Feature-based hand detection in visual images

Ruud Mattheij (R.J.H.Mattheij@tilburguniversity.edu)
Tilburg center for Cognition and Communication (TiCC) Warandelaan 2, 5037 AB Tilburg
Dante Building, Room D 336, The Netherlands

Eric Postma (E.O.Postma@tilburguniversity.edu)
Tilburg center for Cognition and Communication (TiCC) Warandelaan 2, 5037 AB Tilburg
Dante Building, Room D 343, The Netherlands

Abstract

Recent developments in image processing and machine learning techniques facilitate the automatic coding of human behavior. This paper proposes an efficient and effective classification method for the automatic coding of hands in still images and in image sequences. The method combines an efficient and effective feature-extraction method with a powerful machine-learning algorithm. The evaluation of the detector on a challenging database of natural images of human hands in a large variety of poses results in a performance that is comparable to state-of-the-art detectors, while being able to perform in real-time. This leads to the conclusion that our feature-based method performs state-of-the-art hand detection and offers a promising starting point for the efficient automatic coding of gestures.

Keywords: Hand detection; automated gesture annotation; Viola-Jones detector; Haar features; random forests

1. Introduction

Advanced image processing and machine learning techniques facilitate the automatic coding of human behavior. In the domain of human gestures, the availability of low-cost visual and depth sensors such as Microsoft's Kinect device boosts the performances of computational body-part trackers (Shotton et al., 2013) that are required for the coding process. In previous work (Mattheij, Postma, van den Hurk, & Spronck, 2012), a combination of an efficient and effective feature-extraction method with a powerful machine-learning algorithm (Breiman, 2001; Criminisi, Shotton, & Konukoglu, 2012) was found to outperform the state-of-the-art pixel-based method of Shotton et al. (2011) on the automatic detection of faces in depth images. This method extracted features by calculating the difference in depth contours, similar to the well-known face detector for visual images that was proposed by Viola and Jones (Viola & Jones, 2001).

This paper explores the application of the feature-based method for the automatic detection of human hands in visual images. The main goal is to establish if the feature-based method can detect hands in natural scenes at a satisfactory level. The criterion for "satisfactory performance" is set by the best hand detector currently available. This is the detector proposed by Mittal, Zisserman, and Torr (2011). Their method proceeds in two stages. In the first stage, three detectors generate proposals for bounding boxes enclosing the hand region. These detectors take hand shape, spatial context (e.g., the region of the arm), and skin color as their inputs. In the second stage, the three proposals are integrated into a single hand detection estimate. The method is endowed with some additional tricks to optimize the hand-detection performance. Being tailored to the detection of hands inevitably limits the generalizability of the method to the detection of other body parts.

The feature-based method proposed described in this paper offers three main advantages.

1. The computation of visual features and classification are fast and can be executed in real time,

2. the classification is based on decision trees and is therefore fully transparent, and

3. the method is generic and can be applied to any body part.

Extension to the automatic detection and classification of arbitrary body parts is straightforward and allowing for automatic gesture coding. Although the current study is confined to the detection of hands in visual images, in future studies depth images will be included as to allow for reliable detection of hands and gestures in three-dimensional space.

2. The feature-based method

The feature-based method for hand detection consists of two parts: feature computation and classification. The feature computation proceeds using the Viola-Jones features (Viola & Jones, 2001), that are widely applied to face detection tasks. The classification is performed using the random forest classifier (Breiman, 2001; Criminisi et al., 2012).

Feature computation

The feature computation employs Haar features and integral image representations. Below, both are briefly outlined.

Haar features Haar features are two-dimensional filters (or masks) that respond to vertical, horizontal, or diagonal intensity transitions in images, such as oriented contours, edges, or bars. The features are based on the well-known Haar wavelets (Guf & Jiang, 1996) and are defined in terms of adjacent square regions in an image. In the context of the detection of hands, Haar features may respond to visual characteristics, such as the (partial) contours of hands or fingers. The orientation and size of the visual characteristics to which features respond depends on the nature of the features (see below).

A feature f for position P in an image I can be computed by calculating the sums S of the pixels enclosed by two square

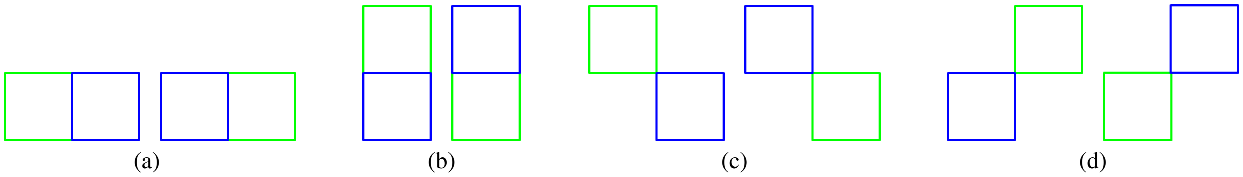


Figure 1: The feature types used in the feature-based method: (a) horizontal features, (b) vertical features, (c) diagonal, and (d) anti-diagonal region features.

areas and subtracting these sums. This results in a single feature value $f(I,P)$ that provides an indication of the direction and magnitude of an intensity transition within an image region.

In what follows, we describe the computation of the feature values in more detail. Feature values depend on

- the parameter r^2
  defining the size of the individual square regions in pixels, and
- the i feature types
  defining the relative orientation of the two constituent square regions of the feature.

The sizes of the square regions define the area over which the intensity difference is calculated. Employing larger squares for the features results in a feature value that describes the image transition over a larger area in the image. By calculating the sum of the square areas for all possible square sizes r^2 (which can be achieved very efficiently using the integral image representation, see below), scale-invariant responses are obtained.

The feature type describes the locations of the two constituent square regions in relation to each other, thereby providing an indication of the direction of the intensity transition. Figure 1 illustrates the four pairs of feature types employed in the feature-based method (i.e., $i = 8$), which allow for the detection of horizontal, vertical, diagonal and anti-diagonal intensity transitions.

In the figure, the green square represents region $S_{\mathrm{i}}(x_{\mathrm{a}}, y_{\mathrm{a}}, r^2)$ and the blue square represents region $S_{\mathrm{i}}(x_{\mathrm{b}}, y_{\mathrm{b}}, r^2)$. For all possible combinations of r^2 and i at pixel position P in the image, the feature value is computed using equation (1). The coordinates of the (upper left corners of the) two square regions P are defined by (x_a, y_a) and (x_b, y_b), respectively. The maximum feature response is selected.

$$f(I,P) = \sum_{r=1}^{r_{max}} \sum_{i=1}^{8} S_i(x_a, y_a, r^2) - S_i(x_b, y_b, r^2), \qquad (1)$$

Integral image representations The Haar features can be computed rapidly using an alternative image representation called the integral image representation (Viola & Jones, 2001). Adopting the integral image representation allows for a considerable speedup in the computation of the Haar features.

Classification

A special type of decision forest classifier (Breiman, 2001), called the random decision forest classifier (Criminisi et al., 2012) is used to perform the binary classification of the feature representations of the instances in our dataset. Randomized decision forests are fast and effective classifiers that employ an ensemble of decision trees for prediction. Each individual tree consists of binary split- and leaf nodes. Individual split nodes compare single features from the feature vector with a threshold, branching left or right depending on the outcome of the comparison. The leaf nodes of the trees contain the prediction results (HAND DETECTED or NO HAND DETECTED). The predictions of all decision trees are averaged over the ensemble of trees, yielding the final classification decision by taking a majority vote.

3. Experimental method

Below, the dataset of hand images and the evaluation procedure are described.

The hand dataset

The experiments are performed using the database of Mittal et al. (2011) that contains annotated visual images of human hands. Using these annotations, square image regions containing the hand were selected and labeled as a positive instances. Negative examples were selected by selecting non-annotated square regions that did not overlap with the regions containing a hand. As in (Mittal et al., 2011), the positive instances were standardized by rotating them until the fingers of the hand pointed upwards. The resulting dataset contains 1600 positive examples and 1600 negative examples, each with a resolution of at least 30 × 30 pixels. Figures 2a and 2b display examples of positive and negative instances, respectively. In addition, each of these figures shows two examples of features.

Evaluation of the feature-based method

The aim of the experiment is to determine to what extent the feature-based method can detect hands in natural scenes at a satisfactory level. The criterion for a satisfactory level is taken from Mittal et al. (2011). As stated in the introduction, Mittal et al. (2011) relied on three detectors. Using their hand

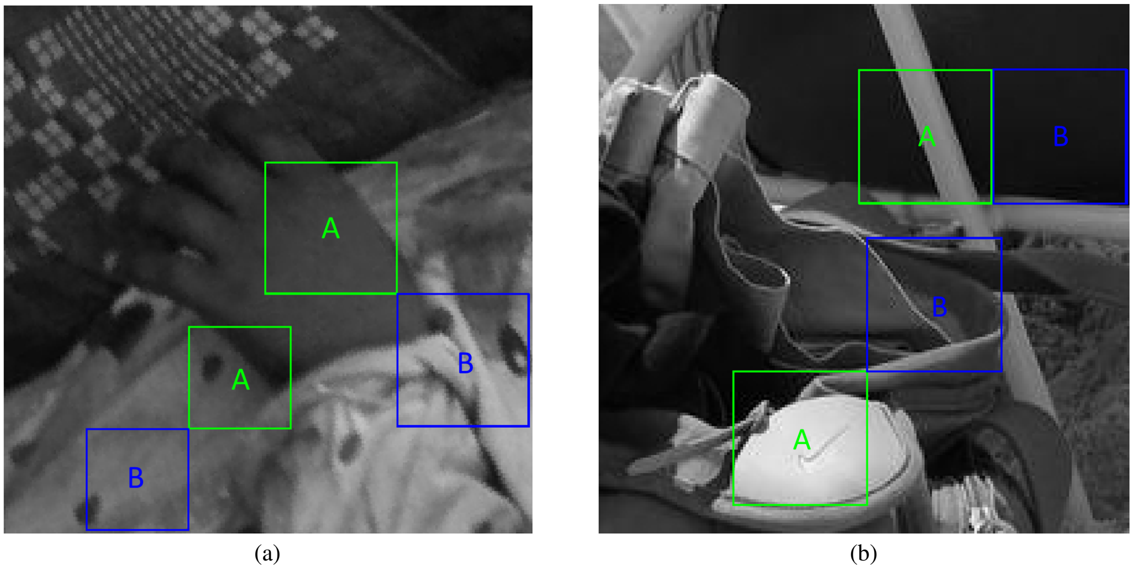


Figure 2: Two representative images of the dataset: (a) a positive example image containing a hand, and (b) a negative example image not containing a hand. Overlaid on both images are two features, each composed of two square regions labelled A and B. The feature-based method relies on the values obtained by subtracting B from A.

detector only, they achieve an average recall of 74.09% and for their combined three detectors 85.3%. The detection performance of the feature-based is called satisfactory when a recall of at least 75% is obtained on a set of hand images as used in their study. To assess the effect of the number of Haar features extracted per image, the experiment is repeated for subsets of N features, with $N \in \{1,2,\dots,50\}$.

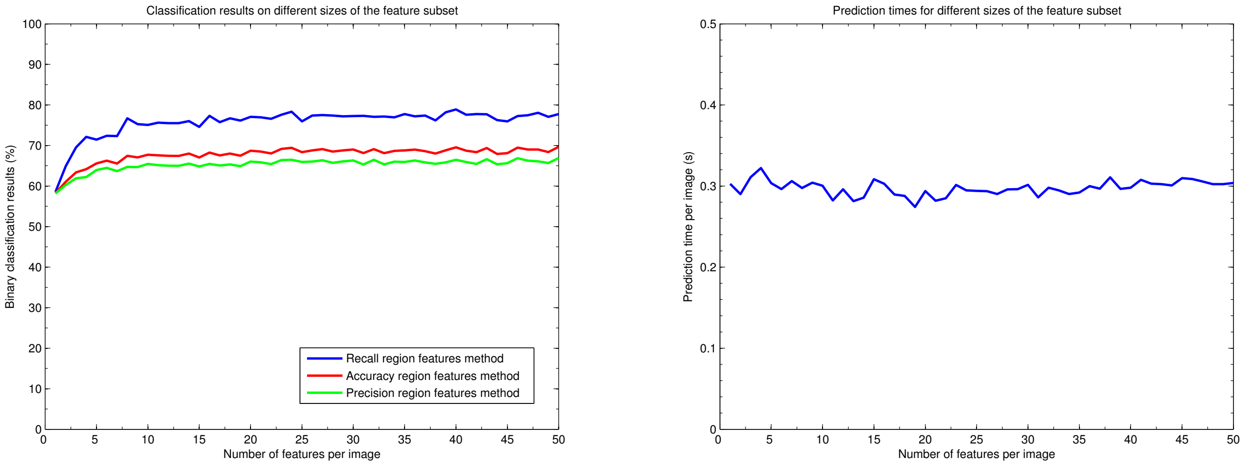
The evaluation of our detector was performed using 10-fold cross-validation. For every fold, the evaluation is performed on the same set of training and test images. Each fold consists of 2880 training examples (1440 positive and 1440 negative examples) and 320 test images (160 positive and 160 negative examples). The training examples are used to train a random decision forest consisting of 50 trees, while the test images are used to evaluate the performance of the detector. The entire training- and evaluation sequence took approximately 5 hours on a 24-core Linux computation server.

4. Results

The results of the experiment indicate that the feature-based method performs well above chance level (50%). Figures 4a and 4b show a plot of the classification results (recall, accuracy, and precision) and prediction time (expressed in seconds) as a function of the number of Haar features extracted per image. If at least five features are extracted, the performances are almost at the level achieved with more than five features. The optimal performance is obtained using 40 features per image. Figure 3 shows a confusion matrix containing the classification performance averaged over all test folds for this value of the number of extracted features. Expressed in percentages, the following performances are obtained: accuracy 69.5% (σ = 2.87%), recall 78.9% (σ = 4.83%), and precision 66.6% (σ = 2.90%). The corresponding average prediction time per image is 0.298 seconds (σ = 0.006 seconds). This indicates that our detector is able to detect the majority of the positive examples (resulting in a high recall), at the cost of the percentage of correct positive predictions (resulting in a lower precision), while able to classify images in real time. Moreover, the performance of the feature-based method is on a par with the state-of-the-art hand shape detector and may therefore be called satisfactory.

		Predicted value	
		positive	**negative**
Actual value	**positive**	126.2 (7.73)	33.8 (7.73)
	negative	63.7 (9.40)	96.3 (9.37)

Figure 3: Confusion matrix for the average classification results for a subset of 40 features per image. The values represent the average number of classifications over over all folds; the standard deviation is shown between brackets.



(a)

(b)

Figure 4: (a) Average detection performance (accuracy, recall and precision, expressed in percentages), and (b) the average prediction time (expressed in seconds) the region-features detector on subsets of various sizes.

5. Discussion and conclusion

The results obtained on the hand-detection task suggest that effective and efficient automatic coding of hand gestures is feasible at real-time speed using the feature-based method.

The training and evaluation sequence was performed on our database with visual images of human hands. The images in the dataset were standardized by rotating them until the fingers in each image pointed upwards. This differs from the approach of Mittal et al. (2011), who solely rotated the images of the training set and evaluated their method on the test set using different image angles from which the detector selected the best detection. We intend to include a sliding window technique that enables our feature-based method to search through an image using multiple image rotations.

Improving the detection performance of the feature-based method by increasing the number of feature types and by incorporating (skin) color, is likely to enhance the performance. Including the third dimension using depth images acquired by the Kinect is straightforward (see, e.g., (Mattheij et al., 2012)) and is expected to result in a reliable detection and tracking of hands. Given image sequences (videos) of human gestures, the feature-based method can be readily used for the automatic coding of hand gestures. The generic nature of the feature-extraction method and the robust machine learning algorithm enables extension to other body parts.

It is concluded that the feature-based method performs fast and state-of-the-art hand detection and offers a promising starting point for versatile and efficient automatic coding of gestures.

Acknowledgements

We thank the anonymous reviewers of this paper for their insights and contributions to this paper.

References

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Criminisi, A., Shotton, J., & Konukoglu, E. (2012, February). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, *7*(2-3), 81–227.

Guf, J., & Jiang, W. (1996). The haar wavelets operational matrix of integration. *International Journal of Systems Science*, *27*(7), 623-628.

Mattheij, R., Postma, E. O., van den Hurk, Y., & Spronck, P. (2012). Depth-based detection using haar-like features. In *Proceedings of the 24th benelux conference on artificial intelligence.*

Mittal, A., Zisserman, A., & Torr, P. (2011). Hand detection using multiple proposals. In *Proceedings of british machine vision conference.*

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., … Blake, A. (2011). Real-time human pose recognition in parts from single depth images. *Computer Vision and Pattern Recognition*, *2*, 3.

Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., … Blake, A. (2013). Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *99*(PrePrints), 1.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition (CVPR)*, *1*, 511-518.