SERGIO CARRATO, GIOVANNI RAMPONI
DIA, University of Trieste
via A. Valerio 10, 34127 Trieste, Italy
email {carrato, ramponi}@units.it

Assistive technologies based on image processing for people with visual impairments: a tool to help social interaction of the blind

Social interactions involve verbal and non-verbal communication cues, which most of the time are so connected that we hardly notice the difference. Non-verbal communication comes in a variety of forms such as physical movements (hand and eyes movements, posture, face expressions), appearance (clothes, accessories, make-up) and the distance between communicators. Visually impaired people cannot directly access visual information and are disadvantaged in daily communications. They could also feel uncomfortable asking other people to report non-verbal information. The most important non-verbal cues that visually impaired people may need to access are the number of people, where a person is directing her attention, identity, appearance, if the physical appearance of a known person has changed since the last time the user encountered her, hand and body motions, illicit behaviour [Little et al. 2005, Khrishna et al. 2008].

The University of Trieste, also with the help of a private donation, has recently started a research project that aims at developing user-friendly vision-based techniques that may assist the social interaction of a person affected by a very severe visual impairment or a total blindness, by providing her with some of the information quoted above. A system view of the required information acquisition and processing chain has been devised, and the properties of its various components have been determined: from the acquisition of video data, to data pre-processing and analysis, the extraction of non-verbal information that the blind person cannot perceive, and its summarization and communication to the user through an audio or haptic channel [Bonetto et al. 2015, Carrato et al. 2015].

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 (students and personnel of the University) and university 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 the blind, who in this way is able to manifest needs and preferences that will make the final devices suitable for a practical usage. The Ethics Committee of the University has also been involved, and is following step by step the evolution of the project.

After a series of meetings meant to determine common goals for the project and the ways to achieve them, the Users' Group took charge of the acquisition of a dataset of video sequences acquired by blind people, in collaboration with the researchers of the project. Some example video frames are shown in a page of the Website of the project (http://www2.units.it/ipl/fra_2015/work.html). Users who did the acquisition 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 often affect the blind and are typically acquired at an early age [Mollow/Rowe 2011, Fazzi et al. 1999]. It is obvious that such mannerisms drastically affect the quality of the acquired video. This is the reason why we did not opt for a single acquisition modality: the final choice will be user-dependent.

Two commercial devices have been used to record the scene at the same time: in one case, the camera was mounted on the bridge of a pair of sunglasses, in the other case 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. Of course, these devices will be replaced by similar equipment able to stream video via wire, WiFi or Bluetooth.

The video sequences were acquired in different contexts, selecting conditions in which the user could be interested in detecting the presence of some of her acquaintances, and in approaching them in a most natural way. The selected ambients are a university library, a coffee shop, the hall of a public building, the neighborhood of a bus stop. In all cases, proper procedures were followed to comply with the normative about privacy protection for all the people present in the scene.

Performing this task permitted to make some observations:

- The scene conditions are very different in the different contexts, and can change rather abruptly with time.
- Since the user of course lacks any feedback about the subjects in the field of view, faces can be partially occluded or 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) is prone to failure due to the mentioned lack-of-feedback issue. However, wide angle causes geometrical distortions to appear. Compensating for them is theoretically simple, but implies computational costs that may be not compatible with the available hardware and processing time.
- 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 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 necklace camera, and tightly held hair and dress can be unpleasant to wear.
- Sudden, fast and wide subjective movements are present, especially in the glasses-mounted 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 extract video shots that, by inspection, were deemed to contain events that are valuable for the goals of the project. Faces in the selected shots are presently being annotated. Particular care was placed in selecting the features which need to be annotated, and the way in which the annotation has to be done. Indeed, it is known that the way in which annotation is performed can modify the outcome of an experimental comparison of face detection methods. The annotations involved faces that indicated the possibility of an immediate social interaction with the user. For this purpose, 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. A maximum estimated distance between the subject and the User of 5 meters was considered and, anyway, the longest side of the rectangle should be at least 20 pixel long. 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 centers of the two eyes and of the mouth were also annotated.

Some preliminary experiments on face detection are being performed on the annotated data. Processing blocks performing face recognition among a set of acquaintances of the User's, and vocal synthesis of the name(s) of possibly recognized individuals, will soon permit to complete the basic chain of the system.

Keywords: visual impairments, image processing, computer vision, non-verbal communication, social interactions.

References

G. Little, J. Black, S. Panchanathan, S. Krishna, *A wearable face recognition system for individuals with visual impairments*, 7th international ACM SIGACCESS conference on Computers and accessibility, pp.106–113, 2005.

S. Krishna, V. Balasubramanian, D. Colbry, S. Panchanathan, T. McDaniel, *Using a haptic belt to convey non-verbal communication cues during social interactions to individuals who are blind*, HAVE 2008, IEEE International Workshop on Haptic Audio Visual Environments and Games, pp.13–18, 2008.

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*, 9th Int.Symp. Image and Signal Processing and Analysis, ISPA 2015, Zagreb, Croatia, September 7-9, 2015.

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

A. Molloy, FJ. Rowe, *Manneristic behaviors of visually impaired children*, Strabismus, Volume 19, Issue 3, pp.77-84, 2011.

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, pp. 522–528, 1999.