

Face Detection using 3-D Time-of-Flight and Colour Cameras

Jan Fischer, Daniel Seitz, Alexander Verl
Fraunhofer IPA, Nobelstr. 12, 70597 Stuttgart, Germany

Abstract

This paper presents a novel method to apply standard 2-D face detection methods on 3-D data. The procedure uses a sensor setup consisting of a 3-D time-of-flight camera and a colour camera. At first, face detection is performed on the less structured and low-resolution 3-D range image. Only for those areas regarded as faces on the 3-D range image, processing continues on the corresponding high-resolution 2-D colour image areas. This enables a pre-filtering of the visible area prior to the actual face detection step on selected colour regions.

1 Introduction

In the area of service robotics, it is inevitable to give a robot like Care-O-bot[®] 3 [1] the awareness of humans within its vicinity in order to perform communication and interaction. Therefore, reliable and robust face detection and recognition are two of the most important software components for service robots.

Over the last decades, research was primarily focused on 2-D face detection [2]. With the upcoming of 3-D time-of-flight sensors new approaches to real-time 3-D face detection are attracting more and more attention. However, most methods focus on matching computationally expensive 3-D face models against 3-D image data. Recently, Böhme et al. [5] proposed a face detection approach by solely using image data from a time-of-flight sensor. This paper carries the idea of Böhme et al. on by combining the information of a time-of-flight sensor and a colour camera.

The main idea of the paper is to use the advantages of the 3-D time-of-flight sensor and perform at first face detection on the 2-D range image. This enables a pre-filtering of the visible areas by focusing on face contours prior to the actual face detection step on selected colour regions. Within the range image, depth is encoded with ordinary gray values. So, it is possible to apply standard 2-D face detection methods on the range image. It was decided to apply the well-known method of Viola and Jones [3] to guarantee robust and real-time face detection.

2 Hardware Setup

The sensor setup for the proposed face detection approach is shown in figure 1. It consists of a colour camera as well as the 3-D time-of-flight sensor SwissRanger 4000 [7]. The other colour camera on the image has not been used for the given task.

Compared to the colour camera, the time-of-flight sensor outputs range data instead of colour information. The 3-D sensor emits an amplitude modulated near-infrared light,

which is reflected by the illuminated scene. Each pixel of the sensor demodulates the reflected light and determines the range by the measured phase shift. Based on the reconstructed signal, an intensity image and a range image with depth information is created. Using a time-of-flight sensor instead of using a common stereo camera system possesses several advantages. At first, the time-of-flight sensor gives 3-D image data with a frame rate of 30 Hz, a rate hardly achieved by state-of-the-art stereo systems. Furthermore, the acquired range images are dense and not sparsely populated. Compared to stereo cameras, no triangulation must be performed to compute depth information. Therefore, even unstructured image areas have an assigned depth value. The disadvantages of having a significantly lower image resolution and lesser accuracy of the range data are not relevant for the given problem of face detection.

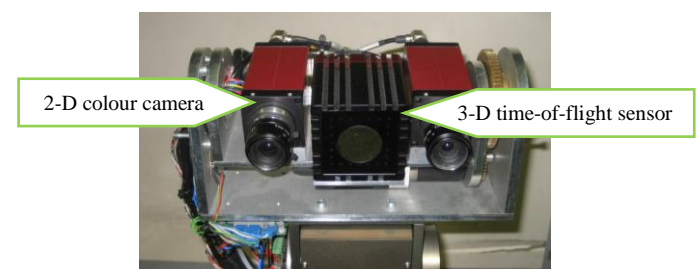


Figure 1: Sensor setup for face detection consists of a 2-D colour camera and a 3-D time-of-flight sensor

2.1 Sensor Fusion

Through calibration it is possible to allow a mapping of the 3-D range data to the corresponding colour values from the colour camera to get a coloured 3-D image of the scene. Initially, both cameras are calibrated to estimate their intrinsic and extrinsic parameters using a standard calibration tool like Bouguet's Matlab calibration toolbox [6]. With the determined intrinsic parameters both images are undistorted. Using the extrinsic parameters of the camera

pair, a 3-D translation vector T_{tof}^{col} and a 3×3 rotation matrix R_{tof}^{col} are calculated to map the 3-D time-of-flight data directly to the corresponding undistorted image data of the colour image. Using the results of the intrinsic and extrinsic calibration, each pixel of the 3-D time-of-flight camera is assigned the corresponding colour information from the colour camera. Given a 3-D coordinate \mathbf{x}_{tof} relative to the time-of-flight sensor, the corresponding 3-D coordinate \mathbf{x}_{col} relative to the 2-D color camera is computed as follows

$$\mathbf{x}_{col} = R_{tof}^{col} \mathbf{x}_{tof} + T_{tof}^{col}$$

To compute the corresponding 2-D color image coordinate \mathbf{u}_{col} [pixels] from the 3-D coordinate \mathbf{x}_{col} [meters], \mathbf{x}_{col} is normalized by dividing it through its z-coordinate before applying the intrinsic matrix M as follows

$$\mathbf{u}_{col} = M \mathbf{x}_{col}$$

The procedure is repeated for each pixel of the 3-D time-of-flight camera. In order to take advantage of the 1388×1038 high resolution colour image, the 204×204 low resolution range image from the time-of-flight camera is resized by a factor of 3 using bilinear interpolation prior to the sensor fusion process. The result is an image of size of 612×612 pixels containing 3-D coordinates and colour information for each pixel. By artificially increasing the image size of the 3-D range image, more color information is preserved during sensor fusion. This related to the fact that each interpolated range value is assigned a colour values from the native colour image.

There are significantly more elaborated methods for sensor fusion of time-of-flight and colour cameras, where especially the problem of false colour matchings near edges is targeted. The proposed methods rely either on the incorporation of several camera views [8], or on noise-aware filters to upsample the low resolution range image to the dimensions of the high resolution colour image [9]. However, the proposed method is sufficient in its simplicity for the given application as stated below. Most importantly, it meets real-time requirements and therefore does not limit the speed of the application.

3 Method

The proposed algorithm is based on the well-known Viola and Jones object detector [3], applied to the problem of detecting faces. However, the main difference to the original method is the fact, that face detection is initially performed on range images from the 3-D time-of-flight sensor. Those regions of the range image that have been labeled to be faces, are subject to further processing, by computing their corresponding colour values and performing face detection on the coloured image regions again. Being labeled as a face region on both, the range and colour image, an image area is considered as showing a face. The proposed algorithm is structured in two stages. Initial-

ly, two classifiers are trained, one for detecting contours of heads on range images and one for detecting faces on colour images. In a second stage, the classifiers are applied to the image data for face detection on the 3-D range image data and face detection on selected regions of the 2-D colour image.

3.1 Classifier Training

Training is performed to create two classifier cascades, one to operate on colour images and the other to operate on the range images of the 3-D time-of-flight sensor. The training procedure for both classifier cascades is the same, with the distinction that the manual labeled training data of face and non-face regions is taken either from the range image or from the colour image. An excerpt of the training data for the range image classifier is shown in figure 2.



Figure 2: Excerpt of the training images based on 3-D image data from the time-of-flight sensor

The Viola and Jones object detector consists of a cascade of weak classifiers, each trained with the AdaBoost algorithm. To perform classification, an image region is successively passed through the weak classifiers, as long as none of the weak classifiers has rejected it. Once a weak classifier has rejected an image region, it is considered as not showing a face and classification stops. Only when an image has passed all weak classifiers without being rejected, it is classified as showing a face. The described control flow is visualized in figure 3.

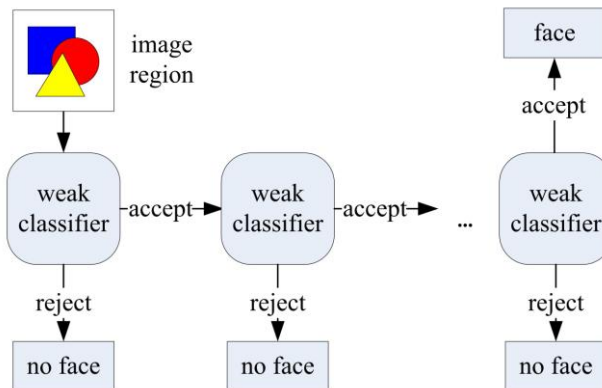


Figure 3: Control flow of the Viola and Jones classifier cascade composed of several weak classifiers

The main idea of using a cascade of weak classifiers is that the first weak classifiers are trained to reject the majority of image regions not containing faces. This enables the classifier to quickly process entire images, by successively extracting image subregions and applying the classifier cascade. Each weak classifier in the Viola and Jones approach uses a composition of rectangular features, so called Haar-like features, for classification. These features are placed within the considered image subregions and the underlying pixels are subtracted from each other. To obtain a classification decision, the difference is compared against a threshold obtained during training. In the basic approach three types of Haar-like features are distinguished. The first type consists of two horizontal or vertical adjacent rectangles whose associated image pixels are subtracted from each other. The second type consists of three vertical or horizontal adjacent rectangles and subtracts the pixels of the exterior two rectangles from the middle one. The third Haar-like feature type is composed of four rectangles, arranged like a chessboard pattern. The difference is computed by subtracting the main from the off diagonal elements. All three feature types are visualized in figure 4.

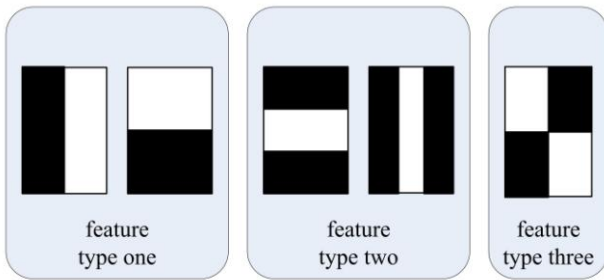


Figure 4: The three basic Haar-like feature types. Pixels of an image subregion covered by the black areas are subtracted from pixels covered by the white areas.

The usage of Haar-like features has a huge computational advantage. Through the introduction of integral images, Viola and Jones proposed an efficient method to compute the sum of pixel values from a rectangular region in linear time. The integral image is computed only once for the entire source image. Each pixel value of the integral image holds the sum of all pixel values above and to the left of the pixel. The sum of pixels within a rectangle is computed by referring to the corresponding integral image values of the rectangles' corner pixels and performing three elementary operations on them.

Through scaling the features to different sizes and translating the features to different positions, more than 160.000 possible features are created for a given image region with a size of 24×24 pixels. It is the task of the AdaBoost training algorithm to select the most distinguished features for each weak classifier to best detect faces. Training begins with the first weak classifier of the classifier cascade. All labeled face and non-face regions are subject to training. When the desired false-positive rate and the desired detection rate have been reached, AdaBoost stops training and proceeds with the next weak classifier. In order for the next weak classifier not to pro-

duce similar classification results than its predecessor, training is performed only on those labeled faces and non-faces that have been falsely classified by the preceding classifiers. Training continues until the cascade achieves the desired false-positive rate r_f given by the product of the individual false-positive rates from each weak classifier

$$r_f = \prod_{j=1}^n r_f^j$$

The described training procedure is executed separately for faces and non-faces on the 2-D color image and the range image from the 3-D time-of-flight camera.

3.2 Face Detection

Face detection is performed by repeatedly sliding a subregion of a given size across the source image and applying the cascade of weak classifiers on it. In order to achieve scale invariance, the subregion as well as the Haar-like features are progressively scaled up after a complete scan of the image by the subregion. The procedure is repeated until the subregion has reached the size of the source image. The proposed face detection approach applies the outlined detection procedure in two phases. At first, the classifier cascade, trained on the range image from the 3-D time-of-flight camera, is applied to detect the contours of heads on the range image data. After processing the 2-D range image with the classifier cascade, all classified face regions are assigned their corresponding color values using the described sensor fusion method. All other image pixels not classified as faces are filled with black colour. Afterwards, the second detection phase applies the classifier cascade, trained on 2-D colour image data, on the resulting colour image. The selective assignment of colour values greatly improves the performance of the algorithm by significantly reducing the false-positive rate and computational complexity as illustrated in section 4. Those image subregions labeled as containing faces by both, the range image classifier cascade and the color image classifier cascade, are considered as faces. The detection procedure is visualized in figure 5.



Figure 5: Detection results on the 3-D range (left) and colour image (right)

4 Results

Within experiments, a set of 120 faces from 10 persons and 424 non-face regions were used to train the classifiers. It has been taken care to capture different viewing angles and different mimics with the training set. The different face positions for training are schematically outlined in figure 6.

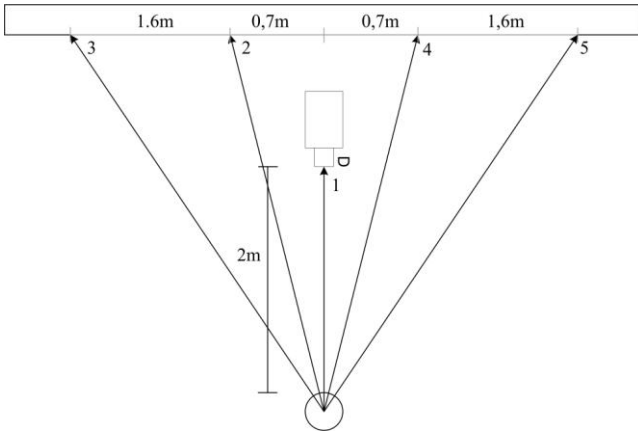


Figure 6: Different face positions for the training data

All weak classifiers have been trained with a target detection rate of 0.995 and a false-positive rate of 0.4.

The classification performance of the proposed method has been measured by processing 360 images each containing one face. The classification results are compared with the classification performance of the Viola and Jones algorithm applied on colour images, only. The results are shown in figure 7.

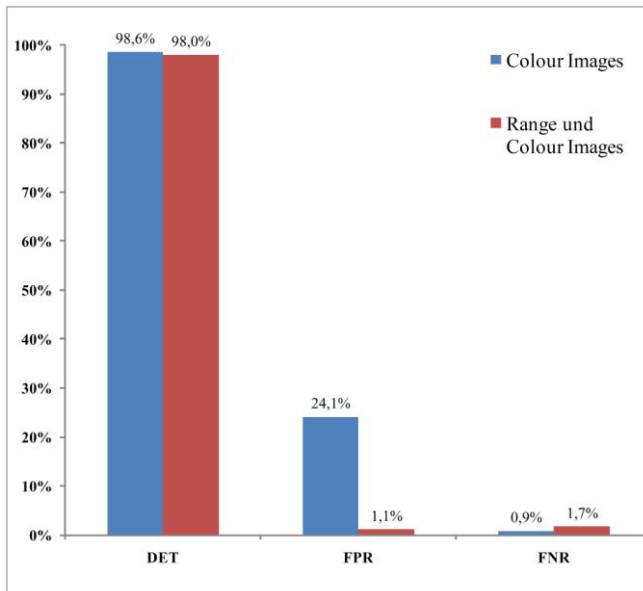


Figure 7: Detection rate (DET), false-positive rate (FAR) and false-negative rate (FRR) of the proposed algorithm compared to the application of the original Viola and Jones algorithm on colour images, only.

The most significant improvement is achieved with a reduction of the false-positive rate from 24.1 % with the original method to 1.1 % when using the proposed algorithm. Detection rate and false-negative rate of both methods are almost similar. The significant improvement of the false-positive rate originates from the initial processing of the range image from the 3-D time-of-flight sensor. The classifier cascade for the range image captures the geometric shape of a head, an aspect that could not be incorporated by the original method working only on 2-D colour images.

The overall computation time has been reduced significantly compared to applying the Viola and Jones face detector on colour images, only. This may sound surprisingly at the first glance, as two images have to be processed. The reason for this significant speedup relies on the fact, that the range images usually provide less structured data compared to colour images and enable the classifier cascade to reject most of the image areas within its first stages. However, all face areas detected on the range image are processed twice. At first, they are processed on the range image and afterwards on the corresponding colour image. This additional processing time is strongly dependent on the number of detected image locations within the range image. Within our scenario, on average less than 30 % of the image data remains after processing the range image, what does not eliminate the outlined benefit in computation time. The proposed method additionally offers the possibility to limit the distance of possible faces by executing range segmentation and invalidating all pixels with a distance greater than a specified threshold. The affected pixels could be set to black and further improve processing time.

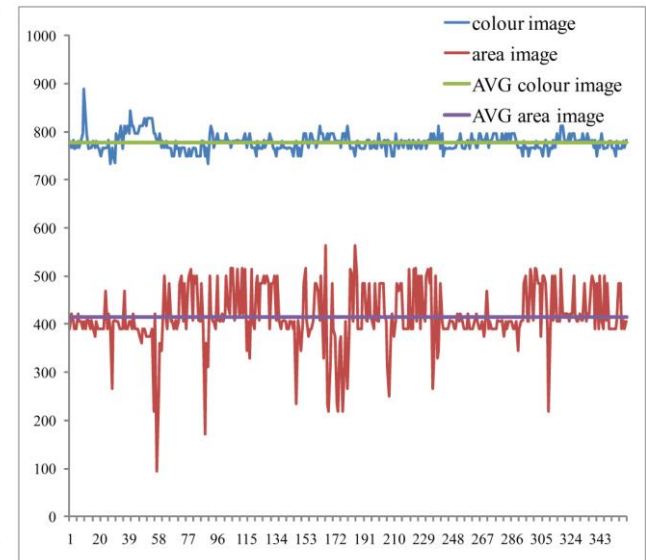


Figure 8: Computation time of the proposed algorithm compared to the application of the original Viola and Jones algorithm on colour images, only. The horizontal axis enumerates the processed images, the vertical axis the processing time in ms.

Figure 8 compares the processing time of the proposed algorithm with the processing time, when using the Viola and Jones detector on colour images, only. The strong dependence of the proposed method on the number of detected image locations within the range image, is visible in the high variance of the measured computation time (red line). In extreme cases the computation time could be reduced by a factor of 8. On average computation time is reduced by a factor of 2.

5 Conclusion

The purpose of the paper was to show the potential of extending classical 2-D image processing techniques to range image from a 3-D time-of-flight sensor. The paper presented a new approach to reduce the false-positive rate for an object detection process. In experiments the false-positive rate could be minimized from 24.1 % with the original object detector to 1.1 % with the proposed method. This reduction is possible, because the proposed algorithm uses two different detection processes: detection on 3-D range image to capture face contours and detection on 2-D colour image to capture colour information.

The first detection process on range images excludes most image areas for further processing on colour image data. The second detection process is able to detect faces on the colour image, but only in the areas where contours of heads have been detected on the range image. False-positives, which are detected by the first process, can be eliminated by the second detection process. The proposed method not only reduces the number of false-positives, but also the total detection time is decreased by about 30 % in our experiments.

Future experiments will target the incorporation of range and colour values from the 3-D time-of-flight sensor and the colour camera into the one classifier cascade as proposed by Böhme et al. to further improve the detection performance.

6 Literature

- [1] Reiser, U.; Connette, C.; Fischer, J.; Kubacki, J.; Bubeck, A.; Weissshards, F.; Jacobs, T.; Parlitz, C.; Hägele, M.; Verl, A.: Care-O-bot 3 - Creating a product vision for service robot applications by integrating design and technology. In IEEE/RSJ International Conference on Intelligent Robots and Systems, USA: St. Louis, Oct. 11-15, 2009, pp. 1992-1197
- [2] Zhao, W.; Chellappa, R.; Phillips, P. J.; Rosenfeld, A.: Face recognition: A literature Survey. In ACM Comput. Surveys (CSUR) Archive 34 (4) 2003, pp. 399-458
- [3] Viola, P.; Jones, V.: Rapid Object Detection using a Boosted Cascade of Simple Features. In Proc. Computer Vision and Pattern Recognition, 2001, pp. 511-518
- [4] Freund, Y.; Schapire, R. E.: A decision-theoretic generalization of on-line learning and an application to

- boosting. Computational Learning Theory: Eurocolt '95, Springer-Verlag, 1995, pp 23-27
- [5] Böhme, M.; Haker, M.; Riemer, K.; Martinetz, T.; Barth, E.: Face Detection Using a Time-of-Flight Camera. In Lecture Notes in Computer Science, Volume 5742, 2009, pp 167-176.
- [6] Bouguet, J.: Camera Calibration Toolbox for Matlab, http://www.vision.caltech.edu/bouguetj/calib_doc/
- [7] Oggier, T.; Büttgen, B.; Lustenberger, F.; Becker, G.; Rüegg, B.; Hodac, A.: SwissRanger™ SR3000 and first experiences based on miniaturized 3-D-TOF cameras. In Proc. 1st Range Imaging Research Day, 2005, Zürich, Switzerland, pp 97-108
- [8] Kim, J. M.; Theobalt, J.; Diebel, J.; Kosecka, J.; Matusik, B.; Thrun, S.: Multi-view image and ToF sensor fusion for dense 3-D reconstruction, 2009, In Proc. of 3-DIM 2009, ICCV
- [9] Chan, D.; Buisman, H.; Theobalt, C.; Thrun, S.: A noise-aware filter for real-time depth upsampling. In: Proc. of ECCV Workshop on multi-camera and multi-modal sensor fusion algorithms and applications, 2008, pp 1-12