Automatic Eye Gesture Recognition in Audiometries for Patients with Cognitive Decline

A. Fernandez[1], M. Ortega[1], M.G. Penedo[1], B. Cancela[1], and L.M. Gigirey[2]

[1] VARPA Group. Department of Computer Science. University of A Coruña, Spain
{alba.fernandez,mortega,mgpenedo,brais.cancela}@udc.es
[2] Audiology Unit-University School of Optics and Optometry, USC, Spain
luz.gigirey@usc.es

Abstract. This paper provides a specifically adapted methodology for supporting the audiologists when testing the hearing of patients with cognitive decline or other communication disabilities. These patients can not interact with the audiologist conventionally, but they often express gestural reactions when they perceive the auditory stimuli typically associated to the eyes region. From a video sequence captured during the hearing evaluation, we analyze the movements in the area of the patient's eyes, so we can detect these gestural reactions. We define a set of different gestures for classification, based on the expert knowledge. The proposed method achieves an accuracy of the 90.65% when classifying these movements, showing their separability, and therefore, the possibility of interpreting them with high-level information as positive reactions to the auditory stimuli.

Keywords: Hearing assessment, eyes detection, optical flow, movement analysis.

1 Introduction

Hearing loss and communication disorders may have a negative impact on the state of emotional, physical and social well-being [1]. Impaired hearing results lead to greater isolation by restricting the patient's social life and, therefore, impacting negatively in his psychosocial well-being. According to this, we can observe that hearing plays a key role in the process of "active aging" [2]. This impact on the aging process makes necessary to conduct regular hearing checks. Liminar Tonal Audiometry (LTA) is the "gold standard" in the evaluation of the hearing capacity and prevalence of hearing impairments. The standard procedure for a LTA consists on sending auditory stimuli to the patient, and asking him to raise his hand when he perceives this stimuli. However, the widespread use of audiometric tests involves some operational constraints, especially among some population groups with special needs or disabilities, like individuals with cognitive decline.

In this last case, the standard protocol becomes unenforceable since no interaction audiologist-patient is possible. The audiologist needs to focus his

M. Kamel and A. Campilho (Eds.): ICIAR 2013, LNCS 7950, pp. 27–34, 2013.

attention on unconscious facial reactions, since a typical interaction question-answer is not possible with this type of patients. The expert's attention is mainly focused on the eye region where he tries to detect changes in the gaze direction, eyes opening as a reaction to the perception of an auditory stimulus, and other particular expression changes that could indicate some kind of perception by the patient. Proper detection and interpretation of these gestural reactions requires broad experience by the audiologist as each patient may present different gestures. Furthermore, it is an entirely subjective procedure so it becomes an imprecise problem, very prone to errors, difficult to compare between experts and difficult to reproduce.

In [3] an automatic system was proposed to provide automatic solutions for patients without cognitive impairments whom you can establish a typical audiologist-patient interaction. Therefore, it is needed to develop an objective screening method that provides support to the experts when evaluating the hearing capacity in those situations with operational limitations. To that end, we want to automatically classify the patient facial reactions. Furthermore, the possibility on having a tool of this type, will allow the training of other audiologists to evaluate this type of patients. This contribution could be of great interest for the audiologist community, since no previous automatic approaches were developed for this task.

This system must be useful to be integrated in the audiometric domain which has a very particular setup (an example can be seen on Fig. 1): the patient is located in front of the camera at a determined distance, face always in frontal position, subtle movements in the eye region, etc. Given the nature of the patients whom this system is aimed, there is a need for a method based on movement in the interest region (the eye region) rather than based on facial features since it will not always be possible to properly locate these features or the patient has an inability to show expressions associated to that feature.

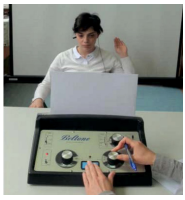


Fig. 1. Typical setup of the video sequences

The scope of this paper is to use image processing techniques applied on this particular domain in order to locate and characterize the facial gestures, more specifically, eyes gestures, since eyes were identified by the experts as the more expressive facial feature for this kind of patients. Once the interest region around the eyes is located, we want to detect and analyze the movements or expression changes that occur within this region. To analyze these global movements, we apply the iterative Lucas-Kanade [4] optical flow method with pyramids, and

then we train and apply different classifiers that allow us to differentiate the different movements that occur. This initial classification will serve as the basis for later providing them of a meaning, and being able to properly interpret those that represent a response to an auditory stimuli.

The remainder of this paper is organized as follows: Section 2 is devoted to explain the clinical protocol for a LTA. Section 3 presents the methodology. Section 4 shows the experimental results and their interpretation. Final conclusions are presented in Section 5.

2 Clinical Protocol for LTA

As discussed before, it is important to evaluate the hearing capacity to detect evidences of hearing problems. The standard test for checking the hearing capacity is the Liminar Tonal Audiometry (LTA). LTA checks a person's ability to hear the loudness and pitch of sounds. The results are charted on a graph called audiogram. A pure tone audiometry test measures the least audible sound that a person can hear. Normal hearing is expected to be between -10dB(HL) and 15dB(HL). The inability to hear pure tones below 25dB indicates hearing loss.

During the test, the patient will be wearing earphones connected to an audiometer, that will deliver the tones at different frequencies and intensities to the patient's ear, one ear at a time. The typical behavior is that the patient is asked to raise his hand when he perceives the auditory stimuli. This interaction was previously analyzed in [3], for measuring the response times and identifying patients with response times abnormally slow. As mentioned in the introduction, we are focusing now on patients with cognitive decline or "special needs". In this particular population group, certain gestural reactions can be considered as a response to assess the hearing.

In next Section we explain the methodology to evaluate this kind of patients.

3 Analysis of Eye-Based Gestural Reactions

For patients with cognitive decline or other communication disabilities, eye-based gestural reactions are going to be analyzed. According to the experts, patients can change the gaze direction when they perceive an auditory stimuli, in other cases, for instance they will react by opening their eyes exaggeratedly. Therefore, our movements of interest can be grouped into several classes such as: eye opening, eye closure or changes in the gaze direction.

To reach a proper gesture classification from the patient we propose a methodology as depicted in Fig. 2. Each step is explained in more detail next.

Face Location. Proper face location will serve as a first step to facilitate the eye area location, making this second search faster and less error prone. Since we are working in a stable domain where the conditions in which the test is performed are already know, we can ensure that faces will always be in frontal position. A Viola-Jones [5] approach can be applied. The Viola-Jones object detector

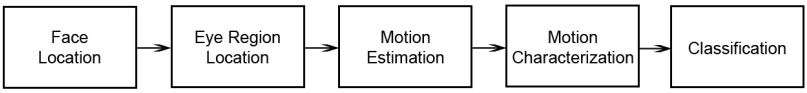


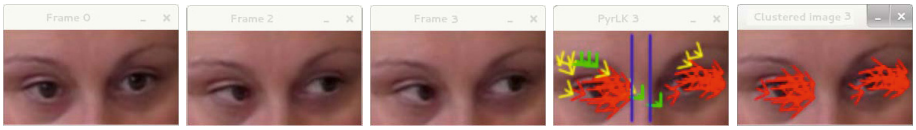
Fig. 2. Schematic representation of the methodology

framework is a general detector that has been trained and optimized for the face detection task (among others). In particular, a classifier for the detection of frontal faces is available in the OpenCV repository. This face detector is not as flexible as other approaches, but it is a low computational solution and its robustness is clearly demonstrated.

Eye Region Location. Once we succeeded in locating the face within the global image, we are going to search for the eyes region within the detected face area. To that end, we specifically trained a new Viola-Jones detector. This eye detector was trained by us with more that 1000 images of the eye area, manually selected (since there was no image database for this region). These training images where cropped from different face images coming from different face databases. This eyes detector was trained and tested obtaining excellent results (greater than 90%) for both open and closed eyes, and with different expressions.

Motion Estimation. Now, within the detected eyes region, we need to detect and analyze the movements or expression changes that occur therein. It must be noted that in out domain, people behavior can be erratic or lacking part of the typical movements associated to a particular gesture. Moreover, each individual may present different particularities in this regard. Therefore, classical solutions based on feature point registration (such as [6]) or template analysis (i.e. [7,8]) can not be applied here. We propose a novel approach aimed to this domain based on global movement analysis for description. This way the system can take advantage of any movement the patient does paralleling the analysis from the expert. Assessing the domain and the features of images to be treated, we propose to analyze the optical flow between eye region images. Optical flow was previously applied for similar general tasks as in [9], where an affective-responsive interactive photo-frame analyzing activity and facial expressions was developed. In this case, we applied the iterative Lucas-Kanade [4] optical flow method with pyramids. We want to remark that our video sequences have a frame rate of 25 FPS. Considering this, in order to have expression changes notable enough to be detected, we perform comparisons with a tree frame space, i.e. optical flow is calculated between frame i and frame $i+3$. A comparison between consecutive frames would not make sense since no significant expression changes can happen so quickly. However, a space too large would not make sense either, since short movements would be lost. Considering that in our domain significant movements occur during 4-5 frames, a 3 frame window is a good compromise that provides the expected results. Furthermore, after several experiments, we decided not to follow the classical use of the optical flow generating the features to track in the first frame and using them

all along the sequence. In our case, we generate these features for the first frame in each comparison; since changes in face position could affect very significantly to these features. In Fig. 3 we can see a sample of the optical flow calculation. Figures 3(a), 3(b), 3(c) represent consecutive frames. As mentioned before, we leave a gap of three frames, comparing Figures 3(a) and 3(c). The results of the optical flow are shown in Fig. 3(d). Since we are interested on the most significant changes, we will consider only the stronger flow vectors (empirically, those with a length greater than 7 pixels, although it depends on the sequence resolution), these stronger vectors are shown in Fig. 3(e).



(a) (b) (c) (d) (e)

Fig. 3. Sample optical flow images. Optical flow is calculated between (a) and (c). (c) Is the frame *i+3*. Optical flow results in (d). We represent in green vectors with softer movements, yellow for the intermediate and red for vectors with stronger movements. Stronger vectors in (e) after filtering the rest.

Motion Characterization. Taking this flow information as basis, we want to characterize a discrete set of movements, since every patient is going to react differently to the auditory stimuli. That is the reason why a classical methodology of global characterization of facial expression changes is not applicable in this domain. So, what is proposed is to rely on very simple gestures in our ROI. After reaching consensus with the experts, we have identified five typical movements considered the more relevant: eye opening, eye closure, gaze shift to the left, gaze shift to the right and global movement of the region. Whereupon, the goal is to analyze the entire sequence of optical flow and to detect where the significant movements occur.

In this initial approach we want to test whether it is possible to distinguish reliably the various patient's movements to try to characterize them. To that end, what we propose is to characterize the movement when it occurs. When no significant movement is present, classification task does not take place, but when movement exists we classify it into one of the five possible categories. In order to classify significant movement into a window of frames, we are going to extract some descriptors from the optical flow.

One of the movement features is the orientation. The vector orientation provides information about possible changes in the gaze direction or if eyes are in a motion of opening or closure. Similarly, it is important to know the vector's magnitude, because this feature provides information about the intensity of the movement. Finally, knowing the dispersion of the optical flow vectors with the same orientation allows to discriminate between more focalized and more global movements. According to this, our descriptor is comprised of a vector of 24

values. The first 8 values are related to the orientation. The orientations of the optical flow vectors are classified according to the 8 classes represented in Fig. 5, on these bases, the orientation histogram is computed and it corresponds to the first 8 values. The second 8 values correspond to the average length for each orientation. Finally, the last 8 values try to represent the dispersion of the optical flow vectors, to that end, we compute the average of the distances between vectors for each orientation. A sample of these histograms can be seen in Fig. 4.

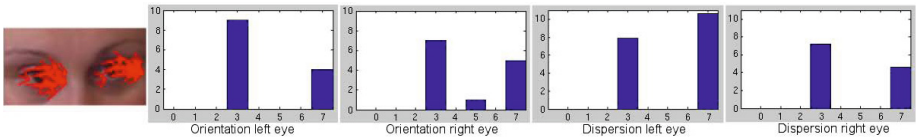


Fig. 4. Sample image with histograms for orientations and histograms for the dispersion

Classification Once we have the descriptors, we conducted a study of different classifiers, which are very common in this type of pattern recognition problems. The results obtained for some of these classifiers are shown in Section 4. Also, in this next section we discuss a pair of optimizations that were applied and which do not affect the quality of the obtained results.

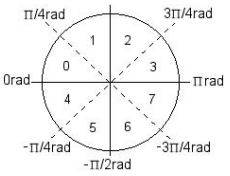


Fig. 5. Classification of the angles of the optical flow

4 Experimental Results

Several classifiers were trained and evaluated for this task. To perform this experiment different video sequences recorded during the audiometries were manually labeled. For this experiment, we worked with full HD video sequences consisting on more than 2500 real frames from audiometry recordings. Due to the preliminary nature of this study, and especially because of the difficulty of obtaining images from this specific type of patients, we only could work with images from a single patient who had these special conditions. For those frames where the optical flow detected a significant movement, this movement was classified into one of the possible categories (i.e., eye opening (class EO), eye closure (class EC), gaze shift to the left (class GL), gaze shift to the right (class GR) and global movement (class GM)). In the general case, changes in the gaze direction are not very common. It is for this reason that the training samples were at the beginning very unbalanced, having many samples for eye opening, eye close

and global movement, but only a few for the gaze shifting. The first training tests with this whole training set (comprising 850 samples) were not successful, since the classifiers try to maximize the global accuracy, hereby penalizing those classes with few elements. Considering this, new trainings were performed balancing the number of samples per class. The goal was to find the maximum number of possible elements that allow to maximize the global accuracy without penalizing classes with few elements. With 23 samples for Class GL and 21 samples for Class GR, the maximum number of samples for the other classes that maximized the global accuracy without penalizing the classes with few elements were 75 samples. In Table 1 we show the accuracy for some of the classifiers that we tested. Trainings are performed randomly taking 75 samples for classes EO, EC and GM, in 10 different experiments, and applying a cross validation with a 10-fold. The classifiers here represented are: Naive Bayes, Logistic Model Trees (LMT), Multilayer Perceptron, Random Forest and Random Committee.

Table 1. Accuracy of the classifiers

	Naive Bayes	LMT	Perceptron	Random Forest	Random Comm.
Class EO	0.3973	0.6409	0.6935	0.7144	0.7052
Class EC	0.1908	0.7475	0.7474	0.7921	0.7841
Class GM	0.6619	0.8196	0.8264	0.8447	0.8434
Class GL	1	0.884	0.912	0.84	0.872
Class GR	0.891	0.5407	0.609	0.5498	0.5726
Global	**0.5244**	**0.7338**	**0.7582**	**0.7702**	**0.7698**

According to the results showed in Table 1, we select Random Forest as the best suited for this task. So, we took this model and applied it to our entire dataset obtaining the results depicted in the confusion matrix Table 2(a). A global accuracy of the 76.71% was obtained. Although these results are quite acceptable, it can be observed that we have a lot of confusion especially between the first three classes.

Having into account the domain knowledge we know that a continuity along the movement must exist. I.e., if we have an eye closing for 3 consecutive frames, and in the other eye we detect two frames classified as eye closure, and an intermediate frame as global movement, it is very likely that we had a miss-classification and that eye should be also classified as eye closure. Applying a voting system based on the requirement of this continuity, we correct some

Table 2. Confusion matrices. (a) Initial. (b) After first optimization. (c) After second optimization

(a)

	EO	EC	GM	GL	GR
EO	80	9	12	8	5
EC	45	177	7	2	11
GM	10	7	91	1	2
GL	0	0	0	23	0
GR	0	0	0	0	21
Global accuracy: 0.7671					

(b)

	EO	EC	GM	GL	GR
EO	104	2	5	3	0
EC	37	185	7	2	11
GM	7	7	96	0	1
GL	0	0	0	23	0
GR	0	0	0	0	21
Global accuracy: 0.8395					

(c)

	EO	EC	GM	GL	GR
EO	101	2	5	2	0
EC	3	53	2	0	2
GM	6	6	90	0	1
GL	0	0	0	18	0
GR	0	0	0	0	19
Global accuracy: 0.9065					

miss-classifications and the results are improved obtaining the confusion matrix showed in Table 2(b) and a global accuracy of 83.95%.

Finally, considering the domain and according to the experts opinion, isolated movements of one only frame are discarded, because a movement without continuity does not represent a significant movement, and even the expert is not able to detect it. After this second consideration, the results were improved significantly (as showed in Table 2(c)) and the final global accuracy reaches the 90.65%.

5 Conclusions

In this paper a new approach to analyze facial expression changes is presented in order to support the audiologists when they are testing the hearing of patients with cognitive decline or other disabilities. It is important to remember that with this kind of patients no standard interaction is possible, making inapplicable other previous solutions. This initial study shows that it is possible to reliably separate the different movements of these patients. The obtained results of an 90.65% of accuracy, allow us to confirm the possibility of distinguishing between the five classes of interest movements defined by the experts. The main contribution of this work consists on providing a novel method for the global interpretation of the movements or gestural reactions that present this specific group of patients, which are completely different from the reactions of normal people, and are also different for each patient. This singularity made inapplicable the use of classical techniques or other general solutions. For future works we want to conduct more complex studies in order to corroborate the information given by the experts.

References

1. Davis, A.: The prevalence of hearing impairment and reported hearing disability among adults in great britain. Int. J. Epidemiol. 18, 911–917 (1989)
2. Espmark, A., Scherman, M.: Hearing confirms existence and identity-experiences from persons with presbyacusis. Int. J. Audio 42, 106–115 (2003)
3. Fernández, A., Ortega, M., Cancela, B., Penedo, M., Vazquez, C., Gigirey, L.: Automatic processing of audiometry sequences for objective screening of hearing loss. Expert Syst. Appl. 39(16), 12683–12696 (2012)
4. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the IIJCAI 1981, pp. 674–679 (1981)
5. Viola, P., Jones, M.: Robust real-time object detection. Int. J. Comput Vision 57, 137–154 (2004)
6. Geetha, A., Ramalingam, V., Palanivel, S., Palaniappan, B.: Facial expression recognition - a real time approach. Expert Syst. Appl. 36(1), 303–308 (2009)
7. Kumano, S., Otsuka, K., Yamato, J., Maeda, E., Sato, Y.: Pose-Invariant Facial Expression Recognition Using Variable-Intensity Templates. Int. J. Comput. Vision 83(2), 178–194 (2009)
8. Akakin, H.C., Sankur, B.: Robust classification of face and head gestures in video. Image Vision Comput. 29(7), 470–483 (2011)
9. Dibeklioglu, H., Ortega, M., Kosunen, I., Zuzanek, P., Salah, A., Gevers, T.: Design and implementation of an affect-responsive interactive photo frame. Journal on Multimodal User Interfaces 4, 81–95 (2011)