Face Locating and Tracking for Human–Computer Interaction

Martin Hunke *Alex Waibel*

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Effective Human-to-Human communication involves both auditory and visual modalities, providing robustness and naturalness in realistic communication situations. Recent efforts at our lab are aimed at providing such multimodal capabilities for human-machine communication. Most of the visual modalities require a stable image of a speaker's face.
In this paper we propose a connectionist face tracker that manipulates camera orientation and zoom, to keep a person's face located at all times. The system operates in real time and can adapt rapidly to different lighting conditions, cameras and faces, making it robust against environmental variability. Extensions and integration of the system with a multimodal interface will be presented.

1 Introduction

Today's computers are largely incapable of recognizing and using many human conversation signals. Artificial speech synthesis and speech, gesture, and handwriting recognition significantly underperform human abilities. Further developments in this field would simplify human-computer interaction and highly improve the robustness of communication.

Recent research is increasingly concerned with the visual aspects of communication. For instance, the simultaneous evaluation of lip movements and acoustic signals improves recognition rate in the presence of background noise [2]. Eye movement tracking enables computers to calculate where an individual is looking [1], offering alternative interaction capabilities. Recognition of user identity and gestures can further enhance information and command interpretation.

Many of the above developments require a stabilized image containing the speaker's face of predefined size and position. Without the ability to track moving speakers, the use of these systems is limited to stationary faces, seriously limiting the system's flexibility.

In this paper a face tracking system will be described, that removes this limitation. For this purpose a standard camcorder is installed on a pan tilt unit which controls horizontal and vertical position. In addition to moving the camera the zoom lens is adjusted to maintain the individual's face of a relatively constant size within the image.

The system is able to locate persons in an arbitrary environment. Of all individuals found the one nearest to the camera is automatically selected for further tracking. For this purpose the position of the camera and the zoom lens are frequently adjusted to maintain a centered position of the face at a desired size within the camera image. Further interactive systems, which make use of the visual information of the observed face, are supplied with the subimage of the camera picture containing the face in predefined position and size. All operations are performed in real time, 11 times per second on average.

2 System Structure

The system operates in two main modes: locating and tracking. Beginning in the locating mode the system searches for arbitrary faces in the camera image. Once a face is located the system will proceed in the tracking mode which is essentially locating of a known face in a restricted area around the last location. During tracking the system learns features of the observed face and adjusts to changing lighting situations. The features used for locating faces are primarily shape and color. In addition, if movement of an object is detected, this information is used additionally for distinguishing between foreground and background.

The system structure is illustrated in figure 1. In the first step all areas containing face–like colors are

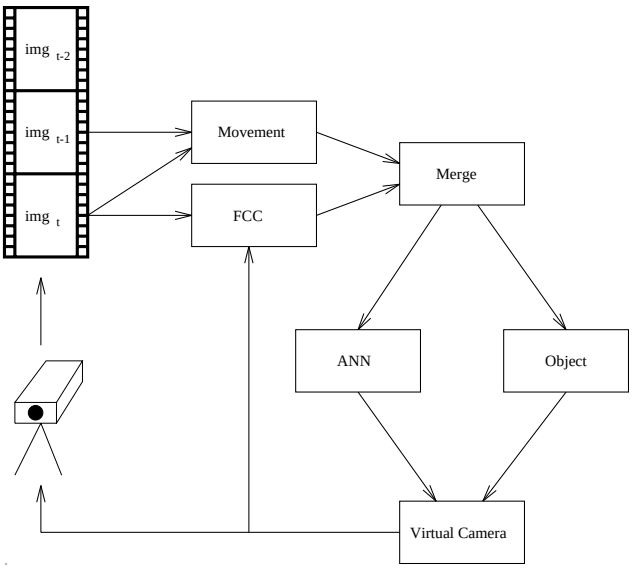

Figure 1: System structure

extracted using a Face Color Classifier (FCC), which is able to classify colors as face colors or background colors. If movement is detected, only those face–like colored areas are considered, which contain movement. The second step detects objects using the combined color and movement information from the previous step. For this purpose we compared a simple algorithm detecting connected objects (*object*) as well as an artificial neural network (ANN), which additionally considers shape to detect faces.

Once a face is located it can be assumed to be at a similar position and, in addition, of similar size in the next camera image. Therefore only an area with twice the dimensions of the face around the last position is investigated in the next image, drastically reducing computational requirements. This area is called virtual camera, since it is adjusted in position and size (corresponding to the physical zoom lense) like a real camera. If the position of the virtual camera moves close to one of the image boundaries, the physical camera is moved, so that the virtual camera and with it the observed face becomes centered again. Similarily, if the size of the virtual camera becomes too large or too small, the zoom lense is adjusted to maintain a optimal object resolution. This method avoids unnecessary movements of the physical camera.

In the initial locating mode the system locates faces in the entire image requiring movement as essential feature, because the appearance of the faces is unknown. Movement is detected by determining the difference of corresponding pixels in two following images, requiring the camera to be stationary during the locating phase. Of all located faces the largest one (which is assumed to be nearest to the camera) is selected for further tracking and the system switches into the tracking mode, adjusting the virtual camera to size and position of the located face. The FCC is adjusted to those colors actually occurring in the face. By repeating this adjustment in each image, the system is capable of rapid adjustments to changing lighting conditions. Because position and appearance of the face are known in the tracking mode, movement is no longer required, but determined as additional feature. A more detailed description of the entire system is given in [4].

2.1 Color

To handle the dependencies of colors on the hardware and lighting conditions the colors are normalized and classified by the FCC into face–like colors and background colors. For this purpose the brightness information of the pixels $Q = (R, G, B)$ transmitted by the camera is normalized, so that the sum of all three values is constant. Normalized colors $q = (r, g)$ therefore are two–dimensional and can be arranged in a triangle, the chromatic diagram (figure 2).

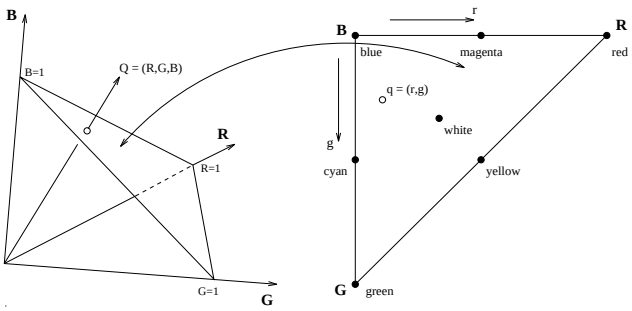

Figure 2: Color normalization

The FCC is based on color distributions in faces and is trained with individuals of different skin color. Because the normalized colors occurring most frequently in the faces used for training cover only a considerably small part of all possible colors, they can be used as a highly significant feature in locating faces. The value $N_{r,g}$ is defined as the frequency, how often a color (r, g) occurs in an image. A color distribution is defined as the set of all normalized frequencies $\overline{N}_{r,g}$:

$$\overline{N}_{r,g} = \frac{N_{r,g}}{\max_{i+j \leq 255} N_{i,j}}$$

Figure 3 shows the color distribution of the marked area in the face given in figure 4(a), giving the value

$\overline{N}_{r,g}$ to each color (r, g) of the two–dimensional chromatic diagram, which is located in the upper left half of the square. Dark pixels indicate high values $\overline{N}$ and therefore colors occurring frequently in the original image. Colors occurring less then 256 times as seldom as the most frequent color ($\overline{N}_{r,g} < \frac{1}{256}$) are represented with white. The right part of the figure shows a magnification of the interesting color range. The colors occurring most frequently are marked by a rectangular and defined as face–like colors. In figure 4(b) only those pixels with face–like color are shown, demonstrating the usefulness of the feature color. Because the eyes, nostrils, and lips are colored with colors occurring more seldomly in a face, they are not considered as face colors. As the illustration suggests, the FCC can be used to locate lips or eyes in faces.

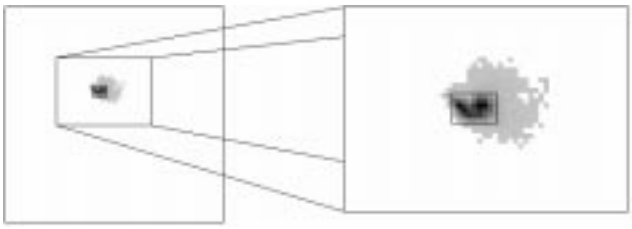

Figure 3: Color distribution

(a) (b)

Figure 4: Face and application of FCC

Because in the locating mode the face–colors of individuals are unknown, an a–priori classifier is required, considering all human skin–colors. Figure 5 shows a color distribution obtained by images of 30 individuals (asian, black, and white) with different skin–color. The range of occuring colors is much larger compared to the distribution in figure 3, but the set of the most frequent colors is still a very small subset of all possible colors. The seemingly strong differences between different skin–colors are mainly based on brightness of reflected colors, allowing the use of normalized colors as a feature, even if the appearance of a face is unknown.

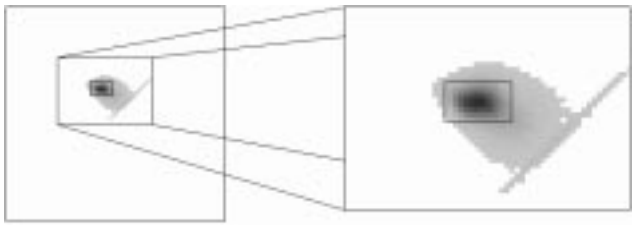

Figure 5: Color distribution of 30 faces

During the initial locating mode the system uses a general FCC, trained with low–pass filtered color distributions of persons of different skin type to recognize all human skin–colors. As few as five individuals have been found sufficient for this training process. In the tracking mode the FCC is adjusted to the actually occurring colors by determining a color distribution of the located face. By low–pass filtering the distribution each color is considered as a representative of similar colors, resulting in a generalization of the color distribution. If the color dependencies of the system change, for example by exchanging the camera or framegrabber, only this general FCC must be adjusted. All other system parts (especially the neural networks) are not affected.

2.2 Shape

The combined movement and color information is used to locate faces by locating coherent areas with face–like shape. A simple routine detecting the largest coherent area of pixels with face–like color (and motion, if present) obtains good results in most situations. However, if a person moves an arm to the face, the routine will mistake the combined face and arm as a face.

To distinguish between faces and face–like colored objects like hands and arms the shape of an object must be considered. For this purpose the combined color and movement information is fed into two neural networks as illustrated in figure 6. Both networks consist of a two–dimensional input retina containing the preprocessed image of the virtual camera, a hidden layer, and a output layer using a distributed output representation. This network structure has been successfully used for automatic vehicle driving in the ALVINN–project [5]. In this representation several output units represent different values, so that the actual output is determined by the neuron with the highest activation.

The first network determines the position of an object with a face–like shape, using an input retina of

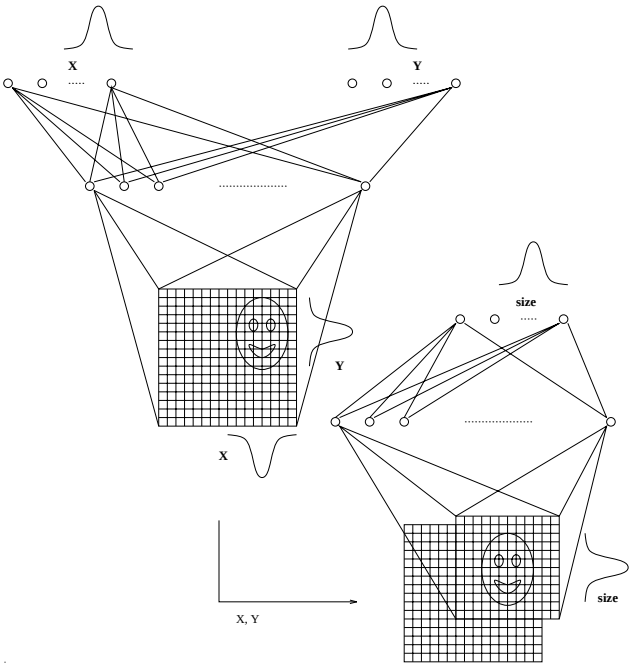

Figure 6: Neural network for face locating

16x16 pixels. The extracted subimage of 12x12 pixels around this position is fed into the second network to determine the size of the centered object. By splitting the position and size extraction into two following procedures, the size estimation could be reduced to centered objects. The output of the network gives precise information about the position and size of the located face, which is used to manipulate camera orientation and zoom. The subimage containing the observed face is automatically extracted and fed into systems for lip–reading or investigation of other visual aspects. Because of using preprocessed data it is not necessary to retrain the networks for different lighting situations, skin types or hardware, though the network makes use of color information.

Two databases are used to generate images to train the neural networks. One database contains faces of constant position and size recorded in front of a blue background, allowing automatic resizing and shifting of the face to given values and superimposing the image with a new background from the second database. With this method an unlimited number of training examples, containing an example input image with varying position and size and the desired output can be created. The desired output is given by the position and size of the face. The networks were trained by backpropagation on 5000 scaled and shifted example images generated with a database containing 72 images of 24 faces of different sex, age, hair style, skin color, etc. The best results were obtained with a hidden layer of 40 neurons.

(a) (b)

Figure 7: Examples of artificial images
(a) Artificial images, **(b)** Application of the general FCC

Figure 7 shows in (a) several artificial example images shrinked to the resolution of the network's input retina and in (b) the same images after preprocessing with the general FCC, which were used for the training of the networks. The network performance was evaluated using a test set containing images from different individuals than in the training set. An image is correctly classified, if the difference of the network output and the actual position and size of the face is less than 10% of the size of the face. The networks classified the position correctly in 93.6% and the size in 95.2% of the images. Another pair of networks were trained on grayscale images without preprocessing. The achieved performace of only 68% in estimating the position and 32% in estimating the size shows the usefulness of the color classification.

3 Results

The performance of the entire system was evaluated on test sequences containing more than 2000 images. Seven subjects of different skin s were recorded in front of different backgrounds. All subjects were asked to perform arbitrary movements in front of the camera, such as sitting down in a chair, standing up, changing the distance to the camera, etc. In each image the

position and size of the face were manually marked and compared with the output of the tracking system. The sequences differed in background and speed of the movements.

Depending on the sequence, the face was located in 96% to 100% of all images in the sequence. The average difference of the actual position of the face and the output of the system were less than 10% of the size of the head.

4 Conclusion and Future Directions

We show that face locating and tracking can be performed on a single workstation in real time, resulting in an average of 10 to 12 extracted faces per second. The achieved reliability and accuracy of the system are sufficient for the combination with lip–reading and similar systems [3].

The use of color has proved to be extremely useful for the task of locating faces. The developed face color classifier is a powerful abstraction of colors which is able to adapt online to different lighting situations and skin colors. We also show that neural networks are capable of classifying shapes as face or non–face.

Ongoing research combines this system with a lip–reading system and a microphone array. The stabilized output of the tracking system containing the face of an observed moving head is used for lip–reading. The information about the location of the speaker can be used for beamforming with a microphone array to reduce competing acoustic signals from different locations. Conversely the array is able to give locations of detected speech signals which could be used in the face tracking system. Applications for videoconferencing will also be considered.

Acknowledgements

This research was sponsored by the Department of the Navy, Office of Naval Research under Grant No. N00014–93–1–0806.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

[1] S. Baluja and D. Pomerleau. Non–Intrusive Gaze Tracking Using Artificial Neural Networks. In *Advances in Neural Information Processing Systems*, volume 6. Morgan Kaufmann, 1993.

[2] C. Bregler, H. Hild, S. Manke, and A. Waibel. Improving Connected Letter Recognition by Lipreading. In *International Conference on Acoustics, Speech and Signal Processing*, 1993.

[3] P. Duchnowski, U. Meier, and A. Waibel. See Me, Hear Me: Integrating Automatic Speech Recognition And Lip–Reading. In *ICSLP*, 1994.

[4] M. Hunke. Locating and Tracking of Human Faces with Neural Networks. Technical Report CMU–CS–94–155, School of Computer Sience, CMU, Pittsburgh, U.S.A., August 1994.

[5] D. Pomerleau. *Neural Network Perception for Mobile Robot Guidance*. PhD thesis, School of Computer Sience, CMU, February 1992.