Hand Pose Detection for Vision-based Gesture Interfaces

Luis Antón-Canalís
Institute of Intelligent Systems and Numerical Applications
Campus Universitario de Tafira
35017 Gran Canaria, Spain
lanton@iusiani.ulpgc.es

Elena Sánchez-Nielsen
Department of E.I.O. and Computation
University of La Laguna
38271 S/C de Tenerife, Spain
enielsen@ull.es

M. Castrillón-Santana
Institute of Intelligent Systems and Numerical Applications
Campus Universitario de Tafira
35017 Gran Canaria, Spain
mcastrillon@iusiani.ulpgc.es

Abstract

Vision-based applications designed for human-machine interaction require fast and accurate hand detection. However, previous works on this field assume different constraints, like a limitation in the number of detected gestures, because hands are highly complex objects to locate. This paper presents an approach which changes the detection target without limiting the number of detected gestures. Using a cascade classifier we detect hands based on their wrists. With this approach, we introduce two main contributions: (1) a reliable segmentation, independently of the gesture being made and (2) a training phase faster than previous cascade classifier based methods. The paper includes experimental evaluations with different video streams that illustrate the efficiency and suitability for perceptual interfaces.

1. Introduction

Machine Vision is increasingly introducing new commercial applications, offering a progressively broader variety of services. Among these applications, there are some that may become a new generation of human-machine interfaces. Commercial applications like Sony Eye Toy [9] demonstrate the strength of this concept. Hand gesture detection and posterior classification become decisive in this kind of applications in order to support visual aspects of interaction [5]. In this paper, we introduce a variance in the hand location problem which leads to a fast and accurate hand detection, suitable for human-machine interfaces.

1.1. Previous Work

Most hand segmentation approaches have been developed based on previous works in face detection. Human skin color modeling is one of the most revisited methods [4] due to its simplicity, but it must be aided by structural features like edges or motion. For example, skin color and elliptic shapes have been used to detect faces in [3], while a watershed algorithm on the skin-like coloured pixels in collaboration with a condensation algorithm was applied in [1] for segmenting a specific set of hand gestures.

Several classification methods for view-independent hand posture recognition were investigated in [14]. In [12] an 86.2% accuracy rate was achieved using elastic graph matching techniques with different feature types.

Cascade classifiers are currently considered the fastest and most accurate pattern detection method for faces in monocular grey-level images [8]. Recent works show their successful application in a wide range of conditions for face detection [2]. 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. It is possible, however, to train a different classifier for each recognizable gesture [10], or a single classifier for a limited set of hands [6], but that leads to the detection of a low number of gestures. Most previous works usually follow that premise, so only a certain amount of gestures can be recognized by the classifier.

A detailed analysis of rotational bounds for training and detection of hand appearances [7] reveals that only 15º rotations can be efficiently detected with a Viola-Jones detector. Most importantly, the training data must contain rotated hand sample images within these limits.

Since training a detector for every possible hand gesture is prohibitively expensive and training a single classifier reduces the number of gestures detected, we propose and evaluate a cascade classifier which is trained to detect wrists. 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. Aditionally, there is no limitation in the number of gestures being detected, as long as wrists are

not concealed. Moreover, the training time for the cascade classifier is greatly reduced.

This paper is organized as follows: Section 2 describes the approach, from the training of the classifier to the final hand extraction. Experimental results are presented and discussed in section 3. Finally, section 4 provides conclusions and our future trends.

2. Methodology

The main issues related to the training set, number and size of sample images and their importance in the detection stage are addressed.

Once a classifier has been trained, it is applied to a video stream. Our approach is focused on two main steps: First, we use a Viola-Jones detector [8] to find frontal faces, in order to reduce the search space, and afterwards we apply our wrist detector approach. Finally, a patch image containing the hand is extracted.

2.1. Training Stage

The Viola-Jones cascade classifier method combines increasingly more complex classifiers in a cascade, allowing background regions of the image to be quickly discarded while spending more computation on promising object-like regions [8].

The underlying problem with hand shapes in the training stage is that they are not self-containing objects. Therefore, large portions of background might be seen between fingers and around the palm, as shown in Figure 1a and 1.c. It is much easier to train objects that can be described using only internal features, like a book cover. However, in the case of hands, some background is usually involved in the structure of the target object.

Three main issues need to be addressed in order to maximize the efficiency of the classifier. The first one is shape variation of the trained object. In our case, hands are collected from different people performing a wide range of gestures. The second one is that sample images must show dissimilar light conditions: nature, direction, intensity. Also, light color should change within samples. The last aspect is related to diversity of backgrounds, so the classifier can infer which part of the sample images belong to the object of interest and which part must be ignored.

A large amount of samples is necessary in order to consider every possible background and light condition, and of course every possible gesture, which results in a strong variation among samples. This situation leads to huge training sets and higher computational cost of the training stage. Moreover, no classifier could be granted under those conditions.

Instead of using whole complete hands, we simplify the problem using wrists. With this purpose, we employ just the lower half part of pictures of a training set that contained 4213 different hands sized 20x20 pixels under different conditions, previously built for hand detection with our own samples and others selected from available datasets [12]. This way, the trainer only takes into account images that 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 1b. Variation among samples is much lower, and the portion of background included in each sample is smaller. Therefore, the complexity is reduced, speeding up the training step.

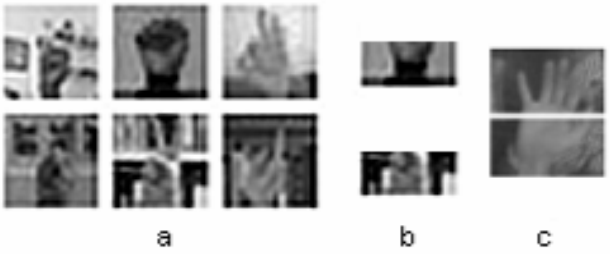

Figure 1. Positive sample images: a) whole hand, b) lower part of hands, used by our wrist classifier, c) detail of a sample image, divided in two sections.

2.2. Detecting Appearance of Hands

The search area can be reduced if people are first located. For that purpose, faces are detected using a cascade classifier as described in [2]. According to human body regular proportions [11], we define the area image of user's body belonging to each face occurrence. Human arm length is usually around three times the length of a head, so a boundary of the distance that a hand can reach knowing the location of the 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 during the detection process after a convenient analysis of each head's position and size. If no faces are detected, the search space problem is aimed to the original image’s dimension.

Then, a second cascade classifier is applied and wrists are detected. When a wrist is located, it will be contained in a rectangle which width coincides with the hand width in the image. As a simplification, just vertical gestures are used, so the rest of the hand will be directly above a detected wrist.

The next step of the hand detection process involves growing the detection area to enclose the whole hand. Taking again into account human proportions, the height of a hand is usually between 2.8 and 3.2 times the width of its wrist, so the region of interest can be resized to include the hand, and some extra space around it. A

scheme of the hand detection process is illustrated in Figure 2.

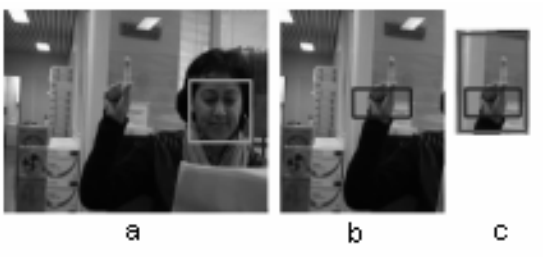

Figure 2. Hand location process: a) face detection, b) wrist detection in the reduced search area, and c) region growing according to general wrist-hand properties.

3. Experimental Results

The first advantage of our wrist detector over the whole hand detector is the time needed for training. Using the same amount of training images (5653 negative samples and 4130 20x20 positive samples) it takes less than 24 hours on a PIV 2.8Hhz 1Gb RAM to train a 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. There is also a reduction in the size of positive sample images: from 20x20 to 20x10 pixels as seen in Figure 1a and 1b. Furthermore, we trained both classifiers and we found that our system reduces three times the false detection rate.

The feature based wrist detector was applied on 12 different videos with an average of 1500 frames each one, 25 frames per second. These videos contain 12 different people with different backgrounds and light conditions, making more than 20 different vertical hand gestures. Figure 3 illustrates different results using the wrist detector approach with diverse people, background and light conditions.

Figure 4 shows detection rates achieved in each video, which represents the rate of frames where the classifier locates a wrist in relation to the total amount of video frames. An average rate of 0.88 was achieved. The main reason for not reaching a better performance is that our training set was not created specifically for wrist detection, so wrists are not homogeneously distributed among samples. That is, some of them show a larger portion of the arm and other samples show too few palm, reducing the performance of the classifier.

We measured the amount of detected wrists in the total amount of analyzed frames in each video, and from those detections we calculated false positive error rates. Results are shown in Figure 5.

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

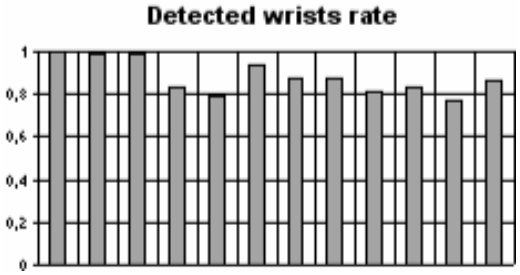

Figure 4. Wrists detection rate for each analyzed video.

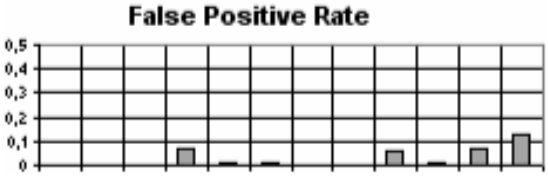

Figure 5. False positives rate for each analyzed video.

The average false positive rate computed is 0.03. Analysis of false positive results reveals that the use of wrists instead of hands can sometimes confuse the classifier, because it does not know anything about what is lying above the detected wrist. This situation, added to the low resolution of the training set, leads to a *stump effect*: there may be no hand above appearance of hand as seen in Figure 6a. However, a detection is obtained in this situation because it coincides with the structure of training samples like those shown in Figure 1b and Figure 1c.

Some other samples of false positives are illustrated in Figure 6b and Figure 6c.

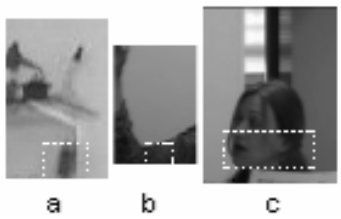

Figure 6. False positive samples: a) vase neck under a crystal shelf, b) L-bended arm, c) face.

False positive results of figure 6 show that further analysis of hand pose detections is required in certain situations. This analysis should be focused on rejecting detections if they do not lead to hands. The area above detections could be analyzed in order to check continuity: if there is a hand above a located wrist, there should not be significant difference between the detection area and the space right above it, as it happens in Figure 6a and 6b (there is nothing similar to the white rectangle right above it). Faces like figure 6c only become a false positive if they are not located with an appropriate face detector.

4. Conclusions and Future Work

We propose a cascade classifier trained to detect hands based on wrists as a simplification of the problem of finding hands in still images or video streams for vision based gesture interfaces. Our approach reduces the search to those areas surrounding detected faces, if they are found, and encloses hand poses resizing the area of a wrist detected by the classifier.

We have tested our approach in different experiments which cover diverse people, backgrounds and light conditions. The number of different detected gestures is larger than previous classifier based methods because wrists are highly independent from the gesture being made. However, in certain situations, further analysis of hand pose detections is required in order to remove false positives.

Given an appropriate training set, this kind of classifiers can be used for detecting wrists. In our case, the training set was not created specifically for our purposes. In spite of the positive samples used, promising results were found.

Future research will be focused on an improvement of the training set, collecting images of hands specifically for training the wrists detector and increasing the resolution of positive sample images, so more details could be used in the training stage, thus reducing afterwards the amount of false positives. Also, the possibility of using synthetic images in order to avoid the arduous task of gathering new hands, users, lights and backgrounds is being researched.

Acknowledgments

This work was supported in part 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.

References

[1] Brethes, L., Menezes, P., Lerasle L. and Hayet J. “Face tracking and hand gesture recognition for human robot-ineteraction”, In *IEEE 2004 International Conference on Robotics and Automation*. New Orleans, April 26- May 1, 2004.

[2] H. Kruppa, M. Castrillón Santana and B. Schiele. “Fast and Robust Face Finding via Local Contenxt”. In *Joint IEEE International Workshop on VS_PETS*, Nice, France, 2003.

[3] Hsu, R.L., Abdel-Mottaleb M. and Jain, Anil K. “Face Detection in Color Images”. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(5):696-706, May 2002.

[4] Jones, M. and Rehg, J.M. “Statistical Color Models with Application to Skin Detection”. In *International Journal of Computer Vision*, 46(1):81-96, January 2002.

[5] Matthew Turk. “Computer Vision in the Interface”. In *Communications of the ACM*, 47(1):61-67, January 2004.

[6] Mathias Kösch and Matthew Turk. “Robust Hand Detection”. In 6th *IEEE International Conference on Automatic Face and Gesture Recognition*. May 17-19, 2004, Seoul, Korea.

[7] 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.

[8] Paul Viola and Michael J. Jones. “Rapid object detection using a boosted cascade of simple features”. In *IEEE Conference on Computer Vision and Pattern Recognition*, Volume 1, pp.511-518, December 2001.

[9] ©2003-2004 Sony Computer Entertainment Europe, Developed by London Studios.

[10] Stenger B., Thayananthan A., Torr P. and Cipolla R. “Hand Pose Estimation Using Hierarchical Detection”. In *ECCV Workshop on HCI 2004*, Lecture Notes in Computer Science, Springer-Verlag, vol. 3058, pp. 102-112.

[11] S. Rogers Peck. “Atlas of Human Anatomy for the Artist”. *Oxford University Press* Inc, USA, 1982. ISBN:01950309858.

[12] Triesch, J. Hand Posture Database I, II. http://www.idiap.ch/~marcel/Databases/gestures/main.html

[13] Triesch, J. and C. von der Malsburg. “A System for Person-Independent Hand Posture Recognition against Complex Backgrounds”. In *IEEE Trans.* on *Pattern Analysis and Machine Intelligence*, 23(12):1449-1453, December 2001.

[14] Wu, Y. and Huang T.S. “View-independent Recognition of Hand Postures”. In Proc. *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 84-94, 2000.