Head Pose Estimation and Motion Analysis of Public Speaking Videos

Rinko Komiya, Kyushu Institute of Technology, Iizuka, Japan

Takeshi Saitoh, Kyushu Institute of Technology, Iizuka, Japan

Miharu Fuyuno, Kyushu University, Fukuoka, Japan

Yuko Yamashita, Shibaura Institute of Technology, Tokyo, Japan

Yoshitaka Nakajima, Kyushu University, Fukuoka, Japan

ABSTRACT

Public speaking is an essential skill in a large variety of professions and also in everyday life. However, it can be difficult to master. This paper focuses on the automatic assessment of nonverbal facial behavior during public speaking and proposes simple and efficient methods of head pose estimation and motion analysis. The authors collected nine and six speech videos from a recitation and oration contest, respectively, conducted at a Japanese high school and applied the proposed method to evaluate the contestants' performance. For the estimation of head pose from speech videos, their method produced results with an acceptable level of accuracy. The proposed motion analysis method can be used for calculating frequencies and moving ranges of head motion. The authors found that the proposed parameters and the eye-contact score are strongly correlated and that the proposed frequency and moving range parameters are suitable for evaluating public speaking. Thus, on the basis of these features, a teacher can provide accurate feedback to help a speaker improve.

KEYWORDS

English Oration Contest, English Recitation Contest, Facial Feature Point, Head Pose Estimation, Image Processing, Motion Analysis, Speech Video

1. INTRODUCTION

The ability to communicate in social and public environments can influence an individual's career prospects, help build relationships, and resolve conflict. Public speaking performance is characterized not only by the content presented but also by the presenter's nonverbal behavior, such as gestures and facial expressions. Nonverbal communication expressed through various types of behavior is a key aspect for successful public speaking and interpersonal communication. However, public speaking skills can be difficult to master and require extensive training. Moreover, in reality, the evaluation of public speaking can be subjective as it tends to heavily rely on human judgment. Thus, a system for automatic assessment of public speaking is needed for training.

Strangert and Gustafson (2008) presented via a political speech dataset the concept that vocal variety is correlated to human perception of a good speaker. Koppensteiner and Grammer (2010) used videos of political speakers to investigate different complex motion features and identified a

DOI: 10.4018/IJSI.2017010105

correlation between gesturing and personality ratings. Scherer, Layher, Kane, et al. (2012) used a large publicly available dataset to investigate the effect of audiovisual features on the perception of speaking style and the performance of politicians. They conducted a human perception experiment using eye-tracker data to evaluate human performance ratings and behavior through two separate media: audiovisual and video only. They concluded that several statistically significant features such as pausing, voice quality measures, and motion correlate strongly positively or negatively with certain human approval ratings for speaking style. Fuyuno, Yamashita, Kawase et al. (2014) collected multimodal data of English public speaking by Japanese EFL (English as a foreign language) learners, and analyzed speech-pause distributions and facial movement patterns. In this research, characteristics facial movement patterns were found in their datasets. However, the facial movement was obtained by a feature point set on the speaker's nose.

Recently, some interactive virtual audience systems for public speaking training have been proposed (Pertaub, Slater, & Barker, 2002; Batrinca, Stratou, Shapiro et al., 2013; Tudor, Poeschl, & Doering, 2013; Chollet, Stefanov, Prendinger et al., 2015). Batrinca, Stratou, Shapiro et al. (2013) developed a public speaking skill training system, Cicero, using a combination of advanced multimodal sensing and virtual human technologies. In this system, three kind of sensors; Microsoft Kinect sensor, two webcams, and a lapel microphone were used. Chollet, Stefanov, Prendinger et al. (2015) developed an interactive virtual audience platform for public speaking training. In their system, a depth sensor, an audio sensor, a video camera, and a physiological sensor were integrated, and these multimodal sensors were used to detect different types of behavior. However, these systems require several special devices, such as a head-mounted display (HMD), Microsoft Kinect sensor, and various physiological sensors. Therefore, an efficient but technologically simple training system is needed. Takahashi, Takayashiki, and Kitahara (2016) proposed a support system for improving speaking skills during job interviews, focusing on the skills needed for this specific type of presentation. Although this system was not related public speaking, it only comprised a web camera and a microphone.

Chen, Leong, Feng et al. (2015) proposed an automated scoring model for evaluating public speaking using multimodal cues. In their research, data on two types of public speaking tasks, informative and impromptu presentations, were collected using a Kinect sensor. They calculated the Kinect features, head pose, eye gaze, facial expression, lexical features, and speech features as multimodal features. The calculated values were then fed into three regression models: a support vector machine, glmnet, and random forest. Ramanarayanan, Leong, Chen et al. (2015) also used a similar approach. From the viewpoint of developing an automatic scoring system, these methods are useful; however, from a teaching viewpoint, it is still difficult to give feedback on how to improve public speaking performance. Any system that gives detailed feedback regarding this will be the best for the learner.

This research primarily focuses on the automatic assessment of nonverbal facial behavior during public speaking. A related research has used a Kinect sensor, but our method requires just a standard video camera. This paper proposes some simple and efficient methods for head pose estimation and motion analysis from speech videos.

The remainder of this paper is organized as follows. Sections 2 and 3 describe the proposed head pose estimation and motion analysis methods, respectively. Our speech dataset and experimental results are described in Section 4. This paper concludes in Section 5.

2. HEAD POSE ESTIMATION

Human head movement has three degrees of freedom (DOF) characterized by pitch, roll, and yaw, as shown in Figure 1. In this estimation, the 'target' person for analyzing head movement is engaged in public speaking, and in this scenario, there is often little movement in terms of pitch. Therefore, we focus on roll and yaw movements and angles. Many studies have performed head pose estimation using an automated visual technique that comprises eight conceptual approaches: appearance template