Combining Detectors for Human Hand Detection

Antonio Hernández*, Petia Radeva*+ and Sergio Escalera*+

* *Computer Vision Center, Edifici O, Campus UAB, 08193 Bellaterra, Barcelona, Spain*
E-mail:ahernandez@cvc.uab.es

+ *Departament de Matemàtica Aplicada i Anlisi, Universitat de Barcelona*
Gran Via de les Corts Catalanes 585, 08007 Barcelona, Spain
E-mail:petia@maia.ub.es, sergio@maia.ub.es

Abstract

In this paper we present a hand detector system used for the Person Layout competition at PASCAL VOC Challenge 2010, in conjunction with an external head detector module. HOG features are extracted from the training set, and a clustering is performed in order to categorize the different poses that hands can have. A cascade of classifiers is trained for each one of the discovered hand subclasses, and a sliding-window approach is used for the detection process, followed by a filtering step. Results are shown on the corresponding Person Layout dataset from PASCAL VOC 2010.

Keywords: Object recognition, Person layout, Hand detection, Cascade.

1 Introduction

The person layout problem, as defined in the PASCAL VOC Challenge, consists in the recognition of three body parts. Given an image of a person, we have to predict the bounding boxes of his/her head, hands and feet. In this work we focus only in the hand detection problem, which is really challenging for many reasons. The most influential fact is the high intra-class variability of hands, since they can appear in a wide range of poses and orientations. Moreover, the appearance of hands -texture and color- is also highly variant due to the different skin colors of people and the possible clothes they can wear, like gloves for example. Furthermore, the appearance of different people in the same image introduces the problem of data association, since we do not have only to detect hands, but also to associate them to the correct person.

Previous work in object detection have achieved high performance rates as in the case of face detection [1]. In this case, a cascade of simple classifiers is trained using Haar-like features, and the detection is performed using a sliding-window approach. Although the results they obtain are good, the main limitation of their method is the intolerance to rotations, as analysed in [2]. The same cascade of classifiers based methodology has also been applied to hand detection [3, 4].

In this paper we present our hand detection system, which is also based on a cascade of classifiers approach, but using HOG [5] instead of Haar-like features. HOG features encode the occurrences of gradient orientations in portions of an image, so it is expected from them to describe characteristic edges regarding the shape of the hands, as the straight edges between fingers, for example, in the

Figure 1: Clustering of positive patches in 4 different classes

case of an open hand. As we will see in further sections, we also use information returned by an external head detection module.

The rest of the paper is organised as follows: Section 2 presents our hand detection system, section 3 explains the validation of the system, and finally section 4 concludes the paper.

2 Methodology

This section is divided in two main subsections, the first one detailing the training process of the system , and the second explaining the detection step itself over an input image.

2.1 Training

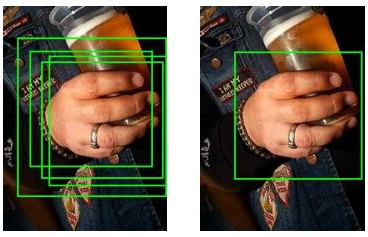
In order to build our training set we extract positive and negative patches from a labeled dataset of images. In the case of the negative patches, we randomly sample the images of the dataset at different scales and check that they do not contain any hands. A normalization step is applied in the extraction of positive and negative patches by rotating them following the orientation of maximum gradient magnitude -as in [6]-, and resizing them to the same size.

Once our hand training set is built, we extract HOG features from the patches, which will be the input of the classifier. Nevertheless, before starting training the classifier, we apply K-means clustering to the HOG features extracted from the positive samples (Figure 1). Thanks to this clustering we can make a first categorization of different kinds of hand poses, so we can treat the classification as K semi-independent problems and reduce the intra-class variability. Using this first classification given by the clustering we can start training our K independent cascades of classifiers, each one of them composed by one SVM at each level. At the first level we train each one of the K SVMs with their corresponding positive examples, and the same set of negative examples for all of them. Then, we look for false positives (FPs) by extracting random negative patches from the image dataset and classifying them with the SMVs from the current level. If one patch is classified as positive by the SVM_k, then we have found a FP for it. This process is repeated until we reach a desired number of FPs for the K SVMs. After that, each set of FPs will be used as the negative training set for the corresponding SVM at the next level, and the positive training set remains the same as the one used at the first level. This algorithm can be repeated until we reach a desired number of levels of the cascade.

2.2 Detection

For the detection step, we scan a given input image of a person with a sliding window at different scales, apply the corresponding normalizations to the patch, and test it with the K previously trained cascades of classifiers. In this way, we will consider that a hand is detected if at least one of the K cascades gives a positive answer.

After the sliding window has scanned the whole image, we apply some filtering steps in order to refine the output of the classifiers, which not only still return a large number of FPs, but also return many true positives (TPs) for a given specific hand. The first filtering step consists in applying an agglomerative clustering to the bounding boxes returned as positive by the cascades. With this agglomerative clustering we can reduce the number of TPs as shown in Figure 2.



(a) (b)

Figure 2: TPs reduction by means of agglomerative clustering, (a) shows the detections given by the cascades and (b) the resulting detections after applying agglomerative clustering

Once we have applied agglomerative clustering, we sort the remaining bounding boxes computing the following score for each one of them:

$$Score_i = SVMScore_i + Colorscore_i \quad (1)$$

where $SVMScore_i$ is the margin returned by the SVM classifier at the last level of the cascade, and $Colorscore_i$ is computed as the intersection between the HSV histogram of the i-th patch and the HSV histogram of the detected head bounding box given by an external head detector applied over the same image in which we are looking for hands:

$$Colorscore_i = HSVhist_i \cap HSVhist_{head} \quad (2)$$

Finally, we assume that in the image of the person 2 hands will be present, and then return the 2 bounding boxes with higher score.

3 Results

For the validation of our system, we have used the following configuration:

◦ *Data*: For the training step of our system we have used the training set for the Person layout competition in the PASCAL VOC Challenge 2010 -composed by 296 images of people, and the Human limb dataset presented in [7], formed by 227 additional images. For the validation of the system, the corresponding validation set from the PASCAL dataset has been used, which consists of 206 images.

◦ *Methods*: The initial K-means clustering over the positive patches in the training step has been applied with $K = 4$,so 4 different cascades of classifiers have been trained, each one of them composed by 2 levels, and one Linear SVMs at each level. The In order to train the following level of the cascade, 1000 FPs have been searched.

◦ *Validation measurement*: For the validation of our system we have computed the mean Average Precision using the PASCAL VOC framework. This framework computes the overlapping factor O between our detected bounding boxes and the ones from the ground-truth:

$$O = \frac{B_{detected} \cap B_{GT}}{B_{detected} \cup B_{GT}} \quad (3)$$

A correct match will be considered if the overlapping is greater than a certain threshold (0.5 by default).

Table 1 shows the Mean Average Precision using the default threshold $O = 0.5$, and $O = 0.05$. The choice of the second threshold on the overlapping is because usually hands are very small regions, and it is difficult to obtain an overlapping higher than 0.5. Moreover, ground-truth annotations are not always square meanwhile our outputs are always square. Looking at these results, we can see that many of the detections lie around the actual hands labeled in the ground truth.

Many reasons could be the factors of obtaining such a low mean AP. The first reason could be

that we are assuming that there are always 2 hands in all images, a dangerous assumption that can highly damage the final results. Secondly, maybe the deepness of the cascade of classifiers is not enough to reduce the number of FPs appropiately. Another reason could be given by the data association problem earlier commented in this paper. In images where more than one person is present, maybe there are correct hand detections -they are actually hands-, but may belong to another person. Finally, part of the error could be due to an error in the external head detector that we use to learn the skin color model and rank the detections. Figure 3 shows some qualitative results.

Overlap threshold	Mean AP
0.5	0.95%
0.05	8.54%

Table 1: Mean Average Precision for different overlapping thresholds.

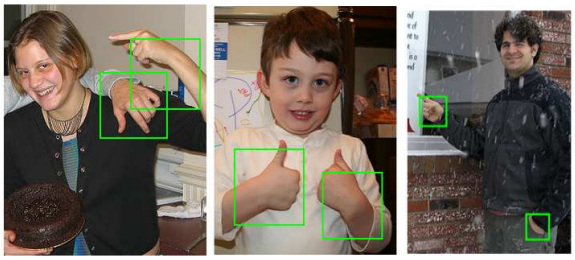


Figure 3: Qualitative results from hand detection

4 Conclusion

We have proposed a hand detection system based on HOG features. At training time, a clustering of the positive hands is performed, and a different cascade of Linear SVM classifiers is trained with the corresponding samples belonging to each cluster. At detection time, a sliding-window approach has been implemented, filtering the output of the classifiers by means of agglomerative clustering and a ranking of the detections using SVM margin and color information.

As future work, different SVM kernels as RBF could be used in order to improve the results. Moreover, at detection time, different aspect ratios could be used in order to find more accurate detections.

References

[1] Paul Viola, Michael Jones, "Robust Real-Time Face Detection", *International Journal of Computer Vision*, 57(2):137-154, 2004.

[2] Mathias Klsch, Matthew Turk, "Analysis of rotational robustness of hand detection with a viola-jones detector", *International Conferece on Pattern Recognition*, 107-110, 2004.

[3] Mathias Klsch, Matthew Turk, "Robust hand detection", *In International Conference on Automatic Face and Gesture Recognition*, Seoul, 614-619, 2004.

[4] EJ Ong, R Bowden, "A Boosted Classifier Tree for Hand Shape Detection", *Face and gesture recognition*, 2004.

[5] N. Dalal and B. Triggs, "Histogram of Oriented Gradients for Human Detection", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2:886-893, 2005.

[6] David G. Lowe, "Object Recognition from Local Scale-Invariant Features", *International Conference on Computer Vision*, 1999.

[7] A. Hernandez, M. Reyes, S. Escalera, P. Radeva, "Spatio-Temporal GrabCut human segmentation for face and pose recovery", *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, vol., no., pp.33-40, 13-18 June 2010