

Multi-User Natural Interaction System based on Real-Time Hand Tracking and Gesture Recognition

Alberto Del Bimbo, Lea Landucci, Alessandro Valli
University of Florence, Italy
Dipartimento di Sistemi e Informatica
Via Santa Marta 3, I-50139 Firenze, Italy
alberto.delbimbo@unifi.it, lea.landucci@unifi.it, valli@micc.unifi.it

Abstract

We present a Computer Vision based system that enables multiple people to interact naturally with a large display table using their own bare-hand gestures. The display presents and supports a particular multimedia application that can be used at the same time even by remote users. Finally we describe two different applications designed for didactic and entertainment scenarios.

1. Introduction

Multi-user systems are becoming increasingly common in different scenarios such as museums, entertainment, information on demand, didactics and they have many advantages such as supporting collaboration on shared data. Studies show that larger displays enable users to create and manage large number of different data, as well as to involve more people in same tasks such as visual manipulation. As large displays become more affordable, researchers are investigating also techniques for making the large-display user experience more effective[1]. There are relatively few tools that directly address the needs of users simultaneously interacting with a large display and they are typically dedicated to users with a high level of experience in traditional interfaces, sometimes expecting them to be provided with specialized tools such as 3D pointing devices or data gloves. We believe that such devices may limit the system's usefulness. Therefore, our goal is to design a multiuser easy-to-use system where people can interact simultaneously with multimedia contents through their own bare-hand [2] gestures. Clearly the system needs to be more oriented towards natural interaction: in order to implement these kinds of improvements, we used non intrusive and invisible input sensors, exploiting advances and results of pattern recognition, and image understanding. We decided to work on

Tabletop interaction [3] because of its great affordance [7] to the unconscious knowledge of any civilized person that would use it: tables are an ideal environment for sharing information and digital objects with others. Since hands are the most communicative part of human body[4], several works address hand gesture commanded interfaces. 3D techniques have been used in systems where robust recognition of complex hand postures is needed (i.e. for 3D Environments[5], 3D objects manipulating[6]). While accurate, 3D techniques have high price in term of computational technical complexity. Cheaper solutions work with 2D data and use vision based methods to perform color or motion detection[8], and recognize gestures through neural networks[9], or Hidden Markov models[10]. By using 2D simple vision based techniques, we created a real-time system based on a single PC station able to track users' hands, recognize their position, shape and motion and make the interface react to them. The gesture recognition method developed, using simple background subtraction combining with the use of infrared techniques, has been designed in order to have low computational cost so that it allows the system to have enough computing resources to be controlled by more users.

2. System design

2.1. Setup features

We built a wooden table with a mat glass surface of 150x100cm on which a multimedia user interface is visualized through a projector fixed to the ceiling close to a webcam; a single standard webcam working at 20 fps is enough to capture the desk surface and providing it with an IR filter we made the system impervious to visible light and projected images. A 2.3 GHz computer is connected to acquire images from the webcam and to visualize the real-time updated interface through the projector; finally an infra-red

illuminator is placed next to the webcam to flood the desk surface.

2.2. Video Processing

The processing chain develops from video processing phase that consists in a segmentation done through background subtraction using a dynamic model of the scene in order to cluster pixels into blobs through a connected components algorithm [11]. Each blob (corresponding to one hand) is described with a collection of statistical and morphological measures, including moments (up to the third), bounding box, and color histogram.

2.3. Hand pose recognition and tracking

In the second processing step, each blob is analyzed evaluating its direction, size and perimeter in order to eventually associate to it a particular *hand pose*. Our system is able to recognize three different poses: open and pointing hand and an hybrid pose. In order to get that, once defined two different circular finding windows sized in accordance with standard human-hand size, we developed the following algorithm.

1. Let \mathbf{P} be the outmost point of blob \mathbf{B} (according to the arm's direction) and let w_1 and w_2 be two circular finding windows centered in \mathbf{P} respectively sized 20 and 3 cm in ray (see fig.1).
2. Let us consider set $\mathbf{S}(w) = \{(x, y) | (x, y) \in \partial\mathbf{B} \cap w\}$, and let $\mathbf{N}(\mathbf{S}(w))$ be the number of elements in $\mathbf{S}(w)$;
 - if $\mathbf{N}(\mathbf{S}(w_1)) \geq \hat{N}$, where \hat{N} is a fixed bound empirically determined, then we consider the shape of \mathbf{B} as an *open hand pose*.
 - otherwise, let us denote by $\mathbf{D}(d_1, d_2)$ the distance between d_1 and d_2 , where $d_1, d_2 = \{\mathbf{S}(w_2) \cap \partial w_2\}$;
 - if $\mathbf{D}(d_1, d_2) \leq \hat{M}$, where \hat{M} is a fixed bound empirically determined, then the shape of \mathbf{B} is recognized as a *pointing pose*;
 - otherwise the shape of \mathbf{B} is recognized as an *hybrid pose*.

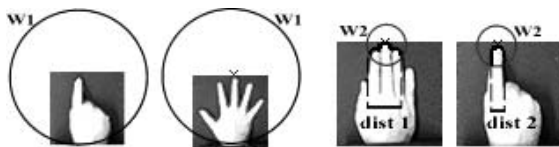


Figure 1. Recognizing poses.

Once recognized the pose, we associate to it a sensible point (the fingertip in case of pointing pose and the center of the palm in case of open hand) and a steady time, both useful data for the interaction dynamics.

The last processing step includes hand tracking in order to recognize and analyze gestures of each individual user. In order to track all the hands of people who are interacting on the table, in every frame we have to find for each blob its sensible point position. First of all we define a circular tracking window within which a point can reasonably move in 1/20 second (that is the time gap between two different frames acquired at 20 fps). Considering a generic blob \mathbf{B} , the tracking window associated is centered in its centroid, defined as follow.

First of all let us take into consideration the expression of the statistic moments of \mathbf{B} :

$$m_{pq} = \sum_x \sum_y x^p y^q \rho(x, y) \quad (1)$$

$$\text{where } \rho(x, y) = \begin{cases} 1 & \text{if } (x, y) \text{ are into } \mathbf{B} \\ 0 & \text{otherwise} \end{cases}$$

Then, the centroid coordinates are defined by the first order moments as follow:

$$\bar{x} = \frac{m_{10}}{m_{00}} \quad \bar{y} = \frac{m_{01}}{m_{00}}. \quad (2)$$

1. Let $\mathbf{P}_B(\hat{f})$ correspond to the centroid of blob \mathbf{B} at frame \hat{f} and let w_B be the circular tracking window centered in $\mathbf{P}_B(\hat{f})$.
2. For every point $\mathbf{O}(\hat{f} + 1)$ s.t. $\mathbf{O}(\hat{f} + 1) \in w_B$
 - if $\mathbf{O}(\hat{f} + 1)$ is the centroid of a blob, then $\mathbf{O}(\hat{f} + 1) \equiv \mathbf{P}_B(\hat{f} + 1)$
 - otherwise blob \mathbf{B} got out of the scenario.

Larger tracking windows allow faster gestures but increase risk of mistakes in case of multiuser mode. In order to obtain an effective hands tracking in such situations, after several testing sessions, we fixed the circular tracking window ray at 8 cm.

3. Multiuser System

When multiple users act at the same time in the same interactive scenario there are many nodal points to take into consideration: first of all the need of different simultaneous hand tracking even in presence of particular hands overlapping (i.e. accidental bump cases). In order to maintain an effective hand tracking we have to distinguish in every frame each acting hand from the others: this means to study a good labelling technique. We chose to associate to each

of them their own centroid 'id' as their label: choosing a proper tracking window we avoid the risk of wrong label assignments. This labelling technique works as long as there are no total or partial blob overlaps, or rather when we assist to a blob merging. In fact, in case of temporary overlapping (i.e. when a "bump" between different users occurs), the blobs involved merge into one blob and the system can be subjected to errors in hands tracking and therefore in the association of the observed gestures to the right users. In order to obtain good results, we have employed some heuristics to detect such cases. We define the following accidental bump model :

- there are two blobs involved;
- bump is an occasional accident;
- minimal portions of each blobs are merged.

Once defined the blob surface as the zero order statistic moments (m_{00} , according with def.1), the system detects a merging case as follow:

- blobs in the current frame are less than them in the previous one;
- the larger one of them has a surface equal to at least 1.5 times the bigger one in the previous frame.

Once such merging is detected, frames are not processed until the two involved blobs detach so that there are no restrictions about the duration of the overlapping. The system predicts the positions of the sensible points associated to the involved blobs using a linear one-step-ahead predictor of the following form:

$$\hat{x}(f) = x(f - 1), \hat{y}(f) = y(f - 1).$$

where $(x(f), y(f))$ are the coordinates of the sensible point at frame f (detach frame). Under the reasonable hypothesis made in our accidental bump model, a one-step-ahead predictor is enough to solve merging cases according with the following three step algorithm.

1. Predict the two blobs' sensible point positions they will have at the frame in which they detach;
2. evaluate the prediction error associated to each blobs in term of distance between predicted and real sensible points (see fig.2);
3. associate to each blob the 'id' of that one which minimizes the prediction error.

4. Two interface prototypes

The two interfaces proposed address two different scenarios in which we believe that a natural interaction multi user tabletop would be exploited: entertainment and didactic/educational. Common goal was to build easy-to-use interfaces where involved people learn to use them only by

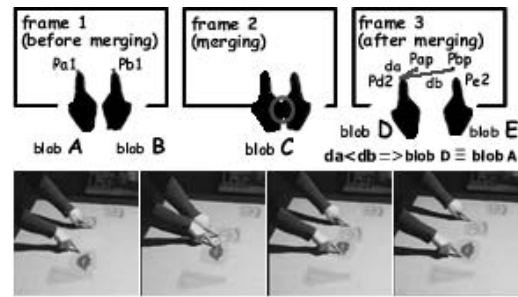


Figure 2. Merging management and testing.

observing other people working on. This meant for us to develop simple visual languages adapted to the different contexts.

4.1. Manipulating multimedia objects

The first interface proposed offers to its user the chance to manipulate some multimedia contents like videos and images by changing their visualization features. In order to develop a visual interaction language as natural as possible, we tested this interface with up to 20 persons to see which gestures would be the appropriate. Finally, The associations chosen between gestures and actions were the following:

Static actions.

selection/deselection: persistence of pointing pose on the object chosen (steady time at least of 700 milliseconds) (see fig.3, a));

play/pause/stop: persistence of open hand on the video object chosen (steady time at least of 500 milliseconds, see fig.3, b).

Dynamic actions.

drag & drop: pointing pose moving through the interactive scenario after an object selection (see fig.3, c));

roto-translation/resizing: two different pointing poses on the object chosen (steady time at least of 800 milliseconds)(see fig.3, d));

clear: open hand pose moving from one side to the opposite side of the table.

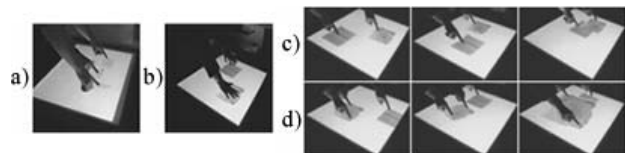


Figure 3. Static and Dynamic actions.

4.2. Big image exploring

This experimental system is part of the project VICOM (Virtual Immersive COMMunications), co-financed by the Italian Ministry of Education, Universities and Research (MIUR). This interface is designed for professor tutorship to students that interact with the collaborative table asking questions about some didactic materials: the professor answers them through a remote computer able to command the interface as well. The associations chosen between gestures and actions has been the following:

Static actions.

buttons selection: persistence of pointing pose on the button chosen (steady time at least of 500 millisecc);

zoom in/out: persistence of pointing pose in the position chosen after correspondent button selection (see fig.4).



Figure 4. Static actions.

Dynamic actions.

exploring image: alternation of open hand and hybrid pose moving from one side to the opposite side of the table (see fig.5, a); **drawing:** pointing pose moving on the table after the correspondent button selection (see fig.5, b).



Figure 5. Dynamic actions.

5. Experimental results and conclusions

This work addresses a working implementation of a 2D hand gesture recognizing system based on images segmentation through background subtraction. We have developed an interactive environment in which more users can interact simultaneously with multimedia objects. In order to solve problems due to hands' shapes overlapping, we have employed a predictor based method that is able to distinguish (under reasonable hypotheses) two hands of distinct users even when they accidentally bump. Finally we have tested the system with two different interfaces prototypes

with more than fifty persons differently aged and experienced during a whole year, and we observed very good results (see fig.6).

Purpose	Failure %	Success %
Recognizing open hand	1	99
Rec. pointing hand	4	96
Rec. hybrid pose	6	94
Solving overlapping	6	94
Rec. cleaning gesture	20	80

Figure 6. Testing results.

References

- [1] K. Dempski, B. Harvey, "Natural Support for Multi-User High Definition Visualization and Collaboration", Accenture Technology, Chicago, USA, 2002.
- [2] C. Berard, "Bare-Hand Human-Computer Interaction", Technische Universitt Berlin, Germany, 2001.
- [3] S.D. Scott, K.D. Grant, R.L. Mandryk, "System Guidelines for Co-located, Collaborative Work on a Tabletop Display", Proceedings of ECSCW'03, European Conference Computer-Supported Cooperative Work, Helsinki, Finland, September 14-18, 2003.
- [4] A. Mulder, "Hand Gestures for HCI - Research on human movement behaviour reviewed in the context of hand centred input", Simon Fraser University, Canada, 1996.
- [5] R. O'Hagan, A. Zelinsky, "Visual gesture interfaces for virtual environments", Australian Nat. Univ., Australia, 2000.
- [6] B. Leibe, T. Starner, W. Ribarsky, Z. Wartell, D. Kru, B. Singletary and L. Hidges, "The Perceptive Workbench: Toward Spontaneous and Natural Interaction in Semi-Immersive Virtual Environments", Georgia Inst. of Tech., USA, 2000.
- [7] D. A. Norman, *Psychology of Everyday Things*, 1988.
- [8] T. Kapuscinski, M. Wysocki, "Hand gesture recognition for man-machine interaction", Rzeszow Univ., Poland, 2001.
- [9] J. Isaacs, S. Foo, "Hand pose estimation for American sign language recognition", FAMU-FSU, Tallahassee, FL, 2004.
- [10] T. Starner, J. Auxier, D. Ashbrook, M. Gandy, "The gesture pendant: a self-illuminating, wearable, infrared computer vision system for home automation control and medical monitoring", Georgia Inst. of Technol., Atlanta, USA, 2000.
- [11] C. Colombo, A. Del Bimbo and A. Valli, "A real-time full body tracking and humanoid animation system", Parallel Computing, North-Holland 2004.