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/© 2015 Elsevier Inc. All rights reserved.

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 bilevel 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