Computer vision for the blind: a dataset for experiments on face detection and recognition

S. Carrato, S. Marsi, E. Medvet, F. A. Pellegrino, G. Ramponi and M. Vittori
DIA, University of Trieste, Italy
e-mail: {carrato, marsi, emedvet, fapellegrino, ramponi}@units.it, michele.vittori@gmail.com

Abstract – We present a video dataset created for the needs of a research project that aims at developing vision-based techniques that assist the social interaction of a blind person. Two totally blind users have acquired the sequences, using at the same time a glasses-mounted camera and a necklace-mounted one. The video sequences were acquired in different environments, selecting conditions in which a user could be interested in detecting the presence of some of his/her acquaintances, to approach them in a most natural way. The sequences have been temporally cropped to extract video shots that, by inspection, were deemed to contain events valuable for the goals of the project. The shots are presently being annotated, and some preliminary experiments on face detection have been performed on the annotated data. We also present some observations about the specific application that is being considered.

I. Introduction

We present some results from an ongoing research project devoted to image processing and computer vision techniques that can facilitate a blind user in his/her social relations [1,2]. We aim at enabling the user to detect and recognize one of his/her acquaintances among the people that are present, as a normally-sighted person would do. A fundamental characteristic of the project is the involvement, from its early stages to its end, of a Users' Group that includes people affected by visual impairments and personnel whose professional role is to take charge of the assistance of people with impairments. For example, video data that are necessary to perform the experiments in the detection of faces and the recognition of the expressions are being acquired directly by a blind person, who in this way is able to manifest needs and preferences that will make the final devices suitable for a practical usage.

This paper presents a video dataset that was acquired with the purpose of performing experiments of face detection and recognition in realistic (and thus difficult) conditions. We make some observations about the filmed scenes, and we illustrate the annotation procedure by which we generate a ground truth for the analysis of the sequences. A section of the paper is then devoted to the hardware requirements imposed by our goals, deducing some solutions for a suitable architecture. On these bases, a few algorithms for face detection that can satisfy our needs are briefly described, and some preliminary results we have obtained on our sequences are presented.

II. Related Work

We consider a system which may assist the social interaction of a blind person. The assistance may tailor several use cases. The issues that are perceived as the most important, according to the project "Social Interaction Assistant" [3], are due to the lack of some non-verbal cues that are present in the personal interaction; they are especially related to person recognition, eye gaze direction detection, and facial expression recognition. To approach these problems, the detection of faces in the frame is a fundamental preliminary step: to this end, we annotated the video sequences in order to enable insightful assessment of face detection algorithms.

Several datasets have been proposed for face detection: they differ in the level of detail of the annotation (e.g., a simple bounding box or several facial landmarks) and in the kind of data which is annotated (e.g., acquired in studio or collected from "the wild"), but most of them are built on images rather than video sequences. The use cases considered in our project require working on video sequences and hence we needed to cope with a much larger amount of data, resulting in a trade-off between how much data to annotate and the detail of annotation. We were driven by the goal of providing a dataset which can foster research beyond the mere detection: for example, including eye-related landmarks and considering consecutive frames permits to reason on pose modifications which may suggest an intention to communicate.

The MALF dataset proposed in [4] consists of 5,250 images and approximately 12,000 faces. The aim of the authors was to provide a large dataset for fine-grained evaluation of face detectors on so-called "in the wild" images. Beyond a well-defined bounding box, the annotations consist, for each face, of several categorical attributes (e.g., w/ or w/o glasses). As in our dataset, the annotations contain information about face relative size and pose, which allows for an assessment tailored to specific subset of faces (e.g., profile faces, far faces, and so on).

A frequently used dataset was presented in [5] which consists of 2,846 grey and colour images with approximately 5,100 faces: the annotations are given in form of elliptical regions which enclose faces. According to the authors, elliptical regions better capture faces; yet, they require special care when used to assess detectors which output rectangular bounding boxes.

Finally, in [6] both a model and a dataset for face detection are proposed which are tailored specifically to detailed landmark localization. The dataset contains 205 cluttered images with 468 faces. For each face, 6 landmarks, a discretized viewpoint and a bounding box form the annotation.

III. THE VIDEO DATASET

Many video sequences have been acquired by members of the Users' Group of the project. We show some examples recorded by two of them; these users are fully blind from birth, but they are determined to behave as far as possible in an autonomous way. They suffer only slightly from head- and body-posture modifications and from mannerisms, like body rocking, that are typically acquired at an early age by visually impaired people [7,8]. It is obvious that such mannerisms drastically affect the quality of the acquired video, and we expect that further sequences we will add to our dataset will provide other interesting case studies. Also for this reason, we opted for acquiring the scenes in two modalities: the final choice will be user-dependent.

Two commercial devices have been used to record the scene at the same time: in one case, which we will refer to as GL, the camera was mounted on the bridge of a pair of sunglasses; in the other case, labelled as NL, on a light support held by a short necklace. The glasses-mounted camera had a resolution of 1280 x 720p pixel and an angle of view of 135 deg.; the resolution of the necklace-mounted camera was 1920 x 1080p pixel and its angle of view was 124 deg. The GL sensor proved to provide slightly poorer images in terms of detail sharpness. The positioning of the videocameras is shown in Fig. 1. In a real deployment, these devices might be replaced by similar equipment able to stream video via wire, WiFi or Bluetooth; however, the results discussed here represent influential factors in the continuation of the project.

The video sequences were acquired in different environments, selecting conditions in which the user could be interested in detecting the presence of some of his/her acquaintances, to approach them in a most natural way. The selected locations are a university library, a coffee shop, the hall of a public building, the neighbourhood of a bus stop. In all cases, proper procedures were followed to comply with the normative about privacy protection. More precisely, signs providing reference to specific national laws were posted in the area in advance and during the acquisition, to inform passersby about the shooting of video sequences for scientific purposes. Written informed consent was obtained from all the people who actively participated in the operations.

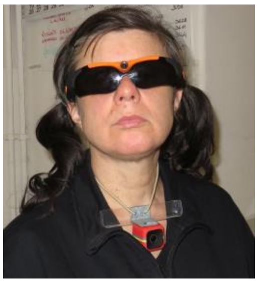

Figure 1. Positioning of the two videocameras for the acquisitions

Some sample frames of the sequences are shown in Figs. 2 and 3, which permit to notice their typical characteristics. Several observations can be made.

- Since the user of course lacks any feedback about the subjects in the field of view, faces can be partially occluded or can be partially outside the frame.
- The wide angle of view of the acquisition devices is a necessity: acquiring with standard optics (e.g. with a camera such as the ones typically mounted on smartphones) would make worse the above mentioned problem of faces outside the frame. However, wide angle causes geometrical distortions to appear. Compensating for them is theoretically simple, but implies computational costs which may make an actual realization quite complex, and may require proper dedicated resources.
- The scene conditions are very different in the different contexts, and can change abruptly with time. In particular, the automatic exposure control of the camera can be unable to comply with the range of the illumination. Back-lighting of the people in the scene is particularly critical, even if many face detection and recognition algorithms are by design relatively robust to this disturbance.
- People in the scene often tend not to look straight towards the user; this is an instinctive behaviour, due to politeness.
- The field of view of both the GL and the NL cameras can easily be partially occluded, by a tuft of hair or by a lapel of the dress respectively. A firmly placed camera, especially the NL one, or tightly held hair and dress can be unpleasant to wear.
- Sudden, fast and wide subjective movements are present, especially in the GL sequences. Some of the users move their body and in particular their head towards perceived sounds; other users are much more static.

The available sequences were temporally cropped to

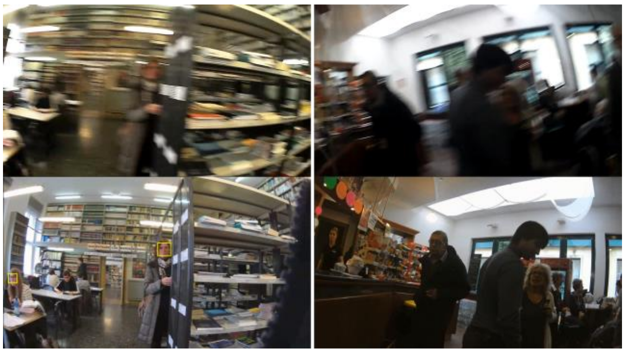

Figure 2. Two examples of motion blur induced by a sudden movement of the head (GL camera, top), and the corresponding frame filmed with the NL camera (bottom)

extract video shots that, by inspection, were deemed to contain events that are valuable for the goals of the project. They are a significant sample, which comprises 6,652 frames containing a grand total of 11,513 faces.

Details about the sequences can be seen in Tab. 1.

IV. Annotation of the Dataset

Faces in the selected shots were then annotated using a freely available tool, ViPER-GT [9]. Particular care was placed in selecting which features to annotate, and exactly how. Indeed, it is known that the way in which annotation is performed can modify the outcome of an experimental comparison of face detection methods [10].

We annotated, in each frame, every face whose estimated distance from the user was less than 5 meters, which resulted in a bounding box whose longest side was at least 20 pixel long: the rationale for this criterion was to annotate all faces that indicated the possibility of an immediate social interaction with the user. A rectangle was traced, 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. Yaw (rotation of the head around a central vertical axis) was constrained to +/-90 deg.; pitch (horizontal, left-right axis) and roll (horizontal, antero-posterior axis) were not constrained; however, 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 centres of the two eyes and of the mouth were also annotated.

V. Requirements for the Implementation Platform

The particular characteristics of the project require special features of the platform to be used in the system implementation: the system must be wearable and have a low power consumption, even if an autonomy of several hours can be considered sufficient for a typical usage. A further constraint is related to the different type of processing to be performed: indeed, the complete system requires both low-level video pre-processing to condition the input stream, and high-level analysis to extract and process the significant information from the video sequence.

The critical conditions in which the sequences are acquired, such as the presence of uncontrolled light dynamics, the distortion introduced by the wide-angle lenses and the possible shaking of the camera, require the adoption of suitable pre-processing algorithms to minimize these drawbacks and to control the input frames characteristics. These algorithms must be able to work in real time directly on the input video stream, thus they can greatly benefit from a suitable parallel processing that is available in processing systems like GPUs or FPGAs.

The selection of the appropriate hardware device must also take into account constraints related to portability and to power consumption; these features lead to a significant limitation in the type of suitable devices. It should be noted in fact that most of the GPU devices are employed in systems like console or desktop computers where low power consumption is not a priority, while very few GPU families [11] have been designed to be used inside mobile systems. Moreover, even within the same low-power GPU family, only the devices with more limited performances are integrated in mobile devices such as smartphones or tablets, while devices with higher performances are typically available only in consoles or notebooks.

TABLE I. Details about the sequences chosen to be annotated

	Length (frames)	**Camera**	**Location**	**Features**
A	379	NL	Outside a coffee shop	High contrast. Partially occluded faces. Tilted camera.
B	1138	NL	Bus stop	Many faces. Good soft light.
C	1380	GL	Inside a library	*Motion blur* induced by head movements. Pretty good light.
D	697	NL	Outside	Partially occluded faces. Very tilted camera.
E	1200	NL	Inside a coffee shop	Low light. Fast movements. Many partially occluded or lateral faces.
F	838	NL	Inside a cafeteria	High contrast. Pretty good light. Slow movements.
G	360	GL	Inside a coffee shop	High contrast. Pretty good light.
H	300	GL	Inside a coffee shop	Low light. Fast movements. Many lateral faces.
I	360	GL	Bus stop	Good light. Hard shadows.

With these premises, using an FPGA seems the most promising solution. In fact, current FPGAs make available, beyond programmable hardware resources such as memories, logic blocks, interconnection networks etc., also some hardware processors as well as communication systems, together with many other functional blocks [12]. These complex architectures allow the designer to realize in a single chip a highly advanced processing system with many customizable functions.

A feasible solution for the realization of the present project is therefore an embedded a system that adopts a high performance FPGA as processing unit. A quite low-priced board with characteristics consistent with the system constraints could be [13]. This board centralizes all the processing inside a single chip of the Xilinx Family Zinq-7000. This chip is equipped with a Dual Core ARM Cortex-A9 processor, several embedded memory blocks, interface blocks for an easily access to any external memories and many other functional blocks. It includes programmable logic that can be exploited for the realization of custom blocks to speed up the specific algorithm.

The complete processing system will implement a real-time operating system on one of the two ARM processing core. This core will handle the acquisition of images from the camera, together with any other data that may be sourced from other sensors that may be integrated in the system such as gyroscopes and accelerometers, suitable to improve the image stabilizer algorithms. These data will be transferred in real time to a shared memory. The video stream, using the FPGA resources, will be processed in real time through suitable image correction algorithms to minimize the drawbacks and the artefacts introduced by the camera motion, the geometric distortion of the lens and the high dynamic range of the input signal. The processed images will then be made available to the second ARM core through an appropriate shared memory. In the second ARM processor core a Linux operating system will be installed. This part of the system will be devoted to high-level signal processing operations such as face detection and face recognition, as well as any access to communication systems. This processor could even benefit from the FPGA hardware to speed up repetitive procedure exploiting the parallel processing facilities [14,15]. Eventually, this processor can also be used to perform the task of estimating the quality of the final results; thanks to these data, through a feedback control it can tune the parameters of the pre-processing algorithms to optimize the performances.

VI. Selected Face Detection Algorithms

The development of new algorithms for face detection is out of the scope of this project, which is more focused on our specific application (i.e., to develop an effective assistive technology for the blind); consequently, we analysed the state of the art in face detection algorithms and selected, based on the characteristics of the acquired video and on the computational performances that can be provided by a mobile platform, those which better fit our needs; they are briefly described in this section.

We point out that, at least in the first phase of our project, we aim at detecting faces that are almost frontal with respect to the user. This of course makes life simpler for the detection tools, in terms of both performances and computational load, and it is also a sensible first choice to detect those people in the scene who may be willing to interact with the user. Further developments of the algorithms and hardware will permit to include in the detection also subjects whose head is turned away.

The Normalized Pixel Difference (NPD) algorithm [16], very recently proposed by Liao et al., is based on the difference to sum ratio between couples of pixels, and uses a deep quadratic tree for the learning algorithm. It should be very effective in detecting faces with arbitrary size and pose, also in presence of occlusions and in cluttered scenes; a Matlab implementation has been made available by the authors in [17].

The so-called Viola-Jones (VJ) algorithm is older, as it has been proposed in 2001 [18] for generic object detection and in 2004 [19] specifically for face detection. It is based on Haar-like features and a cascade of simple classifiers, and should be very fast, also thanks to the use of a new image representation, the *integral image*, which permits the computation of the features at any scale in constant time; part of its wide popularity is probably due to its availability in the OpenCV framework [20].

Pixel intensity comparison is also used in the PICO algorithm [21]. Similarly to the VJ case, a cascade of rejecters is used as a decision tree; no integral image is used, and high processing speed is given by the simple structure of the tree nodes, which are just binary tests. The full source code has been made available by the authors in [22].

The three mentioned tools for face detection have been used in our experiments as provided by their authors; no attempt has been made to re-train them and make them more suitable to our dataset.

VII. Estimation of the Quality of the Face Detection Algortithms

The evaluation of the performances of the face detection algorithms is not trivial. From one side, their behaviour typically depends on several parameters, and it may be difficult to find the best combination of their values. To this extent, we decided (with one exception, described below) to use the default values suggested by the authors, assuming they performed an exhaustive series of tests. Actually, the best parameters combination can significantly vary according to the characteristics of the video data (e.g. blurriness, lighting conditions and backlighting, dimensions of the faces…), so that a more careful algorithm comparison should include the tuning of the parameters using our video database.

Performances cannot be evaluated based only on the number of detected faces in each frame. If for a certain frame the ground truth reports e.g. two faces, an algorithm may correctly detect only one face (this is a *true positive*, TP), miss the second one (this is a *false negative* (FN) error) and classify as face another object (a *false positive* (FP) error). Thus, it is common to use two quality parameters, *precision* and *recall*, defined as

$$precision = \frac{tp}{tp + fp} \tag{1}$$

$$recall = \frac{tp}{tp + fn} \tag{2}$$

being *tp*, *fp* and *fn* respectively the number of TPs, FPs and FNs.

It has also to be noted that the evaluation of *tp*, *fp* and *fn* requires some care, for two reasons. The first is that, since these algorithms typically provide the *x* and *y* coordinates of the detected face bounding box (BB), some tolerance has to be allowed with respect to the coordinates of the BB in the ground truth: a misplacement of e.g. a couple of pixels in a 50 pixel size face should not be interpreted as a wrong detection. The second is related to the fact that, in case multiple faces are present in the image, the algorithm will likely *not* find the faces in the same order in which they appear in the ground truth; an automatic procedure based on a mere one-by-one position matching would erroneously yield only few, if any, TPs. It is then necessary to apply a minimization algorithm, which solves the assignment problem of finding the best match between the positions found by the algorithm under test and those reported in the ground truth: we used the Hungarian algorithm [23].

VIII. Preliminary results

Trials are still in progress, but we can provide some first numerical results and some preliminary considerations derived from the direct observation of the comparison results.

First of all, we observe from Tab. 2 that the NL videocamera permitted to obtain far better results than those of the GL camera. This is due to several factors, including the slightly poorer quality of the GL camera in terms of resolution and sharpness, and the better stability of the NL solution.

TABLE II. Average *precision* and *recall* for each camera type

	Average Recall (%)	**Average Precision** (%)
NL	45.69	20.06
GL	29.85	13.92

TABLE III. Average *precision* and *recall* for each algorithm

	Average Recall (%)	**Average Precision** (%)
NPD	33.29	27.27
VJ	28.98	17.25
PICO	43.70	24.87
PICO alt	53.77	12.62

Then, Tab. 3 shows that the algorithm PICO [21] offered the best results when considering both precision and recall. Its recall values have a +10% margin on the second best algorithm and a modest -3% drop in precision with respect to NPD [16], which is the best algorithm precision-wise.

It is very interesting to analyze the worst average results, as seen in Tab. 4. Tests on videos A, D, E and H gave significantly worse results, which can be explained by some common characteristics in the videos. As seen in Tab. 1, A, E and H present harsh light conditions, being shot in fairly dark and highly backlit areas, while videos A and D are both shot with a tilted camera. It has to be noted that in those four worst cases Viola-Jones [19] had better results than NPD, while still being outperformed by PICO both in the recall and in the precision results. Actually, the same four videos coincide with the cases where PICO had the best precision values.

Our preliminary experiments then indicate that in good conditions, good light, straight necklace-mounted camera, NPD has the best precision values and stands almost on a par with PICO's recall values. In difficult conditions, bad light, high contrast, and tilted camera, PICO performs significantly better than the other tested methods.

Comparing the positions of the TPs with those of the FPs, it is evident that the former are relatively stationary in adjacent frames, while the latter tend to appear in different positions of the image with no apparent continuity. Thus, it has been decided to focus on the recall results and give less importance to the precision of the tests. It is indeed reasonable that simple methods that exploit temporal correlation can be devised, which will be able to detect and reject many FP cases.

We then decided to modify a parameter in the PICO options to further improve its recall values, at the cost of worse precision and longer computing times. The results of these alternative tests can also be seen under the name "PICO alt" in the previous tables. They highlight how, at the cost of a halved precision, the recall values further improve: about 10% more faces are recognized. Even if presently we have no quantitative data, it should be noticed that this alternative method will be significantly affected in terms of the time taken to make the detection.

TABLE IV. *Precision* percentages for the different tested methods, on each video

	NPD	**VJ**	**PICO**	**PICO alt**
A	4.65	2.01	4.91	2.27
B	55.25	40.20	51.14	27.35
C	10.58	17.07	12.31	6.77
D	8.47	3.58	14.48	7.82
E	12.31	8.43	13.04	7.76
F	35.40	13.39	29.11	18.43
G	22.70	27.92	22.43	11.47
H	13.21	8.78	23.36	17.19
I	47.21	46.29	54.83	29.88

IX. Conclusions

The previous analysis of the face detection results on our dataset, and specifically the consistency in results variation among algorithms on the different video sequences, confirm that the dataset is a valid benchmark for testing face detection algorithms. We can observe that even in the best case scenario the recall results obtained never pass the 80% mark, and average at little more than 50% even when considering the algorithm which performed best in this area.

This confirms the usefulness of a specific dataset to measure algorithm performances relative to the scope of a social aid to visually impaired people, instead of evaluating such performances on generic datasets such as FDDB [5] or “Faces in the wild” [6]. A comparison of the figures of merit available for the latter databases and the results we obtained on our dataset clearly indicates that the notion of data acquired “in the wild” has to be modified: real world problems such as the one we are addressing imply the need to cope with even harsher environments.

Acknowledgments

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).

The authors are grateful to the Users' Group for their precious collaboration. They also thank Igor Pandžić, Nenad Markus and Visage Technologies for providing the PICO software, and Anil K. Jain and Shengcai Liao for the NPD software.

References

[1] M. Bonetto, S. Carrato, G. Fenu, E. Medvet, E. Mumolo, F.A. Pellegrino, G. Ramponi, "Image Processing Issues in a Social Assistive System for the Blind," in 9th International Symposium on Image and Signal Processing and Analysis, ISPA 2015, Zagreb, Croatia, September 7-9, 2015.

[2] S. Carrato, G. Fenu, E. Medvet, E. Mumolo, F.A. Pellegrino, G. Ramponi, "Towards More Natural Social Interactions of Visually Impaired Persons," in Int. Conf. on Advanced Concepts for Intelligent Vision Systems, ACIVS 2015, Catania, Italy, Oct. 26-29, 2015.

[3] S. Panchanathan, S. Chakraborty, T. McDaniel, “Social Interaction Assistant: A Person-Centered Approach to Enrich Social Interactions for Individuals with Visual Impairments”, IEEE Journal of Selected Topics in Signal Processing, to be published, 2016 (IEEE Early Access Articles).

[4] B. Yang et al. "Fine-grained evaluation on face detection in the wild," Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. Vol. 1. IEEE, 2015.

[5] V. Jain and E. Learned-Miller, “FDDB: A Benchmark for Face Detection in Unconstrained Settings,” Technical Report UM-CS-2010-009, Dept. of Computer Science, University of Massachusetts, Amherst. 2010.

[6] X. Zhu and D. Ramanan. "Face detection, pose estimation, and landmark localization in the wild," Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012.

[7] A. Molloy and F. J. Rowe, "Manneristic behaviors of visually impaired children," Strabismus, Volume 19, Issue 3, 2011, pp.77–84.

[8] E. Fazzi, J. Lanners, S. Danova, O. Ferrarri-Ginevra, C. Gheza, A. Luparia, U. Balottin, and G. Lanzi, "Stereotyped behaviours in blind children," Brain and Development, vol. 21, no. 8, 1999, pp. 522–528.

[9] Language and Media Processing Laboratory, University of Maryland: “A Video Metadata Markup Tool: ViPER-GT,” 2005. (v4.0) (2016, Feb 19) [Online] Available:
http://viper-toolkit.sourceforge.net/products/gt/

[10] M. Mathias, R. Benenson, M. Pedersoli and L. Van Gool, “Face detection without bells and whistles,” ECCV 2014, European Conference on Computer Vision, Zurich, CH, Sept. 6-12, 2014.

[11] Whitepaper NVIDIA Tegra 4 Family GPU Architecture (v1.0) (2013, Feb 22) [Online] Available: https://www.nvidia.cn/content/PDF/tegra_white_papers/Tegra_4_GPU_Whitepaper_FINALv2.pdf

[12] Xilinx inc. “Zynq-7000 All Programmable SoC Overview,” datasheet DS190 (v1.9) (2016, Jan 20) [Online] Available: http://www.xilinx.com/support/documentation/data_sheets/ds190-Zynq-7000-Overview.pdf

[13] Dave Embedded System: “Bora: Xilinx Zinq XZ7Z010/ZC7Z020 CPU Module,” (2016, Feb 20) [Online] Available: http://www.dave.eu/sites/default/files/files/bora-leaflet.pdf

[14] S. Jin, D. Kim, T. T. Nguyen, D. Kim, M. Kim, J. W. Jeon, "Design and Implementation of a Pipelined Datapath for High-Speed Face Detection Using FPGA," in Industrial Informatics, IEEE Transactions on, vol. 8, no. 1, 2012, pp. 158-167.

[15] C. Kyrkou, C. S. Bouganis, T. Theocharides, M. M. Polycarpou, "Embedded Hardware-Efficient Real-Time Classification With Cascade Support Vector Machines," in Neural Networks and Learning Systems, IEEE Transactions on , vol. 27, no.1, 2016, pp. 99-112.

[16] S. Liao, A. K. Jain and S. Z. Li, “A Fast and Accurate Unconstrained Face Detector,” in IEEE Trans. on PAMI, vol. 38, n. 2, 2016.

[17] S. Liao, A. K. Jain and S. Z. Li, “A Fast and Accurate Unconstrained Face Detector,” (2016, Feb 19) [Online] Available: http://www.openpr.org.cn/index.php/107-NPD-Face-Detector/View-details.html

[18] P. A.Viola and M. J. Jones, “Rapid Object Detection using a Boosted Cascade of Simple Features,” accepted conference on computer vision and pattern recognition, 2001.

[19] P. A. Viola and M. J. Jones, “Robust real-time face detection,” in International Journal of Computer Vision, vol. 57, no. 2, 2004, pp. 137-154.

[20] Open Source Computer Vision: “Face Detection using Haar Cascades,” (2016, Feb 19) [Online] Available: http://docs.opencv.org/master/d7/d8b/tutorial_py_face_detection.html

[21] N. Markus, M. Frljak, I. S. Pandzic, J. Ahlberg and R. Forchheimer, “A method for object detection based on pixel intensity comparisons,” unpublished.

[22] Nenad Markus: “pico,” (2016, Feb 19) [Online] Available: https://github.com/nenadmarkus/pico

[23] H. W. Kuhn, “The Hungarian Method for the Assignment Problem,” in Naval Research Logistics Quarterly 2, 1955, pp. 83–97.