Signal Processing 89 (2009) 2425–2434

Contents lists available at ScienceDirect

Signal Processing

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

Automatic lip localization under face illumination with shadow consideration

Meng Li, Yiu-ming Cheung *

Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China

ARTICLE INFO

Article history:
Received 12 November 2008
Received in revised form
11 May 2009
Accepted 21 May 2009
Available online 6 June 2009

Keywords:
Lip-reading
Lip localization
Face illumination
Shadow

ABSTRACT

Lip-reading has potential attractive applications in information security, speech recognition, secret communication and so forth. To build an automatic lip-reading system, one key issue is how to locate the lip region, particularly under the changing illumination condition. Empirical studies have shown that the recognition rate of a lip-reading system greatly relies on the accuracy of the lip localization. Unfortunately, to the best of our knowledge, lip localization under face illumination with shadow consideration has not been well solved yet. Moreover, this problem is also one of the major obstacles to keeping an automatic lip-reading system from the practical applications. This paper therefore concentrates on this problem and proposes a new approach to obtain the minimum enclosing rectangle surround of a mouth automatically based upon the transformed gray-level image. In this approach, a pre-processing is firstly made to reduce the interference caused by shadow and enhance the boundary region of lip, through which the left and right mouth corners are estimated. Then, by building a binary sequence based on the gray-level values along with the vertical midline of mouth, the top and bottom crucial points can be estimated. Experiments show the promising result of the proposed approach in comparison with the existing methods.

1. Introduction

Motivated by human ability to lip-read, the useful information on speech content can be obtained through analyzing the subtle cue conveyed by lip movement of speakers [1]. The intimate relation between the audio and the visual sensory modality in human recognition can be demonstrated with audio-visual illusions such as the “McGurk effect” [2]. It suggests that speech perception is multimodal involving information from more than one sensory modality. In 1984, the first automatic lip-reading system was presented by Petajan [3,4]. From then on, lip-reading has received considerable attention from the community because of its potential attractive applications in information security, speech recognition, secret communication, and so forth [5,6]. For example, we can utilize the lip-reading technique as a visual password complementary to the popular character-based password to enhance the security level in banks, e-business, home security, and so forth.

In lip-reading, one key issue is the lip localization, i.e. how to obtain the accurate position of lip or mouth from image. Paper [7] demonstrates that the error rate of this automatic visual speech recognition (AVSR) system in studio environment, i.e. ideal light condition without shadow, is 37.3%. In contrast, the visual-only word error rate will reach 76.2% when an AVSR system is utilized in real world. One main reason is that the shadows will be generated when the face illumination comes from the different directions, i.e. parts of mouth region of a speaker will be covered by shadows. Under this situation, it is non-trivial to localize the lip precisely from a face image.

* Corresponding author.
E-mail address: ymc@comp.hkbu.edu.hk (Y.-m. Cheung).

0165-1684/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2009.05.027

Fig. 1. The vertices marked by dot along the vertical and horizontal axis of a mouth.

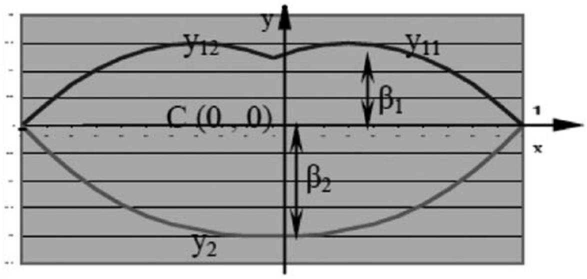

Fig. 2. The illustration of lip shape model proposed in [11].

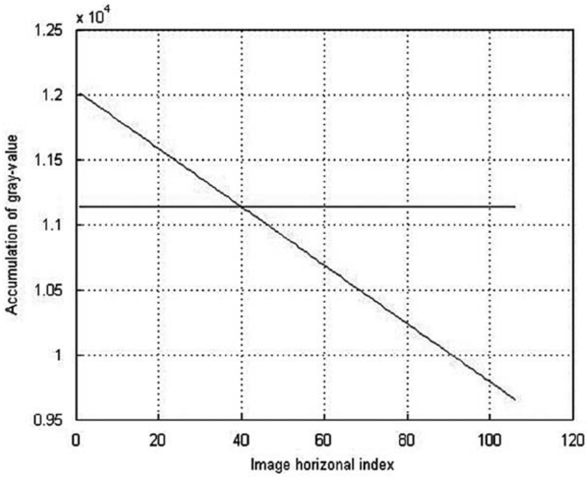

Fig. 3. The linear fitting of accumulation of gray-level value for each column and the mean value. The intersection of the two curves is corresponding into the boundary of shadow.

Consequently, the imprecise localization of lip will mostly lead to the degraded performance of an AVSR system. This implies that the accuracy of lip localization is one of the most important factors to determine the recognition rate of lip-reading. In this paper, we will therefore concentrate on studying the lip localization only.

Thus far, several methods have been proposed to enhance the performance of lip localization for AVSR system. For instance, paper [8] presents an approach that employs the hue-filter to distinguish lip and surrounding skin region. The papers in [9–12] utilize the information of red component and saturation to localize the lip region. Also, paper [13] utilizes a gradient-based Canny edge detector to locate the mouth corner. In [14], the input image is projected into YUV color coordinate system and the accumulations of V value in each row and column of the image are utilized to estimate the crucial points (i.e. the top, bottom and two mouth corners) of a lip.

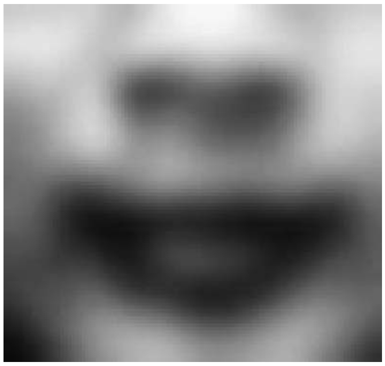

Fig. 4. The filtered image composed of two filtered sub-images.

Fig. 5. The image using the pre-processing that is made up by the two components: I_{sl} and I_{sr}. The vertical crease is corresponding into the shadow boundary.

Furthermore, a class of widely used methods is active shape model (ASM) [15] or active appearance model (AAM) [16]-based ones [17–23]. They build a deformable model for lip by learning the patterns of variability from a training set of correctly annotated images. The shape of modal can be adjusted by a parameter set so as to match and locate the lip in test image. Empirical studies have shown their success, but they need to label some landmarks manually for training. Alternatively, optical flow-based methods give an effective way to locate the lip region [24]. They utilize the apparent velocity distribution of the brightness patterns in an image to obtain the boundary of lip. The optical flow-based methods can give the important information regarding the spatial arrangement of the viewed objects and the rate of change in this arrangement, but sensitive to the translation, scaling, rotation, and the change of illumination condition [25] in particular.