The Face in Activity Analysis and Gesture Interfaces

Alejandro Jaimes

FXPAL Japan, Corporate Research Group, Fuji Xerox Co., Ltd., Japan

ABSTRACT

In this paper we discuss two systems we have developed and raise research questions on the role of the face in their implementation for different scenarios. The first system allows a user to define hotspot templates for gesture-based interaction. A camera points at any space, in which the user can define configurations of 2D hotspot rectangles to define 2D gestures. The rectangles respond only to motion in one direction, making them highly robust. The second framework is an activity and user posture alarm and summarization system. As camera placed on top of a computer screen can be used to monitor a user's posture and activities as he uses his computer. We will describe the two systems and discuss the advantages, disadvantages, and motivation for using or (not) using face processing algorithms.

Categories and Subject Descriptions

I.4.9 [Image Processing and Computer Vision]: Applications; H.5.2 [User Interfaces]: Ergonomics

General Terms

Algorithms, Measurement, Human Factors

Keywords

Ergonomics, Computer Vision, posture, ergonomics, gesture, interaction.

1. INTRODUCTION

Lower hardware costs and higher computational power make camera-based interaction techniques promising. Interaction may be explicit (e.g., user-given commands), or implicit (e.g., computer reacts to natural non-command user actions), and frameworks can have varying ranges of complexity, using simple motion detection or computationally intensive algorithms for complex tasks. As in all systems, there are always tradeoffs between complexity, robustness, and functionality, and in dealing with human-computer interfaces, the question of whether, or how face-processing algorithms may be used is an important one.

In this paper we describe two systems for camera-based human-computer interaction which are based on two different models, but which share a philosophy of adaptability and simplicity. Both systems are highly flexible in that the user can make many decisions about what the systems do, and both systems use simple algorithms that can be implemented on low-cost hardware. Although some experiments have been carried out using face-processing algorithms in the context of the two systems, the benefits and drawbacks of face-processing have not been explored in depth. The purpose of this paper, therefore, is to raise some questions about the use of face-processing algorithms in the context of similar systems.

First we briefly describe each of the systems and then we outline some of the research issues related to face-processing.

2. HOTSPOTS

Humans use gestures to communicate naturally, but defining a set of meaningful and computationally recognizable 3D gestures can be difficult. Two-dimensional gestures, on the other hand, can be easily defined, assigned application or user-dependent meanings, and recognized using simple computer vision techniques.

The hotspot system, presented in [2] is a camera-based, adaptable user interface system that uses hotspot components for 2D gesture-based interaction. A camera points to the user's desktop. The user defines an interaction area within the desktop by pointing the camera to a desired location, defines new commands by configuring hotspot areas, and executes them by moving his hand across them. For example, to move to the previous page in a document, the user moves his hand from left to right across a hotspot area (Figure 1). The framework has many applications, as the desk can become a large interactive space in which a rich set of hotspot configurations can be created with different meanings. For instance, imagine an instrument in which each hotspot or hotspot configuration represents a different sound; a desk with multiple displays so that different parts of the desk can be used to simultaneously interact with multiple documents; hand gestures similar to "mouse gestures" for browsing; or special gestures for people with disabilities.

In the basic setup the camera is placed above the desktop (for example, on top of the computer monitor) and the corresponding video captured by the camera appears in a resizable window on the computer screen (Figure 1). The user defines the interface area by moving the camera to point at a desired physical location: in Figure 1 (left) the main capture area is behind the keyboard. A user might prefer to point the camera elsewhere (e.g., in front of the keyboard as in Figure 1, right), and the system can be used in the context of interactive spaces such as meeting rooms (see Figure 3 and Figure 4).

ACM CHI '06, April, 2006, Montreal, Canada.

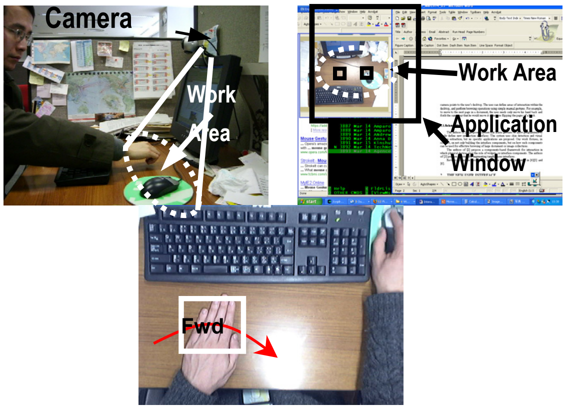


Figure 1. System setup. The camera (left) points at a physical work area on the desktop. The interface image as viewed by the user (middle) can be resized as desired and shows the captured video images and hotspots (right image).

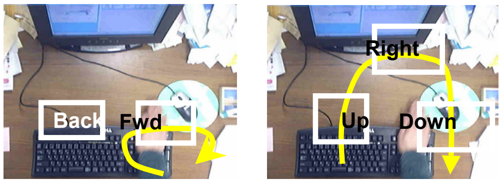


Figure 2. Screen shot of captured image. On the right image a composition of hotspots represents a single command (e.g., re-load a webpage), whereas on the left each hotspot represents a different command (back, forward).

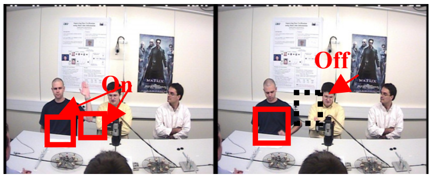


Figure 3. Meeting room scenario. Note that hotspots could be defined relative to the face.

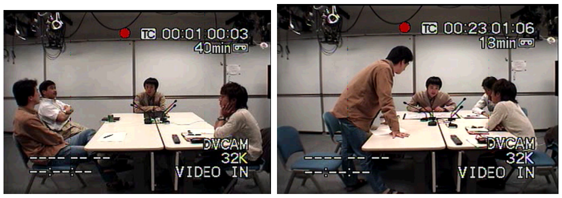


Figure 4. Scenes from meeting room scenario. Note possible difficulties in applying face processing techniques.

3. POSTURE AND ACTIVITY ALARM AND SUMMARIZATION SYSTEM

Many people spend long periods of time in front of their computer and often suffer back, shoulder, and neck pains. Ergonomics for computer users has therefore gained importance, on one hand because every year companies loose millions of dollars due to injuries sustained at the workplace by "information workers," and on the other hand because injuries and discomfort are so common. Problems occur because of many factors, such as *environment* (e.g., inadequate equipment or equipment arrangement), *activities* (e.g., typing for too long without a break, etc.), or *bad user habits* (e.g., inadequate posture, etc.).

It is well known in the medical field that depression affects gait, posture, and of course, productivity. An individual that is not productive because he is depressed may sit in unhealthy postures, focus on the wrong activities, or limit the range of activities that he performs. One of the problems, however, is that most of us are not consciously aware of how much time we spend on different activities, and of our body postures while we work. At the same time, it is widely recognized that user attention is a limited resource, so computing devices should negotiate rather than impose the volume and timing of their communications with the user. It would be extremely useful, therefore, to have a system that helps the user increase his own awareness of his habits and activities, and that lets the user decide how that information might be used in other contexts such as interruption management.

With this motivation in mind, we created a system for monitoring a computer user's posture and activities in front of the computer. In our system, a camera is placed on top of the computer screen and the computer user is monitored by the system as he works. The system uses the camera to measure the user's posture and determine his current activity (e.g., speaking on the phone, stretching, etc.). Feedback is given to the user, in real time, on the goodness of his upper body posture. In addition, input from the camera and a microphone are used to classify the worker's activities and give him summaries of what he has been doing for a determined period of time. The system can be used in the context of attentive interfaces and for interruption control (e.g., switch off my e-mail and/or phone if I am reading).

Our approach uses background subtraction to obtain silhouettes. From the silhouettes we extract geometric features to classify activities, and obtain vertical projections to separate head from torso and measure head and shoulder angles (see Figure 5 and Figure 6). We use input from a microphone to determine audio activity (someone speaking, silence, keyboard being used) to differentiate activities that are visually similar.

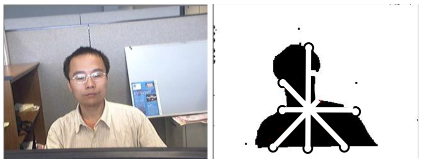
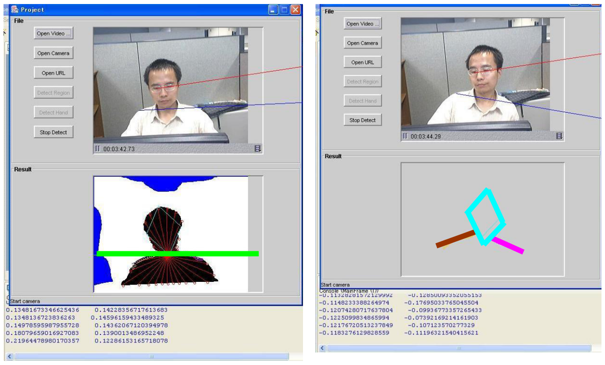


Figure 5. The endpoints of the lines define a polygon used for activity classification.



```
0.1348167334662543б    0.14228356717613683
0.134813672383б263    0.14596159433489325
0.14978595987955728    0.14362067120394978
0.18079659016927083    0.13900134869522248
0.21964478980170357    0.12286153165718078
```

```
-0.11328281572129992    -0.12850093352055153
-0.1148233388264974    -0.17695033765045504
-0.12074280717637804    -0.09936773357265433
-0.1225099834865994    -0.07392169214161903
-0.12176720513237849    -0.107123570277329
-0.1183276129828559    -0.11196321540415621
```

Figure 6. Extraction of head, and shoulder angles.

4. USE OF FACE

The hotspot system can be used in many different scenarios. For instance, the camera may be pointed at any interactive space in a room. In that case, the hotspots could be made to function in relation to the location of the person or persons that may interact with the system. Face detection could be used in this context to find people, and using that information constraint the approximate location of their limbs. For instance, in a meeting room scenario actions such as raising hands could be recognized using face detection: the location of the hotspots would not be fixed, but rather be relative to the location of the face.

In the meeting room scenario, several challenges for face detection remain, including the following:

- Occlusion
- Lighting
- Orientation (do we need 3D face information?)
- Scale
- Obstruction (where to place the cameras?)
- Robustness

Although the meeting room scenario is highly constrained, there are many possible variations that make application of face detection or face processing (e.g., recognition) algorithms difficult. The benefits, on the other hand, could be increased robustness and more functionality. Identifying people in the meeting, for instance, could be highly beneficial.

In contrast, in the second application, it is not difficult to assume that the main "object" appearing in front of the computer (and therefore camera) is a person. Therefore, the question is whether face detection is necessary or if algorithms to perform other tasks can be performed efficiently. In considering face processing, for instance, it is possible to look into many possibilities, including recognition tasks and visible affective states or activities. For instance, the system could be used for security (e.g., identify the person in front of the computer is the owner), to determine if the person is tired or energetic, or to detect finer activities such as yawning or sleeping.

The challenges in the second application, can therefore be summarized as follows:

- Computational complexity
- Robustness (e.g., for identification)
- Calibration (e.g., for gaze or other more specific processing)

The main benefits would probably be added functionality. As pointed out in [1], several of the errors in the current, simple implementation, can be eliminated using audio input. The current implementation for head-torso separation fails if the user has long hair. However, simple heuristic constraints or color-based face detection are likely to be sufficient, raising the question of whether face processing is really necessary. Furthermore, if it is implemented in any of the two systems, other considerations must be taken into account including speed and complexity.

5. CONCLUSIONS

This paper briefly describes two systems that do not currently use face processing algorithms, but that could possibly benefit from applying face processing algorithms at different levels. Many questions are raised in the context of such applications, including whether algorithms specific for face processing (e.g., detection, recognition) are needed and the level of analysis required (e.g., face detection, gaze, 3D location, etc). In addition, we must consider efficiency, and robustness. If the domain is highly constrained, as in the posture system, why do we need face processing and is it worth the effort given the range of open research issues?

6. REFERENCES

[1] A. Jaimes, "Posture and Activity Silhouettes for Self-Reporting, Interruption Management, and Attentive Interfaces," *2006 International Conference on Intelligent User Interfaces (IUI '06), Sydney, Australia, Jan. 28-Feb. 1, 2006.*

[2] A. Jaimes and J. Liu, "Hotspot Components for Gesture-Based Interaction," *Interact 2005,* Rome, Italy, Sept. 12-14, 2004.