User Head Movement Recognition and Interpretation System for Computer Interaction

Ciprian Ovidiu UNGUREAN
„Stefan cel Mare" University of Suceava

***Abstract*—The aim of the paper is to describe a system for the head gesture recognition developed in the frame of INTEROB project[1]. The goal of this project consists in developing an interaction based on gestures with information on robotic systems. In the paper we discussed a method for controlling the mouse pointer movements on the screen by recognizing the operator head movements captured by a video camera. In the second part of the paper it is described a fast and accurate method for hand posture recognition in video sequences.**

***Index Terms*—human computer interaction, gesture recognition, image processing, face recognition, Haar-like features, camshift algorithm**

Some time ago there wasn't any significant difference between computer users (all being programmers). As the computer becomes more and more necessary, we will have to use different computing systems, miniature work stations, mobile and omnipresent. GUI interface types, especially peripherals like mouse and keyboard won't be able to stand up to future HCI requirements.

Lately, the opportunity to create post-WIMP (window, icon, menu, pointer)[1] interfaces that can be represented through gesture interaction techniques, multimodal interfaces, tactile interfaces, virtual reality or augmented reality is being studied. These appear out of the necessity to have support for a flexible and efficient interaction that can be easy to learn and use in a natural and intuitive way.

Theoretically, the more entrances and the more diverse they are, the more efficient transmitting information is. For example, a system that receives both vocal and gesture commands, in case one of the commands isn't recognized the other can substitute.

The general tendency is to optimize the entry for a certain number of applications. This isn't a satisfying solution because it leads to the designing of new gesture entries for each application. Gestures must be complementary to each other for to be used without causing confusion or adverse effects during the interaction with the system. If data instill is difficult to use, to understand or it is tiresome, the users will loose their motivation [2].

A non-contact interface system implementation method consists in gesture utilization. Because gestures are a natural method of non-verbal communication, their implementation as a command system could bring efficiency in usage [3][4][5]. A method through which the user could easily communicate, free and efficient can be done by attaching a digital camera to provide the information for the system to process [6]. We propose to realize such an equipment to allow computer interaction using only head movement to replace the mouse and keyboard.

The article is organized as follows:

I. A short description of the user head movements recognition and interpretation system

II. Utilization of the movement history in an image stream

III. Presentation of a performance testing method for the proposed approach and the obtained results

IV. Application evaluation in 2 games

V. Conclusions

I. A SHORT DESCRIPTION OF THE USER HEAD MOVEMENT RECOGNITION AND INTERPRETATION SYSTEM

In our approach we used the Adaboost algorithm because it is fast and efficient in the detection of human head in different illumination and shadow conditions, but due to the used classifier, only in video sequences which contain frontal images of the face. Due to low calculus cost and efficiency, the Camshift tracking method was used after face detection. Knowing the user's head position in every moment for movement recognition, we implemented the method that refers to "movement forms". The proposed approach has as effect the successful detection, tracking and interpretation of human head movements in different light and shadow conditions of the surroundings or the face.

Viola et al.[7][8] have used Haar features and a faster method for computing them in images at any scale within the integral image. Haar-like features are efficient in image classification because they encode spatial relations between different image regions. In face detection, these regions are defined by the contrast difference between the eyes and the cheeks or using contrast scales across the nose area and the eyes area.

[1] The work on this paper was funded by the INTEROB project, through contract 131-CEEX II03/02.10.2006

Digital Object Identifier 10.4316/AECE.2008.01011

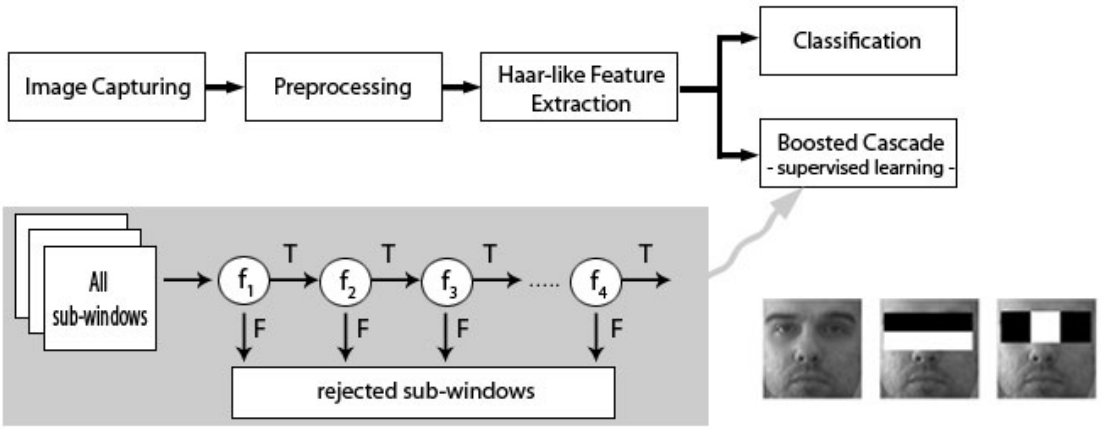

Figure 1. Block diagram of AdaBoost face detection stage.

The selected features are sorted according to their importance in order to be used as a cascade of classifiers. A cascade of classifiers is a degenerated decisional tree in which at each stage a classifier will reject some background patterns and accept the front faces from the image. Each classifier stage is trained using the machine learning algorithm Discrete Adaboost. It builds a strong classifier using a large set of weak classifiers. The speed of feature evaluation is important because the frontal face detection algorithm consists in sliding windows at all scales over the input image. Using this algorithm, user front face is detected in various varying illumination condition, occlusions or image noise.

Theoretically we could trace a face by applying the face detection algorithm proposed by Viola and Jones to the successive frames received from the web camera. Because the purpose of our application is to interpret gestures caused by head movement and because the classifier used by the Viola and Jones algorithm doesn't detect the user's face unless the user is looking at the webcam, we implemented a tracking algorithm based on color blends – Camshift (Continuously Adaptive MeanShift) [9] which is an extension of the MeanShift algorithm proposed by [10]. Camshift is a tracking algorithm that uses a combination of colors (user face skin color, obtained when Adaboost was applied) thus being able to detect head position in video frames, no matter its orientation. Camshift can be applied in dynamic changing distributions by resizing the next frame's search window, considering the initial moment (zero-moment) the current frame. This algorithm can be used successfully in anticipating the object's position in a sequence of video frames. Thus, the face position detected through the Adaboost method could also be known in case the user is not looking directly towards the webcam, like when the head is bowing, tilting or twisting.

After the front face is detected, we create the user's skin color histogram in Hue Saturation channels from HSV color space [11]. The skin color histogram is used as a model for converting incoming pixels to a corresponding probability of skin image.

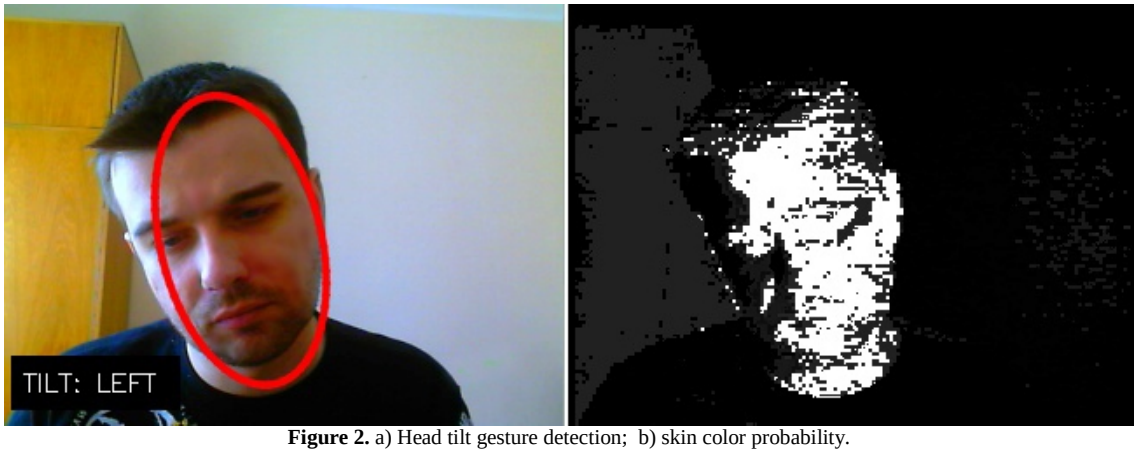

Figure 2. a) Head tilt gesture detection; b) skin color probability.

a. Knowing the ellipse angle orientation the head tilt gesture is detected. b. The brighter pixels have the highest probability to be skin pixels.

After the face is detected through Adaboost, Camshift is used in tracking. Camshift is not very exact, but it can tell us in which area the users face is. The search area on which Adaboost can be applied again, for an eventual search will be smaller than the entire frame received from the webcam, but larger than the window returned by Camshift. Using this approach, we benefit from the advantages offered by the two methods. (efficiency in face detection, tracking speed).

Metodă	Time cost per frame
AdaBoost	57ms
AdaBoost + Camshift	18ms

These results were statistically obtained for a period of 5 minutes on a Windows XP system with an Intel Pentium M 1.7 GHz CPU.

The conditions for which Adaboost will search for a face in the whole frame are the following:

- the tracking window is too narrow or too tall
- the condition imposed by the ratio between height and width of the tracking window is not met, as to represent a face
- inclination of the ellipsis that represent the face is horizontal.

After the first front face detection in the video frame, I'm using anthropometric human head dimensions [12][13] (width, height) as a constant. Using this approach, gestures like moving towards and back from the webcam can be recognized.

II. UTILIZING MOVEMENT HISTORY IN A STREAM OF IMAGES

The main methods of movement recognition and interpretation consist of trajectory analysis of the moving parts [14][15][16], using the hidden state-space Markov model [17] or the active-passive model [18].

In [19] there exists an alternative for the gesture recognition methods through frame superposing that form a specific movement for the human behavior motion patterns. Looking at movement in an unclear video sequence two reference points are obvious. The first is the spatial area in which the movement takes place. This reference point is given by the pixel area where something is significantly modifying, no matter the way in which it moves. The second reference point is the way in which the movement evolves within this area (i.e. an element which expands or which rotates in a certain location). The human visual system has developed its one ways to exploit these notions of how and where, and it captures sufficient movement proprieties to be used in recognition.

The last N frames are stored in a temporary circular vector. Each new frame is transformed in gray shades and is saved at the new location in the temporary vector. The absolute difference between two successive elements of the vector is thresholded to represent silhouette movement in the current frame. Movement history in the image is calculated as follows:

```
mhi(x,y) =   timestamp     if silhouette(x,y)!=0
               0           if silhouette(x,y)=0
and mhi(x,y)<timesamp-duration
             mhi(x,y)       altfel
```

The way in which a movement was performed is represented through MHI, and MEI represents where the movement took place. By use of the opening morphological filter on movement energy within the image the noise produced by small movements in the frame are eliminated.

Movement direction for each component is calculated as follows:

$$f(x,y) = \arctan \frac{F_y(x,y)}{F_x(x,y)} \quad (1)$$

In which $F_x(x, y)$ and $F_y(x, y)$ are spatial derivatives of the x,y directions from MHI.

Knowing the user's head position in every moment, the MHO algorithm can be applied. The advantage consists in the fact that the movement within a number of frames can be resumed in a single gradient image, and using this model is not time related. The main problem in using this method consists in segmenting and labeling the silhouette that performs the movement. In our case the silhouette labeling is made using the Adaboost-Camshift method. Most of the silhouette extraction approaches use background, optic flow or stereoscopic extraction. We chose the frame difference method. We calculate the direction of those components which define the movement, only if they lie within the face area. Using movement direction detection from the gradient (motion gradient operation – MGO) [19] and the ellipsis which frames the face area applying Camshift allows us to recognize translation, rotation and inclination head gestures.

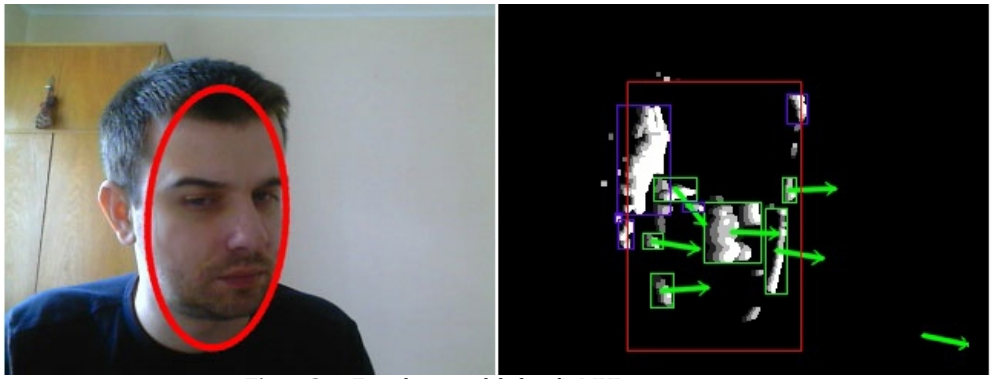

Figure 3. a. Face detection-labeling b. MHI representation.

In this Fig. the red ellipsis successfully marks the user's face area due to applying the detection and tracking algorithm. Next to it is an example of the frame with MHI applied. The red square marks the face area. Direction and magnitude are calculated only for components within the face area (green marked). The current frame movement

direction, after summing the shift vectors, is represented in the lower right area.

III. PERFORMANCE TEST

To test the head movement gestures recognition rate we implemented a test application. At application start, a movie with a virtual character performs a series of directional gestures using its head, and the user must reproduce them. The test application receives the webcam's frame flow and recognizes user gestures. Accuracy correctly detected gestures is 74.2 %.

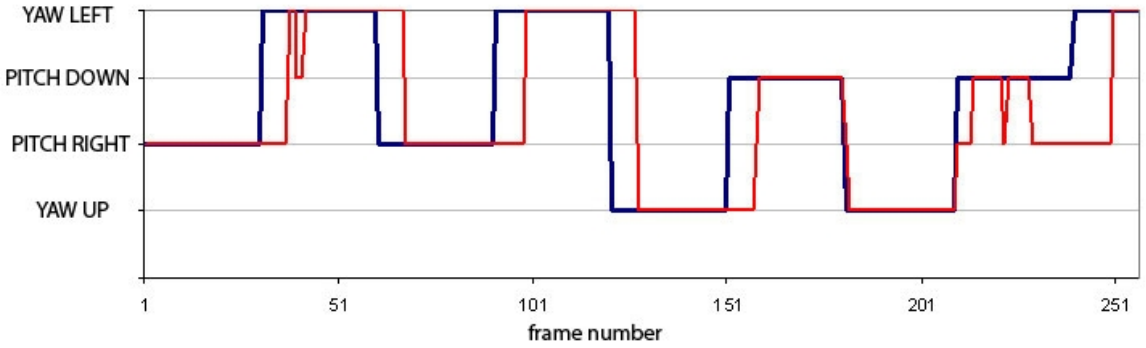

Figure 4. Results of the head tracking system.

In this graphic, the virtual character's gestures are blue marked. Red marks gestures recognized by the test application. In this case, in which the user must reproduce the virtual character's movements, as noticed in the representation, a delay appears due to the necessary user time to recognize the virtual character's gesture.

Figure 5. a) virtual head; b) user head with recognized gesture.

IV. INTEGRATION IN APPLICATIONS

We also tested this approach's efficiency in 2 computer-games: Pac-Man and Counter Strike 1.6. In the arcade game Pac-Man the user's head movements direct de main character through the labyrinth to gather points and to avoid the ghosts. One of the users specified that after the test/evaluation period of the implementation, when he wanted to play the game at home, on his one PC, he would have wanted to be able to control the Pac-Man character using head gestures instead of directional keyboard buttons.

Counter Strike is a FPS (first person shooter) game, in which movement through the virtual space is done using the directional keys and the point of view is modified through mouse movement. In our test application, a user head movement leads to a change in the point of view. I.e. when moving the head to the right, the point of view will move to the right.

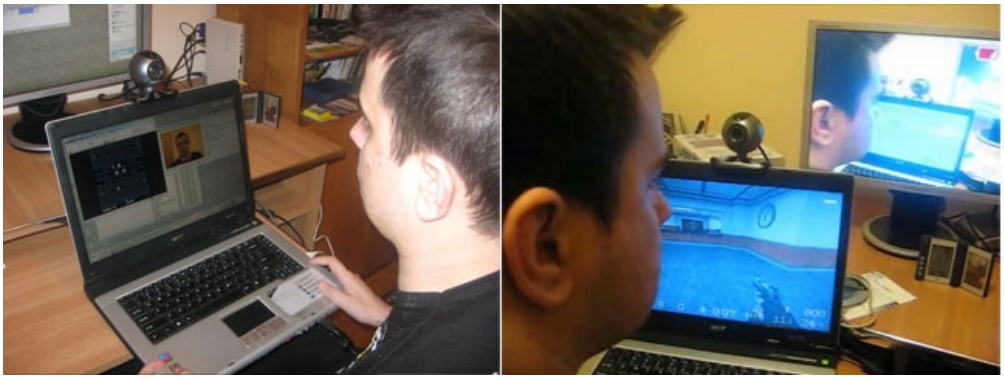

Figure 6. Interacting with Counter-Strike game using head moves.

V. CONCLUSION

When the users need to select an object of interest by pointing it by using physical interfaces as mouse, keyboard, tracking or sensors, the action is carried out indirectly. A software tracker which can follow the user's face, without any specialized equipment and in a completely passive and non-interfering way is a great success. Our tests have found that most users were able to control sample applications after a few minutes of practice.

Perceptual user interfaces are inspired from real world and human-to-human interactions. Research in this field will allow people to use technology more efficient, natural and easy to learn.

ACKNOWLEDGMENTS

The presented work is supported by the research grant 131-CEEX-II03/02.10.2006. The research consortium of this project is composed by: University of Suceava, Technical University of Iasi, University "Alexandru Ioan Cuza" Iasi, SC Genpro Iasi, SC CAOM Pascani, and the European partners Universite des Sciences et Technologies de Lille and "KaHo Sint Lieven" Gent.

REFERENCES

[1] A. van Dam, "Post-WIMP user interfaces". Communications of the ACM, Vol. 40, No. 2, Pages 63-67, Feb. 1997.

[2] S. Oviatt and W. Wahlster (eds.), Human-Computer Interaction (Special Issue on Multimodal Interfaces), Lawrence Erlbaum Associates, Volume 12, Numbers 1 & 2, 1997.

[3] Radu Daniel VATAVU, Ștefan-Gheorghe PENTIUC, Christophe CHAILLO, On Natural Gestures for Interacting in Virtual Environments Advances in Electrical and Computer Engineering, Suceava, Romania ISSN 1582-7445, No 2/2005, volume 5 (12), pp. 72-79.

[4] George MAHALU, Radu PENTIUC (2001) Acquisition and Processing System for the Photometry Parameters of The Bright Objects Advances in Electrical and Computer Engineering, Suceava, Romania, ISSN 1582-7445, No 1/2001, volume 1 (8), pp. 26-31.

[5] Pentiuc, S.G., Vatavu, R., Cerlinca, T.I., and Ungureanu, O. "Methods and Algoritms for Gestures Recognition and Understanding". The Eighth All-Ukrainian International Conference, UkrOBRAZ'2006, pp. 15-18, Ukraine, August 2006.

[6] Keates S, Perricos C "Gesture as a means of computer access". Communication Matters. 10. 1. 17-19

[7] P. Viola and M. J. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", Proceedings of IEEE Computer Society's Computer Vision and Pattern Recognition (CVPR 2001), Vol. 1, pp. 511-518, 2001.

[8] R.Lienhart, J.Maydt, "An Extended Set of Haar-like Features for Rapid Object Detection", Intel Labs,2002, Intel

[9] G. R. Bradski. "Computer video face tracking for use in a perceptual user interface". Intel Technology Journal, Q2 1998.

[10] D. Comaniciu and P. Meer, "Robust Analysis of Feature Spaces: Color Image Segmentation," CVPR'97, pp. 750-755.

[11] T. S. Caetano, D. A. C. Barone, "A probabilistic model for human skin color", IAP Conf. 2001, pp. 279-283.

[12] Leslie G. Farkas, Jeffrey C. Posnick, Tania M. Hreczko, "Anthropometric Growth Study of the Head", In: The Cleft Palate-Craniofacial Journal: Vol. 29, No. 4, pp. 303–308, 1992.

[13] http://en.wikipedia.org/wiki/Vitruvian_Man

[14] Polana R, Nelson R, "Low level recognition of human motion". In:Workshop on Motion of Nonrigid and Articulated Objects. pp 77–82, 1994.

[15] Black M, Yacoob Y, "Tracking and recognizing rigid and nonrigid facial motions using local parametric model of image motion". In: Proceedings International Conference Computer Vision, pp 374-381. 1995.

[16] Madabhushi A, Aggarwal J, "A Bayesian approach to human activity recognition". In: Proceedings of IEEE Workshop on Visual Surveillance, pp 25–32. 1999.

[17] Chen, F.S, Fu, C.M., Huang, C.L., "Hand gesture recognition using a real-time tracking method and hidden Markov models", IVC(21), No. 8, August 2003, pp. 745-758.

[18] Cutler R, Davis L, "Robust real-time periodic motion detection, analysis, and applications". IEEE Trans Pattern Anal Mach Intel 22(8):781–796, 2000.

[19] Gary R. Bradski & James W. Davis Motion segmentation and pose recognition with motion history gradients Machine Vision and Applications, 2002, 13: 174–184.

[20] Marius CERLINCA, Adrian GRAUR, Ștefan-Gheorghe PENTIUC, Simulation of Switch Box Routing in FPGA, Advances in Electrical and Computer Engineering, Suceava, Romania, ISSN 1582-7445, No 1/2002, volume 2 (9), pp. 86-90.