RAPID STEREO-VISION ENHANCED FACE DETECTION

Sergey Kosov, Kristina Scherbaum, Kamil Faber, Thorsten Thormählen, Hans-Peter Seidel

Max-Planck-Institut Informatik[1] , Saarbrücken, Germany

ABSTRACT

This paper presents a real-time face detection algorithm. It improves state-of-the-art 2D object detection techniques by additionally evaluating a disparity map, which is estimated for the face region using a calibrated stereo camera setup. First, faces are detected in the 2D images with a rapid object classifier based on haar-like features. In a second step, falsely detected faces are removed by analyzing the disparity map. In the near field of the camera, a classifier is used, which evaluates the Eigenfaces of the normalized disparity map. Thereby, the transformation into Eigenspace is learned off-line using a principal component analysis approach. In the far field, a much simpler approach determines false-positives by evaluating the relationship between the size of the face in the image and its distance to the camera. This novel combination of algorithms runs in real-time and significantly reduces the number of false-positives compared to classical 2D face detection approaches.

Index Terms— Image analysis, object detection, stereo vision

1. INTRODUCTION

Face detection is often the first step in complex image processing applications, like, face recognition, visual surveillance, or human-machine interaction. This explains the high interest of the research community in this topic. In 2001, Viola and Jones [1] presented a widely used machine learning approach for visual object detection, which is capable to detect faces in images in real-time. It employs a coarse-to-fine strategy, where a classifier is trained that selects a few critical haar-like features from a large set and then combines increasingly more complex ones in a cascade. These features are applied to detect faces in the image domain whereby consecutively smaller image patches become relevant as the complexity of the features increases. In order to achieve real-time performance, the core idea is to eliminate non-face patches at early stages of the cascade, to compute the more complex features only for the most relevant image patches. In 2002 Lienhart et al. [2] improved this method by introducing a novel set of rotated haar-like features, which are more powerful, but still easy to calculate. In combination with a superior post optimization procedure based on gentle AdaBoost, they could decrease the false alarm rate significantly at given hit rates. Followed by an empirical analysis [3] they also provided an implementation of their method as part of the OpenCV library.

In contrast to those methods, our approach does not work on monocular image sequences but processes synchronized stereo images instead. The approach can be divided into an off-line training phase and a real-time detection phase. For each stereo pair of the training images we train two different classifiers, one classifier on monocular frontal face images and one on the disparity maps, which we estimated from the stereo images. The classifier for monocular images follows the approach by Lienhart et al. with rotated haar-like features, whereas the classifier that works on the disparity maps is trained by generating Eigenfaces [4] using principal component analysis. During real-time detection, we first apply the classifier that is based on monocular images and allows for a slightly higher rate of false positives, which is then checked by the disparity map classifier that eliminates the falsely detected positives. If the image patch containing the face is too small, we revert to a simpler approach that analyzes the size of the face in the image in relation to its estimated distance to the camera.

This paper is not the first to look into the combination of 2D image appearance and 3D depth maps from passive stereo. Especially, for the application of face recognition the benefit of additional 3D information has been shown before [5, 6]. Very related to our approach is the work by Wang et al. [7, 8]. They also present a real-time face detection system, which uses passive stereo and can additionally track and recognize faces. However, their approach uses a morphological filter in combination with some heuristics to detect the closest face to the camera in the depth map. As a result, only one face can be

[1]This work was partially funded by the Cluster of Excellence on Multimodal Computing and Interaction and the Max Planck Center for Visual Computing and Communication (BMBF-FKZ01IMC01).

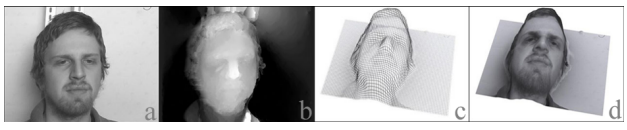

Fig. 1. Quality of the disparity maps: (a) left frame of a stereo image, (b) reconstructed disparity map, (c) corresponding depth-map on a 3D grid, (d) reconstructed 3D head.

detected, whereas our multi-stage machine learning approach can detect multiple faces in the same image.

2. STEREO SETUP AND ALGORITHM

Our stereo setup consists of two cameras with an image resolution of 384×288 pixels and baseline distance of 20 cm. After off-line calibration with a calibration pattern [9], we rectify the input images to standard stereo geometry and estimate a disparity map. Let the left picture be denoted by $I_l(x, y)$ and the right picture by $I_r(x, y)$. We then minimize the functional: $E(z(x, y)) = \iint_\Omega |I_l(x, y) - I_r(x - d(x, y), y)| \, dS$, where $d(x, y)$ is the disparity at pixel $(x, y)^\top$. To find the minimum, we use a variational approach for disparity map estimation,which combines powerful tools such as regularization, automatic tracing and controlling the convergence of the iteration process with the help of Monte Carlo-based prediction technique. Though, regularization of the disparity maps is applied, depth discontinuities are preserved. Recently, a fast implementation of this approach was presented [10]. Figure 1 shows an example of a disparity map estimation. As can be verified by visual examination, the approach generates a disparity map of high quality.

3. NEAR-FIELD STEREO-ENHANCED FACE DETECTION

It is known that boosted cascades of simple feature based detectors rapidly achieve high detection performances when applied to monocular images [1, 2] . However, when applied to images that contain not only faces but also pictures of faces or simple face-like line structures, these detectors still show some false alarms, while on the other hand sometimes rejecting actual faces. One reason is that each stage of the cascade consists of a weak classifier that is based on haar-like features. These are well known to be sensitive to edges, bars, and simple image structures. To overcome these problems, we employ an additional classifier, that is not based on the appearance but evaluates the disparity map. First, we trained a boosted classifier cascade, which uses rotated haar-like features as introduced by Lienhart et al. [2]. Using discrete AdaBoost, we trained on a sets of 4700 positive samples (taken from [11]) and 3300 negative images (office scenes without faces). This classifier cascade performs as shown in the receiver operator characteristic (ROC) in Figure 3.

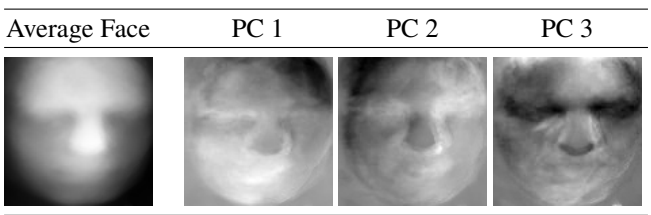

Fig. 2. Results of the PCA on disparity maps: The average face (on the left side) and the first three Eigenfaces.

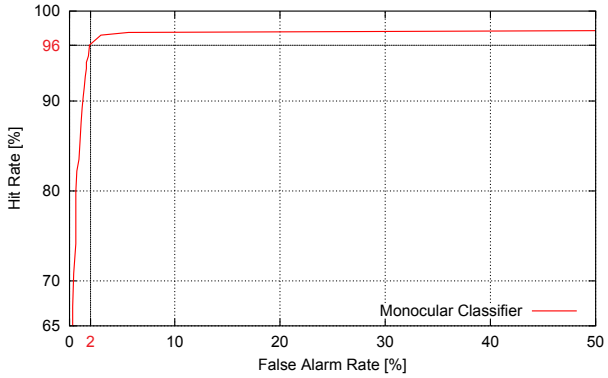

Fig. 3. Excerpt of the receiver operator characteristic (ROC) for the monocular face detection classifier. At 2% false alarm rate, our trained classifier achieves a hit rate of 96%.

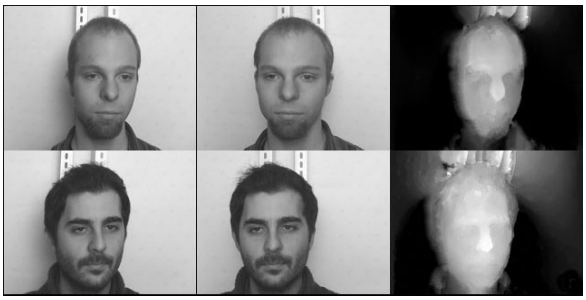

Fig. 4. Stereo image pairs and generated disparity maps from the PCA *training* set of faces.

To detect faces in 2D images, we apply the trained classifier on the left image of all stereo pairs and such identify several image patches that are potential face candidates. To evaluate these candidates in 3D, we perform a second classification step on the disparity maps. This PCA classifier is trained off-line on 30 facial regions, as shown in Fig. 4, which were cropped out of the normalized disparity maps. The principle component analysis estimates the probability distribution of facial disparity maps around their average, as represented by the so called Eigenfaces (cf. [4]), which are shown in Fig. 2. During online detection, we evaluate whether a potential candidate, as selected by the 2D classifier, is a veridical face by projecting its signal into PCA space. Considering the variances of the learned probability distribution, we calculate the Mahalanobis distance of the current test signal to the average face. By choosing a threshold of $\sigma = 3.29$ in terms of standard deviation σ, we accept all face candidates that lie within the range of $\approx \ 99,9 \ \%$ of all training samples. Face candidates that lie above this threshold will be rejected as non-faces.

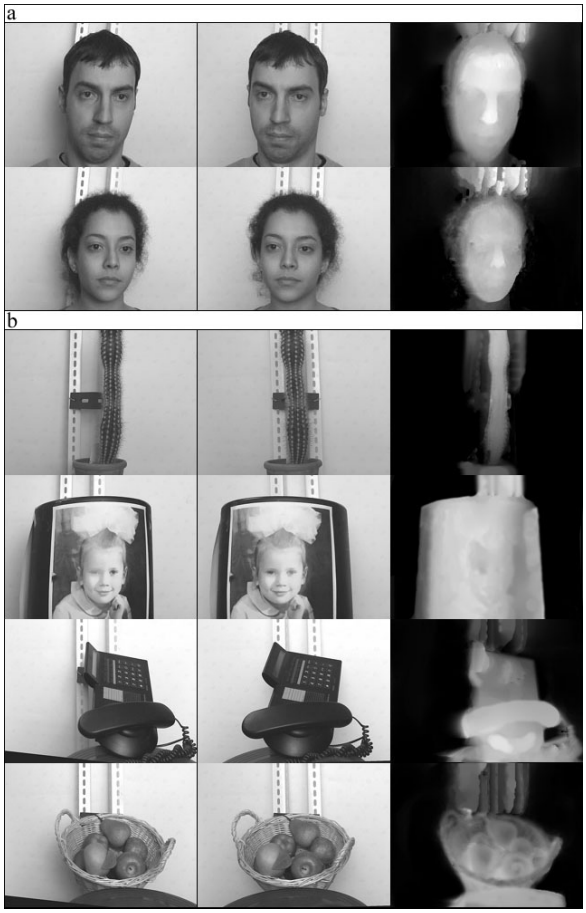

Fig. 5. Stereo image pairs and generated disparity maps from the *test* set: (a) examples of faces, (b) examples of non-faces, e.g., only an picture or a sketch of a real face.

4. FAR-FIELD STEREO-ENHANCED FACE DETECTION

If a face is detected in a small image patch, the face detection algorithm described above fails because a reliable disparity map can not be estimated. Therefore, if the patch size of a face is smaller than 40×60 pixels, we revert to a simpler alternative to detect false positives generated by the classifier for monocular images. For each detected face region, we detect feature points and estimated their disparity to the right frame with the Kanade-Lucas-Tomasi approach [12]. Afterwards, the median disparity is calculated within the detected face region, and the corresponding distance to the camera is calculated with the known calibration data of the cameras. From the size of the face in the image and the distance, the actual size of the face in 3D can be determined. The size is then checked against an interval of sizes for veridical faces, which is learned from a database of 3D face scans. If the face size is smaller than 15.78 cm or larger than 24.16 cm, the face is marked as a false-positive and is dismissed.

5. EXPERIMENTS AND RESULTS

We tested our stereo-enhanced face detector on a large number of input samples, which were different from the samples used during training. Thereby, 40 samples contained one or more faces and 19 images contained some other non-face object. Fig. 5 shows a few examples of face and non-face samples used in the test set. Tab. 1 compares the detection rates of the monocular detection approach with those obtained with our stereo-enhanced method. All false-positives could be removed by taking the additional 3D information into account. The number of true-positives and false-negatives stayed the same. Figures 6 and 7 illustrate the shortcomings of the classical monocular approach because it can not distinguish a real face from a photo print. Our improved detection approach dismisses the false detection with the near-field stereo-enhanced classifier, Fig. 6, and the far-field stereo-enhanced classifier in Fig. 7. In Tab. 2 timings for the different step of our algorithms are given. For a scene with a single face a frame rate of 11 fps for the far-field approach and 6 fps for the near-field approach can be achieved.

	40 Faces			19 Non-Faces	
	FP	TP	FN	FP	TN
monocular	7	38	2	5	14
stereo enhanced	0	38	2	0	19

Table 1. Comparison of the detection rates of the monocular detection approach with our stereo enhanced method. The two shaded columns show that the number of false-positives (FP) decreases using our method, and the number of true-positive (TP) and false-negative (FP) stays the same.

Step	FF [msec]	NF [msec]
Run monocular detector	27	27
Estimate disparity map	-	78
Transform into Eigenspace	-	42
Estimate sparse disparity map	63	-
Total	90	147

Table 2. Timings for an Intel® Core™ 2 CPU with 2,66 GHz. The monocular detector must be run once per image, the other algorithms must be run once per detected face. Here FF and NF denote far-field and near-field approach, respectively.

6. CONCLUSION AND DISCUSSION

In this paper a widely used monocular face detector based on a trained haar-feature cascade is extended by an additional classifier that evaluates the disparity map of a passive stereo camera. The algorithms runs in real-time and significantly reduces the the number of false-positives compared to the

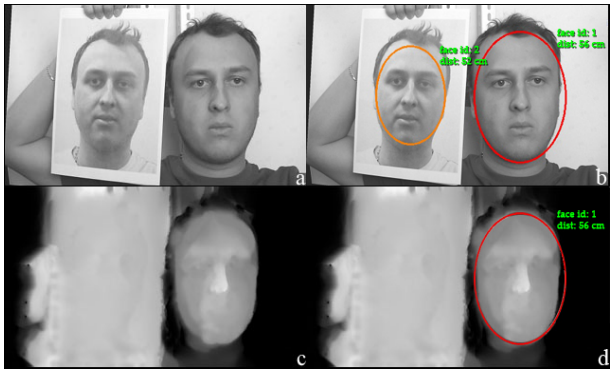

Fig. 6. Detecting false positives with the near-field stereo-enhanced classifier: (a) input image, (b) two detected faces of the monocular classifier (true positive and one false positive), (c) calculated optical flow, (d) the stereo-enhanced classifier dismisses the false positive ($\sigma_{TP} = 3.27$, $\sigma_{FP} = 6.35$).

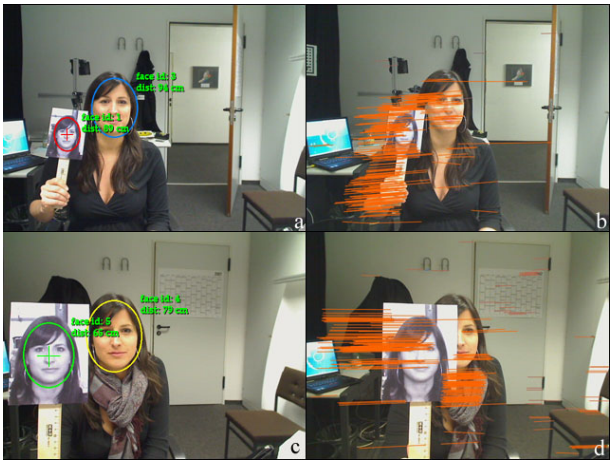

Fig. 7. Detecting false-positives (marked by crosses) with the far-field stereo-enhanced classifier: (a) left frame with detected faces, (b) right frame with overlayed feature point disparities, (c)+(d) second example, where the faces have similar image sizes but different depth.

monocular approach. In fact, as our test set is rather small (40 faces and 19 non-faces samples), all false-positives could be removed in our experiment. We are still working on the extension of our training and test set and are planning to make the data available to the research community.

Currently, the system has the limitations that only frontal faces are detected. A possible solution is to train both the monocular and the stereo extension on different face orientations, resulting in a different detector for each orientation, and then run all detectors in parallel. This is left for future work.

7. REFERENCES

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Computer Vision and Pattern Recognition*, 2001, p. 511.

[2] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Proc. IEEE Intern. Conf. on Image Processing*, 2002, pp. 900–903.

[3] R. Lienhart, E. Kuranov, and V. Pisarevsky, "Empirical analysis of detection cascades of boosted classifiers for rapid object detection," in *DAGM 25th Pattern Recognition Symposium*, 2003, pp. 297–304.

[4] M.A. Turk and A.P. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE Computer Vision and Pattern Recognition*, 1991, pp. 586–591.

[5] F. Tsalakanidou, D. Tzovaras, and M. G. Strintzis, "Use of depth and colour eigenfaces for face recognition," *Pattern Recogn. Lett.*, vol. 24, no. 9-10, pp. 1427–1435, 2003.

[6] T.-H. Sun, M. Chen, S. Lo, and F.-C. Tien, "Face recognition using 2D and disparity eigenface," *Expert Syst. Appl.*, vol. 33, no. 2, pp. 265–273, 2007.

[7] J.-G. Wang, H. Kong, E. Sung, W.-Y. Yau, and E.K. Teoh, "Fusion of appearance image and passive stereo depth map for face recognition based on the bilateral 2DLDA," *J. Image Video Process.*, vol. 2007, no. 2, pp. 6–6, 2007.

[8] J.-G. Wang, E.T. Lim, X. Chen, and R. Venkateswarlu, "Real-time stereo face recognition by fusing appearance and depth fisherfaces," *J. VLSI Signal Process. Syst.*, vol. 49, no. 3, pp. 409–423, 2007.

[9] R.Y. Tsai, "A versatile camera calibration technique for high-accuracy 3-d machine vision metrology using off-the-shelf cameras and lenses," *IEEE Transaction on Robotics and Automation*, vol. 3, no. 4, pp. 323–344, 1987.

[10] A. Bruhn, J. Weickert, T. Kohlberger, and C. Schnörr, "A multigrid platform for real-time motion computation with discontinuity-preserving variational methods," *Int. J. Comput. Vision*, vol. 70, no. 3, pp. 257–277, 2006.

[11] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Proc. IEEE Computer Vision and Pattern Recognition*, 2005, pp. 947–954.

[12] J. Shi and C. Tomasi, "Good features to track," in *Proc. IEEE Computer Vision and Pattern Recognition*, 1994, pp. 593 – 600.