Real-Time Scale-Invariant Face Detection on Range Images

Maurício Pamplona Segundo, Luciano Silva, Olga Regina Pereira Bellon
IMAGO Research Group, Universidade Federal do Paraná
81531-980 Curitiba, Brazil, PO Box 19092
http://www.imago.ufpr.br
{mauricio,luciano,olga}@inf.ufpr.br

***Abstract*—We present a scale-invariant face detection approach based on boosted cascade classifiers using range images as input. The detector was developed to be employed as a preliminary stage for any real-time 3D face recognition system. The required computation time for this task was considerably reduced by eliminating the need for scanning an input image in multiple scales. Our experiments were performed using two well-known databases, and the proposed approach was favorably compared against a state-of-the-art face detection approach. We achieved a detection rate of 99.9% with only 0.2% of the images presenting false detections. We also evaluated the detector performance in face images presenting large pose variations and obtained detection rates as high as when using frontal face images.**

***Index Terms*—Face detection, boosted cascade classifiers, range images, real-time, 3D face recognition**

I. Introduction

Face recognition based on 2D images was the main focus of researches regarding face biometrics for many years [1]. The potential of the 3D face recognition to solve classical 2D face recognition problems that have endured for all those years [2], such as illumination and pose, led to an increased interest in the development of 3D face recognition systems.

Despite the advantages, hardware technology for 3D data acquisition presents limitations: laser scanners are slow and expensive; stereo-based sensors are inaccurate and demand high computational resources; structured light-based sensors have specific environment requirements; and time-of-flight cameras only capture low resolution images. Recent advances in 3D acquisition devices based on structured light using patterns invisible to the naked eye, such as the PrimeSensor[1] and the Kinect sensor[2], present a good combination of speed, price, accuracy and usability. This new generation of 3D sensors is able to capture up to 60 frames per second of 3D data. Thus, computer vision systems (*e.g.* face recognition) have a higher demand for real-time 3D processing techniques.

Face detection is critical in fully automatic face recognition systems [1], [3], [4]. This problem has been solved using or adapting 2D face recognition techniques. Some works employed 2D face detection techniques to locate the face when the color information was also available, such as skin color segmentation [5] and boosted cascade classifiers [6], [7]. Other works adapted 2D techniques to directly locate the face in the 3D image, such as horizontal and vertical projections [8], ellipse detection [9], [10], eigenfaces [11] and boosted cascade classifiers [12], [13].

In this work we also adapt the 2D detection technique based on boosted cascade classifiers [4] to directly locate faces in 3D images, as in [12], [13]. We chose this technique because it is one of the most successful techniques for face detection in the literature, and it presents a good cost-benefit regarding computational cost, detection rates and amount of false positives, even for images containing unknown faces and environments (*i.e.* not included in the training set).

The main contribution of this work is the reduction of the complexity in the search stage, which is achieved by eliminating the need to look for faces in different scales. To this end, we created a projection image for an input range image that ignore the focal length parameter and obtain faces with the same size as result. By estimating the face size, it is possible to locate all the faces in a single image scan and in real-time.

We also propose an approach for face detection over pose variation using the same boosted classifier employed for frontal face detection. In order to do that, we first create projection images from different viewpoints for each input range image. Then we apply the frontal face classifier in all projection images and locate faces even when they present self-occluded parts.

In our experiments, we used the Face Recognition Grand Challenge (FRGC) database[3] and the Binghamton University 3D Facial Expression (BU-3DFE) database [14], since they have been extensively used for researches regarding 3D face analysis [5], [9], [15]–[18] and contain faces with different locations, facial expressions and resolution. We also used images acquired by the Kinect Sensor to show the applicability of our approach to real-time low-cost low-resolution devices.

The remainder of this paper is organized as follows: first we discuss related works and introduce the proposed face detection approach in Section II; details regarding experiments and results are discussed in Section III; finally, we present our conclusions in Section IV, followed by the references.

[1]http://www.primesense.com/
[2]http://www.xbox.com/kinect
[3]http://face.nist.gov/frgc/

978-1-4577-0653-0/11/$26.00 ©2011 IEEE

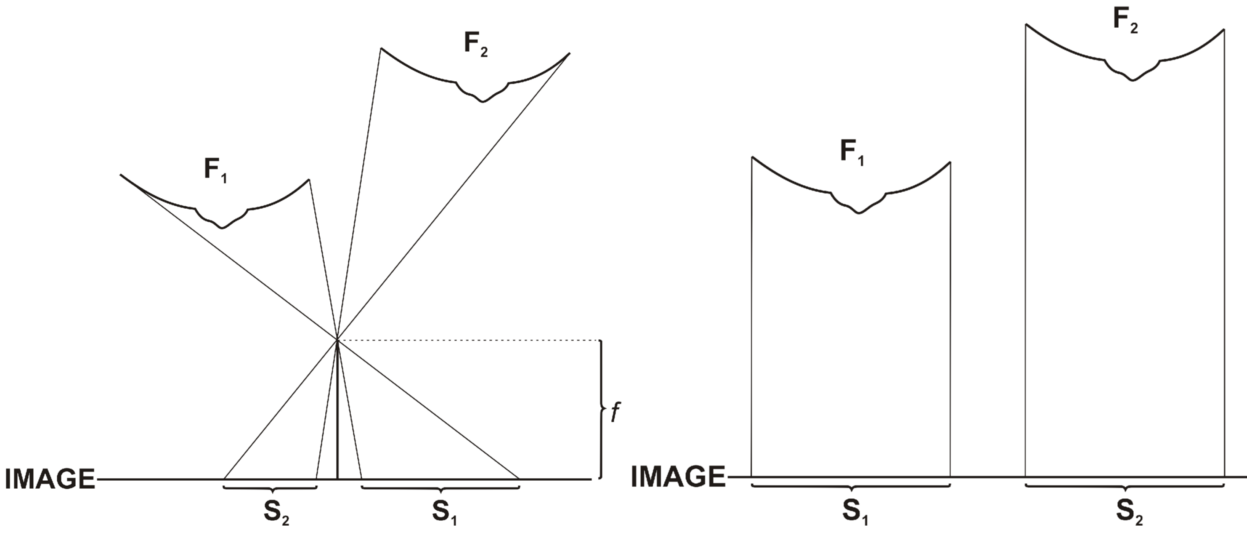


Fig. 1. Illustration of the perspective projection influence in the face size (a) when compared with direct projection (b).

II. 3D Face Detection using Boosted Cascade Classifiers

A. Face Detection using Boosted Cascade Classifiers

There are two main stages in a face detection system: training and detection. In the training stage, a classifier is designed to distinguish between face and non-face images. Then, the obtained classifier is applied to classify image regions as face or non-face in the detection stage.

The technique proposed by Viola and Jones [4] is based on very simple features known as Haar features, which are rectangular masks with different size, shape and location. When a Haar mask is applied over images of a same pattern, such as faces, their values tend to be close. However, a single feature is not enough to discriminate complex patterns, so several features are combined in order to obtain a strong classifier. To find discriminative features, a training set containing face and non-face samples is required. All possibilities of Haar features are tested using this training set, and the Adaboost algorithm [19] is employed to select the most discriminative features.

Since almost all sub-regions of a regular image are non-face sub-regions, the features selected for face detection are divided in weak classifiers and organized in cascade to speed up the detection process. The cascade is organized in a way that the first weak classifiers contain fewer features than the following ones. To be selected as a face, an image subregion must be accepted by all weak classifiers; otherwise it will be considered as non-face.

The cascade classifier is then used to detect faces in an input image by scanning the image with a sub-window and applying the obtained classifier on it. To detect faces in any scale, the sub-window and the features in the cascade classifier are scaled in small steps until they reach the input image size.

B. Related Works

Böhme *et al.* [12] and Fischer *et al.* [13] employed the previous technique to detect faces in range images. In both works distinct cascade classifiers are created for each intensity and range images and combined in different ways to improve detection results. There were no changes in the Viola and Jones' technique, they just trained a classifier and performed face detection using different input information.

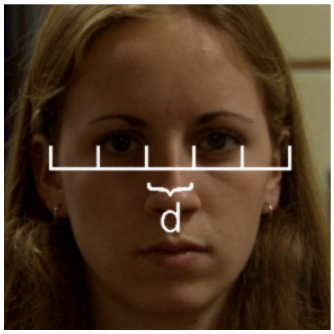
In this work, we do not just use range images as additional information to the original detection technique. We also used such information to improve the detection process in a way that it is only possible when the range data is available. We propose a scale-invariant representation of the range image that eliminates the need for scanning an input image in multiple scales. By doing so, it considerably speeds up the detection stage. We also use the range data to adjust and detect faces with pose variation using a single cascade classifier.

C. Improving Face Detection

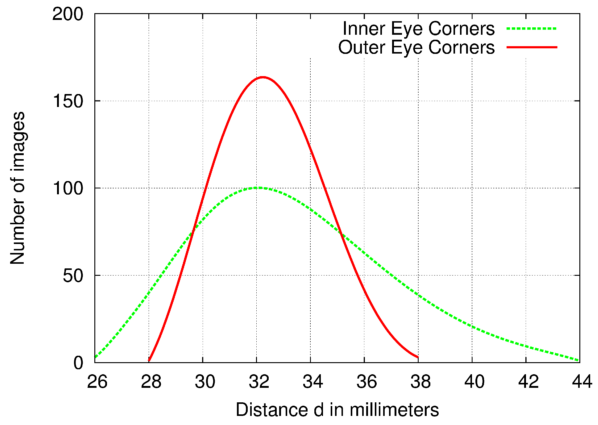
As shown in Figure 1, the camera perspective changes the real size of the objects. Two frontal faces $\mathbf{F}_1$ and $\mathbf{F}_2$ with the same size are shown in Figure 1(a), and their respective sizes $\mathbf{S}_1$ and $\mathbf{S}_2$ in the **IMAGE** are different when captured using a focal length **f**. For this reason, even knowing that faces have similar sizes in the real world, a face detector must scan an input image with different sub-window sizes to be able to detect all faces contained on it.

As a way to solve this problem, we build a projection image that maps X and Y axes of the range data into the columns and rows of the projection, respectively, as illustrated in the Figure 1(b). As such, objects of similar sizes in the real world have similar sizes in the projection image. So there will be no need for scanning the projection image looking for faces in multiple scales.

The first step to estimate the face size in a projection image is to determine the face size in the real world. As it may be seen in Figure 2(a), the face size is about five times the distance **d** between inner eye corners. We computed the distance value **d** for all images in the training set of the FRGC database using inner eye corners and outer eye corners, as shown in Figure 2(b), and in both cases the mean value of **d** is about 33 millimeters. So, the mean face size is about 165 millimeters. This value is not critical and may be applied to other databases, since it is an attribute of the human face and is not peculiar to the FRGC database.



(a)



(b)

Fig. 2. Illustration of the face size employed in this work (a) and the histogram of the distance between inner and outer eye corners of all FRGC training images (b).

The second step is to stipulate the resolution value. The face size in the projection image is proportional to the resolution value. Thus stipulating the resolution means choosing the desired face size in the projection images. In this work we set the resolution value in a way that all faces in the projection image have the size of the faces used for training. In the literature [4], [20], efficient cascade classifiers were successfully obtained during training stage using face images with sizes ranging from 18×18 to 24×24 pixels, so we used face images with 21×21 pixels. A face with 165 millimeters projected in a 21×21 image results in a resolution value r close to 0.13 pixels per millimeter.

The projection image g is then created by mapping each pixel p in the input image into the projection image using the following Equations 1-3:

$$row_p = \frac{\max Y - Y_p}{r} \tag{1}$$

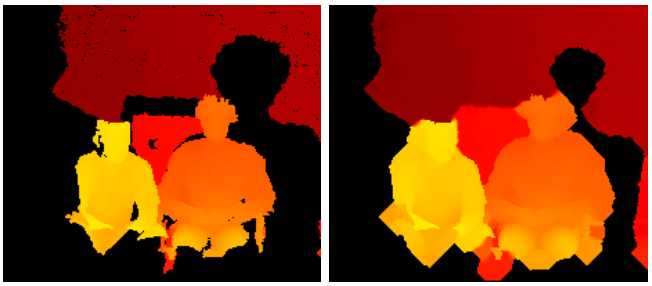
$$column_p = \frac{X_p - \min X}{r} \tag{2}$$

$$g(column_p, row_p) = \frac{Z_p - \min Z}{r} \tag{3}$$

where X_p, Y_p and Z_p are the X, Y and Z coordinates of p, $\max Y$ is the maximum Y value allowed in g, and $\min X$ and $\min Z$ are the minimum X and Z values allowed in g.

By applying Equations 1-3 to all pixels of an input image, we obtain a projection which may contain holes that may substantially affect the detection results, as reported by Colombo *et al.* [11]. To fill the holes, we assign the nearest neighbor pixel value to blank pixels which are close to regions with valid data, as shown in Figure 3. We also apply a 3×3 mean filter in the resulting projection to reduce the noise.

(a)



(b) (c)

Fig. 3. Original range image (a) and the resulting projection image before (b) and after (c) hole filling.

The detection stage is performed by scanning the projection image using a 21×21 sub-window. In order to deal with multiple detections, we perform two steps similar to [4]. First all detections are divided into disjoint groups of detections by analyzing the overlapping area between detections. Then we compute the median location for each group and use it as the final detection of this group.

D. Detecting Faces over Pose Variation

One of the advantages of using range data for face detection is the possibility of using pose adjustments in the detection process. To this end, we create other projection images (**IMAGE'** in Figure 4) with different viewpoints for each input range image. Then, we apply the frontal face cascade classifier to all projection images and deal with detections in different projection images as being multiple detections.

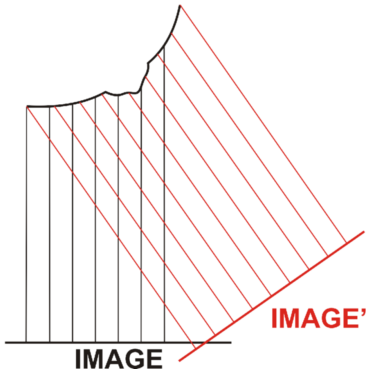


Fig. 4. Illustration of two projection images **IMAGE** and **IMAGE'** with different viewpoints.

The same scheme may be used to improve face detection for frontal faces only. In order to achieve that, projection images with slightly different viewpoints are created in order to overcome small pose variations.

III. Experimental Results

We used the FRGC and BU-3DFE databases to evaluate the performance of our face detection approach. The FRGC database is divided into training and validation sets containing 943 and 4,007 images from 275 and 466 subjects, respectively. Images have the size of 640×480, with an average of about 97,000 valid points and about two points per millimeter. Eye corners were manually marked in all FRGC images to provide a ground truth location of the face to be able to extract faces of the training set and to evaluate the detection results of the validation set. The BU-3DFE database contains 2,500 images from 100 subjects, and its face images present high facial expression variations. This database provides surface meshes with 20,000 to 35,000 polygons and the ground truth location for the eye corners, and the models are not necessarily frontally posed.

In this work, a detected face is considered correct if its location error is smaller than 15% of the face size. This error threshold, which is empirically defined, is employed to guarantee that both eyes are inside the detected face square, as may be seen in Figure 2.

We compared our detection approach against the Viola and Jones' detector available in the OpenCV Library[4] because it has been employed or suggested by 2D and 3D face recognition works in the literature [1], [6], [15]. This baseline approach is the state-of-the-art for face detection on intensity images with the best cost-benefit regarding detection accuracy, false detections and computation time. The FRGC database provide color and range images, so we used the range images to evaluate the proposed approach and their respective color images to evaluate the baseline approach[5].

A. Training results

The same set of Haar features used for training in the baseline approach [20] was employed in this work. The cascade classifier was trained using 943 faces extracted from the FRGC training set as face samples. Other image parts from this same set and a set of 3,019 intensity non-face images[6] are used as non-face samples. The intensity images were included as non-face samples to make the classifier robust against unknown patterns.

The target detection and false-positive rates were 0.999 and 10^{-6} respectively, and the training stage reached the target after adding only six weak classifiers to the cascade classifier, with a total of 32 Haar features. The small size of the obtained cascade classifier also contributes to a gain in speed, and it is achieved because range images are more invariant (*i.e.* illumination and pose) than intensity images.

B. Detection results

In our first detection experiment, we used the FRGC validation set to compare the proposed approach against the baseline approach. As presented in Table I, we obtained an equivalent detection rate even with a considerably lower false detection rate and a computation time of 97% faster.

TABLE I
Face detection results for the FRGC validation set using the baseline approach and the original and enhanced versions of the proposed approach.

Method	Det. rate	False det. rate	Time (sec)
Baseline	99.4%	13.5%	0.295
Proposed approach	99.3%	0.0%	0.008
Enhanced approach	99.9%	0.02%	0.038

The huge difference between computation time of baseline and proposed approaches is due to two reasons. The first one is the cascade classifier, which is considerably simpler when range images are used for training. The second reason is the optimization of the scanning stage, since the baseline approach has to include the scale parameters during the search, while the dimensions of the object to be detected are already known in the proposed approach.

As far as complexity is concerned, the scanning stage for an image with N pixels is $O(N\sqrt{N})$ in the baseline approach (*i.e.* each scan is $O(N)$ and it is done for up to $\sqrt{N}$ different scales). In the proposed approach, the complexity is only $O(M)$, where M is the amount of pixels in the projection

[4]http://opencv.willowgarage.com/wiki/FaceDetection

[5]Results were obtained with the following configuration: cascade = haarcascade_frontalface_alt.xml; scale_factor = 1.1; min_neighbors = 2; min_size = 40×40; flags = CV_HAAR_DO_CANNY_PRUNING

[6]http://tutorial-haartraining.googlecode.com/svn/trunk/data/negatives/

image. The creation of the projection image is $O(N)$, so the final complexity of the proposed approach is $O(M+N)$, which is still less complex than the baseline approach.

After that, we tested the scheme proposed in subsection II-D to ease small pose variations when detecting frontal faces. In this enhanced version of the proposed approach, five projections are created for each input image and then used for face detection: one projection using the original viewpoint, and four new projections using viewpoints with -5 and $+5$ degrees of inclination in Y and Z axes. X-axis variations were not considered because they do not considerably affect the projections (*i.e.* face symmetry is preserved). The results for this enhanced approach (Table I) show an improvement in detection rate, false detection rate and computation time in comparison to the baseline approach.

Then we evaluated the performance of the proposed approach using large pose variations. To this end, we applied a random rotation in the Y-axis between -45 and $+45$ degrees in 200 images of the validation set and removed self-occluded points, as exemplified in Figure 5. Then, we employed the same scheme of using different viewpoints to overcome the pose variation problem. In this case, we created nine projections of the original image using -40, -30, -20, -10, 0, $+10$, $+20$, $+30$ and $+40$ degrees of inclination in the Y-axis. In this experiment we were able to detect the face in all images and there were no false positives either, with an average time of 0.062 seconds.

(a) (b) (c)

Fig. 5. FRGC image rotation: original image (a), rotated image (b) and rotated image in the original pose (c).

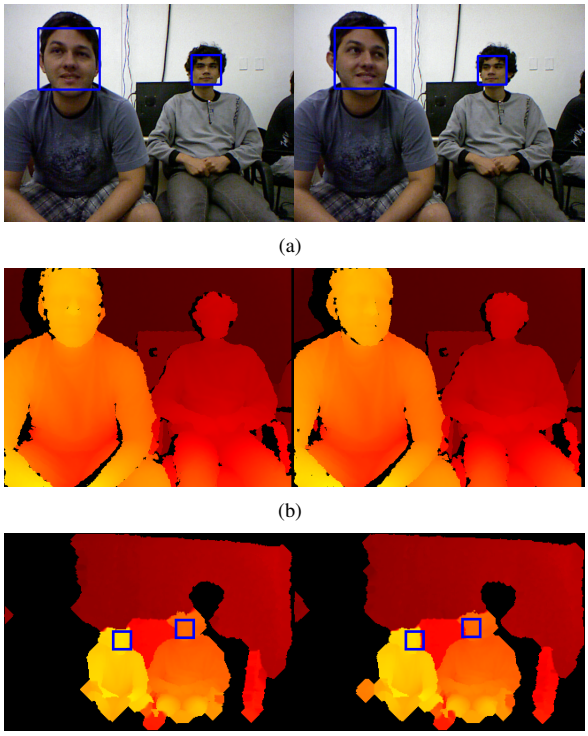
As it may be noticed, we obtained detection rates with rotated face images as high as with frontal face images. As a result, we concluded that it is possible to detect faces presenting large pose variations using a single cascade classifier. Viewpoints with inclination higher than 45 degrees were not considered because almost half of the face is self-occluded and will require further handling.

Since the pose variation of the BU-3DFE models is unknown, we used the proposed approach to detect faces presenting variations in multiple axes in our final experiment. We considered inclinations of -10, 0 and $+10$ degrees in all axes, totaling 27 different viewpoints. The results are shown in Table II, and we obtained a considerable improvement in the detection rate when using the enhanced version to deal with pose variation. This experiment shows that the proposed approach may be used to detect faces in arbitrary poses (*i.e.* if there is no self-occlusion issues). Our detector may also be employed to head pose estimation by saving the viewpoint each face was detected.

TABLE II
FACE DETECTION RESULTS FOR THE BU-3DFE DATABASE USING ORIGINAL AND ENHANCED VERSIONS OF THE PROPOSED APPROACH.

Method	Det. rate	False det. rate	Time (sec)
Proposed approach	21.5%	0.1%	0.007
Enhanced approach	99.9%	0.1%	0.180

Finally, our face detection approach may be employed to images acquired from different devices with no major changes, perhaps requiring additional background range images to create a stronger classifier. Figure 6 shows an example of applying the obtained cascade classifier to detect faces in images acquired by the Kinect sensor. As shown, there are faces with different sizes in the intensity image, and they have the same size in the projection image.



(a)

(b)

(c)

Fig. 6. Examples of color (a) and range images (b) acquired by the Kinect sensor, and results of face detection using the baseline approach (a) and the proposed approach (c).

The experiments were performed in an Intel Core 2 Duo 2.40GHz processor, and the computation time of all variations of the proposed approach allows its use in real-time applications. Since the enhanced versions consist of multiple runnings of the original approach, they may be parallelized on multiple processors. The projection image creation and face detection stages may also be parallelized on graphics processing units [21], and the time required for our detectorto be performed may be considerably reduced.

IV. Conclusion

We have presented our approach for face detection using boosted cascade classifiers of Haar features, by adapting a well-known approach for 2D face detection [4] to range images. We improved the detection stage by eliminating the scale parameter during the search, and also proposed a scheme for detecting faces presenting large pose variations.

We used the FRGC and BU-3DFE databases in our experiments to evaluate the proposed approach. Although none of the two databases contain images with multiple faces or complex background patterns, they help to prove the concept of using range information to detect faces in a faster and more reliable way. We also used images acquired by the Kinect sensor to show that our detector works with background data and multiple faces.

We compared our results against a state-of-the-art face detector and obtained equivalent detection rates. However, we obtained a considerably smaller amount of false detections and a 97% faster computation time. We also used range information to adjust small pose variations and obtained better face detection rates. Then we used pose adjustment to deal with images presenting large pose variation, and obtained a detection rate as high as when using frontal face images.

The computation time of all variations of the proposed approach allows their use in real-time applications. This detector may also be employed to detect other patterns in range images, such as facial features, the upper body or even the full body, only requiring estimates of the object dimensions in advance.

Finally, our object detector using range images presents significant advantages: 1) it is faster, since input images are no longer scanned in multiple scales; 2) it is more reliable, since range images are more invariant than intensity images and a single scale scan reduces the amount of non-object image parts to be tested; 3) it is robust to pose variation; 4) it may be used to pose estimation; and 5) it is highly parallelizable.

Acknowledgment

The authors would like to thank Dr. J. Phillips and Dr. P. Flynn for allowing them to use the FRGC images, Dr. L. Yin and The Research Foundation of State University of New York for allowing them to use the BU-3DFE images, and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Financiadora de Estudos e Projetos (FINEP), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for the financial support.

References

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[2] K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition," *Computer Vision and Image Understanding*, vol. 101, no. 1, pp. 1–15, 2006.

[3] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[4] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[5] K. I. Chang, K. W. Bowyer, and P. J. Flynn, "Multiple nose region matching for 3d face recognition under varying facial expression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1695–1700, 2006.

[6] X. Lu and A. K. Jain, "Multimodal facial feature extraction for automatic 3d face recognition," Department of Computer Science, Michigan State University, Tech. Rep., 2005.

[7] Y. Wang, J. Liu, and X. Tang, "Robust 3d face recognition by local shape difference boosting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1858–1870, 2010.

[8] X. Lu and A. K. Jain, "Automatic feature extraction for multiview 3d face recognition," in *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, 2006, pp. 585–590.

[9] M. Pamplona Segundo, L. Silva, O. R. P. Bellon, and C. C. Queirolo, "Automatic face segmentation and facial landmark detection in range images," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 5, pp. 1319–1330, 2010.

[10] F. Tsalakanidou, S. Malassiotis, and M. G. Strintzis, "Face localization and authentication using color and depth images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 152–168, 2005.

[11] A. Colombo, C. Cusano, and R. Schettini, "3d face detection using curvature analysis," *Pattern Recognition*, vol. 39, no. 3, pp. 444–455, 2006.

[12] M. Böhme, M. Haker, K. Riemer, T. Martinetz, and E. Barth, "Face detection using a time-of-flight camera," *Lecture Notes in Computer Science*, vol. 5742, pp. 167–176, 2009.

[13] J. Fischer, D. Seitz, and A. Verl, "Face detection using 3-d time-of-flight and colour cameras," in *Proceedings of the 41st International Symposium on Robotics and 6th German Conference on Robotics*, 2010.

[14] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, 2006, pp. 211–216.

[15] A. Mian, M. Bennamoun, and R. Owens, "An efficient multimodal 2d-3d hybrid approach to automatic face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 11, pp. 1927–1943, 2007.

[16] H. Tang and T. S. Huang, "3D facial expression recognition based on properties of line segments connecting facial feature points," in *8th IEEE International Conference on Automatic Face and Gesture Recognition*, 2008, pp. 1–6.

[17] I. Kakadiaris, G. Passalis, G. Toderici, M. Murtuza, Y. Lu, N. Karampatziakis, and T. Theoharis, "Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 640–649, 2007.

[18] T. Faltemier, K. W. Bowyer, and P. J. Flynn, "Using a multi-instance enrollment representation to improve 3d face recognition," in *Proc. of the IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2007, pp. 1–6.

[19] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Proceedings of the Second European Conference on Computational Learning Theory*, 1995, pp. 23–37.

[20] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Proceedings of the International Conference on Image Processing*, vol. 1, 2002, pp. 900–903.

[21] H. Ghorayeb, B. Steux, and C. Laurgeau, "Boosted algorithms for visual object detection on graphics processing units," in *Computer Vision – ACCV 2006*, ser. Lecture Notes in Computer Science, 2006, vol. 3852, pp. 254–263.