Abstract

Head pose and gaze direction play significant roles in inferring the focus of attention, and they also help us to design more human-centered computer systems. The use of camera-based techniques for remote sensing of head pose and gaze should lead to a wide range of applications. However, although many methods have been proposed, there are technical limitations in the estimation techniques. Accurate estimation using only a monocular camera is still a difficult task, and existing methods often require calibration prior to estimation. The goal of this thesis is developing a head pose and gaze estimation system with minimal requirements; the methods used in the developed system do not need active calibration or equipment other than a camera.

The first part of this thesis describes a monocular method of tracking 3D head poses and facial actions. Using a multilinear face model that treats interpersonal and intrapersonal shape variations separately, this method estimates the parameters in real time by integrating two frameworks: particle filter-based tracking for time-dependent poses and facial action estimation and incremental bundle adjustment for person-dependent shape estimation. The use of this unique combination in conjunction with the multilinear face model enables tracking of faces and facial actions in real time without having to use pre-learned individual face models.

In the second part of this thesis, an unconstrained gaze estimation method is presented that uses an online learning algorithm and that allows free head movement by the user in a casual desktop environment. The key assumption is that the user gazes at the cursor position whenever s/he presses the mouse button. The user's eye images and 3D head poses are continuously captured on the basis of the head pose estimation method described in the first part of the thesis. By using clicked positions as exemplars of gaze positions, our system collects learning samples for estimating gazes while the user uses the PC, un-

aware of the system's activities. The samples are adaptively clustered in accordance with the head pose, and the estimation parameters are incrementally updated. In this way, our method avoids the lengthy calibration stage prior to using the gaze estimator.

One of the drawbacks of our method is that it cannot be applied to passive displays without user interaction. To solve this problem, we developed a calibration-free gaze sensing framework, and it is presented in the last part of this thesis. It uses visual saliency maps of video frames that are computed in a bottom-up manner. By relating the saliency maps with the appearances of the eyes of the person watching the video frames, our method automatically constructs a gaze estimator. In order to identify gaze points efficiently from saliency maps, saliency maps are aggregated to generate probability distributions of gaze points. Mapping between the eye images and gaze points is then established by Gaussian process regression. This results in a gaze estimator that frees users from active calibration and that can be applied to any type of display device.

Using the methods proposed in this thesis will make head pose and gaze estimation substantially more practical by reducing installation and setup costs. They can be used with commonly available cameras, and estimation procedures without manual initialization can be seamlessly integrated into daily computer interactions. This will enhance the potential for future investigation of attention-based application systems that will enrich our daily lives with ubiquitous computing devices.