Face Expression and Motion Analysis over Monocular Images

Ana C. Andrés del Valle
VICOMTech Research Center, Spain

INTRODUCTION

As computers evolve towards becoming more human-oriented machines, human-computer interfaces, behavior-learning robots, and disable-adapted computer environments will use face expression analysis to be able to react to human action. Furthermore, faces play an essential role in human communication. Consequently, they have been the first objects whose motion has been studied in order to recreate animation on synthesized models or to interpret motion for a posteriori use. The analysis of motion and expression from monocular (single) images is widely investigated because non-stereoscopic static images and videos are the most affordable and extensively used visual media (i.e., Webcams).

Synthetic faces are classified into two major groups: avatars and clones. Generally, avatars are a rough and symbolic representation of the person, and their animation is speaker independent because it follows generic rules disregarding the individual that they personify. Clones are more realistic and their animation takes into account the nature of the person and his real movements. Whether we want to animate avatars or clones, we face a great challenge: the automatic generation of face animation data. Manually generated animation has long been used to create completely virtual characters and has also been applied to animate avatars. Nevertheless, many computer applications require real-time and easy-to-use face animation parameter generation, which means that the first solutions developed using motion-capture equipment prove to be too tedious for many practical purposes. Most applications utilizing Talking Heads aim at telecommunication uses. In such a context, real-time capabilities and low computing cost for both analysis and synthesis are required. Current trends in research tend to use speech analysis or synthesized speech from text as a source of real-time animation data. Although these techniques are strong enough to generate parameters to be used by avatars, they cannot provide realistic data for face animation.

BACKGROUND

Systems analyzing faces from monocular images are designed to give motion information with the most suitable level of detail, depending on their final application. Image input is analyzed in the search for general facial characteristics: global motion, lighting, and so forth. At this point, some image processing is performed to obtain useful data that can be afterwards interpreted to obtain face animation synthesis (see Figure 1).

We have categorized the most performing techniques for facial analysis on monocular images in three groups: "those that retrieve emotion information," "those that obtain parameters related to the Face Animation synthesis used," and "those that use explicit face synthesis during the image analysis."

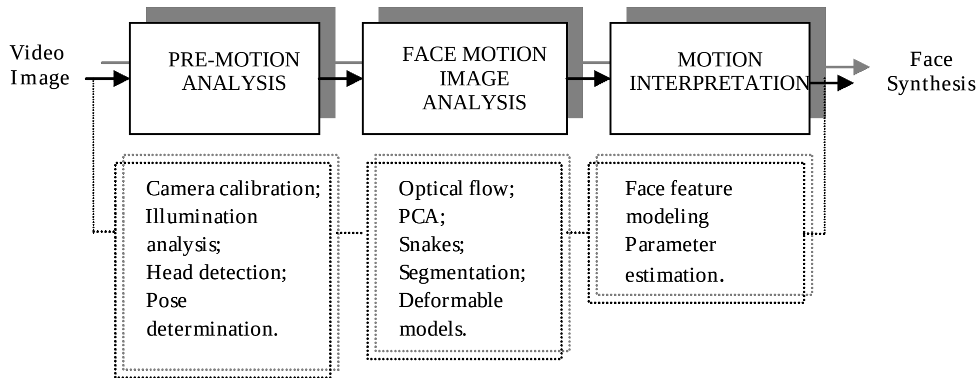
RETRIEVING EMOTION INFORMATION

Humans detect and interpret faces and facial expressions in a scene with little or no effort. The systems we discuss in this section accomplish this task automatically. The main concern of these techniques is to classify the observed facial expressions in terms of generic facial actions or in terms of emotion categories, and not to attempt to understand the face animation that could be involved to synthetically reproduce them.

Summary of the Most Performing Approaches

- The researcher Yacoob (Yacoob & Davis, 1994; Black & Yacoob, 1997) explored the use of local parameterized models of image motion for recognizing the non-rigid and articulated motion of human faces. These models provide a description of the motion in terms of a small number of parameters that are related intuitively to the motion of some facial features under the influence of expressions.
- Huang and Huang (1997) introduce a system developed in two parts: facial feature extraction (for the

Figure 1. Diagram of the general procedure for facial expression analysis



training-learning of expressions) and facial expression recognition. The system applies a point distribution model and a gray-level model to find the facial features. Then, the position variations are described by 10 action parameters (APs).

- Pantic and Rothkrantz (2000) describe another approach, which is the core of the Integrated System for Facial Expression Recognition (ISFER). The system finds the contour of the features with several methods suited to each feature—snakes, binarization, deformable models, and so forth—making it more efficient under uncontrolled conditions: irregular lighting, glasses, facial hair. A neural network (NN) architecture of fuzzy classifiers is designed to analyze complex mouth movements.
- Many systems base their description of face actions on the Facial Action Coding System (FACS) proposed by Ekman and Friesen (1978). The importance granted to FACS is such that two research teams, one at the University of California, San Diego (UCSD) and the Salk Institute, and another at the University of Pittsburgh and Carnegie Mellon University (CMU), were challenged to develop prototype systems for automatic recognition of spontaneous facial expressions.

  The system developed by the UCSD team, described in Bartlett et al. (2001), analyzes face features after having determined the pose of the individual in front of the camera. Features are studied using Gabor filters and afterwards classified using a previously trained hidden Markov model (HMM). In their study, Bartlett et al. claim AU detection accuracy from 80% for eyebrow motion to around 98% for eye blinks.

  CMU opted for another approach, where face features are modeled in multi-state facial components of analysis. They use neural networks to derive the AUs associated with the motion observed. They have developed the facial models for lips, eyes, brows, cheeks, and furrows. In their article, Tian, Kanade, and Cohn (2001) describe this technique, giving details about the models and the double use of NN, one for the upper part of the face and a different one for the lower part. The average recognition rates achieved are around 95.4% for upper face AUs and 95.6% for lower face AUs.
- Piat and Tsapatsoulis (2000) take the challenge of deducing face expression out of images from another perspective, no longer based on FACS. Their technique finds first the action parameters (MPEG-4 FAPs) related to the expression being analyzed, and then they formulate this expression with high-level semantics. To do so, they have related the intensity of the most used expressions to their associated FAPs.
- Other approaches (Chen & Huang, 2000) complement the image analysis with the study of the human voice to extract more emotional information. These studies are oriented to develop the means to create a human-computer interface (HCI) in a completely bimodal way.

Pantic and Rothkrantz (2000) offer overviews and comparative studies of many techniques, including some of those just discussed, analyzed from an HCI perspective.