Interactive Digital Mirror

CARLOS HITOSHI MORIMOTO

Departamento de Ciência da Computação do IME-USP - Rua do Matão 1010, São Paulo, SP 05508, Brazil
hitoshi@ime.usp.br

Abstract. This paper describes some preliminary results of our ongoing project on digital mirror interfaces. Regular mirrors reflect ambient light from the scene towards the observer. Digital mirrors capture the ambient light with a camera, extract information about the scene, and display *appropriate* information to the user, combining real-time computer vision systems with realistic computer graphics. We describe the works on both ends, image processing from the camera, and image output to a computer screen. The computer vision routines developed so far are basically for human face detection, tracking and 3D head pose estimation. The main computer graphics routines are able to render a 3D head model in real-time. We are starting to integrate both ends, and coding very simple behaviors to the virtual head.

1 Introduction

Different than regular mirrors, digital mirrors can change the image of the scene *reflected* on the mirror, but leaving most of the scene, or at least part of it, looking the same. How to segment objects from the input video stream, how to process the segmented objects (deleting them from the scene, modifying them, or substituting them for other objects) and how to realistically render the new scene are the main challenges of this project. Our initial goal is to process faces, before extending the system to more general objects.

There are several examples of interactive graphic systems based on computer vision in the literature used for various applications such as virtual and augmented reality environments, avatars, virtual trainers, and other interactive virtual agents. The general approach is to extract information about the scene using computer vision, and use this information to update a scene model to be rendered using computer graphics. We describe a few systems that detect and track human faces, which is the basis of our current digital mirror.

Some examples of systems that are able to find and track faces in video streams are described in [21, 16, 2, 4]. Yang and Waibel [21] use color information to track face regions and have done rigorous studies to validate the chosen color space for human face detection [20]. Oliver *et al.*[16] present a 2D real-time single person lip and face tracker that uses color information to detect and track the face candidates. Birchfield [2] uses the interior color and boundary gradient of an elliptical region to control a camera to follow a single subject as it moves in a room, and La Cascia and Sclaroff [4] have developed a 3D head tracking technique that is robust to varying illumination conditions. In their technique, the head is modeled as a texture mapped cylinder, and the tracking is formulated as an image registration problem in the cylinder's texture map image.

More interactive systems capable of not only tracking a face but recognizing its pose and/or gestures are presented in [1, 3, 5, 7, 11, 15, 18]. Azarbayejani *et al.*[1] presents a system to manipulate a virtual reality clone using heuristic filters to track feature templates. Heinzmann and Zelinsky [11] and Morimoto *et al.*[15] describe a face tracking systems for head gesture recognition which can be used in human-computer interfaces. Bradski [3] uses a robust nonparametric statistical technique to track flesh tone regions in 3D, and Toyama [18] integrates color, intensity templates and dark features on the face to estimate the full 3D pose of a single head and uses this information to control the cursor in a computer interface.. Colmenarez *et al.*[5] describe a real-time face feature tracking based on an information-based maximum discrimination learning technique.

Of all these systems, the one that most resembles our project was developed by Darrel *et al.*[7], because the detected faces are not simply displayed, but processed to combine various graphical effects, and only on the face region. They extract range data using special stereo vision hardware and combine color information to segment the close skin tone regions. Then a neural network-based static face description is used to detect faces. The complete system uses several processors and its cost is a barrier to many applications. But the face detector and tracker is very robust due to the integration of all these techniques.

The approach described in this paper is a much lower-cost solution, but yet real-time and robust. The next section introduces the digital mirror project, describing the vision, graphics, and behavior modules. Section 3 demonstrate results of the current system prototype, and Section 4 concludes this paper.

[0] We are grateful for the suport from FAPESP (Fundção de Amparo à Pesquisa do Estado de São Paulo) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico).

1530-1834/01 $10.00 © 2001 IEEE

2 Digital Mirror

A digital mirror is composed of several computer vision and computer graphics routines. Figure 1 shows the block diagram of our digital mirror interface.

The scene is captured by a camera and processed by the digital mirror to produce an *appropriate reflection* on the computer monitor. Since we are only processing face objects, the faces are detected, tracked, and their poses estimated. The Behavior Module would record the past history of each face's motion to recognize gestures and expressions, which might be used to trigger different mirror behaviors (an interactive mirror). Depending on the user's state and the current mirror behavior, the internal face models are updated, and rendered. Next we describe these modules in detail.

2.1 Computer Vision Modules

The computer vision modules are responsible to detect and track faces and facial features, and to estimate the head's pose and position. So far we have developed and integrated the face detection and tracking algorithms.

Most solutions proposed for detecting faces are based on templates and other geometrical constraints [10, 22] as well as artificial neural networks [17], color histograms [9, 16, 21], or fusion of several modes or cues [2, 6, 8]. We have used active IR lighting to detect pupils and group them into faces, as described in [13]. We briefly describe the system next.

Pupils candidates can be robustly detected using the bright pupil effect described in detail in [14]. Grouping of pupil candidates into faces can be done using simple heuristic rules, such as temporal and spatial consistency, i.e., they blink at the same time, they move together according to the same rigid motion of the head, etc.

Spatial cues are formed by static properties of the eye such as position, size, and aspect ratio, but can also include color of the iris, skin tone surrounding the eyes, etc. Eyes from the same face are likely to appear in horizontal lines and have approximately the same size. Other constraints are imposed based on the expected size of a face, which is estimated from the properties of the camera (size of the sensor and lens) and the illuminators.

All these rules help grouping the pupil candidates into pairs. A simple consistency check can be implemented due to the fact that the imaginary line segment connecting the eyes of a face cannot cross the inter-eye lines from other detect faces. Once multiple faces are detected, the most salient one is used to initialize the face tracker.

Since the pupil candidates are lost during blinking, a robust tracker must rely on other tracking modes. Our tracker relies on two operation modes. The first mode uses the information from the multi-face detector, and the second is a feature correlation tracker that uses the sum of absolute differences (SAD) as the object function to be minimized. To ensure stability of this process, two zero order recursive estimators are used to combine the information from both modes, similar to [6]. The state of the tracked face is represented by its size and position, which are treated independently by the two recursive estimators (one for the position and another for the dimensions of the face box). A state of each recursive estimator is defined by a two parameter vector (the position (x,y) of the center of the face, and the width and height (w,h) of the face box). Each vector is accompanied by a covariance.

Movements of the subject's face is unpredictable, but assuming the frame rate is much faster than the rigid head motion, the predicted state vector can be considered to be the last updated estimate

$$\check{\mathbf{X}} = \hat{\mathbf{X}} \tag{1}$$

and the predicted covariance matrix is

$$\check{\mathbf{C}} = \hat{\mathbf{C}} + (\Delta t)^2\mathbf{W} \tag{2}$$

where the uncertainty in position and size grows quadratically with the time interval Δt between the observation and the last estimation, and $\mathbf{W}$ captures the precision loss of each component, and depends on the properties of the underlying process.

New face observations $(\mathbf{Y}, \mathbf{C})$ are used to update the state of the estimators as follows:

$$\hat{\mathbf{C}} = [\check{\mathbf{C}}^{-1} + \mathbf{C}^{-1}]^{-1} \tag{3}$$

$$\hat{\mathbf{X}} = \hat{\mathbf{C}}[\check{\mathbf{C}}^{-1}\check{\mathbf{X}} + \mathbf{C}^{-1}\mathbf{Y}] \tag{4}$$

The covariance matrix $\hat{\mathbf{C}}$ is an estimation of the error of the estimated state vector $\hat{\mathbf{X}}$. The face detected closest to the estimated position, and within certain error boundaries, is used to update the state.

It is not possible to get observations from the multiple face detector for every frame because of blinking and failures in the grouping process. When no face is detected, or no face closer than a certain threshold in size and position to the predicted face is detected, the SAD correlation tracker is called to determine the 2D translation of the face last used as measurement. The translated face is then used as the new measurement to updated the position estimator.

The SAD correlation tracker determines the translation (i,j) of the feature point $F_{t-1}(x, y)$ to its corresponding tracked point $F_t(x + i, y + j)$ in the current frame F_t by minimizing the SAD(i,j) function within a search neighborhood defined by the region of support around the feature being tracked.

The SAD correlation tracker uses a small region of support and search window around the left pupil, so that

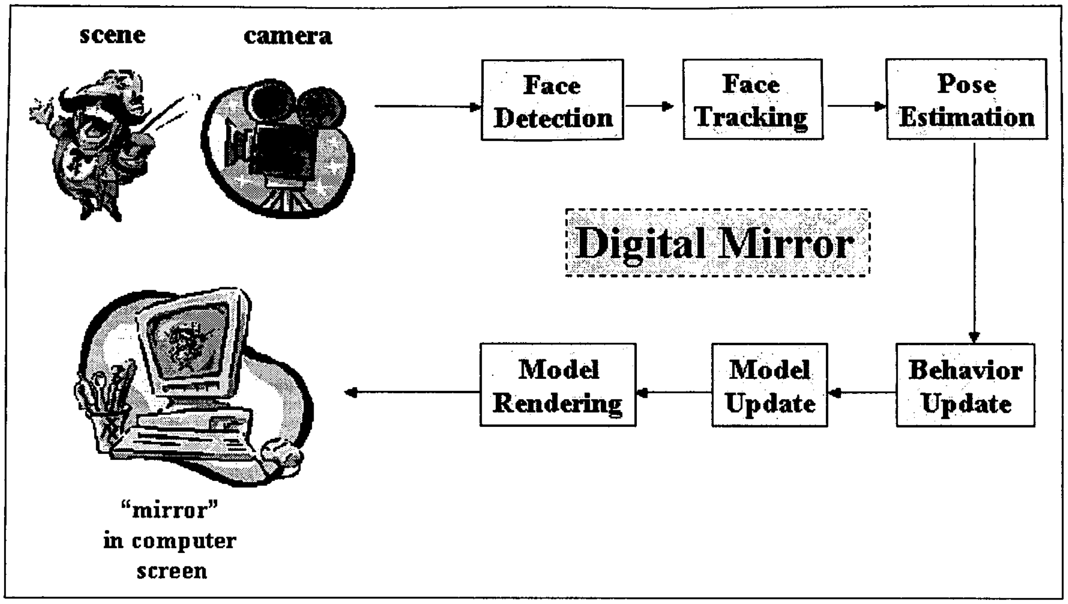

Figure 1: Block diagram of the digital mirror architecture.

it can loose track very easily. To avoid tracking of non-face objects, if the SAD correlation tracker gets called consecutively for more than a certain number of frames, without a face being detected within the predicted region by the multiple face detector, the process is re-initialized with the next most salient face. A surprisingly robust tracker is obtained from the combination of the SAD correlation tracker with the multiple face detector using the recursive estimators, given that neither mode could robustly operate by itself, and one mode has to rely on the other to compensate each others weaknesses.

2.2 Computer Graphics Modules

Computer graphics techniques are used to model and render the faces and facial expressions. In order to achieve realism, we use a muscle-based facial model [12].

There are basically two techniques to represent faces in computer graphics, the first is based on interpolation models, and the second on physics based modeling. The interpolation model is the most common and faster technique used for graphic animation. This method consists of creating a library of polyhedric images with different expressions (anger, surprise, fear, happiness, etc.), generating other intermediate expressions among these, moving the vertices of the polygons along a straight line segment determined by corresponding vertices in different images. So, an infinite number of images can be generated by changing the proportion of the images in the library.

However, this method only allows for facial expressions that belong to the expression library. In order to get good results it is necessary to this library to be big enough, although the generated image set hardly gets near the possible expression range of a human face.

Physics-based modeling consists of simulating the way human faces accomplish local deformations, i.e., through the use of facial muscles. Although this method is slower than the interpolation method, due to larger amount of calculation to determine the movement of the vertexes of the polygons, it is more convenient to our purposes because this can reach the whole range of human facial expressions. The physics-based modeling also allow for a more compact facial expression representation because it does not require reference images, it only need to store a muscle state vector that correspond to the muscular contractions that result in that particular expression.

We started with the DECFace model [19], which source code is publicly available, and made several modifications to increase its realism. The DECFace only models the facial muscles, and we have incorporated eye balls, a lower jaw, and teeth. The code is very efficient and heavily based on OpenGL routines, what is appropriate for real-time applications such as digital mirrors.

2.3 Behavioral Model

The behavior model determines what the *reflectedface* does. We have not yet implemented the texture mapping routines to display a virtual head similar to the user's face (virtual mirror), nor the 3D head pose estimation routines to allow the *reflections* to behave exactly like the user's face. But we have started the integration to test for real-time performance, with very simple behaviors, describe in the next section.

3 Experimental Results

We first show results from the multiple face tracker, and the muscle-based face model later. Figure 2 shows an example result of the multi-face detection and tracking algorithm. Detection works very well for distances of up to 3 meters from the camera for most subjects tested, which is appropriate for an interactive digital mirror application, but further experiments will have to be conducted to determine the factors which contribute to this variance.

Observe that the person closest to the camera has two boxes, one around the face and a second around the left eye, which represents the search window used by the SAD correlation tracker. Figure 2a clearly shows the bright pupils caused by the active IR lighting scheme, and Figure 2b shows a regular dark pupil image with the detected face boxes superimposed. Figure 2c shows the detected pupils within the detected face boxes.

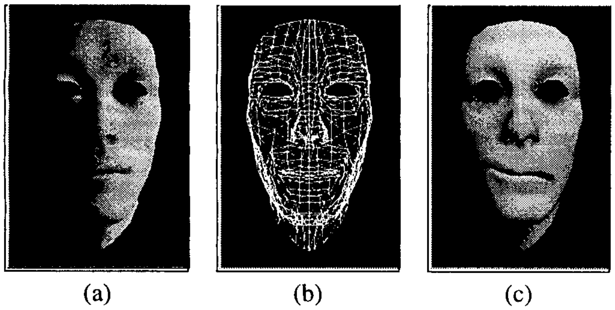

Figure 3: (a) A face from the original DECFace program. (b) Wireframe representation of the face. (c) Shows a muscle contraction.

Figure 3a shows an image from the original DECFace program, and Figure 3b, its corresponding wireframe representation, showing the polygons that constitute that face. Figure 3c shows the face after a muscle contraction. Observe that in this model the mouth does not open, and the eye sockets are empty.

We have extended the model to include an lower jaw, teeth, and eyeballs, hardcoded a few basic facial expressions that can be seen in Figure 4. We have integrated the face tracking algorithm with the face rendering algorithm to simulate eye contact between the *reflectedface* and the user. The 2D position of the user's face in the image is used to control the vergence of the eyes and some neck rotation in the *reflected image*, so that the user has the feeling that the virtual head is always looking back at the user. The whole system is current running on a Pentium III 933MHz with Windows 2000, and is able to track multiple faces and render a realistic face in real-time (over 20 frames per second).

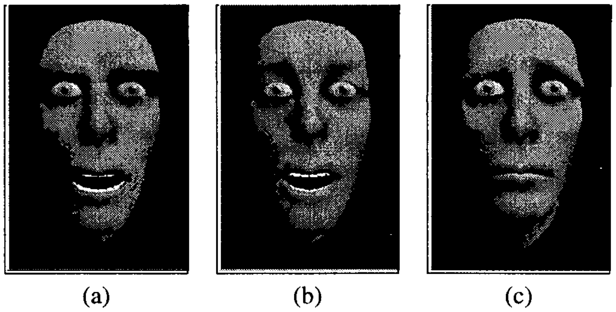

Figure 4: A few expressions using the extended facial model. (a) Anger. (b) Surprise. (c) Sadness.

4 Conclusion

We have presented the current status of our ongoing project on interactive digital mirrors. Different than regular mirrors, digital mirrors can change the image of the scene *reflected* on the mirror, but leaving most of the scene, or at least part of it, looking the same. We have used a robust multiple face detector and tracker based on active IR illumination, and developed a physics based face model to generate realistic graphics output, and tested the integration of both modules using an eye-contact application, that randomly changes facial expressions.

We are currently implementing a pose estimation algorithm, and studying ways to perform robust facial feature tracking, in order recognize head gestures and facial expressions. Once that is accomplished, a complete digital mirror will be implemented, and also with more complex behaviors, such as acting as a mirror most of the time, but showing fear when the user shows an "ugly" face, or other expression after a particular gesture from the user.

References

[1] A. Azarbayejani, T. Starner, B. Howowitz, and A. Pentland. Visually controlled graphics. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(6):602–605, 1993.

[2] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 232–237, Santa Barbara, CA, June 1998.

[3] G. Bradski. Computer vision face tracking for use in a perceptual user interface. Technical Report Q2, Intel Corporation, Microcomputer Research Lab, Santa Clara, CA, 1998.

[4] M. L. Cascia and S. Sclaroff. Fast, reliable head tracking under varying illumination. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Fort Collins, CO, June 1999.

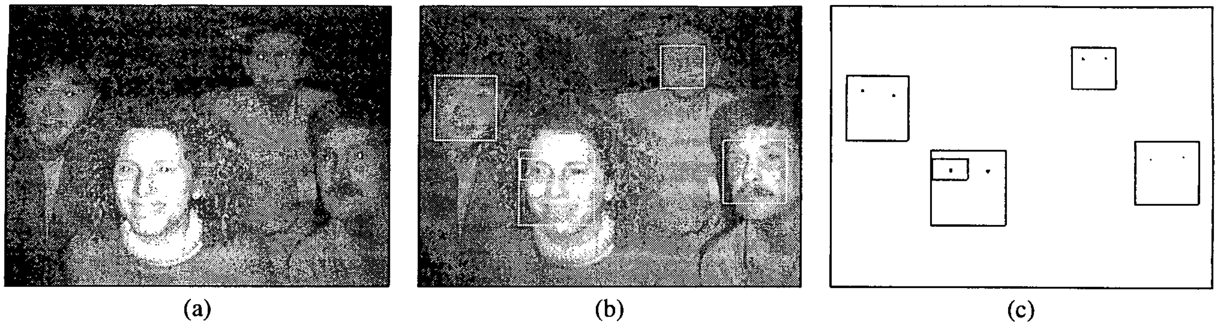

Figure 2: (a) Observe the bright pupils caused by the active IR illuminators, and (b) the detected faces. (c) Shows the detected pupils with the corresponding face boxes. The face with two boxes in (b) is the most salient one. The small box surrounds the feature (eye) to be tracked.

[5] A. Colmenarez, B. Frey, and T. Huang. Detection and tracking of faces and facial feaures. In *Proc. of the International Conference on Image Processing*, Kobe, Japan, October 1999.

[6] J. Crowley and F. Berard. Multi-modal tracking of faces for video communications. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 640–645, Puerto Rico,PR, June 1997.

[7] T. Darrell, G. Gordon, J. Woodfil, and M. Harville. A virtual mirror interface using real-time robust face tracking. In *Proc. of the 3rd Int. Conf. on Automatic Face and Gesture Recognition*, pages 616–621, Nara, Japan, April 1998.

[8] T. Darrell, B. Moghaddam, and A. Pentland. Active face tracking and pose estimation in an interactive room. Technical Report 356, M.I.T. Media Laboratory Perceptual Computing Section, Cambridge, MA, 1996.

[9] P. Fieguth and D. Terzopoulos. Color based tracking of heads and other mobile objects at video frame rates. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–27, Puerto Rico,PR, June 1997.

[10] V. Govindaraju, S. Srihari, and D. Sher. A computational model for face location. In *ICCV*, pages 718–721, 1990.

[11] J. Heinzmann and A. Zelinsky. Robust real-time face tracking and gesture recognition. In *Proc. of the International Joint Conference on Artificial Intelligence, IJCAI*, volume 2, pages 1525–1530, 1997.

[12] Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. *Computer Graphics (SIGGRAPH 95)*, 20(4):56–62, 1995.

[13] C. Morimoto and M. Flickner. Real-time multiple face detection using active illumination. In *Proc. of the 3rd Int. Conf. on Automatic Face and Gesture Recognition*, Grenoble, France, March 2000.

[14] C. Morimoto, D. Koons, A. Amir, and M. Flickner. Pupil detection and tracking using multiple light sources. *Image and Vision Computing*, 18(4):331-336, March 2000.

[15] C. Morimoto, Y. Yacoob, and L. Davis. Recognition of head gestures using hidden markov models. In *Proc. International Conference on Pattern Recognition*, Vienna, Austria, August 1996.

[16] N. Oliver, A. Pentland, and F. Berard. Lafter: Lips and face real time tracker. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 123–129, Puerto Rico,PR, June 1997.

[17] H. Rowley, S. Baluja, and T. Kanade. Rotation invariant neural network-based face detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 38–44, Santa Barbara, CA, June 1998.

[18] K. Toyama. "look, ma - no hands!" hands-free cursor control with real-time 3d face tracking. In *Proceedings of 1998 Workshop on Perceptual User Interfaces*, pages 49–54, San Francisco, CA, November 1998.

[19] K. Waters. A muscle model for animating three-dimensional faces. *Computer Graphics (SIGGRAPH 87)*, 21(4):17–24, 1987.

[20] J. Yang, R. Stiefenlhagen, U. Meier, and A. Waibel. Visual tracking for multimodal human computer interaction. In *Proc. ACM SIGCHI - Human Factors in Computing Systems Conference*, pages 140–147, Los Angeles, CA, 1998.

[21] J. Yang and A. Waibel. A real-time face tracker. In *Proceedings of the Third IEEE Workshop on Applications of Computer Vision*, pages 142–147, Sarasota, FL, 1996.

[22] A. Yiulle, P. Hallinan, and D. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8(2):99–111, April 1992.