Applied Computing and Informatics (2014) **10**, 1–13

King Saud University
Applied Computing and Informatics
www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

Effective and precise face detection based on color and depth data

Loris Nanni [a,*], **Alessandra Lumini** [b], **Fabio Dominio** [a], **Pietro Zanuttigh** [a]

[a] *DEI, University of Padova, via Gradenigo, 6, 35131 Padova, Italy*
[b] *DISI, University of Bologna, via Venezia, 52, 47521 Cesena, Italy*

Received 3 February 2014; revised 31 March 2014; accepted 7 April 2014
Available online 21 April 2014

KEYWORDS

Face detection;
Skin detection;
Depth map;
Viola–jones detector

Abstract In this work an effective face detector based on the well-known Viola–Jones algorithm is proposed. A common issue in face detection is that for maximizing the face detection rate a low threshold is used for classifying as face an input image, but at the same time using a low threshold drastically increases the number of false positives. In this paper several criteria are proposed for reducing false positives: (i) a skin detection step is used to reject a candidate face region that does not contain the skin color, (ii) the size of the candidate face region is calculated according to the depth data, removing the too small or the too large faces, (iii) images of flat objects (e.g. candidate face found in a wall) or uneven objects (e.g. candidate face found in the leaves of a tree) are removed using the depth map and a segmentation approach based both on color and depth data.

The above criteria permit to drastically reduce the number of false positives without decreasing the detection rate. The proposed approach has been validated on three datasets composed of 180 samples including both 2D and depth images. The face position inside samples has been manually labeled for testing.

* Corresponding author. Tel.: +39 3493511673.
E-mail address: nanni@dei.unipd.it (L. Nanni).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.aci.2014.04.001

A Matlab version of the system for face detection and the full testing dataset will be freely available from http://www.dei.unipd.it/node/2357.

1. Introduction

Face detection has attracted the attention of many research groups due to its widespread application in many fields as surveillance and security systems, as human–computer interface, face tagging, behavioral analysis, content-based image and video indexing, and many others (Zeng et al., 2009). Face detection is the first crucial step for facial analysis algorithms (i.e. face recognition/verification, head tracking, and facial expression recognition) whose goal is to determine whether or not faces are present in an image and eventually return their location and extent (i.e. a bounding box). It is a more challenging problem than face localization in which a single face is assumed to be inside an image.

Most of the literature in this field deals with frontal face detection from two-dimensional (2D) images: the problem is often formulated as a two-class pattern recognition problem aimed at classifying each sub-window of a given size of the input image as either containing or not containing a face (Jin et al., 2004). Then the classification is performed by common technologies for 2D facial recognition such as Eigenface, Fisherface, waveletface, PCA (Principal Component Analysis), LDA (Linear Discriminant Analysis), Haar wavelet transform, and so on. The Viola–Jones detector (Viola and Jones, 2001) is probably the most famous approach for frontal 2D detection: it involves exhaustively searching an entire image for faces, with multiple scales explored at each pixel using Haar-like rectangle features boosting classification. Two different face detection strategies based on slightly modified Viola–Jones are proposed in Anisetti (2009). In Küblbeck and Ernst, 2006 boosting has also been used in conjunction with Modified census transform (MCT), to improve illumination invariance. In (Huang et al., 2007) a method able to detect faces with arbitrary rotation-in-plane and rotation–off-plane angles in still images or video sequences is proposed. In (Jianxin et al., 2008) is designed a classifier that explicitly addresses the difficulties caused by the asymmetric learning goal (the minority class is the face class).

Despite the success of these and of several other methods, designed to provide accurate detection performance under variable conditions (Zhang and Zhang, 2010), most of difficulties in precise face detection still arises in the presence of illumination changes and occlusions. One possible way to improve algorithms for face detection is to incorporate models of the image processing that efficiently integrates multiple cues, such as stereo disparity, texture and motion.

For example, Microsoft Kinect is a depth sensing device that couples the 2D RGB image with a depth map (RGB-D) which can be used to determine the depth of every object in the scene. Each pixel in Kinect's depth map has a value

indicating the relative distance of that pixel from the sensor at the time of image capture. Depth information captured by Kinect is not useful to differentiate among different individuals at a distance, due to its very high inter-class similarity, but thanks to its low intra-class variation may be useful to improve the robustness of a face detector by reducing sensitivities to illumination, occlusions, changing of expression and pose. Kinect devices have been extremely popular recently, due to their low-cost and availability, and the first benchmark datasets have been collected for 3D face recognition (Tsalakanidou et al., 2003) or detection (Hg et al., 2012).

Several recent approaches use depth map or other 3D information for face detection. For instance, the classic Viola–Jones face detection algorithm is extended in (Dixon et al., 2007; Burgin et al., 2011) to simultaneously consider depth and color information in face detection. In (Shieh and Hsieh, 2013) Haar wavelets on 2D are first used to detect the human face and then its position is refined by structured light analysis. Other depth-based detectors are proposed in Shotton et al. (2011), Mattheij et al. (2012): the approach by Shotton et al. (2011) employs depth-comparison features defined as pixel pairs in depth images, to quickly and accurately classify body joints and parts from single depth images, a similar method based on square regions comparison is coupled to Viola Jones face detector in Mattheij et al. (2012) for robust and accurate face detection. In (Jiang et al., 2013) biologically inspired integrated representation of texture and stereo disparity information is used for a multi-view facial detection task with the valuable result of improved detection performance and reduced computational complexity. Disparity information extracted from stereo images allows to strongly reduce the number of locations to be evaluated during the search process. In (Goswami et al., 2013) the authors use the additional information obtained by the depth map to improve face recognition: their approach extracts textural descriptors (the Histogram of Oriented Gradients) from four entropy maps corresponding to RGB and depth information with varying patch sizes and uses Random Forest as a classifier. Another face recognition approach designed specifically for low resolution 3D sensors is proposed in Li et al. (2013), which uses efficient Iterative Closest Point method and facial symmetry for estimating a canonical frontal view from non-frontal views.

This work, similar to other approaches proposed in the literature (Anisetti et al., 2008), is aimed at using depth information to reduce the number of false positive detections and improve the percentage of correct detections. In (Anisetti et al., 2008) the authors use a 2D multi-step algorithm to obtain a coarse-to-fine classification, then refine the quality of the face location by a 3D tracking approach.

In this paper an effective and precise face detector designed for upright and frontal faces is presented based on both gray-level image and depth map: depth information is used to filter the regions of the image where a candidate face region is found by the Viola–Jones (VJ) detector Viola and Jones, 2001. A main drawback of VJ is that several false positives occur setting a low threshold for face

classification; in this work several criteria, mainly evaluated on the depth map, are used to drastically reduce the number of false positives:

- the first filtering rule is defined on the color of the region; since some false positives have colors not compatible with the face (e.g. shadows on jeans) a skin detector is applied to remove the candidate face regions that do not contain skin pixels;
- the second filtering rule is defined on the size of the face: using the depth map it is quite easy to calculate the size of the candidate face region, which is useful to discard smallest and largest faces from the final result set;
- the third filtering rule is defined on the depth map to discard flat objects (e.g. candidate faces found in a wall) or uneven objects (e.g. candidate face found in the leaves of a tree). Combining color and depth data the candidate face region can be extracted from the background and measures of depth and regularity are used for filtering out false positives.

Unfortunately, in the literature there are no freely available large datasets for face detection (with difficult images as complex background, the datasets used in (Shieh and Hsieh, 2013; Mattheij et al., 2012) are quite easy) that contains both the color and the depth map. There are several datasets for face recognition using the depth map, but the face detection step in those datasets is easy. Therefore, the proposed approach has been evaluated on a collected dataset that will be made freely available for further comparisons.

2. Base face detector

The proposed method is based on the widely used VJ face detector (Viola and Jones, 2001), characterized by a slow training, but very fast classification. VJ involves a very simple image representation, based on Haar wavelets, an integral image for rapid feature detection, the AdaBoost machine-learning method for selecting a small number of important features, and a cascade combination of weak learners for classification. The detection performance of VJ strictly relies on the threshold used to classify an input region as face, this value defines the criteria to declare a final face detection in an area where there are multiple detections around an object. Groups of candidate face regions that meet the threshold are merged to produce one bounding box around the target object. Increasing this threshold may help suppress false detections by requiring that the target object be detected multiple times during the multiscale detection phase. Since the original VJ implementation is designed for upright frontal image, in order to handle non upright faces in the work, the original images are also rotated of $\{20°, -20°\}$ before detection.

In the complete system, outlined in Figure 1, first the VJ detector is applied to input and rotated images using a low classification threshold, then all the candidate face regions are filtered out according to three criteria (detailed in the following sub-sections) with the aim of reducing false positives:

- skin detection;
- size of the image;
- flatness/unevenness of the image.

Figure 2 shows the result of the three filtering steps on two sample images.

Depth data acquired by the Kinect are projected over the color images containing the faces in order to obtain a set of aligned color images and depth maps. For this purpose, the calibration data for the depth and color cameras of the Kinect are computed using the method proposed in Herrera et al. (2012). This approach computes both the intrinsic parameters of the depth and color cameras and the extrinsic parameters between the two cameras. The depth samples positions of the image in the 3D space are first computed by using the intrinsic parameters of the depth camera and the 3D samples are then reprojected in the 2D color image reference system by using both the color camera intrinsic parameters and the extrinsic ones. By the end of this procedure, a color and a depth value are associated to each sample.

2.1. Skin detection filter

The presence of skin is a good indicator to assert the detection of a face. In this work a skin filter is applied to candidate face regions which is a simplified version of the ensemble proposed in Nanni et al. (2013). It is the combination by the sum rule of the methods proposed in Jones (2002), Ó Conaire et al. (2007) with three other skin detectors based on the idea proposed in Khan et al. (2011); after the

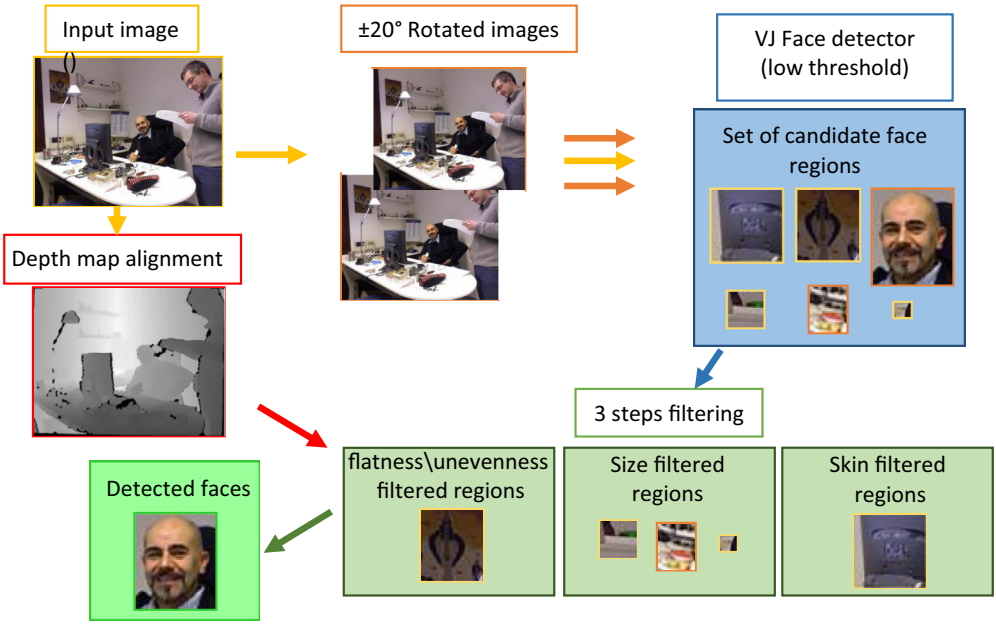

Figure 1 Outline of our complete system.

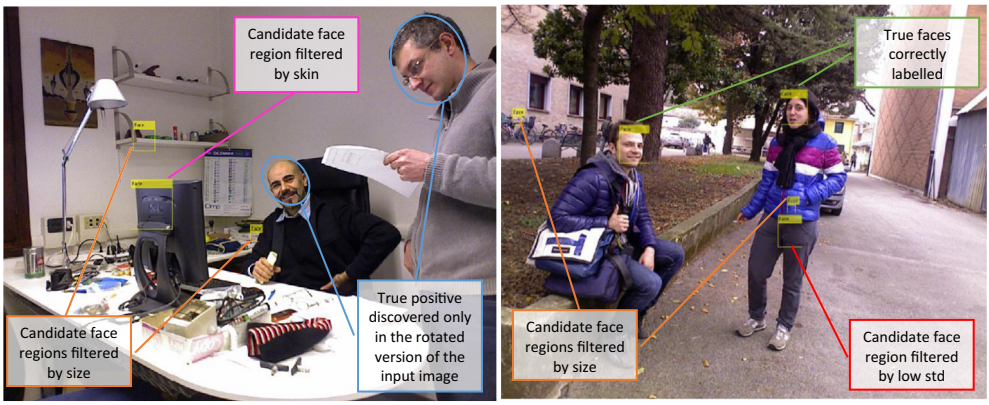

Figure 2 Result of the filtering steps on some sample images.

training phase performed according to the above cited approaches, pixels are classified by a set of lookup tables, built using SVMs trained considering different features (i.e., different pre-processing and color spaces):

- max-RGB color constancy and RGB color space;
- max-RGB color constancy and YUV[1] color space;
- RGB color space.

The color constancy problem is the ability to estimate the unknown light of a scene from an image. The max-RGB color constancy approach is based on the assumption that the reflectance which is achieved for each of the three color channels is equal (van de Weijer et al., 2007).

Since lookup tables[2] are used for the classification task this method allows to classify skin pixels of a given image in real time. Please note that since lookup tables have been calculated from several large datasets (Nanni et al., 2013), the system is not over-trained in the dataset tested in this paper.

2.2. Filter by the size of the image

The size criteria simply remove the candidate faces not included in a fixed range size ([12.5,30] cm). The size of a candidate face region is extracted from the depth map according to the following approach.

Assuming that the face detection algorithm returns the 2D position and dimension in pixels (w_{2D}, h_{2D}) of a candidate face region, its 3D physical dimension in mm (w_{3D}, h_{3D}) can be estimated as:

[1] The color space conversion is performed using the "Colorspace" Matlab toolbox http://www.mathworks.com/matlabcentral/fileexchange/28790-colorspace-transformations.

[2] A lookup table is the result of the pre-calculation by SVM of classification scores of all the combinations of pixel values (2563) from a given color space.

$$w_{3D} = w_{2D}\frac{\bar{d}}{f_x}$$
$$h_{3D} = h_{2D}\frac{\bar{d}}{f_y}$$

where f_x and f_y are the Kinect camera focal lengths computed by the calibration algorithm of Herrera et al. (2012) and $\bar{d}$ is the average depth of the samples within the face candidate bounding box. Note how $\bar{d}$ is indeed defined as the median of the depth samples in order to reduce the impact of noisy samples in the average computation.

2.3. Filter by flatness/unevenness of the image

Another significant information that can be obtained from the depth map is the flatness/unevenness of the candidate face regions. For this filter first a segmentation procedure is applied, then from each face candidate region the standard deviation (*std*) of the pixels of the depth map that belongs to the larger segment is calculated. Regions having a *std* out of a fixed range [0.15, 4] are removed.

The segmentation of both color and depth map is performed according to the approach of Dal Mutto et al. (2012). This segmentation scheme is based on the normalized cuts spectral clustering algorithm (Shi and Malik, 2000) and jointly exploits the geometry and color information for optimal performances.

The basic architecture of the segmentation scheme is shown in Figure 3. The procedure has two main stages: firstly a six-dimensional representation of the scene samples is built from the geometry and color data and then the obtained point set is segmented using spectral clustering.

Each sample in the acquired depth map corresponds to a 3D point of the scene $p_i, i = 1, \ldots, N$. After the joint calibration of the depth and color cameras it is possible to compute the 3D coordinates x, y and z of p_i and to associate to it a 3-D vector containing the R, G, and B color components. Geometry and color then need to be unified in a meaningful way. Color values are converted to a perceptually uniform space in order to give a perceptual significance to the distance between colors that will be used in the clustering algorithm. The CIELab space has been used for this purpose, i.e., the color information of each scene point is the 3-D vector:

$$p_i^c = \begin{bmatrix} L(p_i) \\ a(p_i) \\ b(p_i) \end{bmatrix}, \; i = 1, \ldots, N$$

The geometry is simply represented by the 3-D coordinates of each point, i.e.:

$$p_i^g = \begin{bmatrix} x(p_i) \\ y(p_i) \\ z(p_i) \end{bmatrix}, \; i = 1, \ldots, N$$

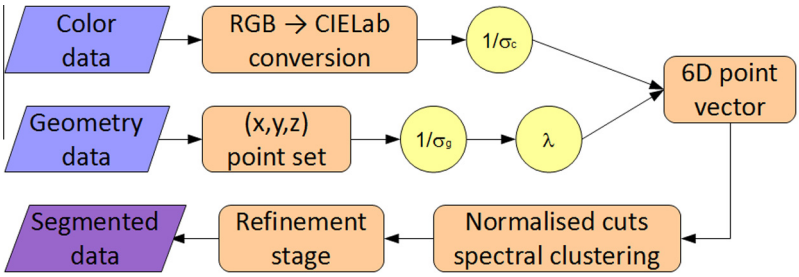

Figure 3 Architecture of the proposed segmentation scheme.

The scene segmentation algorithm should be insensitive to the relative scaling of the point-cloud geometry and should bring geometry and color distances into a consistent framework. Therefore all the components of p_i^g are normalized with respect to. the average of the standard deviations of the point coordinates $\sigma_g = (\sigma_x + \sigma_y + \sigma_z)/3$. The adopted geometry representation is thus the vector:

$$\begin{bmatrix} \bar{x}(p_i) \\ \bar{y}(p_i) \\ \bar{z}(p_i) \end{bmatrix} = \frac{3}{\sigma_x + \sigma_y + \sigma_z} \begin{bmatrix} x(p_i) \\ y(p_i) \\ z(p_i) \end{bmatrix} = \frac{1}{\sigma_g} \begin{bmatrix} x(p_i) \\ y(p_i) \\ z(p_i) \end{bmatrix}$$

In order to balance the relevance of color and geometry in the merging process, the color information vectors are also normalized by the average of the standard deviations of the L, a and b components. The final color representation therefore is:

$$\begin{bmatrix} \bar{L}(p_i) \\ \bar{a}(p_i) \\ \bar{b}(p_i) \end{bmatrix} = \frac{3}{\sigma_L + \sigma_a + \sigma_b} \begin{bmatrix} L(p_i) \\ a(p_i) \\ b(p_i) \end{bmatrix} = \frac{1}{\sigma_c} \begin{bmatrix} L(p_i) \\ a(p_i) \\ b(p_i) \end{bmatrix}$$

From the above normalized geometry and color information vectors, each point is finally represented as

$$p_i^f = \begin{bmatrix} \bar{L}(p_i) \\ \bar{a}(p_i) \\ \bar{b}(p_i) \\ \lambda\bar{x}(p_i) \\ \lambda\bar{y}(p_i) \\ \lambda\bar{z}(p_i) \end{bmatrix}$$

where λ is a parameter balancing the contribution of color and geometry. High values of λ increase the relevance of geometry, while low values of λ increase the relevance of color information. A complete discussion on the effect of this parameter and on how to automatically set it to the optimal value is presented in Dal Mutto et al. (2012).

The computed vectors p_i^f are then clustered in order to segment the acquired scene. Among the various clustering techniques, methods based on pairwise affinity measures computed between all the possible couples of points allows to obtain very accurate and robust results because they do not assume a Gaussian model for the distribution of the points. On the other side they have the drawback that they need to compare all the possible pairs of points and are so very expensive in terms of both CPU and memory resources. Normalized cuts spectral clustering (Shi and Malik, 2000) is an effective example of this family. This method is based on the partition of a graph representing the scene according to spectral graph theory criteria. The minimization is done using normalized cuts and accounts both for the similarity between the pixels inside the same segment and the dissimilarity between the pixels in different segments. The minimization problem is very computationally expensive and several methods have been proposed for its efficient approximation. In the method based on the integral eigenvalue problem proposed in Fowlkes et al. (2004), the set of points is first randomly subsampled and then the subset is partitioned and the solution is propagated to the whole points set by a specific technique called the Nyström method. In order to avoid small regions due to noise a final refinement stage removing regions smaller than a pre-defined threshold is finally applied. In Figure 4 an example of segmented image is reported.

3. Experimental results

The experimental evaluation of the proposed face detection system has been carried out on a dataset composed of three subsets, all containing frontal images:

- Microsoft hand gesture (Ren et al., 2011), it is composed of images of 10 different people performing gestures; each image contains only one face; since the images in the datasets are quite similar to each other we have chosen and labeled for face detection a subset of 42 images.
- Padua Hand gesture (Dominio et al., 2013), it is another gesture dataset composed of images from 10 different people; each image contains only one face; since the images are quite similar we have chosen a subset of 59 images.

Figure 4 Segmentation map, color image and depth map.

Figure 5 Sample images from the dataset which contain faces not detected from the VJ method.

– Padua FaceDec, it is a new dataset collected and labeled for the purpose of this work. It contains 132 images acquired with the Kinect sensor at the University campus in Padova. It includes both outdoor and indoor scenes, framed in different hours during the day, in order to account for the varying lighting conditions. The images capture one or several people performing various daily activities, e.g., working, studying, walking, chatting and so on. Note how most people are not directly looking into the camera, i.e., they did not pose for the frame acquisition but they were doing their activities without being aware of the camera shooting them. Some faces are also partially occluded by objects or other people. For these reasons, this dataset is more challenging than the previous ones.

The three sets have been merged to form a single dataset consisting of 233 images containing 251 faces[3] (only upright frontal faces with a maximum rotation of $\pm 30°$ have been considered). Notice that the parameters of the method have been manually selected and are the same for all the testing images despite their different origin. The dataset is not "easy": in Figure 5 some samples which are not detected by the VJ method[4] are shown.

The aim of the experiment reported in Table 1 is to evaluate the effectiveness of the proposed approach considering the different filtering steps and the use of depth image; the following approaches are compared according to the detection rate (percentage of faces detected), the number of false positives and the *F*-measure[5] evaluated on the whole dataset:

- VJ(k), Viola–Jones detector in the 2D image with a threshold of k;
- VJ(k)-Sz, the above approach filtered considering the size of the candidate face region;

[3] Some images contain more than one face and some no faces.
[4] VJ is executed with a very low recognition threshold ($k = 2$).
[5] The F-measure is the harmonic mean of recall and precision, often used in document retrieval, it is defined as: $2 \times \text{precision} \times \text{recall}/(\text{precision} + \text{recall})$.

Table 1 Comparison of methods in terms of detection rate and number of false positives. Rows corresponding to the optimal setting of the VJ threshold have been highlighted. Bold is the best performance.

	Detection rate (%)	# False positives	*F*-measure
VJ(4)	88.05	193	0.665
VJ(3)	92.03	375	0.539
VJ(2)	95.62	1063	0.308
VJ(1)	95.62	8017	0.056
VJ(3)-Sz	92.03	**123**	0.725
VJ(2)-Sz	95.22	310	0.601
VJ(1)-Sz	95.62	2062	0.189
VJ(2)-SzSk	95.22	196	0.706
VJ(2)-SzSkStd	95.22	165	0.730
VJ(2)-Fin	94.82	143	**0.758**
VJ(2)-Fin-No	94.82	212	0.679

- VJ(k)-SzSk, the base VJ detector filtered considering size and skin;
- VJ(k)-SzSkStd, the base VJ detector filtered considering size, skin and presence of flat/uneven regions (considering the depth map). In this method *std* is calculated on the whole candidate face region (without the segmentation, to reduce the computation time).
- VJ(k)-Fin, the whole approach described in this paper (including segmentation to calculate *std*)
- VJ(k)-Fin-No, as VJ(k)-Fin but without considering the skin filter step.

It is clear that the size is a useful criterion for removing the high number of false positives candidates found by VJ using a low threshold (required to reach a high detection rate). It allows to greatly reduce the number of false positives, the other two filtering criteria allow to further reducing the number of false positives. The proposed approach allows to diminish the number of false positives on the considered dataset from 1063 to just 143 almost without affecting the detection performances.

The depth map allows to remove the false positives in many critical situations. In particular it allows to get the actual size of the candidate face allowing to remove objects too small or too large to be a face. It also aids the segmentation step in the proposed method that is a critical point to ensure proper processing in the remaining steps.

Finally, even if the experiments reported in this paper are related to data acquired by the Kinect, there are several other depth acquisition schemes and sensors that can be exploited. For example stereo vision systems that get the 3D data from two standard cameras, allow to work at big distances by choosing a suitable baseline. Also there is a wide range of 3D sensors that can work at different distances and with different accuracies. The Kinect is one of the most widespread and the cheapest acquisition sensors but not the best.

4. Conclusion

In this work a face detector for frontal faces is proposed. The Viola–Jones face detector is coupled with three heuristic criteria calculated using the depth map with the main goal of obtaining accurate face detection with few false positives.

The proposed system makes use of several criteria for filtering the false positives found by the face detector:

- A skin detection filter is used to remove the candidate face regions that contain enough skin pixels;
- The size of the candidate face is calculated using the depth map to remove regions whose size is out of a fixed range;
- The depth map is used to design a filter rule to discard flat objects (e.g. candidate faces found in a wall) or uneven objects (e.g. candidate face found in the leaves of a tree).

Experimental results show that the proposed system works well in a collected dataset built by color images and depth map.

We are aware that the dataset used for testing is small with respect to those available for 2D face detection, but in our opinion the results clearly confirm that the depth map permits to define criteria for a drastical reduction of the number of false positives. However, it is our intention to collect new images to build a larger dataset.

Future works include collecting a larger dataset and extend the system to deal also with non frontal upright faces. Another future work will be testing different and more performing face detectors, as (Nanni and Lumini, 2012), for reducing the number of false negatives.

References

Anisetti, M., 2009. "Fast and robust Face Detection", Multimedia Techniques for Device and Ambient Intelligence, ISBN: 978-0-387-88776-0, Springer, US (Chapter 3).

Anisetti, M., Bellandi, V., Damiani, E., Arnone, L., Rat, B., 2008. A3FD: accurate 3D face detection. In: Signal Processing for Image Enhancement and Multimedia Processing. Springer, US, pp. 155–165.

Burgin, W., Pantofaru, C., Smart, W.D., 2011."Using depth information to improve face detection". In: Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI '11), pp. 119–120.

Dal Mutto, C., Zanuttigh, P., Cortelazzo, G.M., 2012. "Fusion of geometry and color information for scene segmentation". In: IEEE Journal of Selected Topics in Signal Processing, vol. 6, No. 5, pp. 505–521.

Dixon, M., Heckel, F., Pless, R., Smart, W.D., 2007. Faster and more accurate face detection on mobile robots using geometric constraints. In: IEEE/RSJ International Conference on Robots and Systems (IROS 2007), pp. 1041–1046.

Dominio, F., Donadeo, M., Zanuttigh, P., 2013. Combining multiple depth-based descriptors for hand gesture recognition, Pattern Recogn. Lett., (accepted for publication), available online 24 October 2013.

Fowlkes, C., Belongie, S., Chung, F., Malik, J., 2004. Spectral grouping using the Nyström method. IEEE Trans. Pattern Anal. Mach. Intell. 26 (2), 214–225.

Goswami, G., Bharadwaj, S., Vatsa, M., Singh, R., 2013. On RGB-D face recognition using Kinect. In: IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–6.

Herrera, D., Kannala, J., Heikkilä, J., 2012. Joint depth and color camera calibration with distortion correction. IEEE Trans. Pattern Anal. Mach. Intell. 34, 2058–2782.

Hg, R.I., Jasek, P., Rofidal, C., Nasrollahi, K., Moeslund, T.B., Tranchet, G., 2012. An RGB-D database using Microsoft's Kinect for Windows for face detection. In: Eighth International Conference on Signal Image Technology and Internet Based Systems (SITIS), pp. 42–46.

Huang, Chang, Ai, Haizhou, Li, Yuan, Lao, Shihong, 2007. High-performance rotation invariant multiview face detection. IEEE Trans. Pattern Anal. Mach. Intell., 671–686.

Jiang, F., Fischer, M., Ekenel, H.K., Shi, B.E., 2013. Combining texture and stereo disparity cues for real-time face detection. Sig. Process. 28 (9), 1100–1113.

Wu, Jianxin, Charles Brubaker, S., Mullin, Matthew D., Rehg, James M., 2008. Fast asymmetric learning for cascade face detection. IEEE Trans. Pattern Anal. Mach. Intell. 30, 369–382.

Jin, H.L., Liu, Q.S., Lu, H.Q., 2004. "Face detection using one-class based support vectors". In: Proc. 6th IEEE Int. Conf. Autom. Face Gesture Recog., pp. 457–462.

Jones, M.J. et al, 2002. Statistical color models with application to skin detection. IJCV 46 (1), 81–96.

Khan, R., Hanbury, A., Stöttinger, J., Bais, A., 2011. Color based skin classification. Pattern Recogn. Lett., 157–163.

Küblbeck, C., Ernst, A., 2006. Face detection and track in video sequences using the modified census transformation. Image Vision Comput. 24 (6), 564–572.

Li, B.Y., Mian, A.S., Liu, W., Krishna, A., 2013. Using kinect for face recognition under varying poses, expressions, illumination and disguise. In: IEEE Workshop on Applications of Computer Vision (WACV), pp. 186–192.

Mattheij, R., Postma, E., Van den Hurk, Y., Spronck, P., 2012. Depth-based detection using Haarlike features. In: Proceedings of the BNAIC 2012 conference. Maastricht University, The Netherlands, pp. 162–169.

Nanni Loris, Lumini Alessandra, 2012. "Combining face and eye detectors in a high-performance face-detection system". In: IEEE Multimedia, vol. 19, No. 4, Oct.–Dec. 2012, pp. 20–27. doi:10.1109/MMUL.2011.57.

Nanni, L., Lumini, A., Migliardi, M., 2013. Learning based Skin Classification, submitted to Applied Soft Computing.

Ó Conaire, Ciarán, O'Connor, Noel E., Smeaton, Alan F., 2007. "Detector adaptation by maximising agreement between independent data sources". In: IEEE International Workshop on Object Tracking and Classification Beyond the Visible, Spectrum.

Ren, Z., Meng, J., Yuan. J., 2011. Depth camera based hand gesture recognition and its applications in human–computer-interaction. In: Proc. of ICICS, pp. 1–5.

Shi, J., Malik, J., 2000. Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 22 (8), 888–905.

Shieh, M.Y., Hsieh, T.M., 2013. Fast facial detection by depth map analysis. Math. Prob. Eng., 1–10.

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

Tsalakanidou, F., Tzovaras, D., Strintzis, M.G., 2003. Use of depth and colour Eigenfaces for face recognition. Pattern Recogn. Lett. 24 (910), 1427–1435.

van de Weijer, J., Gevers, T., Gijsenij, A., 2007. Edge-based color constancy. IEEE Trans. Image Process. 16, 2207–2214.

Viola, Paul, Michael, J. Jones, 2001. "Rapid Object Detection using a Boosted Cascade of Simple Features", CVPR.

Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S., 2009. A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. 31 (1), 39–58.

Zhang C., Zhang Z., 2010. "A Survey of Recent Advances in Face Detection", Microsoft Research Technical Report, MSR-TR-2010-66.