Computer Vision for the Blind: a Comparison of Face Detectors in a Relevant Scenario

Marco De Marco, Gianfranco Fenu, Eric Medvet, and Felice Andrea Pellegrino

Department of Engineering and Architecture, University of Trieste, Trieste, Italy
marco.de.marco.ts@gmail.com,{fenu,emedvet,fapellegrino}@units.it

Abstract. Motivated by the aim of developing a vision-based system to assist the social interaction of blind persons, the performance of some face detectors are evaluated. The detectors are applied to manually annotated video sequences acquired by blind persons with a glass-mounted camera and a necklace-mounted one. The sequences are relevant to the specific application and demonstrate to be challenging for all the considered detectors. A further analysis is performed to reveal how the performance is affected by some features such as occlusion, rotations, size and position of the face within the frame.

Key words: face detection, video sequences, blindness, comparison

1 Introduction and related work

In recent years, many contributions have been proposed as "smart" assistants for blind people, with the aim to assist them in the everyday life and in the social interaction. Some of that proposals are based on remote volunteers [1], or on smart devices, as canes [2] or mobile phones [3]. Concerning visual-based smart devices, a challenge consists in building First Person Vision (FPV) systems which can improve the social interactions of visually impaired persons [4]: for instance, such systems should discover the number, the identity, and the emotional state of people around the visually impaired person and communicate that information to him/her. In facts, the *face detection* is the first, essential step of such information flow. Many different approaches have been proposed in the literature, with the aim to detect human faces in images or in still video frames, for different purposes (tracking, recognition, surveillance, safety, human-machine interaction, etc.). The reader may refers to the recent survey [5] and to the references therein to have an overview on this topic. A very debated theme, regarding face detection algorithms, concerns how to compare the performances of different face detection algorithms and how to build proper benchmark datasets (see for instance [6]).

With the aim to offer a benchmark platform for FPV systems assisting blind people, recently a particular video dataset has been collected [7]. An *ad hoc* dataset is needed when a FPV system, or simply a face detection and recognition algorithm, has to be tested for applying in assisting blind people, due some specific features of blind people behavior (for instance mannerism [8]) that are

responsible of disturbances and effects (such as blur, rapidly varying light conditions, occlusions [9]) normally not present in standard video dataset for face detection bench-marking.

Taking into account such features, is it possible to apply a standard face detector to the videos of that specific dataset? What are the performances of well known face detectors, if installed on a wearable smart device, used by a visually impaired person? The present paper deals with this topic, comparing the performances of some classical and some very recent face detectors, applied on the videos introduced in [7] and analyzing the results. In particular, the structure of the paper is as follows: the dataset is described in Section 2, then in Section 3 the face detectors are briefly presented. Section 4 is focusing on the comparison protocol and on the benchmark results, while in Section 5 some remarks and considerations are reported.

2 Dataset

For evaluating the performance of the face detectors, four video sequences have been employed, belonging to the dataset [7]. The mentioned dataset has been acquired specifically for providing realistic sequences for the considered application. More precisely, the sequences have been acquired by a blind person by means of two wearable cameras. Two commercial devices have been used, namely a pair of sunglasses equipped with a camera (*SportXtreme OverLook Gx-9*), and a *Polaroid CUBE* camera, held by a short necklace. The glasses-mounted camera has a resolution of 1280 px × 720 px and field of view of 135°; the resolution of the necklace-mounted camera is 1920 px × 1080 px and its field of view is 124°. Four video sequences, acquired using the different devices in different places (a university library, a coffee shop, an office, the neighborhood of a bus stop) have been selected and manually annotated. The data of the selected sequences, that contain a total of 3699 faces in 1728 frames, are summarized in Table 1.

Table 1: Salient information on the selected sequences.

Name	Resolution	Camera	Location	# frames	# faces
Coffee-shop	1280 × 720	GX9	Indoor	361	809
Library	1280 × 720	GX9	Indoor	361	1074
Office	1920 × 1080	CUBE	Indoor	558	206
Bus-stop	1920 × 1080	CUBE	Outdoor	448	1610

Inspecting the whole dataset some observation can be made [7]:

- faces can be partially occluded, mainly because there is no visual feedback during acquisition;
- the wide angles introduce distortion;

– the scene conditions are very different in the different contexts, and can change abruptly with time, especially the illumination conditions;
– sudden, fast and wide subjective movements occur, especially in the sequences acquired with the glass mounted camera.

As far as the annotation is concerned, each face has been annotated by tracing a rectangle (referred to as *bounding box* in the following). The rectangle is vertically delimited by chin and forehead (normal hairline, independent of the actual presence of hair in the subject), and horizontally delimited by the ears or, for rotated faces, one ear and the opposite foremost point between the tip of the nose and the profile of the cheek. The presence of significant yaw (possibly also with pitch and roll contributions) was denoted by a dedicated flag, set if the farthest eye was not clearly visible. A further flag was set if the face was partially occluded. Beyond the rectangle, the positions of the centers of the two eyes and of the mouth were also annotated. Faces whose resulting bounding box longest size was less than 20 px long were not annotated.

In the following we point out some features of the annotated faces that will be used in Section 4 for a sensitivity analysis:

– normalized bounding box area (NBBA): the ratio between the bounding box area and the frame area;
– normalized distance of the bounding box from the center of the image (NDFC): distance between bounding box center and frame center, divided by the frame circumcircle radius;
– roll angle: the roll angle is estimated as the angle between the x-axis of the frame and the line passing through the eyes;
– root mean square contrast (RMSC) within the bounding box [10], given by $\sqrt{\frac{1}{\#B}\sum_B \left(I_{ij} - \bar{I}\right)^2}$ where I_{ij} is the intensity of the i-th j-th pixel of the bounding box B, $\bar{I}$ is the average intensity of the bounding box and $\#B$ denotes the number of pixel within the bounding box;
– lateral/non-lateral flag (L/NL): a face is labeled as *lateral* when the farthest eye is not clearly distinguishable due to the rotation oh the head w.r.t. the point of view;
– occluded/non-occluded (O/NO): a face is labeled as *occluded* when it is partially occluded.

Each sequence may be characterized by the distribution of the above features related to the faces annotated in the sequence. For space reason, we do not report the distribution of all the features for each sequence. Instead, we fixed a reasonable threshold τ for each of the numeric features (NBBA, NDFC, Roll, and RMSC) and show in Table 2 the percentage of the annotated faces having, for the given feature, a value *below* the threshold: the threshold values (shown in the table) have been chosen manually by looking to the histograms of the occurrences and assuming a bi-modal underlying distribution. For categorical features (L/NL and O/NO), Table 2 shows instead the percentage of annotated faces in which the features assumes the ideal value (non-lateral and non-occluded, respectively) from the point of view of the face detection task.

Table 2: Percentage of annotated faces having a given feature value below the chosen threshold (for numeric features) or equal to the ideal value (for categorical features).

	NBBA	NDFC	Roll	RMSC	L/NL	O/NO
Threshold τ	0.01	0.33	15	0.15	–	–
Coffee-shop	80.34	26.94	86.47	72.93	47.23	76.14
Library	98.13	14.71	67.26	17.51	77.23	98.17
Office	37.86	20.87	95.32	43.69	91.26	92.23
Bus-stop	88.50	52.98	96.58	75.53	94.04	95.71
All sequences	86.70	34.39	87.01	56.34	78.7	92.06

3 Detectors

Viola-Jones The Viola-Jones face detector is well-known and is based on a cascade of simple classifiers. The features are Haar-like and easily computed by means of the integral image [11]. Among the several existing implementations we used the one provided by the Matlab Computer Vision Toolbox.

GMS Vision Google has developed a framework for object detection, integrated within its Google Mobile Service (GMS)[1]. The vision package offers a face detector and a bar-code reader. Because of the limitations imposed by Google about the use of Google Mobile Service, the integration of the face detector GMS Vision within the testing framework was not possible. As a workaround, an Android application was created using an Android emulator environment. Unfortunately Vision libraries available with the Google Mobile Service, do not allow to process a video stream from a file but only from the camera. To overcome this limitation, the OpenCV frame grabber was employed for extracting frames from the video to be analyzed.

NPD The Normalized Pixel Difference (NPD) algorithm [12], is based on the difference to sum ratio between couples of pixels, and uses a decision tree for the learning. A Matlab implementation has been made available by the authors[2].

PICO Pixel intensity comparison is also used in the PICO algorithm [13]. Hence the features are fast to compute and scale independent. The classifier is a random forest with binary decision trees. The full source code has been made available by the authors[3].

Face-Id Face-Id [14] is a face detection framework based on deep learning. It has been developed using Torch Tensor Framework, Lua and C++ languages. The source code was provided by the authors.

[1] `https://developers.google.com`

[2] `http://www.openpr.org.cn/index.php/107-NPD-Face-Detector/View-details.html`

[3] `https://github.com/nenadmarkus/pico`

Visage We used the demo-version of the Face Detect component that belongs to Visage SDK[4]. It is a commercial product that performs the face detection by identifying the facial features in facial images containing one or more human faces. For each detected face it returns the 2D and 3D head pose, 2D and 3D coordinates of facial feature points (chin tip, nose tip, lip corners, etc.), ignored in the present study.

4 Comparison

4.1 Protocol

The above face detectors have been applied to the selected sequences by setting all the parameters to their default values.

Given the output of a detector for a sequence and the corresponding ground truth data, the number of True Positive (TP), False Positive (FP), and False Negative (FN) are determined through the following 3 steps:

1. Calculate the Intersection to Union Areas Ratio (IUAR) index for each pair of a ground truth object and a detection belonging to the same frame of the sequence. For a detection d_i and a ground truth object g_j, $\text{IUAR}(d_i, g_j) = \frac{\text{area}(d_i \cap g_j)}{\text{area}(d_i \cup g_j)}$.
2. Find the best match between ground truth and detections, using Hungarian Algorithm [15]: the best match is the one having the highest cumulative IUAR index.
3. Consider as a True Positive a detection for which the IUAR of the best match is > 0.5, as a False Positive a detection for which the IUAR of the best match is ≤ 0.5, and as a False Negative a ground truth object which is not the best match of any detection.
4. Set TP, FP, and FN, to the counts of True Positives, False Positives, and False Negatives within the sequence.

Given TP, FP, FN, and the number of frames n_f in the sequence, we express the performance of the detector applied to that sequence in terms of three indexes: *precision*, computed as $\frac{\text{TP}}{\text{TP+FP}}$, *recall*, computed as $\frac{\text{TP}}{\text{TP+FN}}$, and *false positive per frame* (FPPF), computed as $\frac{\text{FP}}{n_f}$. Precision and recall are indexes commonly used for assessing the effectiveness in information retrieval tasks, but have also be used in tasks related to computer vision (e.g., image segmentation [16]). The latter index, FPPF, is particularly relevant for the application. Indeed, since the aim of the devised vision-based system is to assist the social interaction of blind persons, it must deliver information "continuously" in time. The number of false detections per frame indicates how frequently, on average, the delivered information is not correct.

[4] `https://visagetechnologies.com/products-and-services/visagesdk/`

4.2 Results and discussion

Table 3 shows the performance for each method and for each sequence, and the average across all the sequences.

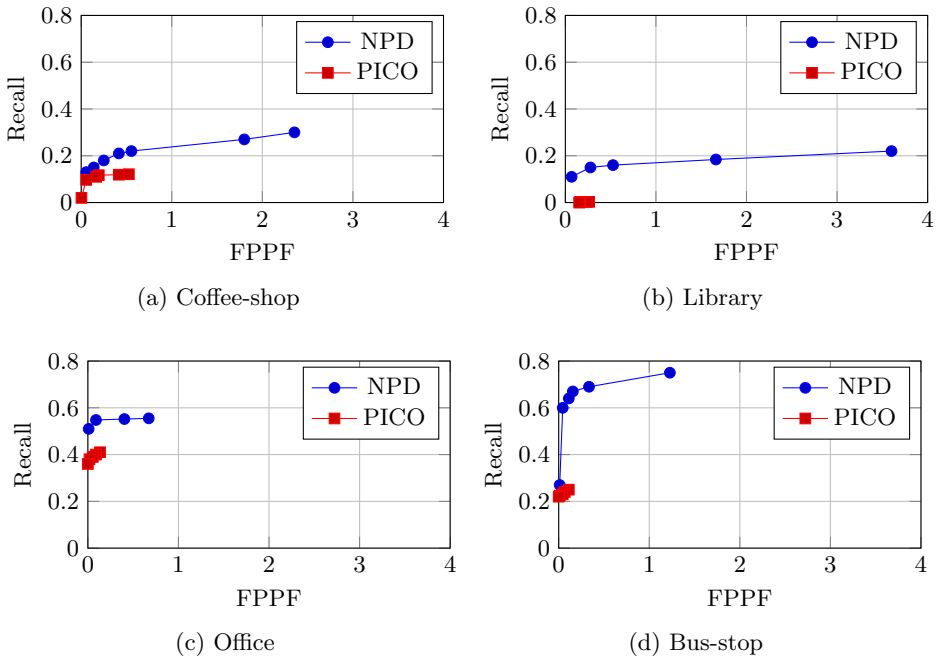
Table 3: Precision, Recall and False Positive Per Frame for the six face detectors and each of the four sequences.

Method	Sequence	Precision	Recall	FPPF
Viola-Jones	Coffee-shop	0.129	0.367	5.543
	Library	0.140	0.267	4.867
	Office	0.031	0.709	8.197
	Bus-stop	0.222	0.725	9.158
	Average	0.132	0.513	7.196
GMS	Coffee-shop	0.364	0.015	0.058
	Library	1.000	0.004	0.000
	Office	0.387	0.141	0.082
	Bus-stop	0.202	0.020	0.290
	Average	0.284	0.021	0.114
NPD	Coffee-shop	0.228	0.305	2.319
	Library	0.159	0.222	3.504
	Office	0.256	0.583	0.625
	Bus-stop	0.687	0.747	1.221
	Average	0.376	0.489	1.735
PICO	Coffee-shop	0.337	0.121	0.535
	Library	0.030	0.003	0.266
	Office	0.538	0.413	0.131
	Bus-stop	0.202	0.020	0.290
	Average	0.589	0.160	0.238
Face-Id	Coffee-shop	0.143	0.001	0.017
	Library	0.889	0.007	0.003
	Office	–	0.0	0.0
	Bus-stop	1.000	0.001	0.000
	Average	0.611	0.003	0.004
Visage	Coffee-shop	0.043	0.002	0.125
	Library	0.045	0.001	0.058
	Office	0.137	0.068	0.158
	Bus-stop	0.072	0.006	0.286
	Average	0.087	0.007	0.163

It is clear by inspecting Table 3 that all the face detectors perform poorly in the considered sequences. The best result seems to be achieved by NPD on the sequence Bus-stop (a recall of 0.747, a precision of 0.687, and 1.221 false positive per frame)—interestingly, this is the only outdoor video sequence. In all

the other cases either the precision or the recall are well below 0.5. It is clear that some detectors, in particular GMS and Face-Id are tuned to avoid false positives (resulting in relatively high precision but low or very low recall) and some other such as Viola-Jones are tuned to avoid false negative (leading to high FPPF). Table 3 shows that the best average recall (0.513) is achieved by Viola-Jones, that leads however to the worst FPFM. The best precision is achieved by Face-Id that leads to the worst recall. The detectors that seem to be tuned halfway between the extrema (such as NPD, resulting in a recall of 0.489 and a precision of 0.376), still do not exhibit a satisfactory performance.

In order to gain deeper insights, we plotted the Receiver operating characteristic (ROC) curves for the detectors PICO and NPD, which are shown in Figure 1 for each of the four sequences. We chose these methods because they provide, along with each detection, a confidence value which can be used to further refine the outcome of the frame processing by discarding the detected objects for which the confidence is low—the other 4 methods do not provide such an information. Figure 1 confirms the results of Table 3 and highlights the obvious trade-off between recall and FPPF.



(a) Coffee-shop (b) Library

(c) Office (d) Bus-stop

Fig. 1: ROC curves of PICO and NPD for the four sequences.

From results shown in Table 3 and Figure 1, it can be argued that the sequences under examinations are particularly challenging. Hence we performed a sensitivity analysis of the performance with respect to some features of the

annotated faces that can possibly explain the poor performance of all the detector. The results of the sensitivity analysis are reported in Table 4, in terms of the average recall achieved by different methods computed on annotated faces having the feature value below or above the reference threshold τ (for numeric features, see Table 2) or equal to a given value (for categorical features)

Table 4: Influence of faces features on recall: average recall achieved by different methods computed on annotated faces having the feature value below or above the reference threshold τ (for numeric features, see Table 2) or equal to a given value (for categorical features).

	NBBA		NDFC		Roll		RMSC		L/NL		O/NO	
Method	$< \tau$	$\geq \tau$	$< \tau$	$\geq \tau$	$< \tau$	$\geq \tau$	$< \tau$	$\geq \tau$	L	NL	O	NO
Face-Id	0.001	0.001	0.001	0.002	0.001	0.000	0.000	0.002	0.001	0.001	0.002	0.000
GMS	0.006	0.039	0.004	0.041	0.041	0.002	0.006	0.039	0.044	0.001	0.043	0.002
NPD	0.304	0.160	0.122	0.342	0.443	0.009	0.277	0.188	0.441	0.024	0.441	0.023
PICO	0.054	0.143	0.046	0.151	0.190	0.005	0.093	0.104	0.196	0.001	0.187	0.010
Viola-Jones	0.364	0.149	0.148	0.366	0.491	0.004	0.325	0.189	0.486	0.027	0.475	0.038
Visage	0.002	0.018	0.002	0.018	0.019	0.000	0.003	0.017	0.019	0.001	0.020	0.000

Regarding the NBBA, it can be observed that some detectors perform better with bigger faces, some the opposite. This is easily explained by the default parameters of each detector[5]. Concerning L/NL, the table shows that all the detectors perform better with non-lateral faces and this is not a surprise, as well as the results obtained for O/NO and Roll, that show that all the detectors perform better with non-occluded faces and with low in-plane rotation (roll angle).

On the contrary, results concerning RMSC and NDFC deserve some comments. Regarding the sensitivity to the RMS contrast, Table 2 shows that the detectors can be divided into two groups: NPD, PICO and Viola-Jones exhibit rather small sensitivity to the contrast, while the remaining detectors perform more poorly at low contrast. A possible explanation is that NPD, PICO, and Viola-Jones are based on ad hoc features consisting of differences of intensity values that are either normalized (NPD), or computed over a normalized candidate window (Viola-Jones) or simply, contribute to the decision function based on the difference sign only; in all the three cases, however, contrast insensitivity is incorporated in the detector, at the level of features. We do not know the details of the other detectors, but we conjecture that none of them is based on contrast-insensitive features.

[5] We did not change the default parameters on purpose, for two reasons: first, some detectors have fixed parameters and second, the choice of default parameters made by the authors of the detector may reflect a compromise between various aspects of performance that we are not aware of.

As far as the normalized distance from center (NDFC) is considered, Table 2 shows that all the detectors perform better with off-center faces. This is surprising because off-center faces undergo major distortions to to wide angle optics and one would expect the opposite result. By looking to the sequences, one may argue that this result is due to the different size of the faces close to the center w.r.t. that far from the center. However, since the detectors have a diverse behavior w.r.t. the size of the faces (see results for NBBA), that explanation must be rejected. We plan to investigate this point in future work, by considering the co-distribution of the NDFC and other features, such as lateral/non-lateral, occluded/non-occluded, roll angle and other (for instance, the sharpness of the bounding box).

5 Conclusions and future work

We considered the problem of FPV systems for the improvement of social interactions of visually impaired persons and, in particular, the task of face detection on video sequences captured by devices worn by the blind person. We evaluated the effectiveness of six face detectors on a set of four sequences which have been captured purposely basing on the considered application: the video sequences exhibit specific disturbances and effects related to the acquisition machinery and scenario. We systematically took into account those disturbances and effects by defining a set of six quantitative features on which we based a sensitivity analysis of the face detectors effectiveness.

Our comparative experimental evaluation shows that the considered detectors perform poorly in the considered application, with figures suggesting that their usage would be hardly practical in the general task of detecting all faces in the frame. Indeed, a possible future expansion of the present study consists in considering a more specific application: for instance, the detection of faces of persons who are approaching the visually impaired user, or of persons who are actively interacting with the user as a premise for a subsequent facial expression recognition.

Acknowledgment

This work has been supported by the University of Trieste - Finanziamento di Ateneo per progetti di ricerca scientifica - FRA 2014, and by a private donation in memory of Angelo Soranzo (1939–2012).

References

1. Online: Be my eyes. Available at http://www.bemyeyes.org.

2. Jin, Y., Kim, J., Kim, B., Mallipeddi, R., Lee, M.: Smart cane: Face recognition system for blind. In: Proceedings of the 3rd International Conference on Human-Agent Interaction. HAI '15, New York, NY, USA, ACM (2015) 145–148
3. Chaudhry, S., Chandra, R.: Design of a mobile face recognition system for visually impaired persons. CoRR **abs/1502.00756** (2015)
4. Carrato, S., Fenu, G., Medvet, E., Mumolo, E., Pellegrino, F.A., Ramponi, G.: Towards more natural social interactions of visually impaired persons. In: International Conference on Advanced Concepts for Intelligent Vision Systems, Springer (2015) 729–740
5. Zafeiriou, S., Zhang, C., Zhang, Z.: A survey on face detection in the wild: Past, present and future. Computer Vision and Image Understanding **138** (2015) 1 – 24
6. Hsu, G.S., Chu, T.Y.: A framework for making face detection benchmark databases. IEEE Transactions on Circuits and Systems for Video Technology **24**(2) (Feb 2014) 230–241
7. Carrato, S., Marsi, S., Medvet, E., Pellegrino, F.A., Ramponi, G., Vittori, M.: Computer Vision for the blind: a dataset for experiments on face detection and recognition. In: Proceedings of the 39th International Convention on Information and Communication Technology, Electronics and Microelectronics, Opatija, Mipro Croatian Society (2016) 1479–1484
8. Fazzi, E., Lanners, J., Danova, S., Ferrarri-Ginevra, O., Gheza, C., Luparia, A., Balottin, U., Lanzi, G.: Stereotyped behaviours in blind children. Brain and Development **21**(8) (1999) 522 – 528
9. Bonetto, M., Carrato, S., Fenu, G., Medvet, E., Mumolo, E., Pellegrino, F.A., Ramponi, G.: Image processing issues in a social assistive system for the blind. In: 2015 9th International Symposium on Image and Signal Processing and Analysis (ISPA), IEEE (2015) 216–221
10. Frazor, R.A., Geisler, W.S.: Local luminance and contrast in natural images. Vision research **46**(10) (2006) 1585–1598
11. Viola, P., Jones, M.J.: Robust real-time face detection. International journal of computer vision **57**(2) (2004) 137–154
12. Liao, S., Jain, A.K., Li, S.Z.: A fast and accurate unconstrained face detector. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(2) (2016) 211–223
13. Markuš, N., Frljak, M., Pandžić, I.S., Ahlberg, J., Forchheimer, R.: Object detection with pixel intensity comparisons organized in decision trees. arXiv preprint arXiv:1305.4537 (2013)
14. Dundar, A., Jin, J., Martini, B., Culurciello, E.: Embedded streaming deep neural networks accelerator with applications. IEEE Transactions on Neural Networks and Learning Systems (2016) to appear.
15. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2) (1955) 83–97
16. Fenu, G., Jain, N., Medvet, E., Pellegrino, F.A., Pilutti Namer, M.: On the assessment of segmentation methods for images of mosaics. In: Proceedings of the 10th International Conference on Computer Vision Theory and Applications (VISIGRAPP 2015). (2015) 130–137