Available online at www.sciencedirect.com
ScienceDirect
Signal Processing 86 (2006) 3549–3558
SIGNAL PROCESSING
www.elsevier.com/locate/sigpro

Multimodal speaker/speech recognition using lip motion, lip texture and audio ☆

H.E. Çetingül*, E. Erzin, Y. Yemez, A.M. Tekalp

College of Engineering, Koç University, Sarıyer, Istanbul 34450, Turkey

Received 1 July 2005; received in revised form 5 December 2005; accepted 1 February 2006
Available online 2 June 2006

Abstract

We present a new multimodal speaker/speech recognition system that integrates audio, lip texture and lip motion modalities. Fusion of audio and face texture modalities has been investigated in the literature before. The emphasis of this work is to investigate the benefits of inclusion of lip motion modality for two distinct cases: speaker and speech recognition. The audio modality is represented by the well-known mel-frequency cepstral coefficients (MFCC) along with the first and second derivatives, whereas lip texture modality is represented by the 2D-DCT coefficients of the luminance component within a bounding box about the lip region. In this paper, we employ a new lip motion modality representation based on *discriminative analysis* of the dense motion vectors within the same bounding box for speaker/speech recognition. The fusion of audio, lip texture and lip motion modalities is performed by the so-called *reliability weighted summation* (RWS) decision rule. Experimental results show that inclusion of lip motion modality provides further performance gains over those which are obtained by fusion of audio and lip texture alone, in both speaker identification and isolated word recognition scenarios.
© 2006 Published by Elsevier B.V.

Keywords: Speaker identification; Isolated word recognition; Lip reading; Lip motion; Decision fusion

1. Introduction

Audio is probably the most natural modality to recognize speech content and a valuable source to identify a speaker [1]. Video also contains important biometric information, which includes face/lip texture and lip motion information that is correlated with the audio. Audio-only speaker/speech recognition systems are far from being perfect especially under noisy conditions. Furthermore, it is a known fact that the content of speech can be revealed partially through lip-reading. Performance problems are also observed in video-only speaker/speech recognition systems, where poor picture quality, changes in pose and lighting conditions, and varying facial expressions may have detrimental effects [2,3]. Hence, robust solutions for both speaker and speech recognition should employ multiple modalities, such as audio, lip texture and lip motion in a unified scheme.

The design of a multimodal recognition system requires addressing three basic issues: (i) Which

☆This work has been supported by TÜBİTAK under the project EEEAG-101E026 and by the European FP6 Network of Excellence SIMILAR (http://www.similar.cc).
*Corresponding author.
E-mail addresses: ertan@cis.jhu.edu (H.E. Çetingül), eerzin@ku.edu.tr (E. Erzin), yyemez@ku.edu.tr (Y. Yemez), mtekalp@ku.edu.tr (A.M. Tekalp).

0165-1684/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.sigpro.2006.02.045

modalities to fuse; (ii) How to represent each modality with a discriminative and low-dimensional set of features; and (iii) How to fuse existing modalities. Speech content and voice can be interpreted as two different, though correlated, information existing in audio signals. Likewise, video signal can be split into different modalities, such as face/lip texture and lip motion. The second issue, representative feature selection, also includes modeling of classifiers through which each class is represented with a statistical model or a representative feature set. Curse of dimensionality, computational efficiency, robustness, invariance and discrimination capability are the most important criteria in selection of the feature set and the recognition methodology for each modality. As for the final issue, that is, the fusion problem, different strategies are possible: in the so-called "early integration", modalities are fused at data or feature level, whereas in "late integration" decisions or scores resulting from each unimodal recognition are combined to give the final conclusion. Multimodal decision fusion can also be viewed from a broader perspective as a way of combining classifiers, which is a well-studied problem in pattern recognition. The main motivation for multimodal fusion is to compensate possible misclassification errors of a certain classifier with other available classifiers and to end up with a more reliable overall decision. Misclassification errors are in general inevitable due to numerous factors such as environmental noise, measurement and modeling errors or time-varying characteristics of signals. A comprehensive survey and discussion on classifier combination techniques can be found in [4].

State-of-art speech recognition systems have been jointly using lip information with audio [5–9]. For speech recognition, it is usually sufficient to extract the principal components of the lip information and to match the mouth openings–closings with the phonemes of speech. Speaker identification using audio and lip information, on the other hand, has been addressed in only few works such as [10–15]. The main challenge is that the principal components of the lip information are not usually sufficient to discriminate between speakers. Non-principal components are also valuable especially when the objective is to model the biometrics. In the speaker/speech recognition literature, audio is generally modeled by mel-frequency cepstral coefficients (MFCC) [16]. However for lip information, there are several approaches reported in the literature such as texture-based, motion-based, geometry-based and model-based [17]. In texture-based approaches, pure or DCT-domain lip image intensity is used as features [8,11,18]. Motion-based approaches compute motion vectors to represent the lip movement during speaking [10,19]. Geometry-based and model-based approaches, in fact, utilize similar processing methods such as active shape models [20,21], active contours [22,23] or parametric models [24] to segment the lip region. They differ in feature selection such that model-based approaches assign the fitted model parameters as features, while shape features such as lengths of horizontal and vertical lip openings, area, perimeter, pose angle are selected for lip representation in geometry-based approaches. In [10], the lip motion is represented by the full set of DCT coefficients of the dense optical flow vectors computed within the rectangular lip region, and then fused with the face texture and the acoustic features for multimodal speaker identification. However, no discrimination analysis and dimensionality reduction are performed in [10]. The speaker recognition schemes proposed in [10,12,13,25,26] are basically opinion fusion techniques that combine multiple expert decisions through adaptive or non-adaptive weighted summation of scores, whereas in [15,27], fusion is carried out at feature-level by concatenating individual feature vectors so as to exploit the temporal correlations that may exist between audio and video signals. In audio-visual speech recognition [18] concatenates audio and lip data, while in [28] unimodal decisions are combined to obtain the fused result. Furthermore, recent works show the success of multistream HMM structures in speech recognition [7–9,17].

In this study, we use the lip motion features that are extracted by a novel discrimination analysis method [19]. Then we integrate lip texture, lip motion and audio features by the reliability-based decision fusion system reported in [11]. The main contribution of this paper is to investigate the fusion of audio modality with the best lip motion and texture representations for two distinct problems, speaker and speech recognition. In this investigation, the performance gain due to the fusion and the optimal modality selection for speaker and speech recognition problems are also discussed. The audio and lip features are presented in detail in Section 2. In Section 3, we describe the probabilistic framework that we use for the speaker/speech recognition problem, and present the reliability weighted

Download English Version:

https://daneshyari.com/en/article/564966

Download Persian Version:

https://daneshyari.com/article/564966

Daneshyari.com