Fast and Accurate Hand Pose Detection for Human-Robot Interaction*

Luis Antón-Canalís[1], Elena Sánchez-Nielsen[2], and Modesto Castrillón-Santana[1]

[1] Institute of Intelligent Systems and Numerical Applications in Engineering
Campus Universitario de Tafira, 35017 Gran Canaria, Spain
[2] Department of S.O.R. and Computation, 38271 University of La Laguna, Spain
enielsen@ull.es

Abstract. Enabling natural human-robot interaction using computer vision based applications requires fast and accurate hand detection. However, previous works in this field assume different constraints, like a limitation in the number of detected gestures, because hands are highly complex objects difficult to locate. This paper presents an approach which integrates temporal coherence cues and hand detection based on wrists using a cascade classifier. With this approach, we introduce three main contributions: (1) a transparent initialization mechanism without user participation for segmenting hands independently of their gesture, (2) a larger number of detected gestures as well as a faster training phase than previous cascade classifier based methods and (3) near real-time performance for hand pose detection in video streams.

1 Introduction

Improving human-robot interaction has been an active research field in recent years in robotics community. A major challenge is based on detecting and interpreting human behaviours in video data, since it is essential for enabling natural human robot interaction. Our attention focuses on the communication with robots via hand gestures, which are a natural means of non-verbal communication for people.

In this paper, a fast and accurate hand pose detection approach that detects hand gestures in video streams for human-robot interaction is presented. In our approach, the hand pose detection problem is formulated in terms of the integration and combination of temporal coherence information and a cascade classifier method.

The cascade classifier method is based on the fasted and most accurate pattern detection approach for faces in monocular grey level-images [1]. This classifier is trained to detect wrists as an issue for locating hands in the first frames where the interaction with the machine takes place and as mechanism for system reinitialization. The main advantage of this approach is that wrists are highly independent from the gesture being made, so hands are detected without taking into account the gesture. Temporal coherence information is supplied by a template tracker with the aim of achieving real-time performance.

* This work has been supported by the Spanish Government, the Canary Islands Autonomous Government and the Univ. of Las Palmas de G.C. under projects TIN2004-07087, PI20003/165 and UNI2003/06.

2 Related Work

Nowadays, there are several obstacles for achieving robust and efficient hand pose detection methods in video data, mainly due to the fact that the different posed difficulties such as variability and flexibility of articulated hand structure, shape of gestures, real-time performance, varying illumination conditions and complex background clutter. Therefore, previous works assume different constraints, like a limitation in the number of detected gestures using for example a watershed algorithm on the skin-like coloured pixel in collaboration with a particle filtering algorithm [2] for segmenting a specific set of hand gestures [3] or a no-real time hand detection against arbitrary background with an 86% accuracy rate through the use of an elastic graph matching technique for robot control [4]. Also, robust initialization and reinitialization must be addressed in order to carry out an effective hand pose estimation approach when a tracking method is used. However, most tracking approaches need to be manually initialized and cannot recover themselves when they lose the tracked target. As a result, some approaches often assume that the template which represents the target object is correctly aligned in the first frame [5]. Other approaches select the reference models by a hand-drawn prototype template, i.e., an ellipse outline for faces [6]. Moreover, the use of dynamic models that characterize hand motion such as particle filtering algorithm [2] requires training using the object moving over an uncluttered background to learn the motion model parameters before it can be applied to the real scene. However, transparent initializations without user participation are required for interactive human-robot communication.

Recent hand pose detection approaches are focused on Viola-Jones [1] cascade classifiers, commonly used for detecting faces. Although frontal faces share common features (eyes, eyebrows, nose, mouth, hair), hands are not so easily described. Their variability and flexibility make them highly deformable objects, so training a cascade classifier for detecting hands is a complex and arduous task. For that reason, a different classifier for each recognizable gesture has been trained [7], or a single classifier for a limited set of hands has been proposed [8]. However, the use of these approaches leads to the detection of a low number of gestures. Furthermore, real-time performance is not achieved with a cascade classifier method such as the one illustrated in [9] and only 15º rotations can be efficiently detected with a Viola-Jones detector [10]. Most importantly, the training data must contain rotated hand samples within these limits. Therefore, our approach changes the detection target to wrists. As a result, hands are detected without taking into account the gesture. Additionally, there is no limitation in the number of gestures being detected, as long as wrists are not occluded. Furthermore, fast computation is achieved incorporating temporal coherence information. And, the training time for the cascade classifier is greatly reduced. In the following sections, the proposed solution will be described and evaluated with experiments.

3 System Initialization

The Viola-Jones based cascade classifier [1] is used in order to automatically initialize the system for detecting hands in the first frames when the interaction takes place.

This cascade method combines increasingly more complex classifiers in order to quickly discard background areas on the first stages while a deeper analysis is performed in areas of high interest. Hands, however, are highly deformable objects, hard to train and classify due to their variability and flexibility. Next subsections describe the steps taken from the training of the classifier to the final hand extraction.

3.1 Training the Classifier

Training samples must be collected in order to train a cascade classifier. There are two categories: negative and positive samples. The first ones are related to non-object images while positive samples correspond to object images. However, the underlying problem with hand shapes in the training stage is that they are not self-containing objects, so patches of non-object images (background) are shown within positive samples. This makes the training stage harder and time consuming. Different hand samples are shown in figure 1.a. Due to the presence of background patches among positive samples, it is necessary a large collection of images showing hands in front of different sceneries. This, added to the variation of light conditions and hand postures in order to include every possible setting, results in a high computational cost of the training stage and an unreliable detection.

We propose a simplification of the classifier method, using wrist images as object samples. Wrists are much simpler objects, so the variability among samples is reduced and thus a faster training stage is achieved. Some used wrist samples are shown in figure 1.b, while figure 1.c illustrates the difference between the lower and upper section of a positive sample, being the former a simpler object. As long as wrists are not occluded, their detection leads to their hand, thus fulfilling the original goal.

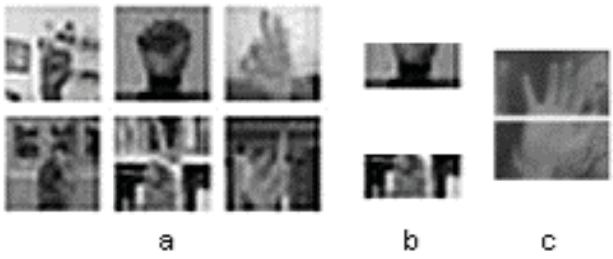

Fig. 1. Positive sample images: a) whole hand, b) lower part of hands, used by our wrist classifier, c) detail of a sample image, divided in two sections.

3.2 Finding and Isolating Hand Pattern

In order to reduce the search space where the wrist cascade classifier is applied, people is first located using a cascade classifier as described in [11]. According to average human body proportions [12], an arm length is around three times the length of a head, so a boundary of the distance that a hand can reach knowing the location of a head can be computed. The result is that, for typical desktop images, more than a half of the original image may be removed from the problem space. If no faces are detected, the search space problem is aimed to the original image dimension.

Once a wrist has been located, its enclosing area, given by the cascade classifier, is resized in order to include the whole hand taking into account natural hand-wrist proportions, where a hand's height is between 2.9 and 3.1 times its wrist's width. This whole hand area is supplied as the tracking pattern, and will be followed in successive frames.

4 Hand Gesture Detection System

Our hand gesture detection system is based on a continuous operation that combines the results computed by the cascade classifier method with a template tracking module. The tracking module is used in order to get benefits from temporal coherence of the hand detection information provided by previous frames.

Robustness to background clutter and low computational costs are the main issues that need to be addressed when a tracker module is used in order to follow hands from previous frames. With this aim, we make use of the tracking algorithm of [13] that has been previously applied to different visual applications such as face and vehicle tracking. This algorithm is focused on the framework of representation spaces based on second order isomorphism [14] that allows the definition of *context objects* notion. The use of this concept allows taking into account similar objects related to the target object and deciding when it is necessary to update the target pattern. Updating hand patterns using this concept avoids confusing the tracked target with clutter and similar objects from background.

Our cyclic operating approach involves four different processing stages. The first process begins when the classifier finds a wrist in the way described in section 3.2. The hand it belongs to is selected as the tracking pattern, so in a second stage the tracker will follow it during the next 30 frames or until the pattern is considered lost. On the next stage, the wrist cascade classifier is applied again. This second time, however, the search space is reduced to an area close to the last tracked pattern position. Once again, the result of the classifier is selected as the new tracking pattern, which will be followed again during 30 more frames or until the pattern is considered lost, as it was achieved in the second stage. Finally, the operation cycle is restarted with a new application of the wrist cascade classifier on the whole search space, as described in section 3.2.

The main assumption underlying this approach is that hands can be frequently expected to enter and exit from view and that a robust reinitialization is required when the tracked hand is lost due to exceptional circumstances based on drastic appearance transformations in the gesture being made. An overview of the different processing stages that take place in our framework is shown in figure 2.

5 Experimental Results

In order to carry out empirical evaluations of the system, 12 different video streams with an average of 1500 frames each one, 320x240 pixels each frame, were acquired at 25 frames per second, and analyzed using a PIV 2.8 GHz. These videos contain 12 different people with assorted background and light conditions, making more than 20 vertical hand gestures. The first two subsections describe the results computed with the classifier method, analyzing the training stage and the performance of the classi-

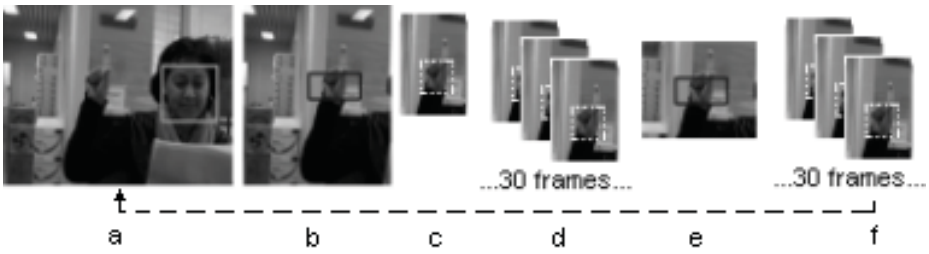

Fig. 2. Hand Gesture Detection System: a) faces are detected, b) wrists detection in the reduced search area, due to faces detected, c) hand used as the tracking pattern, d) hand tracked during 30 frames, or pattern lost, e) new wrist detection, in space around last tracked position f) new detection tracked during 30 more frames, or pattern lost. Finally, the continuous cycle restarts in stage *a*.

fier method using wrists and using the whole hand. The last subsection shows the performance achieved when a tracker module is incorporated.

5.1 Training Stage

The used training set consists of 5653 negative samples and 4130 20x20 positive samples from our own dataset and other samples selected from available datasets [15]. The trainer only takes into account the lower part of those images, 20x10 pixels, which show a hand from its wrist to half the palm, including fingertips of flexed fingers and thumbs (both flexed and stretched), as shown in figure 1.b.

The first advantage of our wrist detector over a whole hand detector is the time needed for training. Using the same amount of training images, it takes less than 24 hours on a PIV 2.8 GHz to train an 18 stages classifier, while the hand classifier needs more than a week to train the same number of stages. Mainly because the variability of the lower half of a hand is much lower than that of a whole hand, so the classifier is able to find similarities among samples much faster.

5.2 Classifier Performance

Besides the lower training time, the wrist detector also reduces three times the false detection rate given by the hand detector. The high amount of gestures, background, people and light conditions present in sample images lead to an unreliable classifier. Using the same positive sample images, but taking into account only the lower part of them, simplifies the problem and therefore reports a false detection rate reduction.

The wrist detector, without the aid of the tracking module, was applied on the test video set. An average detection rate, in relation to the total number of frames, of 0.88 was achieved. This rate is not higher because the training set was originally created having a hand detector in mind, so it is not optimal for the training of wrists.

From the amount of frames where wrists were detected, we measured a 0.97 correct detection. Figure 3 illustrates different results using the wrist detector approach for isolating hand patterns with diverse people, background and light conditions.

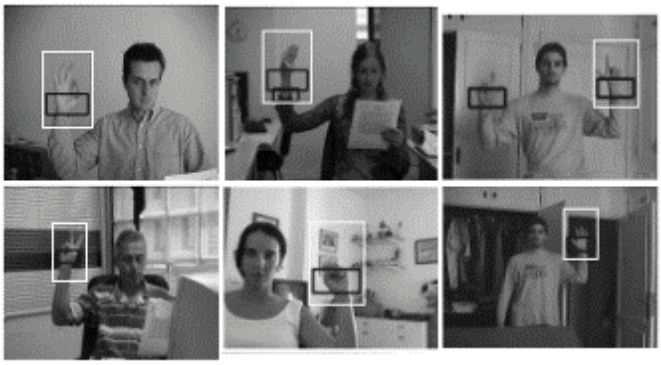

Fig. 3. Hand pose detection results showing wrists detections (dark rectangle) and complete hands (white rectangle).

5.3 Tracking Influence

A second set of tests were performed combining the wrist detector and the pattern tracker, as described in section 4. Even thought the false detection rate raises from 0.03 to 0.06, the amount of frames with detection also raises, from 0.88 to 0.99. There is an absolute increase of true positives of 7%. Figure 4 shows individual results in each video using both techniques. Figure 5 illustrates some frames from a video stream where the tracker follows a hand, which is both moving and changing gestures. The pattern size used for the tracking process is established to 24x24 pixels.

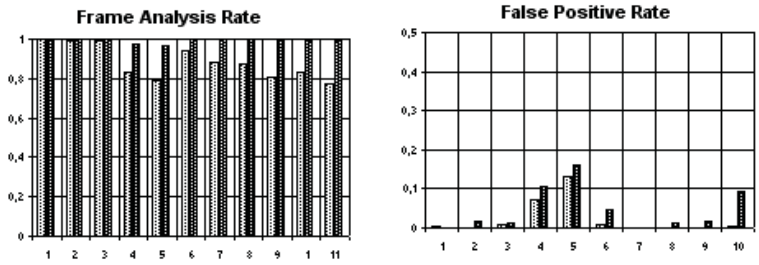

Fig. 4. Frame Analysis and False Positive Rate. Results of the classifier used alone in lighter bars, while darker bars show results for the classifier and tracker cycle.

Fig. 5. Four frames (210, 220, 230, and 240) from a video sequence, where the centre of the tracked pattern is represented by a cross. The rectangle corresponds to the whole hand area computed through the use of hand-wrist proportions from the last wrist detected using the classifier.

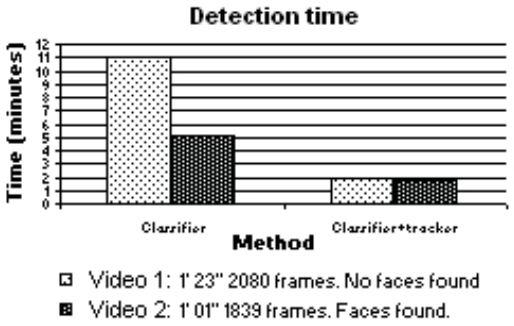

Fig. 6. Time measured for computing the classifier method and the time measured for computing the combination of classifier method and the tracker module for two different videos. Face location is not significant in relation to speed performance, when combining the classifier and the tracker module.

The average processing rate using the hand gesture detection system proposed in section 4 is 16 fps, while the average processing rate using only the classifier method reaches a maximum of 5 fps (2 fps when faces are not found). Figure 6 illustrates the measured time for two different videos using only the classifier method and using the classifier method plus the tracker module. Integrating a tracking module with a classifier based on wrists increases the speed achieved in previous works in relation to real-time performance [7, 8, 9] and also the number of different gestures detected. From these results, we have observed that the influence of face detection and the consequent search space reduction is significant when the classifier is used without the aid of the tracker. If the classifier method and the tracker module are combined, detecting faces is only significant in order to reduce false positives.

6 Conclusions and Future Work

We have developed a fast and robust hand pose detector that integrates temporal coherence information and a wrist detector using a continuously operational system.

We have tested our approach in different experiments which cover diverse people, backgrounds and light conditions. Two major conclusions have been obtained from the experiments: (i) the classifier method based on wrists reduces the false detection rate and the training stage in comparison to a whole hand detector and (ii) combining temporal coherence information and a classifier method based on wrist reduces greatly the hand pose detection time in respect to previous works based on classifier methods [7, 8, 9]. Moreover, the number of gestures detected is also increased.

Future research is focused on an improvement of the training set, estimating transition states of the hand gestures over the time with the purpose of only interpreting a new gesture, when it has taken place.

References

1. Paul Viola and Michael J. Jones.: Rapid object detection using a boosted cascade of simple features. IEEE Computer Vision and Pattern Recognition, Volume 1, pp. 511-518, December 2001.

2. M. Isard and A. Blake: Condensation - Conditional density propagation for visual tracking. International. Journal of Computer Vision, 29(1):5-28, 1998.
3. L. Brethes, P. Menezes, L.Lerasle and J. Hayet: Face tracking hand gesture recognition for human-robot interaction. IEEE International Conference on Robotics and Automation. New Orleans, April 26 – May 1, 2004.
4. J. Triesch and C. von der Malsburg: A System for Person-Independent Hand Posture Recognition against complex backgrounds. IEEE Trans. on Pattern Analysis and Machine Intelligence, 23(12):1449-1453, December 2001.
5. J.M. Rehg and T. Kanade: Visual tracking of high DOF articulated structures: an application to human hand tracking. In 3rd Proc. European Conference on Computer Vision, Volume II, 35-46.
6. M. Spengler, B. Schiele: Towards robust multi-cue integration for visual tracking. Machine Vision and Applications, Springer-Verlag 14: 50-58, 2003.
7. B. Stenger, A. Thayananthan, P. Torr and R. Cipolla: Hand Pose Estimation using Hierarchical Detection. In ECCV Workshop on HCI 2004, Lecture Notes in Computer Science, Springer-Verlag, vol. 3058, pp. 102-112.
8. Mathias Kösch and Matthew Turk: Robust hand detection. In 6th IEEE International Conference on Automatic Face and Gesture Recognition. May 17-19, 2004, Korea.
9. J.Barreto, P. Menezes and J. Dias: Human-Robot Interation based on Haar-like Features and Eigenfaces. IEEE International Conference on Robotics and Automation. New Orleans, April 26 – May 1, 2004.
10. Mathias Kösch and Matthew Turk: Analysis of Rotational Robustness of Hand Detection with a Viola-Jones Detector. In IAPR International Conference of Pattern Recognition, 2004.
11. H. Kruppa, M. Catrillón and B. Schiele: Fast and Robust Face Finding via Local Context. In Joint IEEE International Workshop on VS_PETS, Nice, France 2003.
12. S. Rogers Peck: Atlas of Human Anatomy for the Artist. Oxford University Press, Inc, USA, 1982. ISBN: 01950309858.
13. C. Guerra Artal: Contributions to visual precategoric tracking. Phd thesis, University of Las Palmas G.C, 2002.
14. S. Edelman.: Representation and Recognition in Vision, The MIT Press, 1999.
15. J. Triesch. Hand Posture Database I, II. http://www.idiap.ch/~marcel/Databases