. Along this way, all above derivations are based on the assumption that the videos are captured by static cameras, and this assumption would greatly simplify the lip event detection problem because the mere presence of lip movement provides a strong cue for the motion analysis. However, it is an inherent problem that the relative position of the camera with respect to the speaker is seldom fixed in most image acquisition processes. For instance, the video clips are recorded by a hand-held camera. Under such circumstances, the estimated optical flow may appear drastically different even under moderate change in the position or pose with respect to the camera settings. Therefore, it is imperative to align the image sequence and reduce the impact of the irrelevant motion caused by the moving cameras, featuring on reliable flow calculation.

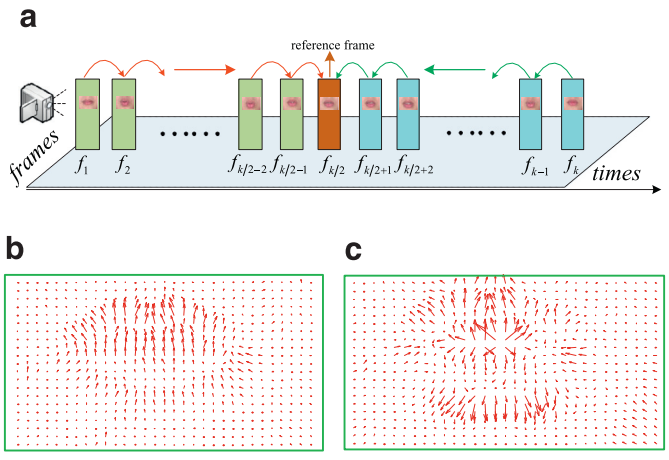


Fig. 1. (a) The image sequence alignment scheme. (b) Optical flow field estimated from the misaligned lip image pairs. (c) Optical flow field estimated from the aligned lip image pairs.

For image sequence alignment, it is reasonable to assume that the misalignment is restricted to the image plane. Under such restriction, the misalignment problem within two adjacent frames can be considered as a domain deformation. More precisely, if $f(x, y, t)$ and $f(x, y, t+1)$ represent two misaligned images at time t and $t+1$, respectively, there exists an invertible transformation τ such that:

$$f(x, y, t+1) = (f \circ \tau)(x, y, t) = f(\tau(x, y, t)) \tag{3}$$

where $f(\tau(x, y, t))$ denotes the tth frame after the transformation parameterized by vector τ. From the practical viewpoint, this kind of misalignment problem can be modeled as a finite dimensional transformation that shares a parametric representation. Specifically, the popular 2D parametric transform [27] of an affine group is always utilized to model the translation, rotation and planar deformation of the background sequence. Within this framework, the lip sequence alignment problem can be intuitively formulated as follows: suppose that $f_1, f_2, \ldots, f_k$ represent k input lip images, but misaligned with each other. Then, there exist a group of domain transformation parameters $\{\tau_1, \tau_2, \ldots, \tau_k\}$ to compensate camera motion, whereby the transformed lip image sequence $\{f_1 \circ \tau_1, f_2 \circ \tau_2, \ldots, f_k \circ \tau_k\}$ is well-aligned at the pixel level.

From the practical viewpoint, it is imperative to specify a reference frame for image sequence alignment, and it is reasonable to take the middle frame of the motion clip into consideration. For reliable flow calculation, as shown in Fig. 1, we select to align the extracted lip region sequences by the following two steps: 1) we utilize the 2D parametric transforms [27] to model the translation, rotation and planar deformation between the neighboring frames of the extracted lip region sequences; 2) we select the robust multiresolution method [28] to compensate for the background motion caused by moving cameras, in which the middle frame $f_{\frac{k}{2}}$ is chosen as the reference frame. That is, as shown in Fig. 1(a), each frame f_j is aligned to the middle frame $f_{\frac{k}{2}}$ before the optical flow computation.

Typical optical flow estimations between the misaligned and aligned lip image pairs (i.e., mouth opening process) are shown in

Fig. 1(b) and Fig. 1(c), respectively. It can be clearly observed that almost all the flow orientations obtained from the misaligned lip image pair point to the top directions. Under such circumstances, the irrelevant pixel motions will affect the interested lip motion flow evidently. As a result, it is very difficult to precisely determine the desired lip event states within this kind of flows, e.g., mouth opening process. In contrast, the optical flows estimated from the aligned lip image pair are able to mark the moving directions of lip pixels perceptually. Within this example, the upper parts of the lip pixels move towards the upward direction, while the lower parts of the lip pixels move towards the downward direction. Therefore, the optical flows estimated from these stabilized image sequences can be well utilized for the subsequent lip motion event detection.

3.2. Oriented histograms of regional optical flow

For lip motion analysis, the motion vector derived from the optical flow is a natural feature to characterize the lip movements. However, as the optical flows are very susceptible to the background noise, scale changes and the directionality of movement, the raw optical flows incorporating the less discrimination power may fail to distinguish the similar motion events. Recently, some researchers have found that the oriented histograms of the optical flow sequence along the temporal axis are able to represent the motion event discriminatively, and this type of descriptor is able to improve the motion analysis performance significantly [29]. This is reasonable because the dynamical patterns of oriented histograms of the optical flow field are able to well characterize the motion appearance distribution globally, which would be less sensitive to the background noise.

Inspired by the recent success of histogram of features in the visual recognition community [29], we select the oriented histograms of optical flow to characterize the lip motion activities. First, we refer to Section 3.1 and align the lip image sequence to reduce the impact of the irrelevant motion caused by the unstable cameras. Next, we substitute frame $f(x, y, t+1)$ in Eq. (1) with $(f \circ \tau)(x, y, t)$ at a time t, and estimate optical flow fields of the aligned image sequence via the Lucas–Kanade algorithm. Mathematically, the magnitude $W(x; y)$ and the orientation $\theta(x; y) \in (-\frac{\pi}{2}, \frac{\pi}{2})$ of the optical flows located at pixel (x, y) are computed by:

$$W(x,y) = \sqrt{u(x,y)^2 + v(x,y)^2}, \tag{4}$$

$$\theta(x,y) = \arctan(v(x,y)/u(x,y)). \tag{5}$$

The motion vector of optical flow is computed at every frame of the video clip, and each flow vector is represented as its primary angle ranging from the horizontal axis. As shown in [29], the orientations of optical flow can be mapped into several angle bins for statistical histogram computation:

$$\begin{cases} -\frac{\pi}{2} < \theta \le -\frac{\pi}{2} + \frac{\pi}{B}, & b = 1 \\ -\frac{\pi}{2} + \frac{(b-1)\pi}{B} < \theta \le -\frac{\pi}{2} + \frac{b}{B}, & 1 < b < B \\ \frac{\pi}{2} - \frac{\pi}{B} < \theta < \frac{\pi}{2}, & b = B \end{cases} \tag{6}$$

where $1 \le b \le B$ is the index of the bin and B is the total number of bins. As a result, the raw histogram of optical flow can be computed as follows:

$$h = \frac{1}{MN} \sum_{x=1}^{M} \sum_{y=1}^{N} \delta[A(\theta(x,y)) - b], \tag{7}$$

$$\delta[\varphi(x)] = \begin{cases} 1, & \varphi(x) = 0 \\ 0, & otherwise \end{cases} \tag{8}$$

where the function $A(\cdot)$ maps the angle $\theta(x, y)$ to its corresponding histogram bin value, and the Dirac delta function $\delta[\varphi(\cdot)]$ is utilized to mark the special bin number particularly.

As the spatially weighted optical histogram not only considers the times of each motion vector appearing in a certain region, but also takes the local characteristics of motion vector into account [29]. For the appearance with almost no temporal motion, the spatial weights tend to zero, whereas these weights are very large within the big motion appearance. Therefore, the utilization of each flow vector weighted according to its magnitude is an effective way to model the motion appearances. Accordingly, the weighted histogram can be formulated as:

$$\hat{h} = C_q \sum_{x=1}^{M} \sum_{y=1}^{N} \delta[A(\theta(x,y)) - b] \cdot W(x,y), \tag{9}$$

where C_q is a constant for histogram normalization.

Evidently, the histogram is inherently a global statistical measure. Nevertheless, the direct utilization of such oriented histogram will not perform a better representation of the lip motions. The main reason lies that the optical flows within the lip region appearance are changing over time and the flow orientations of different parts will mutually affect the probability distribution of the histograms. Under such circumstances, the oriented histograms of such flow vectors are often similar even the lip moving directions are different. Consequently, this type of descriptor would result in a detection failure.

Empirical studies have found that the variations of optical flow vectors within the lip motion appearance are often symmetric because the lip structure is physically symmetric. To avoid the mutual interference within the statistical histogram, it is reasonable to investigate the orientations of optical flow vectors locally. To address this issue, we select to divide the interested mouth region into four separable regions, and propose an oriented histogram of regional optical flow (OH-ROF) to characterize the lip motion appearance. Given a located mouth image of size $M \times N$, the separated four regions R_u, R_d, R_l, R_r, can be mathematically formulated as follows:

$$R_u(x,y) = \begin{cases} y - \frac{N}{M}x \ge 0, \\ y + \frac{N}{M}x - N > 0 \end{cases} \quad \text{s.t. } 0 < x < M, \frac{N}{2} < y \le N. \tag{10}$$

$$R_d(x,y) = \begin{cases} y - \frac{N}{M}x \le 0, \\ y + \frac{N}{M}x - N < 0 \end{cases} \quad \text{s.t. } 0 \le x < M, 0 \le y < \frac{N}{2} \tag{11}$$

$$R_l(x,y) = \begin{cases} y - \frac{N}{M}x > 0, \\ y + \frac{N}{M}x - N \le 0 \end{cases} \quad \text{s.t. } 0 \le x < \frac{M}{2}, 0 < y \le N \tag{12}$$

$$R_r(x,y) = \begin{cases} y - \frac{N}{M}x < 0, \\ y + \frac{N}{M}x - N \ge 0 \end{cases} \quad \text{s.t. } \frac{M}{2} < x \le M, \frac{N}{2} < y \le N \tag{13}$$

where (x, y) denotes the pixel position within the $M \times N$ rectangular lip region. Fig. 2 illustrates the procedures of the proposed regional histogram representation in terms of the four bins, and each OH-ROF

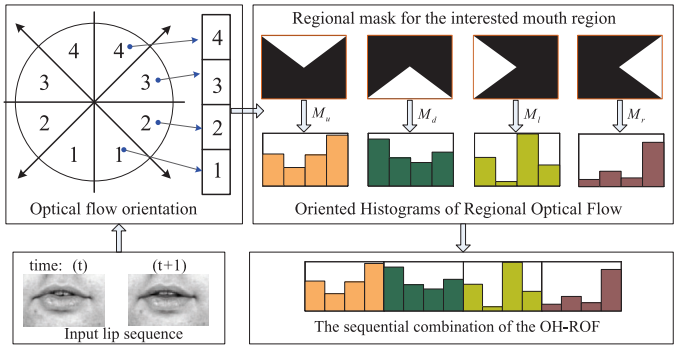


Fig. 2. The pipeline of the proposed OH-ROF visual descriptor.

vector is binned according to its primary angle between the horizontal axis and the vector. According to these region selections, the OH-ROF vectors provide us with four regional histograms at each time instant t:

$$\begin{cases} \hat{h}^t_u &= \hat{h}^t(R_u(x,y)) &= [h^t_{u,1}, h^t_{u,2}, \ldots, h^t_{u,B}] \\ \hat{h}^t_d &= \hat{h}^t(R_d(x,y)) &= [h^t_{d,1}, h^t_{d,2}, \ldots, h^t_{d,B}] \\ \hat{h}^t_l &= \hat{h}^t(R_d(x,y)) &= [h^t_{l,1}, h^t_{l,2}, \ldots, h^t_{l,B}] \\ \hat{h}^t_r &= \hat{h}^t(R_d(x,y)) &= [h^t_{r,1}, h^t_{r,2}, \ldots, h^t_{r,B}] \end{cases}. \tag{14}$$

It can be clearly observed that each OH-ROF vector is independent of the other ones, and these OH-ROF vectors can be efficiently utilized to code the visual appearance of each lip motion frame locally. Since the histogram with a total of B bins is essentially represented as a probability mass function, which should satisfy the following constraint:

$$\sum_{i=1}^{B} \hat{h}^t_{s,i} = 1, \hat{h}^t_{s,i} \geq 0, s = \{u, d, l, r\}. \tag{15}$$

This constraint can also be considered as the normalization operation, and it can make the derived histogram representation scale-invariant. Consequently, the sequential combination of these OH-ROF vectors is able to characterize the lip motions discriminatively, meanwhile the temporal evolution of such a combined vector can be utilized for further lip event analysis.

3.3. Lip motion event detection

Given a video clip, OH-ROF is calculated between every two neighboring frames. As a result, this video clip can be represented by a sequence of OH-ROF descriptors as its signature. Empirical studies have found that the lip motion events can be concretely divided into two patterns: visual silence and speech, in which the lip-dynamic states of speech event can be further comprised of mouth opening and mouth closing. In silent motions, the magnitudes of the optical flows are generally very small, whereas these variations are evidently quite stronger in speech. Therefore, we first select to detect the silent frames and then label the speaking (i.e., opening mouth and closing mouth) frames. Finally, the lip motion events corresponding to visual silence and speech can be well determined.

3.3.1. Silence detection via small flow magnitude

Intuitively, the magnitude of lip pixel flows almost becomes zero when the lip is not moving (i.e., non-speaking period), whereas it has some positive value during the speaking. Therefore, it is natural to detect the silent frames via the optical flow of small magnitudes. Let $G(x, y)$ be the binary mask of small flows at pixels (x, y), the mask corresponding to the flow vector of very small magnitude can be obtained as follows:

$$G(x,y) = \begin{cases} 1, & \text{if} W(x,y) \leq \epsilon \\ 0, & otherwise, \end{cases} \tag{16}$$

where ϵ is a pre-determined threshold utilized for small magnitude determination. If the pixels associated with the very small flow magnitudes almost fill up the whole lip region, this frame can be considered as the silent frame. According to this issue, let $\rho \in [0.9, 1]$ be the tuning parameter to regularize the proportion of the whole region pixels. The silent frames and non-silent frames can be further determined as follows:

$$Event = \begin{cases} Silence, & \text{if} \sum_{i=1}^{M}\sum_{j=1}^{N} G(x,y) \geq \rho(M \times N), \\ speech, & otherwise. \end{cases} \tag{17}$$

3.3.2. Speech state detection via low rank affinity pursuit

After the visual silence detection, the remaining frames with significant flow appearances can be considered as the speaking states. Since OH-ROF vectors are defined and extracted at each frame level, the actual video representation is a time series of histogram descriptors. Therefore, the further lip-dynamic state detection incorporating the mouth opening and closing can be converted to compare these time series equivalently.

In general, the frame descriptors within the same motion event should either share the similar subspace representation that can be grouped together or be highly repetitive. Inspired by this finding, we propose to further detect the lip-dynamic states of non-silent motion event (i.e., mouth opening and closing) by representing the motion sequence as a sequence of subspace clustering based motion subsets instead, in which the frame descriptors within each motion subsets always share the similar low-dimensional subspace representation. Intuitively, the problem of assigning the similar descriptors to its corresponding subspace naturally leads to a challenging problem of subspace clustering, whose goal is to find a multi-subspace representation that best fits the collected data appropriately.

In this paper, we present a low rank affinity pursuit method to detect the similar motion frames within the speaking events. First, we utilize the low rank minimization technique to reduce the effect of noisy motion flows. Then, we construct a local subspace of each flow vector and create an affinity matrix of the whole speaking sequence. Finally, the different lip motion states can be labeled by grouping together all the descriptors sharing the similar subspace representation, meanwhile the labels with respected to the outliers are filtered iteratively to make the motion subsets consecutively and meaningfully. The details are presented as follows:

1:*Low rank minimization:* Given a noise corrupted motion matrix $D = A + E$, where A is an unknown low-rank motion matrix and E is a sparse matrix that represents the noisy components [30]. Consequently, the problem of finding a low rank approximation of D can be formulated as:

$$\min_{A,E} \; rank(A) + \lambda \|E\|_0 \quad \text{s.t.} \quad D = A + E \tag{18}$$

where the parameter λ is a positive value utilized to balance the effects of the two parts. Since this formulation is a highly non-convex optimization problem (i.e., known as NP-hard problem), a common practice is to obtain a tractable optimization by relaxing Eq. (18) as:

$$\min_{A,E} \; \|A\|_* + \lambda \|E\|_1 \quad \text{s.t.} \quad D = A + E \tag{19}$$

where $\|\cdot\|_*$ denotes the nuclear norm, i.e., the sum of the singular values, and $\|\cdot\|_1$ represents the sum of the absolute values of matrix entries. Recently, there have been great progresses on recovering a low rank matrix from the corrupted data and some efficient approaches are available. For computational efficiency, we utilize the accelerated proximal gradient (APG) [31] method to give a solution of Eq. (19) in our implementation.

2:*Local subspace and affinity matrix estimation:* For each frame characterized by an OH-ROF descriptor $\hat{h}_i$, we compute its k nearest neighbors using their Bhattacharya distance as a similarity metric:

$$D(\hat{h}^i, \hat{h}^j) = \sum_{b=1}^{B} \sqrt{\hat{h}^i_b \hat{h}^j_b}. \tag{20}$$

Next, we construct a local subspace $\mathcal{W}_i$ to the current frame based on its k nearest neighbors, which can be achieved by traditional SVD [32]. In this phase, the rank estimation of the local subspace is required in order to truncate the SVD result, which can be accomplished by a Model Selection technique inspired by the work of Kanatani [33].

$$r' = \arg\min_r \left(\frac{\lambda^2_{r+1}}{\sum_{i=1}^{r} \lambda^2_i} + \eta \cdot r \right) \tag{21}$$

where λ_i is the ith singular value, and parameter η depends on the noise level that exists in the local subspace. In general, the higher the noise level is, the larger the value of η should be, and vice versa. As

the local subspace is constructed from the low rank minimization sequence, the noise level is not very high. Therefore, in this step, the value of η should not be assigned at a very large value. Accordingly, we compute an affinity matrix S, in which the affinity $S_{i,j}$ is the inverse of the distance between the local subspace $\mathcal{W}_i$ and $\mathcal{W}_j$ measured in terms of their principal subspace angle θ_{ij}:

$$S_{i,j} = \exp\{-\sin^2(\theta_{ij})\}. \tag{22}$$

3:*Motion state labeling:* As the frame descriptors within each motion subsets always share the similar low-dimensional subspace representation, the motion frames can be labeled by clustering the affinity matrix S and any clustering technique could be used, e.g., spectral clustering, and normalized Cuts.

3.3.3. Outlier filtering

After the silence detection via the small flow magnitudes and speaking frame grouping by low rank affinity pursuit, the whole lip motion sequences can be sequentially labeled into three states: silence, mouth opening, and mouth closing. As the mouth is recorded to be most probably closed during the visual silence, the combination of the successive mouth opening and closing separated by the silent segments can be generally considered as a visual speech interval. Nevertheless, there always exist some outliers within the labeled sequence [34]. That is, these outliers differing from the neighboring frames are always of very small length, e.g., one or two. Since the similar motion frames always appear within the short consecutive frames, these outliers with very limited frame length will affect the lip motion event detection significantly. For example, by the current video capturing device, the frame length of mouth opening process is no less than 4 within a general lip motion event. To tackle this problem, we further utilize a window function to filter out these outliers and replace the label of each outlier with the most frequent item amongst its neighbors. The main steps are summarized as follows:

1:*Outlier detection:* We first detect the label set ℓ of outliers whose local subspace label is not equal to the adjacent ones, i.e., $l_i \neq l_{i-1}$ or $l_i \neq l_{i+1}$.

2:*Neighboring interval marking:* We utilize a window function $w_c[\alpha]$ to mark the neighbor interval of each outlier:

$$w_c[\alpha] = \begin{cases} 1, & \ell(i) - c \leq \alpha \leq \ell(i) + c \\ 0, & otherwise \end{cases} \tag{23}$$

where c is the half size of window function.

3:*Most frequent item filtering:* We select the Most Frequent Item within the marking interval to filter the current outlier until no label value is changed in label sequence L, i.e.

$$\bar{l}_i = \arg\max_s \left(Count[w_c(\alpha) * L == s]\right) \tag{24}$$

where s is the label categories. After these operations, the outliers with very limited frame length can be filtered appropriately and the length of each motion event can be restricted to an appropriate value. As a result, the detected lip event clips would be physically and perceptually meaningful.

3.4. Parameters analysis and tuning

In general, the motion descriptors within visual speech interval can be characterized by a mixture of multiple low dimensional subspaces, and the lip-dynamic states are only comprised of either mouth opening and closing. Therefore, it is reasonable to set the subspace number at 2 for the lip-dynamic state labeling. By a rule of thumb, the number of the nearest neighbors is set at 6 for the local subspace estimation. To reduce the noise impact within the local subspace estimation, the rank of the local subspace can be estimated by a Model Selection technique [33]. Nevertheless, the noise level is unknown and the rank of very small value will more or less impact the realistic motion vectors. To avoid this problem, the rank of the transformed low-dimensional subspace is empirically assigned to the maximum value between 3 and the result generated by Model Selection technique, whereby the motion semantics of flow vectors within the original lip motion sequence can be well maintained. In addition, the spectral clustering algorithm is employed for affinity matrix clustering. Although there exists a short lip motion event, the frame length of such a motion clip would be no less than 4 in general recordings naturally. Therefore, the half size c of the window function is fixed to be 3 for outlier label filtering. As suggested in paper [27,29,31], the number of histogram bin is selected to be 4, while the parameter ϵ, ρ and η are empirically set at 0.53, 0.91, and 1, respectively.

4. Experiments

To evaluate the effectiveness of the proposed lip event detection approach, 60 video clips capturing under different environments were collected for testing. In particular, the mouth regions displayed in these videos were located using the method in [35] and chosen to be predominantly frontal. In the past, some lip motion event detection approaches either extracting the lip shapes or learning a reference model were able to characterize the lip movements. Nevertheless, it is very difficult to obtain the geometric lip parameters with great reliability when the mouth region incorporates the poor contrast between the lip and skin. Meanwhile, it is impractical to establish a training data set and perform a training process to determine the referential lip models in advance. Therefore, the meaningful and fair comparisons with these systems are not presented here. To validate the detection performance, we selected four representative methods (i.e., AOFE [14], MHI-ZM [11], MASF [10], PWPC [15]) and utilized the same parameters as the ones the authors have given to investigate the lip motion events. The main ideas of these competing algorithms are summarized as follows:

AOFE approach: This algorithm utilizes the Affine Optical Flow Estimation to exploit the divergence of the flow field at a coarse scale, whereby the lip-dynamic states corresponding to the mouth opening and closing can be determined.

MHI-ZM method: This approach computes the Motion History Images and employs the Zernike Moment features to segment the isolated utterances, in which the magnitude of Zernike moments corresponding to the uttering frames is much greater than the one of the frames within the period of pause or silence.

MASF approach: This method performs a series of Mouth Area Subtractions of the consecutive frames and selects a smoothing Filtering to achieve the syllable separation visually.

PWPC algorithm: This approach utilizes the Pair-Wise Pixel Comparison scheme between the consecutive mouth images to achieve an isolated word segmentation visually.

In the following sections, we will introduce the experimental setup and then conduct the experiments on different lip motion clips, in which the motion states, detection performances, and empirical comparisons, as well as the related discussions, are included.

4.1. Experimental setup

In order to determine the ground truth, each motion frame has been marked as either visual silence or visual speech via carefully visual inspection, in which the lip-dynamic states of visual speech with respect to the mouth opening and closing were further tagged elaborately. In this paper, we mainly concentrate on detecting the lip motion event in three cases: 1) Mouth opening and closing detection in speech; 2) Speech detection under uniform lighting condition; and 3) Speech detection under mobile platform.

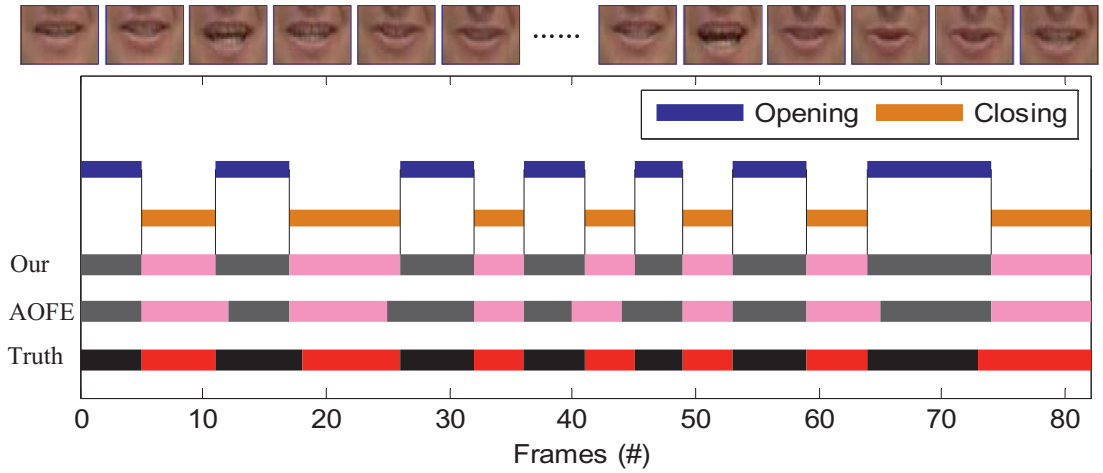


Fig. 3. The lip-dynamic state detection on the VidTIMIT database.

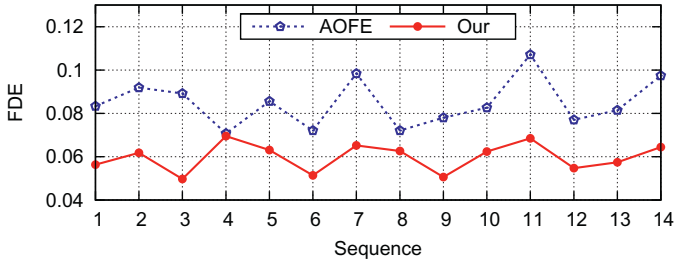


Fig. 4. The FDE values obtained from VidTIMIT database.

4.2. Mouth opening and closing detection in speech

In this case, 14 sequences were downloaded from the publicly available VidTIMIT database [36], and seven speakers (file: 2, 8, 13, 19, 25, 37, 42) phonically recited the short sentences (i.e., sa1 and sx408) quickly and consecutively[1]. Within these continuous speeches, we extracted the speaking sequences and mainly concentrated on the lip-dynamic state detection (i.e., mouth opening and closing) specifically. To the best of our knowledge, few existing algorithms can detect the motion states of mouth opening and closing without extracting the lip shapes and learning the lip model priors, except AOFE method [14]. Therefore, we mainly focus on comparing the proposed approach with this method extensively.

A snapshot of the lip-dynamic state detection result on the VidTIMIT database is shown in Fig. 3. It can be seen that the proposed approach outperforms the AOFE approach visually. Fig. 4 illustrates the frame detection errors (FDE), which is defined as the ratio between the error detected frames and the ground truth, of all the tested sequences. It can be seen that the proposed approach has always generated the smaller frame detection errors. For example, the FDE values obtained by the proposed approach are almost no more than 0.07, while those values obtained by the AOFE approach are always larger than 0.07. It indicates that the proposed approach is able to well determine the lip-dynamic states of the mouth opening and closing, and the segments of these states are closer to the ground truth. The main reason lies that the optical flow vectors of surrounding lips obtained by the AOFE approach will more or less impact the divergence estimation, which may result in a detection failure when the whole flows are directly utilized for motion event analysis. Comparatively speaking, the obtained flow vectors within the proposed approach are weighted according to its magnitude, which is an effective way to reduce the impact of irrelevant motion vectors, i.e., the unordered flows around the lips. Meanwhile, the proposed approach incorporating the regional optical flow is able to degrade the mutual interference between different flow parts. In addition, the proposed low rank affinity pursuit method holds a strong ability to reveal the motion subset sharing the similar low-dimensional subspace representation, which is robust against noise in unconstrained videos. As a result, some ambiguous event decision boundary within this kind of oriented histogram can be well determined.

4.3. Speech detection under uniform lighting condition

In this case, 30 sequences were collected in an office environment with almost uniform lighting conditions. Six speakers phonically uttered the English digits from zero to nine for five times, in which there existed a short silence between the neighboring utterances. As the mouth was recorded to be most probably closed during the visual silence period, the combination of the successive mouth opening and closing can be considered as a short visual speech interval appropriately. Specifically, we detected the lip motion event of visual silence and speech, and compared the proposed approach with MHI-ZM, MASF and PWPC methods.

A snapshot of the speech detection result is shown in Fig. 5. It can be observed that the proposed approach is able to well detect the visual silence and speech, and most of the segmented video clips are close to the ground truth. It indicates that the detection performance obtained by the proposed approach is visually better than the other three competing approaches. In addition, the proposed approach can not only detect the visual silence and speech, but also reveal the lip-dynamic states of mouth opening and closing appropriately.

Further, we utilize the false alarm rate (FAR) and missing alarm rate (MAR) to measure the detection performances quantitatively, where FAR is defined as the percentage of the frames which are detected to be speech but silence actually, MAR is defined as the percentage of the frames which are detected to be silence but speech actually, both relative to the total frames. These two evaluation metrics indicate that the less FAR and MAR values would achieve a higher detection accuracy.

The FAR and MAR values obtained by the different approaches are shown in Figs. 6 and 7, respectively. Table 1 shows the mean values of FAR and MAR, which are, respectively, the average of the corresponding values within all the tested sequences. It can be observed that the proposed approach has always generated the smallest FAR and MAR values in comparison with the other three competing approaches.

[1] http://conradsanderson.id.au/vidtimit/.

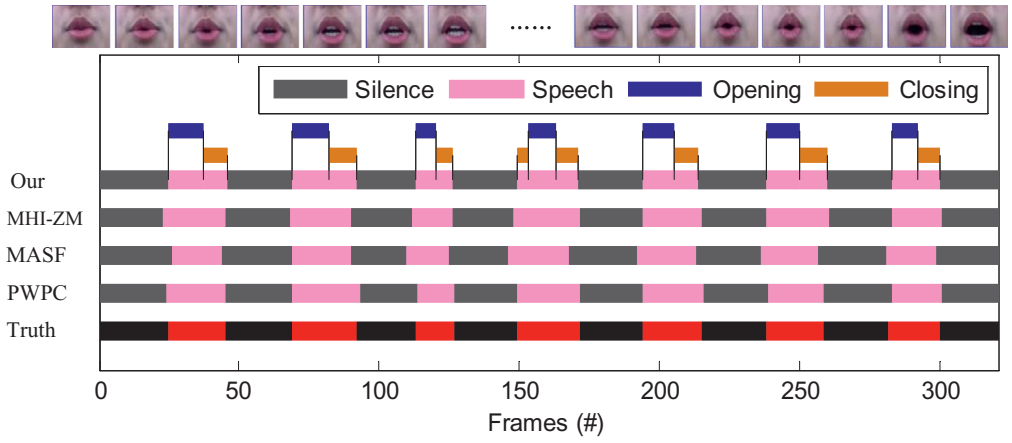


Fig. 5. The speech detection under uniform lighting condition.

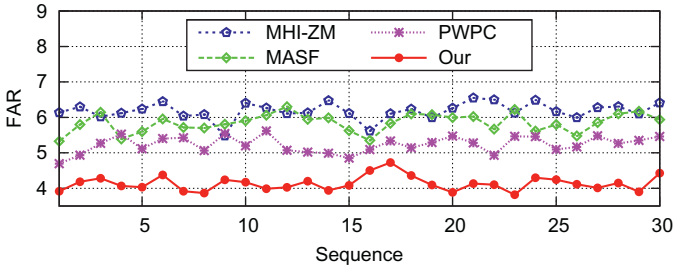


Fig. 6. The FAR values detected under uniform lighting condition.

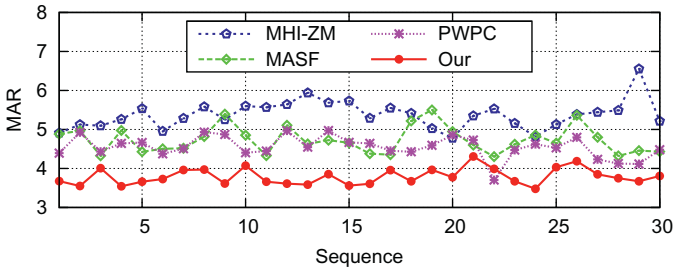


Fig. 7. The MAR values detected under uniform lighting condition.

Table 1
Mean detection errors (FAR and MAR, Office).

Method	MHI-ZM	MASF	PWPC	Our
FAR	6.23	5.88	5.23	4.12
MAR	5.43	4.77	4.53	3.87

The MHI-ZM method [11] utilizing an accumulative image difference technique to detect the changes between consecutive frames is capable of detecting the visual silence and speech in most cases. Nevertheless, there exist a lot of irrelevant motions around the lips, which often degrade the event detection performance seriously. Although the MASF [10] method is able to well extract the mouth areas under uniform lighting condition, the obtained mouth areas are not stable enough, which often fail to give a better representation of the real mouth areas consistently. As a result, the filtered signals of the unstable mouth areas may not exactly determine the event boundaries. PWPC [15] algorithm first utilizes the squared mean difference of gray-scale intensities of corresponding pixels in accumulative frames to characterize the lip movements, and then detects the boundaries of motion event via the step-pulse-shaped representation. Nevertheless, this type of approach incorporating the pair-wise pixel comparing is very sensitive to the irrelevant motion caused by an unstable camera, which often degrades its detection boundaries between the speech and silence.

Table 2
Mean detection errors (FAR and MAR, Mobile).

Method	MHI-ZM	MASF	PWPC	Our
FAR	7.85	7.14	6.34	4.83
MAR	6.74	6.19	5.73	4.38

By contrast, the proposed approach is able to detect the motion event boundaries appropriately and the event segments are not deviated significantly from the ground truth. As shown in Table 1, the values of mean FAR and MAR obtained by the proposed method are always lower than the other three competing methods. That is, the proposed approach has achieved the best detection performances.

4.4. Speech detection under mobile platform

In this case, 16 sequences were captured by a hand-held mobile phone under uneven illuminations. Four speakers were asked to repeat the “hello” word for four times, in which there existed a short silence during each utterance. In this situation, the located mouth sequences were somewhat unstable.

A snapshot of the speech detection result under mobile platform is shown in Fig. 8. It can be observed that the MASF approach often fails to give a better detection of visual silence and speech, in which the event segments are deviated seriously from the ground truth. The main reason lies that the appearance within these mouth sequences capturing under uneven illumination does not have sufficient contrast for precise mouth area extractions, which may lead to an inaccurate detection. Furthermore, the MHI-ZM and PWPC approaches also degraded their performance to some extent. The main reason lies that the extracted lip sequences captured under hand-held mobile platform are always unstable, thereby the computations of motion history image and the comparisons of pair-wise pixels cannot exactly reveal the lip-dynamic states. As a result, the MHI-ZM method often fails to provide an accurate Zernike moments for real lip motion analysis, while the PWPC algorithm cannot provide a stable lip pixel comparison consecutively. By contrast, it can be found that the proposed approach incorporating the sequence alignment is able to well detect the speech event in mobile video visually and the obtained event segments are closer to the ground truth.

Further, the FAR and MAR values of the tested sequences are shown in Figs. 9 and 10, respectively, meanwhile the mean values of FAR and MAR obtained by the different approaches are shown in Table 2. It can be found that the MASF and MHI-ZM approaches

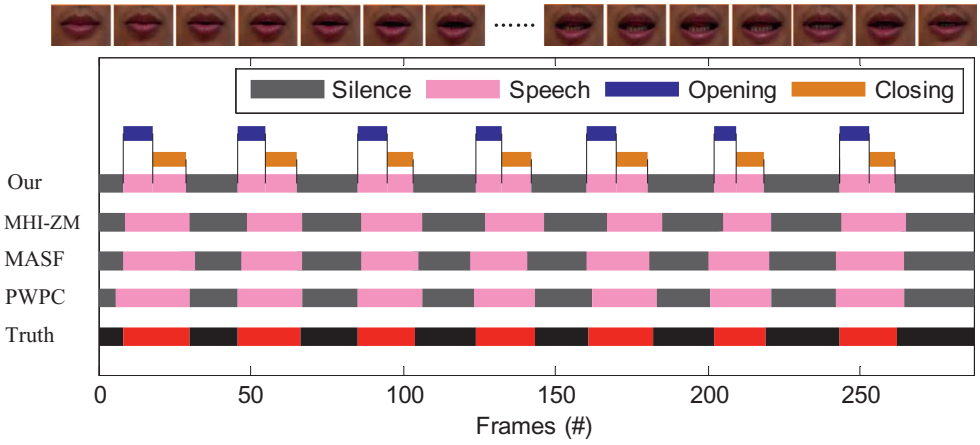


Fig. 8. The speech detection under mobile platform.

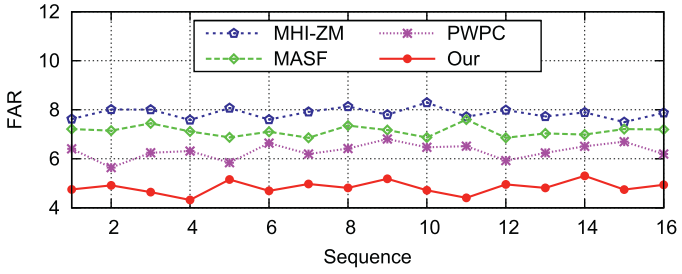


Fig. 9. The FAR values detected under mobile platform.

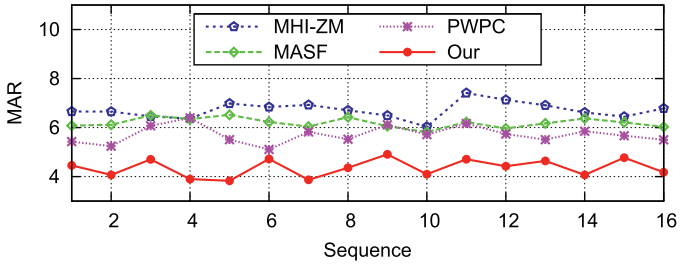
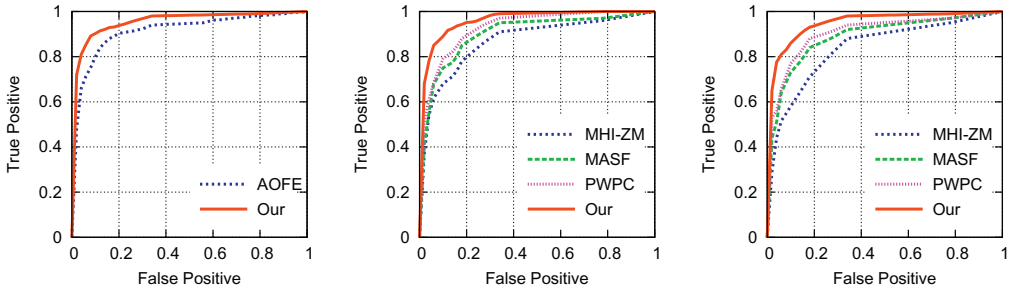


Fig. 10. The MAR values detected under mobile platform.

always generate relatively large detection errors. Although the PWPC algorithm is able to detect the desired lip events in some segments, some motion frames within the neighboring events cannot often be detected accurately. As a result, the degraded performances contributed to a big large detection errors. In contrast, the detection errors obtained by the proposed approach are always smaller than the other three competing approaches. It implies that the proposed approach is able to detect the lip motion events precisely, and the detected event intervals can well appropriate the motion event intervals. Remarkably, the proposed approach could investigate the lip-dynamic states about the mouth opening and closing simultaneously. Experiments have shown the promising results.

4.5. Detection analysis and discussion

We further consider the lip-dynamic state detection and visual speech detection as a binary classification problem, in which the frames of mouth opening and visual speech are marked as the positive labels. Specifically, this kind of detection performance can be evaluated using the ROC curve [37], which graphically demonstrates the changes of true positive rate with respect to the changes of false positive rate in the classification. The ROC curves obtained by the different approaches and tested on the different cases are shown in Fig. 11. It can be seen that the proposed approach has achieved the best detection performances and improved the state-of-the-art



(a) ROC curve in Subsection 4.2 (b) ROC curve in Subsection 4.3 (c) ROC curve in Subsection 4.4

Fig. 11. The ROC curves obtained by different approaches: (a) Lip-dynamic state detection. (b) Speech detection under uniform lighting condition. (c) Speech detection in mobile platform.

results significantly. The main reasons are two-fold: 1) The proposed approach aligns the lip sequences to reduce the impact of irrelevant motions, whereby the reliable flows can be calculated for lip motion analysis; 2) The proposed OH-ROF aiming to reduce the mutually impact existing in the global flows is able to characterize the motion event boundaries discriminatively.

Moreover, since the proposed approach employs an image sequence stabilization to reduce the impact of irrelevant motions caused by the moving cameras, the more computational load is inevitably required. Fortunately, the processing time obtained by the proposed approach is acceptable, e.g., the execution time is around 3.75 s when testing on a lip motion clip of 80 frames (scale size 112×76) and performing in a Matlab coding platform, while the AOFE, MHI-ZM, MASF and PWPC approaches cost 2.18 s, 4.17 s, 3.53 s and 3.41 s, respectively. Except for the AOFE method, which need not a sequence alignment to handle the irrelevant motions, the computation time obtained by the proposed approach is less than the MHI-ZM approach, and is also comparable to the MASF and PWPC methods. Different from the MHI-ZM approach incorporating the stationary wavelet transform on each motion frame to reduce the small variations of the mouth movements, the proposed approach is to reduce the impact of irrelevant motion in the whole motion clip. Accordingly, the less time is needed. Although the image sequence stabilization resulted in a bit more execution time, importantly, the proposed approach has achieved the best detection performance. Furthermore, it is worth noting that the proposed approach does not require any training set on labeled videos or learn the lip motion priors of each visual event in an unconstrained video. With more powerful coding platform, it is expected that the proposed approach would be suitable for real time applications.

5. Conclusion

In this paper, we have proposed an efficient lip motion event detection approach using the oriented histograms of regional optical flow and low rank affinity pursuit. Without any training set on labeled video clips or learning the lip motion priors in unconstrained videos, the proposed approach aims not only to distinguish the frames that depict visual speech from those describing visual silence, but also to investigate the lip-dynamic states of mouth opening and closing simultaneously. Extensive experiments tested on different kinds of video sequences have demonstrated the efficiency of the proposed approach in comparison with the existing counterparts.

Acknowledgment

The work described in this paper was supported by the National Science Foundation of China (No. 61272366, 61300138), the National Science Foundation of Fujian (No. 2014J01239), the Research Foundation (No. 14BS207), the Promotion Program for Young and Middle-aged Teacher in Science and Technology Research (No. ZQN-PY309) of Huaqiao University, and also partially supported by the Faculty Research Grant of Hong Kong Baptist University (No. FRG2/14-15/075, FRG1/14-15/041), and the Research Grants MYRG187 (Y1-L3)-FST11-TYY, MYRG205 (Y1-L4)-FST11-TYY, RDG009/FST-TYY of University of Macau, and Macau FDC Grants T-100-2012-A3 and 026-2013-A.

References

[1] S.L. Wang, A.W.C. Liew, W.H. Lau, S.H. Leung, An automatic lipreading system for spoken digits with limited training data, IEEE Trans. Circuits Syst. Video Technol. 18 (12) (2008) 1760–1765.
[2] H. Cetingul, Y. Yemez, E. Engin, A. Tekalp, Discriminative analysis of lip motion features for speaker identification and speech-reading, IEEE Trans. Image Process. 15 (10) (2006) 2879–2891.
[3] X. Liu, Y.M. Cheung, Learning multi-boosted hmms for lip-password based speaker verification, IEEE Trans. Inf. Forens Secur. 9 (2) (2014) 233–246.
[4] M. Faraj, J. Bigun, Synergy of lip-motion and acoustic features in biometric speech and speaker recognition, IEEE Trans. Comput. 56 (9) (2007) 1169–1175.
[5] M. Bendris, D. Charlet, G. Chollet, Lip activity detection for talking faces classification in tv-content, in: Proceedings of International Conference on Machine Vision, 2010, pp. 187–190.
[6] Y. Tian, T. Kanade, J. Cohn, Recognizing action units for facial expression analysis, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2) (2001) 97–115.
[7] D. Sodoyer, B. Rivet, L. Girin, J. Schwartz, C. Jutten, An analysis of visual speech information applied to voice activity detection, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2006, pp. 601–604.
[8] B. Rivet, L. Girin, C. Jutten, Visual voice activity detection as a help for speech source separation from convolutive mixtures, Speech Commun. 49 (7) (2007) 667–677.
[9] M. Aoki, K. Masuda, H. Matsuda, T. Takiguchi, Y. Ariki, Voice activity detection by lip shape tracking using EBGM, in: Proceedings of International conference on Multimedia, 2007, pp. 561–564.
[10] H. Talea, K. Yaghmaie, Automatic visual speech segmentation, in: Proceedings of IEEE International Conference on Communication Software and Networks, 2011, pp. 184–188.
[11] W. Yau, H. Weghorn, D. Kumar, Visual speech recognition and utterance segmentation based on mouth movement, in: Proceedings of Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications, 2007, pp. 7–14.
[12] V. Libal, J. Connell, G. Potamianos, E. Marcheret, An embedded system for in-vehicle visual speech activity detection, in: Proceedings of IEEE 9th Workshop on Multimedia Signal Processing., 2007, pp. 255–258.
[13] S. Siatras, N. Nikolaidis, M. Krinidis, I. Pitas, Visual lip activity detection and speaker detection using mouth region intensities, IEEE Trans. Circuits Syst. Video Technol. 19 (1) (2009) 133–137.
[14] S. Karlsson, J. Bigun, Lip-motion events analysis and lip segmentation using optical flow, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition Workshops, 2012, pp. 138–145.
[15] A. Shaikh, D. Kumar, J. Gubbi, Automatic visual speech segmentation and recognition using directional motion history images and Zernike moments, Vis. Comput. 29 (10) (2013) 969–982.
[16] S. Taeyup, L. Kyungsun, K. Hanseok, Visual voice activity detection via chaos based lip motion measure robust under illumination changes, IEEE Trans. Consum. Electron. 60 (2) (2014) 251–257.
[17] F. Luthon, M. Lievin, Lip motion automatic detection, in: Proceedings of Scandinavian conference on image analysis, 1997, pp. 253–260.
[18] J. Luettin, N.A. Thacker, S.W. Beet, Visual speech recognition using active shape models and hidden Markov models, in: Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, 2, 1996, pp. 817–820.
[19] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, R. Harvey, Extraction of visual features for lipreading, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2) (2002) 198–213.
[20] P. Liu, Z. Wang, Voice activity detection using visual information, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004, pp. 609–612.
[21] A. Aubrey, Y. Hicks, J. Chambers, Visual voice activity detection with optical flow, IET Image Process. 4 (6) (2010) 463–472.
[22] R. Navarathna, D. Dean, S. Sridharan, C. Fookes, P. Lucey, Visual voice activity detection using frontal versus profile views, in: Proceedings of International Conference on Digital Image Computing Techniques and Applications, 2011, pp. 134–139.
[23] P. Tiawongsombat, M. Jeong, J. Yun, B. You, S. Oh, Robust visual speakingness detection using bi-level HMM, Pattern Recognit. 45 (2) (2012) 783–793.
[24] D. Sun, S. Roth, M.J. Black, Secrets of optical flow estimation and their principles, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2432–2439.
[25] A. Bruhn, J. Weickert, Lucas/kanade meets horn/schunck: Combining local and global optic flow methods, Int. J. Comput. Vis. 61 (3) (2005) 211–231.
[26] S. Baker, I. Matthews, Lucas-kanade 20 years on: A unifying framework, Int. J. Comput. Vis. 56 (3) (2004) 221–255.
[27] Y.G. Peng, A. Ganesh, J. Wright, W.L. Xu, Y. Ma, Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2233–2246.
[28] J. Odobez, P. Bouthemy, Robust multiresolution estimation of parametric motion models, J. Vis. Commun. Image Represent. 6 (4) (1995) 348–365.
[29] R. Chaudhry, A. Ravichandran, G. Hager, R. Vidal, Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1932–1939.
[30] J. Wright, A. Ganesh, S. Rao, Y. Peng, Y. Ma, Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization, in: Proceedings of Advances in neural information processing systems, 2009, pp. 2080–2088.
[31] K. Toh, S. Yun, An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems, Pacific J. Optim. 6 (15) (2010) 615–640.
[32] J. Yan, M. Pollefeys, A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate, in: Proceedings of European Conference on Computer Vision, 2006, pp. 94–106.
[33] K. Kanatani, Motion segmentation by subspace separation: Model selection and reliability evaluation, Int. J. Image Gr. 2 (2) (2002) 179–197.

[34] X. Liu, Y.M. Cheung, S.J. Peng, Z. Cui, B.N. Zhong, J.X. Du, Automatic motion capture data denoising via filtered subspace clustering and low rank matrix approximation, Signal Process. 105 (12) (2014) 350–362.
[35] R. Lienhart, L. Liang, A. Kuranov, A detector tree of boosted classifiers for real-time object detection and tracking, in: Proceedings of IEEE International Conference on Multimedia and Expo, 2003, pp. 277–280.
[36] C. Sanderson, B.C. Lovell, Multi-region probabilistic histograms for robust and scalable identity inference, in: Proceedings of International Conference on Advances in Biometrics, 2009, pp. 199–208.
[37] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874.