Face Detection Using a Time-of-Flight Camera

Martin Böhme, Martin Haker, Kolja Riemer,
Thomas Martinetz, and Erhardt Barth

Institute for Neuro- and Bioinformatics, University of Lübeck
Ratzeburger Allee 160, D-23538 Lübeck, Germany
{boehme,haker,riemer,martinetz,barth}@inb.uni-luebeck.de
http://www.inb.uni-luebeck.de

Abstract. We adapt the well-known face detection algorithm of Viola and Jones [1] to work on the range and intensity data from a time-of-flight camera. The detector trained on the combined data has a higher detection rate (95.3%) than detectors trained on either type of data alone (intensity: 93.8%, range: 91.2%). Additionally, the combined detector uses fewer image features and hence has a shorter running time (5.15 ms per frame) than the detectors trained on intensity or range individually (intensity: 10.69 ms, range: 5.51 ms).

1 Introduction

In this paper, we will examine how a time-of-flight camera can be used for face detection. We will extend the well-known face detection algorithm of Viola and Jones [1] to time-of-flight (TOF) images; as we will show in the results section, the detector trained on the combined range and intensity data not only has a higher detection rate than detectors trained on either type of data alone, but it also requires fewer features and therefore has a shorter running time.

The Viola-Jones face detector is computationally very efficient while at the same time achieving good detection rates. This is due to three important characteristics: (i) The detector is based on image features that can be evaluated quickly and in constant time, independent of the size of the feature; (ii) the detector selects a set of highly discriminative image features using the *AdaBoost* algorithm; (iii) the detector is structured into a cascade of progressively more sophisticated stages. Since most candidate regions in an image are very dissimilar to a face, the early stages of the cascade can discard these regions with little computation; the later stages of the cascade, which require more computation, need to process only a small proportion of candidate regions.

The attractive properties of the Viola-Jones face detector have motivated a large number of researchers to extend this work in various ways, including the use of different features, modifications to the AdaBoost algorithm, and the application to different types of object detection tasks (see e.g. [2,3,4]). The algorithm has also already been applied to TOF data [5]. However, this previous work does not extend the Viola-Jones detector itself to use range features; instead, a standard Viola-Jones detector trained on images from a conventional camera

R. Koch and A. Kolb (Eds.): Dyn3D 2009, LNCS 5742, pp. 167–176, 2009.

is used to find candidate faces in the TOF image; a final detector stage then computes the average distance of each candidate from the range map and rejects candidates whose size does not match the expected size of a face at this distance.

In contrast, the approach we will use in this paper is to include range features as well as intensity features in the set of features used by the detector. As we will show in the results section, the features chosen by the resulting detector consist of an approximately equal number of range and intensity features; the detector has a higher detection rate and shorter running time than detectors trained on the same training samples, but using either the range or the intensity information alone. This underlines previous results (see e.g. [6,7]) showing that it is the combination of range and intensity data that makes the TOF camera a valuable tool for object detection tasks.

2 Method

We use the basic face detection method of Viola and Jones [1] (which we will summarize briefly) but extend the set of features used to both range and intensity features. Since the method was first described, a number of authors have made improvements to the method (see e.g. [2,3,4]), but we use the original algorithm here because we are more interested in the difference made by using range data rather than in absolute performance.

The Viola-Jones face detector consists of a cascade of stages that typically become more sophisticated as one progresses through the cascade (see Fig. 1). The idea is that the overwhelming majority of subregions in an image are nonfaces, and that most of these subregions are "easy", i.e. they can be identified as nonfaces with little computation. Thus, the first stage of the detector contains a computationally efficient classifier that can immediately reject most subregions as being nonfaces; no further processing is carried out on these subregions. Only a small fraction of subregions (both true faces and "hard" nonfaces) are passed on to the next stage for further processing. This next stage performs more computation and, by doing so, can again reject most of the subregions as being nonfaces, passing only a small fraction of subregions on to the next stage, and so on. In this way, the average effort per subregion is kept low because the overwhelming majority of subregions are rejected in the first few stages.

If the detection rate and false-positive rate of the i-th stage (on the input it receives from the previous stage) are d_i and f_i, then the overall detection and false-positives rates of an n-stage cascade are $D = \prod_{i=1}^{n} d_i$ and $F = \prod_{i=1}^{n} f_i$, respectively. A common approach is to train each stage to achieve the same detection rate d and false-positive rate f on its respective input; this results in overall detection and false-positive rates of $D = d^n$ and $F = f^n$.

Each cascade stage is a boosted classifier trained using the AdaBoost algorithm [8]; a boosted classifier combines several *weak classifiers* (each of which performs only slightly better than chance) into a *strong classifier* (which performs substantially better than the individual weak classifiers). The weak

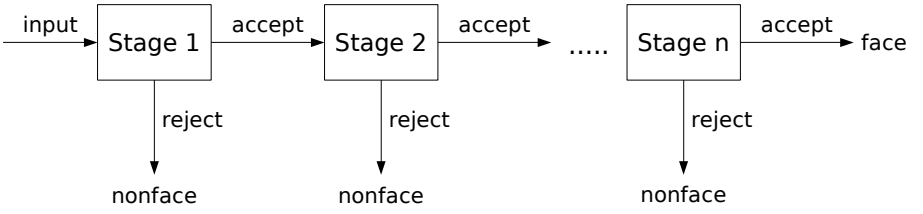

Fig. 1. Cascade structure of the Viola-Jones face detector

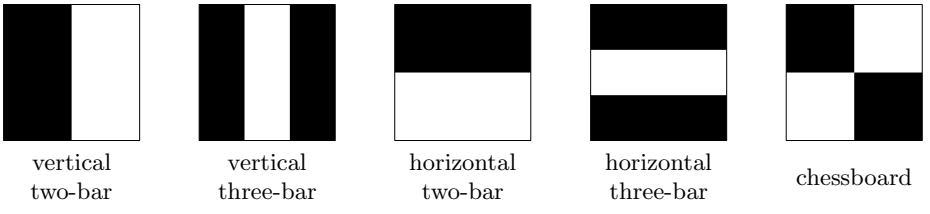

Fig. 2. Haar-like features used by the Viola-Jones face detector. The feature value is obtained by summing the pixels in the white rectangle(s), then subtracting the sum of pixels in the black rectangle(s).

classifiers in the Viola-Jones algorithm are obtained by applying a threshold to an image feature.

The image features, finally, are composed of adjacent rectangles (see Fig. 2); the pixels within each rectangle are summed together, and the resulting values are added or subtracted to obtain the final feature value. For example, the value of the "vertical two-bar" feature is obtained by summing the pixels in the white rectangle and subtracting the sum of pixels in the black rectangle. These features (which are often called *Haar-like features*) have the advantage that they can be evaluated in constant time, independent of their size, using a data structure known as an *integral image*.

The feature set for training the detector is obtained by scaling these features to all possible widths and heights and translating them to all possible positions in the image. Also, each feature (at each size and position) may be evaluated either on the range data or on the intensity data.

Training of the cascade now proceeds as follows. We begin with a training set of face and nonface image patches of constant size. These are used to train the first cascade stage to the desired detection and false-positive rate (evaluated on a validation set). Now, because the next stage will never see those nonface patches that the first stage rejects, we discard all nonface samples rejected by the first stage from the training and validation set, keeping only the false positives. To bring the training and validation set back to their original sizes, we generate new nonface samples by scanning the cascade that has been trained so far across a set of images not containing faces and adding those subregions that the cascade erroneously classifies as faces to the training or validation set until both have

been replenished. We continue adding stages to the cascade in this way until the false-positive rate of the cascade reaches a set target.

Detection proceeds by scanning the cascade across the input image in steps of a certain size. To be able to detect faces of different sizes, the subwindow processed by the detector, along with the features contained in it, is progressively scaled up by a certain factor until it reaches the size of the complete image.

3 Results

The training data for the face detector were recorded using a SwissRanger SR3000 camera [9]. The training set consists of 1310 images (with a resolution of 176 by 144 pixels) showing faces of 17 different persons, in different orientations and with different facial expressions, as well as 4980 images not containing faces. Each face image was labelled by hand with a square bounding box containing the face; some background was included in the bounding box, since previous researchers had reported that this yielded slightly better results than a more tighly cropped bounding box (see the discussion in [1, Sect. 5.1]).

The images were split up into a training, validation and test set, containing 70%, 23%, and 7% of the images, respectively. (The training set is used to select the best weak classifiers for each cascade stage, the validation set is used to evaluate whether the stage has reached its goal detection rate and false-positive rate, and the test set is used to test the final cascade after training is completed.) Face images were cropped to the face bounding box and resized to 24 by 24 pixels (see Fig. 3 for examples). To increase the number of face images in each set, we added versions of each image that were rotated left and right by 3 degrees. After this step, a mirrored version of each face image (including the rotated ones) was also added to the set. The nonface images were full frames of 176 by 144 pixels; to generate examples for training the first cascade stage, subimages of 24 by 24 pixels were cut out of the nonface images. For the second and subsequent stages,

Fig. 3. Examples of intensity images from the face training set

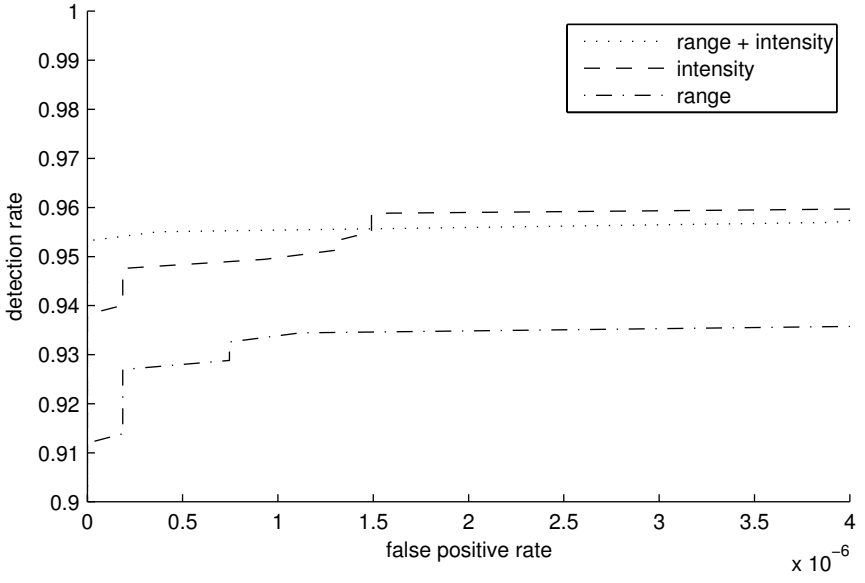

Fig. 4. ROC curves for the detectors trained on the combined range and intensity data as well as both types of data separately

new negative examples were generated by scanning the cascade trained so far across the nonface images and collecting false positives (see also Sect. 2).

In all, there were 5412 faces and 3486 nonface images in the training set, 1752 faces and 1145 nonface images in the validation set, and 534 faces and 349 nonface images in the test set. The data set is publicly available at `www.artts.eu/publications/3d_tof_db`

We trained a detector on the combined range and intensity data as well as on the range and intensity data alone. The target detection rate and false-positive rate for each stage were set to $d = 0.995$ and $f = 0.4$, respectively; the target false positive rate for the complete detector was set to 10^{-8}. The range-and-intensity detector as well as the range-only detector were successfully trained to this target rate. Training of the intensity-only detector did not reach the target rate; training was stopped manually when the detector had added over 1500 features to the cascade stage it was training and the false-positive rate of the stage had stagnated without reaching the goal rate. (This is a typical sign that the detector can no longer generalize from the training to the validation set.) We trained another intensity detector with a lower detection rate per stage of $d = 0.99$; training for this detector did complete, but its performance was consistently worse than that of the detector with $d = 0.995$ whose training was aborted. For this reason, we will only use the latter detector in the tests that follow.

Figure 4 shows ROC curves for the three detectors (computed as in [3]). For false positive rates above $1.5 \cdot 10^{-6}$, the intensity-only detector achieves a slightly higher detection rate than the range-and-intensity detector. Below

Table 1. Cascade structure of the detectors

		detection rate		false-positive rate		number of features		
	Stage	individual	cumulative	individual	cumulative	total	intensity	range
intensity + range	0	1.000	1.000	0.000	1.0e-02	2	1	1
	1	0.995	0.995	0.165	1.6e-03	2	1	1
	2	0.995	0.991	0.138	2.3e-04	2	1	1
	3	0.995	0.986	0.394	8.9e-05	7	4	3
	4	0.995	0.981	0.290	2.6e-05	7	3	4
	5	0.995	0.976	0.385	1.0e-05	12	5	7
	6	0.995	0.971	0.334	3.3e-06	12	8	4
	7	0.995	0.966	0.381	1.3e-06	15	9	6
	8	0.995	0.961	0.380	4.8e-07	20	11	9
	9	0.995	0.956	0.380	1.8e-07	22	13	9
	10	0.995	0.951	0.368	6.8e-08	27	16	11
	11	0.995	0.946	0.365	2.5e-08	44	26	18
	12	0.995	0.941	0.391	9.6e-09	49	26	23
intensity	0	0.999	0.999	0.189	1.9e-01	3	3	0
	1	0.995	0.994	0.367	6.9e-02	15	15	0
	2	0.995	0.989	0.340	2.4e-02	11	11	0
	3	0.998	0.986	0.340	8.0e-03	6	6	0
	4	0.997	0.983	0.391	3.1e-03	10	10	0
	5	0.995	0.978	0.389	1.2e-03	17	17	0
	6	0.995	0.973	0.376	4.6e-04	25	25	0
	7	0.995	0.968	0.384	1.8e-04	23	23	0
	8	0.995	0.963	0.386	6.8e-05	40	40	0
	9	0.995	0.958	0.389	2.6e-05	59	59	0
	10	0.995	0.953	0.392	1.0e-05	62	62	0
	11	0.995	0.948	0.400	4.1e-06	107	107	0
	12	0.995	0.943	0.396	1.6e-06	198	198	0
	13	0.995	0.938	0.399	6.5e-07	117	117	0
	14	0.995	0.934	0.390	2.5e-07	223	223	0
range	0	0.997	0.997	0.071	7.1e-02	2	0	2
	1	0.995	0.992	0.285	2.0e-02	3	0	3
	2	1.000	0.992	0.325	6.5e-03	3	0	3
	3	0.995	0.987	0.373	2.4e-03	11	0	11
	4	0.995	0.983	0.271	6.6e-04	17	0	17
	5	0.995	0.978	0.374	2.5e-04	18	0	18
	6	0.996	0.974	0.381	9.4e-05	10	0	10
	7	0.995	0.969	0.388	3.6e-05	26	0	26
	8	0.995	0.964	0.377	1.4e-05	29	0	29
	9	0.995	0.959	0.396	5.4e-06	56	0	56
	10	0.995	0.954	0.370	2.0e-06	42	0	42
	11	0.995	0.949	0.381	7.7e-07	66	0	66
	12	0.995	0.944	0.396	3.0e-07	114	0	114
	13	0.995	0.940	0.382	1.2e-07	81	0	81
	14	0.995	0.935	0.377	4.4e-08	111	0	111
	15	0.995	0.930	0.380	1.7e-08	139	0	139

this point, the range-and-intensity detector achieves better detection rates. Both detectors are markedly better than the range-only detector over the whole range of false-positive rates shown. All three detectors achieve good detection rates even for a false-positive rate of zero. This is an indication that our test set is relatively "easy" compared to, for instance, the MIT+CMU test set [10], on which the Viola-Jones algorithm produces slightly higher error rates [1]. Whereas the MIT+CMU test set contains images from a variety of sources, including text and line drawings, our test set consists solely of images taken with a single camera. Also, because of the active illumination, the lighting is the same across all images. We believe these factors combine to make the test set "easier".

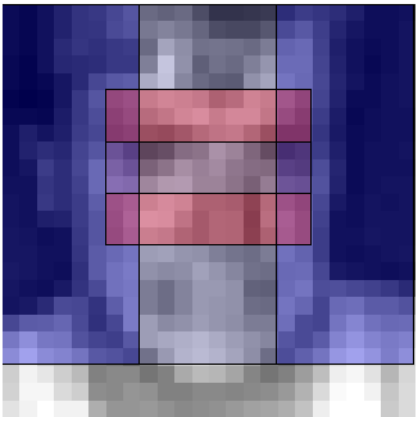

Fig. 5. Features used by the first stage of the range-and-intensity detector. The blue (vertical) feature is a range feature; the red (horizontal) feature is an intensity feature.

We will now examine the cascade structure of the detectors, i.e. the number of features used in each stage together with the detection rate and false-positive rate for each stage (see Table 1).

The first thing that is noticeable is that the first stage of the range-and-intensity detector achieves a false-positive rate of 0 on the validation set, i.e. the false-positive rate was too small to measure on the validation set. When this first stage was run on the set of full-frame test images, its false positive rate was 0.05%. In other words, the first stage already eliminates 99.95% of nonfaces. For comparison, the first stages of the other two detectors had false-positive rates of 18.9% (intensity) and 7.1% (range).

To understand why the first stage of the range-and-intensity detector has such good performance, consider Fig. 5, which shows the features used by this stage: A vertical three-bar range feature and a horizontal three-bar intensity feature. From the sample training image underlayed under the features, it is evident that the range feature responds to the range difference between the face and the background on either side; the intensity feature seems to respond to the difference between the eye region (which is typically darker) and the forehead and cheeks above and below (which are typically lighter).

The fact that the first stage achieves a false-positive rate of zero on the validation set is problematic for computing the false-positive rate of the entire cascade, which is used during training to decide when the detector has reached its performance goal. To be able to compute an overall false-positive rate, we conservatively assumed the false-positive rate for this stage to be 0.01; this assumption is also used in the cumulative rates shown in the table. The assumed rate of 0.01 is probably quite conservative and only affects the overall false-positive rate computed during training, but not the selection of weak classifiers or the false-positive rates computed on the test set.

Table 2. Performance summary of the detectors on the various types of data. Detection rates are given for a zero false-positive rate on the test set. Running times include preprocessing (computation of the integral images).

Detector	Detection rate	Running time per frame
range + intensity	95.3%	5.15 ms
intensity	93.8%	10.69 ms
range	91.2%	5.51 ms

Turning to the number of features per stage, we note that, in most stages, the range-and-intensity detector requires noticeably fewer features to reach its target performance than the other two detectors. Also, note that the range-and-intensity detector uses an approximately equal number of range and intensity features in each stage (with a tendency to use slightly more intensity features in the later stages). This indicates that the range and intensity data contribute approximately the same amount of information to the face detection task.

Finally, we turn to the running times for the various detectors (Table 2). The range-and-intensity detector is more than two times faster than the intensity-only detector and slightly faster than the range-only detector. This reflects the fact that the intensity-only detector uses more features than the other two detectors in the first few cascade stages, which consume the most processing time. Note when comparing the timings that the range-and-intensity detector needs to perform twice the amount of preprocessing (computing the integral images for both range and intensity) but still ends up faster. The table also summarizes the detection rates achieved for a zero false-positive rate on the test set.

4 Discussion

We have shown that a face detector trained on the combined range and intensity data from a TOF camera yields a higher detection rate (95.3%) than a detector trained on either type of data alone (intensity: 93.8%, range: 91.2%). Furthermore, the range-and-intensity detector requires fewer features than the other two detectors. This translates into faster running times: The range-and-intensity detector is over twice as fast as the intensity-only detector and slightly faster than the range-only detector (which misclassifies almost twice as many faces).

The data obtained by the TOF camera is in effect a two-channel image, where one channel contains the range map and the other contains the intensity image. If the TOF camera is combined with a grayscale or RGB camera operating in the visible spectrum (as is the case in the 3DV Systems ZCam [11], for instance), it would be straightforward to extend the method to the additional channels obtained in this way.

The detector we used was a "stock" Viola-Jones face detector. Even better results might be possible using features that are specifically tuned to the type of

structures typically found in range images. One could also investigate the idea of combining range and intensity information in a single feature. Additionally, the many refinements that have been made to the Viola-Jones algorithm since its inception could be incorporated.

However, we are not primarily interested in the maximum absolute performance that a TOF face detector can achieve but rather in the relative difference in performance between face detection on combined range and intensity data versus either type of data alone. We believe that the advantage of the combined range and intensity detector in terms of robustness and speed should be preserved when refinements are made to the underlying algorithms; whether this indeed holds true is a question for future research.

Acknowledgments

We thank the anonymous reviewers for their comments, which helped to improve this paper. This work was developed within the ARTTS project (www.artts.eu), which is funded by the European Commission (contract no. IST-34107) within the Information Society Technologies (IST) priority of the 6th Framework Programme. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

1. Viola, P., Jones, M.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004)
2. Lienhart, R., Kuranov, A., Pisarevsky, V.: Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In: Michaelis, B., Krell, G. (eds.) DAGM 2003. LNCS, vol. 2781, pp. 297–304. Springer, Heidelberg (2003)
3. Brubaker, S.C., Wu, J., Sun, J., Mullin, M.D., Rehg, J.M.: On the design of cascades of boosted ensembles for face detection. International Journal of Computer Vision 77(1–3), 65–86 (2008)
4. Barczak, A.L.C., Johnson, M.J., Messom, C.H.: Real-time computation of Haar-like features at generic angles for detection algorithms. Research Letters in the Information and Mathematical Sciences 9, 98–111 (2006)
5. Hansen, D.W., Larsen, R., Lauze, F.: Improving face detection with TOF cameras. In: Proceedings of the IEEE International Symposium on Signals, Circuits & Systems (ISSCS), vol. 1, pp. 225–228 (2007)
6. Haker, M., Böhme, M., Martinetz, T., Barth, E.: Geometric invariants for facial feature tracking with 3D TOF cameras. In: Proceedings of the IEEE International Symposium on Signals, Circuits & Systems (ISSCS), Iasi, Romania, vol. 1, pp. 109–112 (2007)
7. Haker, M., Böhme, M., Martinetz, T., Barth, E.: Scale-invariant range features for time-of-flight camera applications. In: CVPR 2008 Workshop on Time-of-Flight-based Computer Vision, TOF-CV (2008)

8. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1), 119–139 (1997)
9. Oggier, T., Büttgen, B., Lustenberger, F., Becker, G., Rüegg, B., Hodac, A.: SwissRanger^{TM} SR3000 and first experiences based on miniaturized 3D-TOF cameras. In: Proceedings of the 1st Range Imaging Research Day, Zürich, Switzerland, pp. 97–108 (2005)
10. Rowley, H.A., Baluja, S., Kanade, T.: Neural-network-based face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(1), 23–38 (1998)
11. ZCam: 3DV Systems, Yokne'am, Israel, `http://www.3dvsystems.com`.