Poster: Robust Vision-Based Hand Tracking using Single Camera for Ubiquitous 3D Gesture Interaction

Sergio Rodriguez, Artzai Picón and Aritz Villodas

Infotech Unit, Tecnalia
{ srodriguez , apicon, avillodas } @ robotiker.es

Abstract

Nowadays, the gestures are one of the most natural interaction methods of the Human Being. In fact, a great percentage of the human communication is based on visual communication rather than verbal communication. This work proposes a novel 3D interaction technique based on computer vision for human-computer gesture interaction. The main contribution to the interaction field is that this technique implements and improves on the hedge computer vision algorithms, so it offers a low cost solution and robustness against scenario changes and user changes. Thanks to those characteristics with this system is possible the human-computer ubiquitous interaction.

1. Introduction

In recent years, smart environments are growing faster. As the ubiquitous computing is the main paradigm in those environments, it is clear that new human-computer interaction techniques are needed and those techniques must be more natural than the classical ones.

Researchers all over the world have been searching natural interaction techniques, and the most prominent one is the gestural interaction. Much research has been done [1] [2] [3] [4] [5], but some of them won't be able to be integrated in a smart environment because of they don't meet several ubiquitous computation requirements.

Weiser establishes, in this well-known work on ubiquitous computation [6], that all the components inside an ubiquitous system have to be wireless, and some systems (e.g. [3]) don't fulfil this characteristic. *Weiser* also proposes restrictions in the power consume, so the most haptic systems (e.g. [4]) are not recommended. An alternative to the hardware 3D interaction devices are the interaction techniques based on computer vision. These ones only require a camera and a small computation system. However, some of the vision-based techniques don't meet one of the restrictions imposed by *Weiser*: low computation needed; so systems like [1] aren't recommended for the ubiquitous interaction.

Besides the constraints imposed in [6], there are also some problems related to the use of computer vision for gesture interaction. The illumination is a key point to take into account. In fact, some systems as [5] are not robust enough against illumination changes. These techniques are also affected by the user that is interacting. Both of these factors must be taken into account when designing a 3D gestural interaction system based on computer vision.

2. Proposed Interaction Technique

In this work, the research team was focused on the development of a vision-based 3D interaction technique by developing a system based on just a low-cost camera (a webcam) for 3D interaction. In this way, it's possible to be used in a mobile computing system. The main goal of the work is to fulfil the following requirements: 1) low computational cost, 2) robustness against user changes and 3) robustness against scenario changes. In order to achieve those characteristics, the work has focused on the use of on the hedge computer vision algorithms, improved to obtain a higher robustness ratio.

With this target we have developed a system which lets the user interact freely with its environment, using only the gestures of its hands in a three dimensional space. In the next figure can be depicted this proposed architecture.

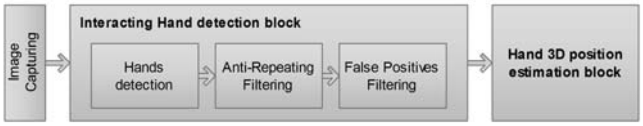
Image Capturing
Interacting Hand detection block
Hands detection
Anti-Repeating Filtering
False Positives Filtering
Hand 3D position estimation block

The first step is related to image capture. The developed method has to detect the hand and transfer this information to the second block that processes mathematically the information given by the first block, in order to estimate the 3D position of the "interacting hand". The following sections describe several aspects of each block.

2.1. Interactive Hand Detection Block

Hand detection is the first step in the proposed interaction system. This is possible to do from different perspectives [1] [2] but in this work we had chosen the AdaBoost algorithm trained with Haar-like features [7] in order to describe and characterize the hand features robustness. Thanks to this algorithm, it is possible to detect the hands that are present in an image. However, this approach still presents some problems that have to be solved as it also detects false positives and repeated hands. In order to avoid these problems two consecutive filters are applied. The first filter removes repetitions on the detected hands and extracts the real hands. The second filter aims to remove the false positives derived from the detection algorithm, and it is based on a Kalman predictor approach.

In next paragraph, the complete hand detection methodology is described in detail:

2.1.1. General Hand Detection

After the image is obtained, the system has to detect the hands in the image. For this purpose, we use one of the most referenced object detection algorithm in the computer vision field: the Viola-Jones detector [7]. This detector uses an AdaBoost algorithm to select a set of Haar-like features which can determine where in the image is an object. This detector is usually used for face detection due to its robustness and low computational cost. Because of these two main characteristics, several works had been done using this algorithm to detect hands [9]. In the present work, we propose to use this approach in order to detect closed hands.

2.1.2. Anti-Repeating Filtering

Because the Viola-Jones detector detects the same hand in different near places. In order to avoid this situation, we propose a first filter. This filter generates unique-detected-hand using information from these repeated hands. This unique hand is generated when the detected hands are near one each other, satisfying the next Euclidean distance condition:

$$\sqrt{\left[C_i(x) - C_j(x)\right]^2 + \left[C_i(y) - C_j(y)\right]^2} < d_{threshold}$$
$$\forall\ i = 0 \ldots N,\ j = 0 \ldots N \,/\, j \neq i$$

Where C_i and C_j are the position of each detected hands, $d_{threshold}$ is the maximum distance between hands and N are the total number of detected hands.

2.1.3. False Positives Filtering

A second problem arises when using the Viola-Jones detector. This is that it is possible that the system detects hands where in the reality doesn't exist (called *false positives*). In order to remove the false positives, this second filter is applied. It defines that a detected hand is a true hand when a) its position is inside a range near the last known hand position and b) its position is near the position predicted by a Kalman filter [8].

2.2. Hand 3D Position Estimation Block

The second block of the work is the 3D position estimation of the hand from the 2D image. The position estimation in the camera plane is quite easy after the previous detection process. The big contribution of this work is the estimation of the third dimension (the depth) of the hand, using only one camera and dealing with the illumination, skin colour and camera noise problems.

For the estimation of the hand depth, we propose to use the information generated during the hand detection process. In this case, we use the estimated hand size obtained after the detection and filtering process. If this size is "relatively" large, it means that the hand is near the camera and if the size is small, the hand will be far from the camera. Problems related to this detection are avoided with an additional mathematical processing. An algorithm assigns a finite number that defines the size, taking into account that the size is variable. The next equation is used to assign that number:

$$\alpha_i = \left\lfloor \frac{\left\lfloor \frac{Size_i}{A} \right\rfloor}{B} \right\rfloor$$

Where *A* and *B* are two empirical determined constants, and this number is only valid number when it accomplish the following condition for a buffer (N):

$$if \quad \alpha_N = \frac{1}{N}\sum_{i=0}^{N} \alpha_i \quad then \quad (valid\,\alpha_N)$$

Thanks to this indicator is possible to know that if the hand is approaching or going away, so this interaction technique can work as a joystick. This feature is a great step forward in the vision-based 3D interaction using only one camera.

3. Preliminary Conclusions and Future Work

In this research, we have proposed a vision-based 3D interaction technique. This technique is the only one that combines factors like robustness against user, scenario and illumination changes, and it is based on the use of one low resolution and low cost camera. All this aspects lets the system to be used by everybody, everytime and everywhere, being suitable for mobile and embedded applications.

This interaction technique is being tested and the preliminary results show that is a perfectly valid system for 3D interaction. The hand detection is quite robust, but we want to make an extremely robust system. For this purpose, we are going to re-train the Viola-Jones detector, using a larger hand database, considering more illumination possibilities and more user hands. The second aspect is to improve the hand depth estimation that could be optimized by the use of automated plane calibration techniques or even the new time of flight cameras.

The future of this work is oriented to gesture recognition in order to make the system capable of distinguishing different gestures and interpret them independently.

Acknowledgements

This work was developed under the project "mIO!", and it is supported partially by the Ministry of Industry, Tourism and Trade of the Government of Spain under the CENIT program.

References

[1] Feng Wang, Ngo Chong-Wah, Pong Ting-Chuen, "Gesture Tracking and Recognition for Lecture Video Editing," 17th International Conference on Pattern Recognition, vol. 3, pp.934-937, 2004.

[2] A. Isasi, T. Bartolomé, A. Picón. "Sistema De Interacción avanzada Utilizando Gestos y Voz Para El Manejo De Un Puente Grúa", Ibero American Conference (CIAWI), Lisbon (Portugal), 2008.

[3]Tao Ni, R.P. McMahan, D.A. Bowman, "Tech-note: rapMenu: Remote Menu Selection Using Freehand Gestural Input", Proceedings IEEE Symposium on 3D User Interfaces, pp.55-58, 2008.

[4] Latoschik, M. E. "A gesture processing framework for multimodal interaction in virtual reality", Proceedings of the 1st international Conference on Computer Graphics, Virtual Reality and Visualisation, Camps Bay South Africa , pp 95-100, 2001.

[5] Nickel, K. and Stiefelhagen, R. "Pointing gesture recognition based on 3D-tracking of face, hands and head orientation", Proceedings of the 5th international Conference on Multimodal interfaces, Canada, 2003.

[6] Weiser, M. "Some computer science issues in ubiquitous computing", *Commun. ACM* 36, 7 ,pp 75-84, 1993.

[7] P. Viola and M. Jones. "Robust Real-time Object Detection", Intl. Workshop on Statistical and Computational Theories of Vision, 2001.

[8] Kalman, R. E. "A New Approach to Linear Filtering and Prediction Problems," Transaction of the ASME-Journal of Basic Engineering, 1960.

[9] Qing Chen, Nicolas D. Georganas, Emil M. Petriu, "Real-time Vision-based Hand Gesture Recognition Using Haar-like Features", Proceedings of Instrumentation and Measurement Technology Conference,Warsaw, Poland, 2007.