Depth-based detection using Haar-like features

Ruud Mattheij [a] Eric Postma [a] Yannick van den Hurk [a] Pieter Spronck [a]

[a] *Tilburg center for Cognition and Communication (TiCC), Tilburg University, P.O. Box 90153, 5000 LE, Tilburg, The Netherlands, R.J.H.Mattheij@uvt.nl*

Abstract

The automatic detection of objects has gained considerable attention over the last few years. Most object-detection approaches rely on visual features that are sensitive to identity-irrelevant variations, such as changes in illumination. Being less sensitive to such variations, depth features may improve detection accuracy. Depth features can be extracted from depth images generated by commercially available depth sensors, such as Microsoft's Kinect device. This paper describes a method for robust and accurate face detection by employing Haar-like region features on the integral image representation of depth images. Our aim is to determine to what extent region-comparison features contribute to effective face detection in depth images, compared to pixel-comparison features. To this end, we present a revision of the recently proposed detector of Shotton et al. [10]. Whereas the detector of Shotton et al. relies on pair-wise pixel comparisons in depth images, our revision compares square regions in a pair-wise fashion. In a comparative evaluation of the original and revised method, we train and evaluate both detectors on our depth images of faces (DIOF) database that is compiled at our lab. The results reveal that the use of region features instead of pixel pairs indeed improves face detection accuracy in depth images. We conclude that employing region features contributes significantly to effective face detection. Future work will address to what extent our results generalize to the detection of body parts and objects in general.

1 Introduction

During the last few years, the automatic detection of objects, such as human body parts, from digital video and image sources has gained considerable attention within the field of image analysis and understanding [7, 12]. Many approaches towards object detection focus on feature-based detection [4, 6]. A well-known example of feature-based detection is the use of Haar-like rectangle features in the state-of-the-art face detector proposed by Viola and Jones [11]. In the Viola-Jones face detector, the rectangle features enable efficient and fast face detection. Despite this success, the detector is sensitive to changes in illumination [14, 15]. Using additional or alternative cues such as depth information [9] may help to overcome such sensitivities by providing illumination-invariant cues, which can potentially make object detectors more robust [2]. Employing depth cues is made possible by commercially available depth sensors like the Microsoft Kinect device.

Shotton et al. [10] proposed a depth-based detector that is able to quickly and accurately classify body joints and parts from single depth images. Their method employs depth-comparison features defined as pixel pairs in depth images. The use of pixel-based features makes their method computationally efficient which allows for real-time operation. The computational efficiency comes at the cost of noise sensitivity. Averaging over larger regions of the depth image reduces the noise and may lead to an improved accuracy.

In this study, we use depth cues for robust and accurate face detection in depth images. Inspired by the work of Viola and Jones [11], we propose a revision of the recently proposed detector of Shotton et al. [10]. Whereas the detector of Shotton et al. relies on pair-wise pixel comparisons in depth images, our revision employs Haar-like features by comparing square regions in a pair-wise fashion. Although for visual images, the use of region features introduces illumination sensitivity, for depth images they may improve upon the noise-sensitivity of pixel pairs. The aim of our study is to determine to what extent region features contribute to effective face detection in depth images. To achieve this aim, we perform a comparative evaluation to determine whether our region-comparison detector yields an improvement with respect to the original pixel-comparison detector.

1.1 Related work

Our region-comparison detector is related to two recent methods for object detection in depth images. The first related method was proposed by Xia et al. [13]. Their method detects head shapes by means of a generic model of the 2D contour and the 3D depth map of the head. The generic model is detected by means of fast convolution. Our revised method differs from Xia et al.'s method in the use of local features, instead of a global shape model. The second related method is due to Plagemann et al. [8] who proposed a method to detect and identify body parts in depth images. Their method identifies points of interest that are based on the differences in geodesic distances, which coincide with salient points of the body. In their method, the shape of the surface meshes is defined by points with similar geodesic distances. A commonality between their method and our detector is that local depth information is used. The main difference is that we apply a series of pre-defined feature types instead of attempting to identify points of interest.

1.2 Outline

The outline of the remainder of the paper is as follows. Section 2 reviews the pixel-comparison detector of Shotton et al. [10] and presents our region-comparison detector. Section 3 describes the experimental methodology used to determine the accuracy of our detector and presents the results of the evaluation. Section 4 discusses results and we conclude on our findings in Section 5.

2 Pixel comparison versus region comparison

In this section, we describe the detector proposed by Shotton et al., henceforth referred to as the pixel-comparison detector (subsection 3.1), and present our revision, the region-comparison detector (3.2). Finally, we outline the randomized decision forest classifier that was trained using the region features (3.3).

2.1 The pixel-comparison detector

The pixel-comparison detector of Shotton et al. [10] employs simple and computationally efficient depth-comparison features to identify different skeletal joints and body parts. Figure 1a shows an overview of the detector. To calculate the features, a subset of random pixel positions is selected from each depth image. (The subset is different for each depth image.) For each position P from this subset, the feature value is computed by comparing the depth value at two offset locations Q and R. The offset locations are defined by the radius and angle with respect to P. The radius is defined to be inversely proportional to the depth value of P. A small depth value results in a larger radius for offset positions P and Q, and vice versa. In this way, a scale-invariant measure of depth is obtained. In our implementation, the angles are defined to be multitudes of 30^o and selected through exhaustive search as to obtain those angles that give the largest difference in the depth values of Q and R. Although an effective approach for feature selection, this might influence the prediction time. Figure 2 shows two example features. The use of pixel pairs as basic features makes this method fast and computationally very efficient at the cost of errors introduced by the use of individual pixel values which may be noisy.

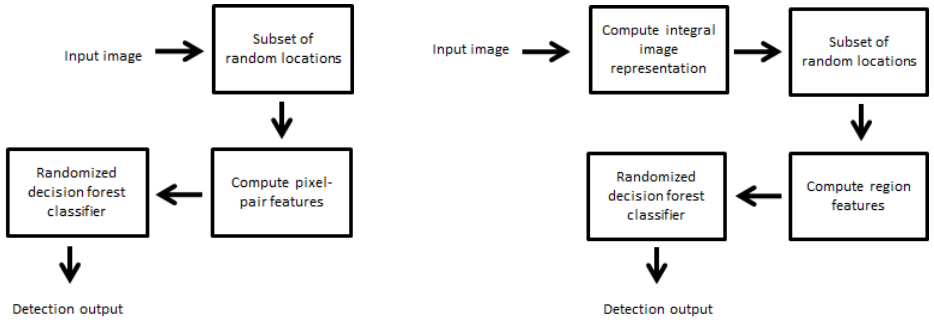
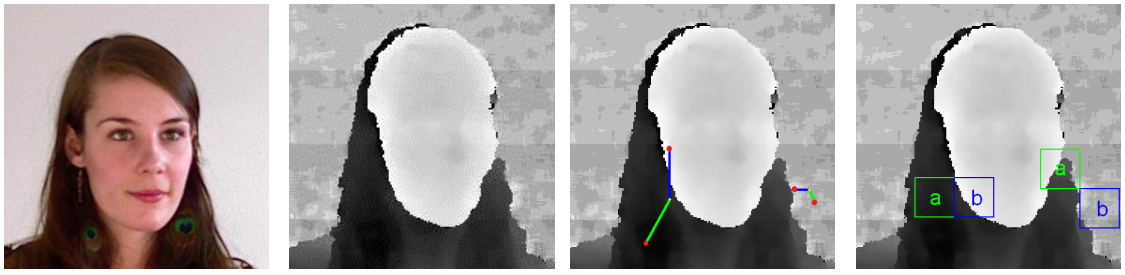


Figure 1: An overview of (a) the pixel-comparison detector that employs pixel pairs as depth features, and (b) our region-comparison detector that employs square regions in a pair-wise fashion, i.e., region features.

For each depth image, the pixel-comparison detector yields a vector with depth features that provide a probabilistic cue about the part of the body sampled. The feature vectors provide the inputs to a randomized decision forest [1] for classification.



(a) (b) (c) (d)

Figure 2: (a) Example of a visual image from our depth images of faces (DIOF) database (described in section 3.1), and (b) the corresponding depth image. (c) Illustration of two pixel-based depth-comparison features, and (d) two Haar-like region features.

2.2 The region-comparison detector

The detector is an improvement of the pixel-comparison detector proposed by Shotton et al. [10] and employs two of the contributions proposed by Viola and Jones [11]: (1) the Haar-like region features, and (2) the integral image representation. Below, we will briefly address these contributions and describe how they are employed to improve the pixel-comparison detector. Figure 1b shows an overview of our detector. As in the pixel-comparison detector, a subset of random pixel positions is selected from each depth image. Region features are two-dimensional filters (or masks) that respond to vertical, horizontal, or diagonal contours and bars in an image. They are based on the well-known Haar wavelets [5]. The features are defined in terms of square regions in an image, hence their name. A region feature f for position P (of one of the pixels in the random subset) in depth image I can be computed by calculating the sums S of the pixels enclosed by two square areas and subtracting these sums from each other. This results in a single feature value $f(I, P)$. This feature value provides an indication of the direction and magnitude of the depth transition over an area in a depth image.

In what follows, we describe the computation of the feature values in more detail. Feature values depend on (1) the parameter r^2 defining the size of the individual square regions, and (2) the configuration i defining the orientation of the constituent square regions of the feature.

The sizes of the square regions define the area over which the depth difference is calculated. Employing larger squares for the region features results in a feature value that describes the depth transition over a larger area in the depth 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), we ensure that the region features capture a large range of head sizes.

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 depth transition. Figure 3 illustrates the four pairs of feature types that we employed in the region-comparison detector, which allow for the detection of horizontal, vertical, diagonal and anti-diagonal depth 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 on position P, we compute the feature value $f(I, P)$ as proposed by [11]. We then select the highest value as the feature value for that position. The addition over feature types is performed to Although region features are less sensitive to erroneous pixel values because they average over many pixels, these features may require more computation than the pixel-pair features employed by the pixel-comparison detector. Fortunately, region features can be computed rapidly using an alternative image representation called the integral image representation [11]. Adopting the integral image representation for depth images allows for a considerable speed-up.

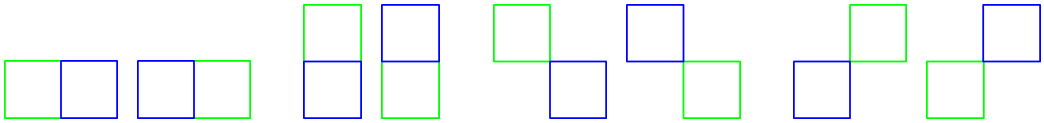


Figure 3: The feature types used in the region-comparison detector: (a) horizontal features, (b) vertical features, (c) diagonal, and (d) anti-diagonal region features.

2.3 The randomized decision forest

The random selection of N pixel positions in an image, results in an N-dimensional feature vector. A randomized decision forest classifier [3] is used to perform the binary classification on the basis of these vectors. 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 each tree contains the prediction result. The predictions of all decision trees are then averaged over the ensemble of trees, thereby giving the final classification.

3 Experiment and Results

In this section we describe the database that was used to train and evaluate our detector (3.1), and the comparative evaluation of the pixel-comparison and region-comparison detector (3.2). We train both detectors on a database with depth images of faces that we compiled in our lab.

3.1 The Depth Images Of Faces database

The database used to train and evaluate the detectors was the depth images of faces (DIOF) database that was compiled at our lab. It contained visual and depth images of human faces under various lighting conditions and distances. Figure 2 shows an example of a visual image and the corresponding depth image of a participant's face. The database assembled images of 100 participants (51 male and 49 female). We employed a Microsoft Kinect device to create visual images with a resolution of 1280×1024 pixels and depth images with a resolution of 640×480 pixels. For every participant, a series of depth images were created on five distances from the Kinect device: 0.5 meters to 2.5 meters, with steps of 0.5 meters. For each distance from the Kinect device, we created depth images under five distinct lighting conditions: dim light, environmental light, fluorescent ceiling light, intense frontal light and intense light from the left side of the participant.

A face detection algorithm was applied to annotate the location of the participants' face in the visual images. The region was selected in the corresponding depth image and labeled as a positive (containing a face) example. Likewise, negative examples were selected by labeling non-annotated regions from the depth images. Given the various distances at which the images were taken, the dimensions of the example depth images images varied between 75×75 pixels and 450×450 pixels. For the final database, we randomly selected 1000 positive and 1000 negative example depth images.

3.2 Evaluating the region-comparison detector

The aim of our experiment is to investigate to what extent region features contribute to effective face detection in depth images as compared to pixel features. We address this aim by training and evaluating the pixel-comparison detector and region-comparison detector on our DIOF database. We repeat the experiment while employing feature subsets of various sizes, starting with a subset of 1 feature per image, up to a subset of 2000 features per image, with steps of 5 features per image.

For the comparative evaluation we employ 10-fold cross-validation. For every fold, the evaluation of both detectors is performed on the same set of training and test images images. Each fold consists of 1800 training examples (900 positive and 900 negative examples) and 200 test images (100 positive and 100 negative examples). The training examples are used to train a randomized decision forest consisting of 50 trees, while the test images are used to evaluate the performance of the detectors. For both detectors,

the average detection performance over all folds is shown in table 1 (the pixel-comparison detector) and table 2 (our region-comparison detector). Both tables report the $Accuracy$ as a performance measure, defined as $(TP + TN)/(TP + FP + FP + TN)$ where TP represents the number of true positives, FP false positives, TN true negatives, and FP false positives. In addition, $Recall$ $(TP/(TP + FN))$, and $Precision$ $(TP/(TP + FP))$, are reported. All measures are expressed in percentages. The entire training- and evaluation sequence took approximately 48 hours on a 24-core Linux computation server.

Number of features	Accuracy (%)	Recall (%)	Precision (%)	Prediction time (s)
1	64.0 (2.29)	55.8 (5.14)	66.7 (2.19)	0.81 (0.04)
5	69.2 (3.74)	71.6 (4.45)	68.3 (3.69)	1.23 (0.06)
15	72.9 (2.29)	82.8 (3.29)	69.1 (2.00)	1.76 (0.09)
25	75.6 (2.71)	86.7 (3.06)	70.9 (2.57)	2.17 (0.15)
2000	77.0 (2.79)	87.7 (3.06)	72.3 (3.03)	105.3 (3.94)

Table 1: Average detection performance on subsets of various sizes, expressed in percentages (accuracy, recall, and precision) or seconds (prediction time) and, between brackets, the corresponding standard deviations for the pixel-comparison detector.

Number of features	Accuracy (%)	Recall (%)	Precision (%)	Prediction time (s)
1	72.4 (2.93)	72.2 (4.34)	72.4 (2.73)	0.84 (0.02)
5	86.0 (1.91)	87.5 (3.06)	84.9 (2.15)	0.90 (0.05)
15	87.5 (2.15)	89.0 (3.37)	86.4 (2.01)	0.92 (0.05)
25	90.3 (3.02)	88.1 (2.06)	90.3 (3.02)	0.88 (0.03)
2000	88.5 (1.80)	90.4 (2.76)	87.1 (1.81)	5.98 (0.27)

Table 2: Average detection performance on subsets of various sizes, expressed in percentages (accuracy, recall, and precision) or seconds (prediction time) and, between brackets, the corresponding standard deviations for our region-comparison detector.

The results of the experiment indicate that the region-comparison detector achieves a significantly higher detection accuracy and precision than the pixel-comparison detector. The recall is slightly better. The results also indicate that the region-comparison detector achieves a considerably shorter prediction time and therefore a higher prediction speed than the original method.

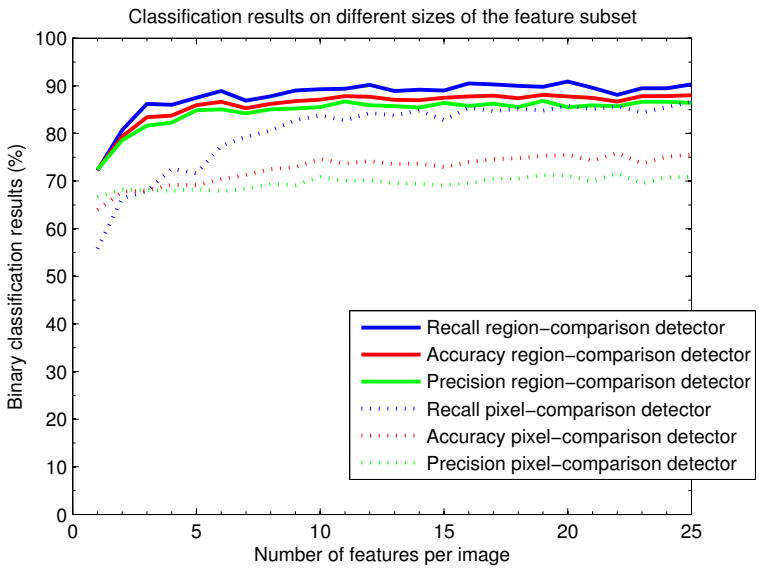


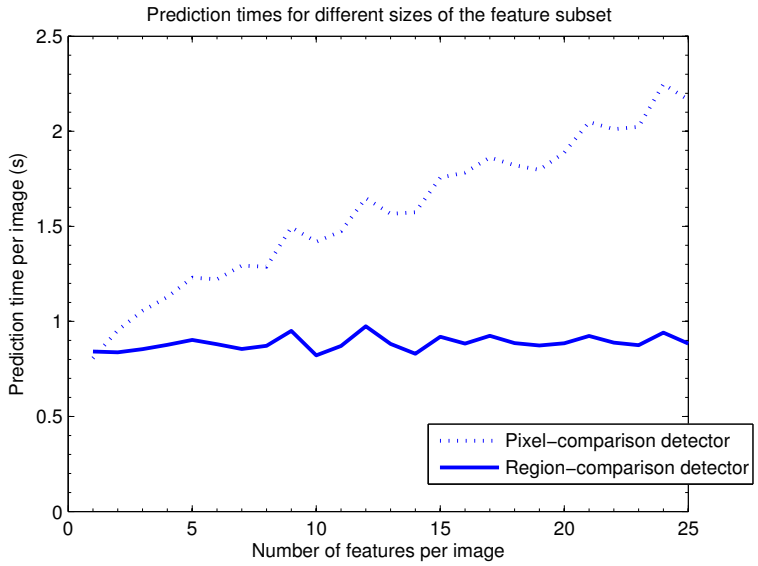
Figure 4: Average detection performance (accuracy, recall and precision, expressed in percentages) for the pixel-comparison detector (dotted line) and the region-comparison detector (solid line) on subsets of various sizes.

Our experiments suggest that employing a small feature subset may already achieve a high detection performance. Figure 5 shows the average detection performance (accuracy, recall and precision) for the pixel-comparison detector and our region-comparison detector for subsets of various sizes. Our experiments indicate that the optimal detection performance on our database is achieved while employing 25 features per image.

For a subset of 25 features per images, the region-comparison detector yields an average accuracy of 90.3% (σ = 3.02%), with an average recall and precision of 88.1% (σ = 2.06%) and 90.3% (σ = 3.02%). The pixel-comparison detector method achieves an average accuracy of 75.6% (σ = 2.71%), with an average recall and precision of 86.7% (σ = 3.06%) and 70.9% (σ = 2.57%), respectively. The prediction time for the region-comparison detector is 0.88 seconds (σ = 0.27 seconds) per image, while the prediction time for the pixel-comparison detector is 2.17 seconds (σ = 0.15 seconds) per image. For this number of features per image, the region-comparison detector yields a significantly higher detection accuracy and precision than the pixel-comparison detector while the region-comparison detector is approximately 2.5 faster than the pixel-comparison detector.

A highly similar pattern of performance results is obtained after employing a subset of 2000 features per image. The pixel-comparison detector achieves an average accuracy of 77.0% (σ = 2.79%), with an average recall and precision of 87.7% (σ = 3.06%) and 72.3% (σ = 3.03%), respectively. The region-comparison detector achieves an average accuracy of 88.5% (σ = 1.80%), with an average recall and precision of 90.4% (σ = 2.76%) and 87.1% (σ = 1.81%). The prediction time for the pixel-comparison detector is 105.3 seconds (σ = 2.76 seconds) per image. The prediction time for the region-comparison detector is 5.98 seconds (σ = 0.27 seconds) per image. For this subset, the region-comparison detector yields a significantly higher detection accuracy and precision than the pixel-comparison detector, while performing approximately 17 times faster than the pixel-comparison method.

Figure 5 shows the prediction times for both detectors on various subset sizes. The results indicate that the prediction time for the pixel-comparison detector increases significantly faster than the prediction time of the region-comparison detector.



(a)

Figure 5: Average prediction times (expressed in seconds) for the pixel-comparison detector (dotted line) and the region-comparison detector (solid line) on subsets of various sizes.

The results of the comparative evaluation of the region-comparison detector on the DIOF database show that the combination of region features and the integral image representation allows for fast and effective face detection in depth images. By employing region features, the region-comparison detector achieves a significantly higher detection accuracy and precision than the pixel-comparison detector.

4 Discussion

The results of our comparative evaluation show that the region-comparison detector achieves a high detection accuracy and precision than the pixel-comparison detector. Below, we briefly discuss three points regarding our findings: the validity of the experiment (4.1), the effect of rectangle features (4.2), and the points of improvement for our face detector (4.3).

4.1 Validity of the experiment

The results of the evaluation suggest that employing region features in combination with the integral image representation improves the detection accuracy of faces in depth images significantly, while maintaining a short prediction time. These results are achieved by training and evaluating the pixel-comparison detector and the region-comparison detector on the DIOF database for a range of feature subset sizes. As no implementation of the pixel-comparison detector was available, we developed our own implementation of the detector proposed by Shotton et al. [10].

Our choice of performing an exhaustive search for selecting appropriate pixel pairs (see section 2.1) has the advantage that it yields the largest difference in depth values, but may impose a computational cost. In future work, we should examine the effect of various realization of the selection algorithm on the performance and speed of the pixel-comparison detector.

The evaluation of the region-comparison detector was performed on the DIOF database. Although the results suggest that employing region features improves the detection accuracy and precision, our results are limited to the task of frontal face detection. Depth images of faces with other orientations or other body parts are not present as distinguishable classes in our database. Future versions of our experiment should adopt a more challenging database with depth images with a larger variety of face orientations and other body parts.

4.2 Effect of region features

The results of our evaluation show that employing region features for detection tasks in depth images results in fast and accurate face detection. The pixel-comparison detector experienced difficulties calculating feature values in noisy depth images. For example, instances in which parts of the participants' faces were too close or too far from the depth sensor of the Kinect device tended to result in erroneous pixel values. Apparently, as the region features employed in our region-comparison detector average over many pixels, these features are less sensitive to background noise in depth images, compared to the pixel-pair features employed by the pixel-comparison detector.

4.3 Points of improvement

We identify three main improvements of our detector and the database that is used for the detector's training and evaluation procedure.

First, our implementation of the detector of Shotton et al. [10] should be validated by applying it to one or more of the data sets employed in their original paper. In this way, we will be able to assess the validity of our implementation.

Second, the region-comparison detector might be improved by employing it for detection tasks that involve faces and other body parts in depth images with other orientations and poses.

Third, the region-comparison detector classifies entire images as containing a face or not a face. The region-comparison detector should be extended by enabling it to investigate points of interest in a given depth image, so it can locate specific areas in an entire depth image that might contain heads.

5 Conclusion

The aim of our study was to determine to what extent the combination of region features contribute to fast and effective face detection in depth images. To achieve this aim, we proposed the region-comparison detector that employs Haar-like region features on the integral image representation of depth images. We trained and evaluated the region-comparison detector and the pixel-comparison detector proposed by Shotton et al. on the depth images of faces (DIOF) database.

The evaluation of the region-comparison detector revealed that employing region features results in fast and accurate face detection in depth images. The results show that the detector yields a significantly higher accuracy and precision than the pixel-comparison detector. Combining region features with the integral image representation results in a short prediction time. The results indicate that a large subset of features per image does not necessarily lead to better detection results. The experiments indicate that a significantly smaller feature subset can also yield a high detection performance.

We conclude that employing region features contribute significantly to fast and effective face detection in depth images and that the region comparison-detector yields an improvement over the detection accuracy of the pixel-comparison detector.

References

[1] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[2] W. Burgin, C. Pantofaru, and W. D. Smart. Using depth information to improve face detection. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 119–120, New York, NY, USA, 2011. ACM.

[3] A. Criminisi, J. Shotton, and E. Konukoglu. 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, February 2012.

[4] G. C. H. E. De Croon, E. O. Postma, and H. J. Van Den Herik. Adaptive gaze control for object detection. *Cognitive computation*, 3(1):264–278, 2011.

[5] J.. Guf and W. Jiang. The haar wavelets operational matrix of integration. *International Journal of Systems Science*, 27(7):623–628, 1996.

[6] M.W. Lee and R. Nevatia. Body part detection for human pose estimation and tracking. In *Proceedings of the IEEE Workshop on Motion and Video Computing*, WMVC '07, pages 23–, 2007.

[7] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In Tomás Pajdla and Jiri Matas, editors, *European Conference on Computer Vision (ECCV '04)*, volume I, pages 69–82, Prague, Czech Republic, 2004. Springer-Verlag.

[8] C. Plagemann, V. Ganapathi, D. Koller, and S. Thrun. Real-time identification and localization of body parts from depth images. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 3108–3113, may 2010.

[9] N. Riche, M. Mancas, B. Gosselin, and T. Dutoit. 3d saliency for abnormal motion selection: The role of the depth map. In J. Crowley, B. Draper, and M. Thonnat, editors, *Computer Vision Systems*, volume 6962 of *Lecture Notes in Computer Science*, pages 143–152. Springer Berlin / Heidelberg, 2011.

[10] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. *CVPR*, 2:3, 2011.

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

[12] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *Int. J. Comput. Vision*, 75(2):247–266, November 2007.

[13] L. Xia, C.C. Chen, and J.K. Aggarwal. Human detection using depth information by kinect. *Workshop on Human Activity Understanding from 3D Data in Conjunction with CVPR (HAU3D)*, pages 15–22, 2011.

[14] C. Zhang and Z. Zhang. A survey of recent advances in face detection. *Learning*, (June), 2010.

[15] W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys (CSUR)*, 35:399–458, 2003.