Chapter XVII
Lip Motion Features for Biometric Person Recognition

Maycel Isaac Faraj
Halmstad University, Sweden

Josef Bigun
Halmstad University, Sweden

ABSTRACT

The present chapter reports on the use of lip motion as a stand alone biometric modality as well as a modality integrated with audio speech for identity recognition using digit recognition as a support. First, the authors estimate motion vectors from images of lip movements. The motion is modeled as the distribution of apparent line velocities in the movement of brightness patterns in an image. Then, they construct compact lip-motion features from the regional statistics of the local velocities. These can be used as alone or merged with audio features to recognize identity or the uttered digit. The author’s present person recognition results using the XM2VTS database representing the video and audio data of 295 people. Furthermore, we present results on digit recognition when it is used in a text prompted mode to verify the liveness of the user. Such user challenges have the intention to reduce replay attack risks of the audio system.

INTRODUCTION

The performance of multimodal systems using audio and visual information in biometrics is superior to those of the acoustic and visual subsystems (Brunelli and Falavigna (1995)), (Tang and Li (2001)), (Bigun et al. (1997b)), and (Ortega-Garcia et al. (2004)) because these systems have a high potential for delivering noise robust biometric recognition systems compared to the corresponding single modalities. This is the general motivation for why there has been increased interest in multimodal biometric iden-
tity recognition. For example in audio based person recognition, phoneme sounds can be acoustically very similar between certain individuals and therefore hard to differentiate. By adding information on lip-motion, the discrimination of identities can be improved.

Speaker recognition using visual information in addition to acoustic features is particularly advantageous for other reasons too. It enables interactive person recognition which can be used to reduce impostor attacks that rely on prerecorded data. Raising antispoofing barriers, known as liveness detection, e.g. to determine if the biometric information being captured is an actual measurement from the live person who is present at the time of capture, for biometric systems is becoming increasingly necessary.

In this chapter extraction of lip-motion features that takes advantage of the spatiotemporal information in an image sequence containing lip-motion is discussed. Motion features are suggested for recognition of human identities and word (digit) recognition which can be used for liveness detection. The discussions include filtering, feature extraction, feature reduction, feature fusion and classification techniques.

Section 2 presents a review of some previous studies relevant to the chapter. The emphasis is on audio-visual systems rather than the massive research body existing in the individual recognition technologies. In particular, lip features suggested previously are discussed in greater detail.

Section 3 presents the theory of three different concepts of motion estimation which is directly relevant to this chapter. The motion estimation techniques based on texture translations and line translations are explicitly contrasted against each other. A further quantification of the speed accuracy of the used motion estimation that assumes moving lines or edges is given. How motion is exploited in other audio-visual recognition studies is also discussed.

In Section 4 we present a discussion on how one can use estimated velocities to produce compact feature vectors for identity recognition and liveness detection by uttered digits. A technique for quantization and dimension reduction is presented to reduce the amount of extracted features. The section also presents the audio and visual features concatenated at the feature level allowing, the integration of different audio and video sampling rates. The visual frames come at one fourth pace of the audio frames do, but contain more data. Yet the final concatenated feature vector must come at the same pace and contain approximately the same amount of data each, to avoid favoring one over the other. The section also presents the performance of visual information as an audio complement feature in speaker recognition and speech recognition using the XM2VTS database. We present a single and multimodal biometric identity recognition system based on the lip-motion features using a Gaussian Mixture Model (GMM) and a Support Vector Machine (SVM) as model builders. Furthermore, we present the experimental test using only one word (digit) to recognize the speaker identity. A discussion on related studies exploiting different techniques for audio-visual recognition is also included.

Section 5 discusses the conclusions of the chapter and presents directions for future work.

REVIEW

In speech recognition, two widely used terms are phoneme and viseme. The first is the basic linguistic unit and the later is the visually distinguishable speech unit (Luettin (1979)). Whereas the use of visemes has been prompted by machine recognition studies, and hence it is in its start stage, the idea of phonemes is old. The science of Phonetics has for example been playing a major role in human language studies. The consonant letters complemented with vocals are approximations of phonemes and the alphabet belongs to greatest inventions of humanity.