Model-Based Eye Detection and Animation

Dr. Aman Bhardwaj

ABSTRACT

In this thesis we present a system to extract the eye motion from a video stream containing a human face and applying this eye motion into a virtual character. By the notation eye motion estimation, we mean the information which describes the location of the eyes in each frame of the video stream. Applying this eye motion estimation into a virtual character, we achieve that the virtual face moves the eyes in the same way than the human face, synthesizing eye motion into a virtual character. In this study, a system capable of face tracking, eye detection and extraction, and finally iris position extraction using video stream containing a human face has been developed. Once an image containing a human face is extracted from the current frame of the video stream, the detection and extraction of the eyes is applied. The detection and extraction of the eyes is based on edge detection. Then the iris center is determined applying different image preprocessing and region segmentation using edge features on the eye picture extracted. Once, we have extracted the eye motion, using MPEG-4 Facial Animation, this motion is translated into the Facial Animation parameters (FAPs). Thus we can improve the quality and quantity of Facial Animation expressions that we can synthesize into a virtual character.

Keywords: Eye motion estimation, MPEG-4 Facial Animation, Feature Points, Facial Animation Parameters, Feature extraction, face tracking, eye tracking.

INTRODUCTION

Nowadays real time communications that enables the sender and the receiver to transmit video contents generated by for example a web camera or a video mobile camera is a very relevant field. This is due to an increasing need for communication over large distance. We can transmit a video conference by using a radio link recorded by a video mobile camera, or via Internet, using a web camera. In both cases, the goal of the communication is transmitting the information in real time and with the lowest bandwidth, which implies a low bit-rate. In order to achieve these purposes, we need to compress the information. The compression of the information can be done according to different concepts: Model-Based Coding and a standard to compress the information, how is MPEG-4. Related with the first concept, many facial animation models have been developed based on Model Based Coding, a concept which will be explained later. Regarding MPEG-4, a part of the standard has been developed in order to deal with the animation of virtual faces: MPEG-4 Facial Animation.

MPEG-4 FA is the standard used to translate, code and transmit the motion efficiently, using a low bandwidth. On the receiver side the motion is synthesized. The Facial Animation models have experienced great evolution in the last years. Animation of human faces has been an active topic of research in computer graphics since the 1970’s. The most important characteristics of these models are that they must be efficient, accurate, and use few parameters to allow low bit-rate(low bandwidth). According to these models, we only transmit some parameters, which will be interpreted at the receiver side of the communication system. This in order to reduce the quantity of information sent over the channel, the parameters should be coded to admit efficient transmission in terms of bandwidth and reconstruct ability at the receiving side of the system. During the 90’s, several advanced face models have been created by different research groups and companies, sometimes using thousands of polygons to describe a face, and sometimes with complex underlying physical and anatomical models. Despite this, the Candide model is still widely used, since its simplicity makes it a good tool for image analysis tasks and low complexity animation.

Figure 1. Candide-1, figure extracted from
Previous work

Eye tracking is a technique used in cognitive science, psychology, human-computer interaction, advertising, medical research, and other many areas. The most widely used current designs are video based eye trackers. In the recent years, various methods of eye detection, iris detection and eye tracking have been proposed. Here, we present some of these approaches. Haro et al. use the physical properties of pupils along with their dynamics and appearance to extract regions with eyes. Infrared lighting is used to capture the physiological properties of the eyes, Kalman trackers are used to model eye/head dynamic and a probabilistic based appearance model is used to represent eye appearance. The camera has two concentric rings of IR LEDs, one a long the camera’s axis, and off-axis, to exploit the red eye effect. They generate two interlaced images for a single frame. The image where the inner ring is on, the bright image, has white pupils. The image where the outer ring is on, the dark image, has dark pupils. The difference image that results by subtracting both images contains valuable information about the iris location. Since this approach uses infrared cameras in order to track the eyes and we are interested in a non-invasive application, we will not consider this approach.

Kashima et al. present an iris detection method that is adaptable to various face and eye movements from a face image. First, the method determines the standard skin color from the previous frame. The face region is divided by the color difference from the standard skin color. Secondly, the eye and the mouth regions are extracted from the face region by hybrid template matching, which consist of the four directional features and color distance features. After, the iris regions are detected by using saturation and brightness from the extracted eye regions in order to separate the eye from the white part of the eye, segmenting the image. Then the edge of the iris is detected from the gray-scale image in iris region by applying the Prewitt’s operator. Finally, the iris positions are determined by the Hough Transform. Thus, this method can detect the iris no matter which gaze direction.

Weiner et al. present a system for gaze redirection in color faces images via eye synthesis and replacement. First, the face is detected via skin detection. Skin is detected using the Cr (chromatic red) and Cb (chromatic blue) component of the YCbCr color space. Following the face detection and extraction, the eye detection is performed working with the red channel of the color image. Edge detection is applied and after the position of the iris is determined by Hough Transform. Then, fine determination of the iris center position and the iris radius are performed by applying another version of the Hough Transform in a rectangular bounding box containing the eye. Finally the image is segmented using the fact that the iris is darker than its surroundings. The intensity image of the eye image is thresholded using an adaptative threshold designed to be bigger than the intensity of the iris pixels but lower than the intensity of the eye white or skin. This system provides good iris segmentation. However it is affected by the lighting conditions.

BACKGROUND

Almost all the solutions to computer vision problems involve manipulation and image processing. So in this chapter I include the basic concepts needed in order to understand this paper.

A visual communication system

A general communication system can be described as follows; the transmitter sends a message to the receiver via a channel. The channel can be a telephone line, a radio link, a satellite link or a storage device. If the input and output of the communication system are images, this system is called a visual communication system. Usually, the capacity of the channel is limited and we need to compress the information in order to achieve a low bit rate, low bandwidth and therefore fast communications. As we commented before, the focus of Model-Based coding is to allow the transmission of the motion extracted from facial images, for instance image video streams containing human faces. So in order to compress the information, a facial animation model is constructed and adapted to the current image sequence. Parameters describing how to transform the wire-frame of the facial model in order to animate a face are extracted, coded and transmitted through the channel. At the receiver side, in which is set up the facial animation model, as well, the wire-frame of the model is adapted according to the parameters received, resulting in an image very close to the original image.

Terminology for Face Image Processing

In Model-Based Coding different kinds of image processing, analysis and synthesis are performed. In order to process the face image, we have to process different stages. The five stages are: Face Detection, Facial Feature Extraction, Face/Facial Feature Tracking, Face Parameter Coding and Face Synthesis. Usually the application of the previous processing benefits the next one, decreasing the quantity of information to transmit. For instance in this thesis, we extract the face from the video sequence in order to reduce the quantity of information and to prepare the image to the next processing steps.
Each stage of the face image processing is described below:

1. **Face Detection**: Given an image, face detection algorithm tries to find the face area in the input image. If there is any face in the input image, the face detector returns the pixels containing the face. Obviously, the performance of the Face Detector depends on the head rotation. In our thesis we have used a face detector which work only with frontal faces without dealing with head rotation.

2. **Facial Feature Extraction**: Given an image and the location of the face, facial feature extraction is related to the process of extracting the coordinates which define the features of the face in the current image. From the extracted facial area, we can calculated the facial model parameters.

3. **Face Tracking or Facial Feature Tracking**: Tracking of a face or facial features is the process of following a detected face or extracted facial features through an image sequences. In this paper we deal with how to track the movement of the eyes in a video stream.

**Structure of the human eye**

The eye is one of the most complex organ in the human body. Briefly, the human eye consists of several basic parts. Among them there are: the retina, the lens, the vitreous body, the cornea and a number of protecting structures. Here the different parts of the eye are described. The information given by them is processed in order to implement the algorithm is extracted and processed. These parts are represented in the figure 2.

![Figure 2. Frontal view of the human eye, figure extracted from](image)

The Iris is the coloured part of the human eye. A simple description of the iris is that it is a coloured diaphragm of variable size whose function is to adjust the size of the pupil to regulate the amount of light admitted into the eye. It does this via the pupillary reflex (which is also known as the "light reflex"). That is, when bright light reaches the retina, nerves of the parasympathetic nervous system are stimulated, a ring of muscle around the margin of the iris contracts, the size of the pupil is reduced, hence less light is able to enter the eye. Conversely, in dim lighting conditions the pupil opens due to stimulation of the sympathetic nervous system that contracts of radiating muscles, hence increases the size of the pupil.

The sclera is the tough white sheath that forms the outer-layer of the ball. It is also referred to by other terms, including the sclerotic and the sclerotic coat (both having exactly the same meaning as the sclera). In all cases these names are due to the the extreme density and hardness of the sclera (sclerotic layer). It is a firm fibrous membrane that maintains the shape of the eye as an approximately globe shape. It is much thicker toward the back/posterior aspect of the eye than toward the front/anterior of the eye.

**IMPLEMENTATION**

There are three main stages in order to detect the position of the iris in a picture or in a frame extracted from a video stream. The first step is to extract the face from the whole image. As we commented before, we can get the image from a web camera or from a video mobile camera. In this system we get the video sequence containing a human face using a web camera. Usually, this picture contains the head facing the camera, and a background. To eliminate the effects of the non-face regions, for each frame from the video sequence, we use a face detector to locate the face (bound by a
rectangle) in the frame. Once we have an image containing the human face, the result from the second stage are two images containing the region of the eyes. Those pictures are processed in the third stage to locate the position of the iris.

**Eye detection and extraction**

After the face region is extracted from the image, the location of the eyes have to be determined. It may be intuitive to threshold the gray-level images to segment the eyes. However, the eyebrows, part of the hairs, the nostrils also display low intensity. Actually, the eyes exhibit strong horizontal transitions (vertical edges) due to the existence of the iris and eye white. The projection of the horizontal transitions is obtained applying the Sobel operator. Looking for the maximum of this projection we can extract the position of the eyes in the image containing the face. Methods for eye detection are suggested in Ref.

**Edge detection: Sobel operator**

In order to detect the vertical edges of the image we apply the Sobel operator into a gray-level color image containing the face, which has been obtained in the previous step. In a nutshell, the Sobel operator calculates the gradient of the image intensity at each point, giving the direction of the largest possible increase from light to dark and the rate of change in that direction. The result after applying the Sobel operator is an image in which the region of constant image intensity are set to zero, and the edges are pointed out by transitions between darker and brighter intensity values. Sobel operator consists of a pair of 3×3 convolution kernels as shown in the figure 3.3. One kernel is simply the other rotated by 90°.

![Figure 3. Face image extracted from the frame image](image)

![Figure 4. Sobel Masks, extracted from](image)

**Definition of MPEG-4 Facial Animation**

According to the ISO/IEC JTC1/SC29/WG11 (Moving Pictures Expert Group - MPEG) has formulated the MPEG-4 standard. SNHC (Synthetic Natural Hybrid Coding), a subgroup of MPEG-4, has devised an efficient coding method for graphic models and the compressed transmission of their animation parameters specific to the model type. MPEG-4 is an object-based multimedia compression standard that allows for encoding of different audiovisual objects (AVO) in the scene independently. The MPEG-4 standard allows using a 3-D face model set up into the transmitter and receiver side, the transmission of the facial parameters in such efficient, compressed and properly way, that the encoder can predict the quality of the presentation of the information at the decoder. According to MPEG-4 specifies a face model in its neutral state, a number of Feature Points (FPs) on this neutral face as reference points and a set of Facial Animation Parameters (FAPs), each corresponding to a particular facial action deforming a face model in its neutral state. Deforming a neutral face model according to some specific FAP values at each time instant generates a facial animation.
sequence. The FAP value for a particular FAP indicates the magnitude of the corresponding action, for example, a deformation of a mouth corner. For an MPEG-4 terminal to interpret the FAP values using its face model, it has to have predefined model specific animation rules to produce the facial action corresponding to each FAP. The FAP values are defined in face animation parameter units (FAPU). The FAPUs are computed from spatial distances between major facial features on the model in its neutral state.

MPEG-4 Facial Animation Parameters

For synthetic faces, the Facial Animation Parameters (FAP) are designed to encode animation of faces reproducing expressions, emotions and speech pronunciation. The 68 parameters are categorized into 10 different groups related to parts of the face. FAPs represent a complete set of basic facial actions, and thus allows the representation of most natural facial expressions. FAPs represent a complete set of basic facial actions including head motion, tongue, and eye and mouth control.

![Feature Points](https://example.com/feature_points.png)

**Figure 5. Feature Points, figure extracted from**

RESULTS

The purpose of this section is to show some results related with the time performance analysis and the accuracy after applying our application to extract the eye motion from a video stream containing a human face. First, we process the input video, in order to extract the current frame. Once, we have an image containing the current frame, we apply the eye tracker application, where, initially we extract the location of the face, obtaining an image in which is stored only the information related with this feature. In this way, we avoid to work with the background of the video, reducing the quantity of information that we need to process. The next steps are detecting and extracting the eyes from the last image, and finally, processing these images in order to extract the position of the iris. The position of the iris is pointed out by placing crosses in the eye centers. Moreover, after processing the current frame and extracting the eye motion, we record this information in a new video stream, which is adjusted for the different frames of the input video stream, in which we have pointed out the position of the eyes. In this way, we can easily follow the movements of the eyes. In this section, we include the results from the time performance analysis from each video. Moreover, we include one of the frames of each video, which the application records in order to show the results once we have extracted the eye motion.

CONCLUSIONS

In this study, a system capable of face tracking, eye detection and extraction, and finally iris position extraction using video streams containing a human face has been developed. The face tracking algorithm is based on the function CV Haardetect Object, which is a function already available in Open CV, Open Source library written in C++. Therefore,
from the frame extracted from the video stream, we extract an image only containing the human face. The detection and extraction of the eyes is based on edge detection. After, the iris center is determined applying different image preprocessing and region segmentation using edge features. The boundary of the iris is found by a circle detection algorithm. After we have located the iris, we can obtain the iris center and thus, the position of the iris in the current frame. Finally, the blinking is detected applying two different thresholds. In order to detect the blinking in a video sequence, two conditions must be fulfilled. The first one is related with the correlation coefficient. We make the correlation between a template containing an open eye and the current eye image extracted from the current frame. If the correlation coefficient is lower than a given threshold, this condition will be true. The second condition is related with the flow movement. By flow movement we understand the displacement up-down of the y-coordinate of the iris center. If this displacement is bigger than a given threshold, also this condition will be assumed true. Both thresholds must be specified by the user through the Graphical Interface

REFERENCES

[6]. Z.Zhou and X.Geng. Projection Functions for Eye Detection. State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China.
[7]. V.Vezhnevets and A.Degtiareva. Robust and Accurate Eye Contour Extraction. Graphics and Media Laboratory. Faculty of Computational Mathematics and Cybernetics, Moscow State University, Moscow, Russia.