Detecting attention through Telepresence

S. Levialdi, A. Malizia, T. Onorati, E. Sangineto, N. Sebe

Sapienza University of Rome - Italy, Universidad Carlos III de Madrid - Spain, Sapienza University of Rome – Italy, Sapienza University of Rome – Italy, University of Amsterdam - The Netherlands
{levialdi@di.uniroma1.it, alessio.malizia@uc3m.es, teresa_onorati@hotmail.com , sangineto@di.uniroma1.it, nicu@science.uva.nl}

Abstract

Attention is the cognitive process of selectively concentrating on one aspect of the environment while ignoring other things. As such, attention is one of the most intensely studied topics within psychology and cognitive neuroscience. In this paper, we present a way to detect attention during a remote communication session. Our approach considers the difference between face-to-face and remote communication. In the first case, participants directly share the physical space, information, and emotional cues, while a remote communication scenario will be hampered by some limitations. To address these, we enhance awareness by adding a cognitive reaction to an event: the system is aware if the user is conscious about the discussion topic and individual/global performance.

In our research, we consider attention as a particular feature of awareness. We measure the attention level, based on user activities and his facial expressions. Activities are performed by participants during a remote video conferencing, trough mouse clicks and keystrokes. Each user's face is analyzed to find the values of seven basic emotional states: neutral, happy, sad, fear, disgust, surprise and angry, following the codification of "universal emotions" suggested in the psychology field.

Keywords: **awareness, attention, activity, facial expressions.**

1 Introduction

Awareness is a key feature in many areas, such as psychology and biology. Generally speaking, awareness comprises a human's or an animal's perception and cognitive reaction to a condition or event. This event can be produced by an internal state (e.g., a feeling), or an external one (e.g., a sensory perception).

In humans, awareness generates many internal and external behaviors. For example, hands' gestures and facial expressions are very important because they express and emphasize an emotion. In fact, one of the main awareness component consists of considering emotional reactions; for example by analyzing physical gestures people's feelings can be interpreted and categorized.

During face-to-face communication, awareness allows members to interact with each other. In this case, awareness and attention are terms used interchangeably, yet they are not synonymous. As mentioned in [5], attention has an intentional focus, directing cognition toward a particular target, differently from awareness.

During a remote communication, people can not see each other, so they can not fully observe their behaviour. Therefore, many problems may arise: turn-taking, interpretation, being passive or attentive, etc.

In this article, we suggest a new approach to fill the gap between a face-to-face communication and a remote one. We base our approach on attention as a particular feature of awareness to be measured during a remote communication.

2 Activity Awareness

In CSCW (Computer Supported Collaborative Work) community [8], the definition of awareness has not yet reached a consensus. During cooperative work, awareness seems to include knowledge and performance of a member interacting in the group work [11].

Turning now to awareness, we distinguish between five different kinds: multimedia, activity, collaborative, social, and action. Carroll et al, [1] introduced the idea of activity awareness. The activity awareness is based on actions that are performed during a work session. The main activity is divided into simple tasks trough a hierarchical structure. Each group member works on a restricted set of simple tasks and his performance is important for the global results of the work.

The activity theory is based on the idea that the activities of a group are more important than common knowledge. In fact, the analysis of each activity can provide some information about the user: his performance against the global performance, the time to complete the task, the relationship with the other members. The shared knowledge

is used to coordinate collaborative tasks that require common decisions.

An example of a CSCW system called Virtual School is presented in [2] (see Figure 1). Using this software, more users can work together on the same topic. It supplies multiple communication channels: chat, e-mail, video conferencing, and a collaborative notebook.

A remote collaborative work system must guarantee both sharing of common resources and awareness. Our analysis of awareness is based on activity theory and on the recognition of physical expressions.

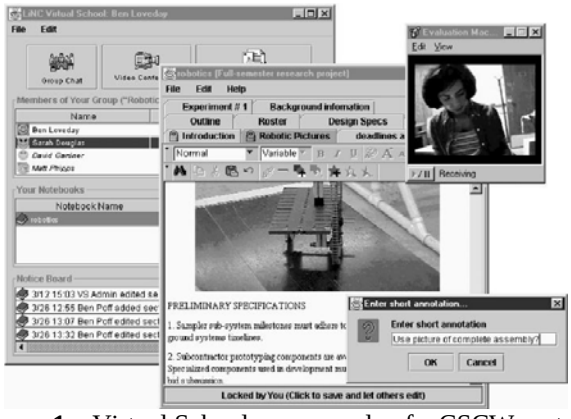

Figure 1 – Virtual School: an example of a CSCW system.

3 Combining facial expression recognition and activity theory for awareness detection

We describe in this section the preliminary experiments we have conducted which aim to combine facial expression recognition [3,4] with users' activity detection in order to allow the system to automatically estimate the level of attention of a user involved in a video conference with other remote users. Although approaches already exist in detecting awareness through emotions [10], we propose to integrate emotions recognition with users' activities.

Our system combines visual data (i.e., facial expression analysis) with other input data (e.g., mouse clicks, keystrokes). Altogether these data are called features and they are input to a statistical framework used to quantitatively estimate the user's involvement in the remote activity.

The statistical classifier we use is based on the well known Naïve Bayes approach in which the statistical distribution of each feature is assumed to be independent from the distributions of the other features [6]. If even the Naïve approach is quite simple it resulted useful for developing a prototype and moreover it usually reveals to be quite effectively for producing experimental applications. The aim of this section is to show how such a classifier has been trained: how data have been collected by means of an experiment with real users not aware of being analyzed by the system; and how such data have been analyzed. The work is in progress and the statistical classifier is currently under construction. Our aim is to automatically estimate the user attention by comparing an input feature vector with its previously learnt statistical model. Even if the system is not complete yet, we believe that the procedure we have adopted to collect and analyze genuine data from users can be interesting and useful for similar studies.

First, we briefly describe the system used to recognize user emotions through the analysis of her/his facial expressions [3,4]. The input of the system is a video sequence grabbed by a common webcam positioned in front of the user (e.g., the same video sequence used in the teleconference for human-to-human communication), see Figure 2.

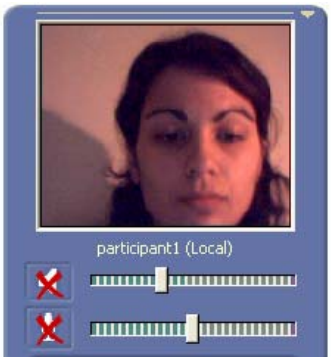

Figure 2 – An example of video frame used by the system.

Such a sequence is analyzed using the system described in [3,4] in order to map facial movements with facial expressions and, then, facial expressions to emotional states. A 3D wireframe model of a generic human face is aligned with the input face contained in the first frame of video sequence. The model is then wrapped in order to fit the particular input face. In the following frames of the video sequence, the wireframe model continues to be wrapped in order to fit possible facial deformations. Such deformations depend on possible facial muscles stretching and bending. By tracking the muscles' movements the model is able to represent the relative displacement modifications of facial features such as eyes, eyebrows or mouth corners, etc. Finally, by comparing the facial movements (better known as Action Units, AUs) with a statistical model previously constructed, the system is able to estimate the current user's most likely emotional state (see Figure 3). The output of this phase is a vector of seven elements, each element representing the probability of the user being in a given emotional state. The seven possible emotional states are: Neutral, Happy, Angry, Disgust, Fear, Sad and Surprise. These emotional states have been defined by Ekman [9] as the universal emotions and are the ones which are shared across different cultures, genders, and ages.

Let us call E_i the emotion estimation vector associated with the i-th frame of the video sequence. Then E_i is a 7-

component vector, where the j-th component gives the probability of the user being in the j-th emotional state. For instance, with reference to Figure 3:

$$E_i = (0, 0.2657, 0.4278, 0.002, 0, 0.3059, 0.004)^T$$

represents a user emotional state in which "surprise" is judged by the system as the most likely emotion (about 42% of probability), even if other states (e.g., fear") have good chances to be valid as well.

Figure 3 – Example of our real-time working system with the emotional state probabilities displayed on the right.

On the other hand, the user activity is represented by two features: the number of mouse clicks and the number of keystrokes measured in a given temporal interval. These features are computed by the Morae system [7]. Let A_i be the 2-component vector representing the above mentioned activity features. For instance:

$$A_i = (0.025, 0.9375)^T$$

represents the number of mouse clicks and keystrokes, respectively, intercepted by the system in a time interval of few seconds and averaged on the number of frames included in the interval. More specifically, the two values above refer to the quantities reported in Table 1 and a time interval composed of 80 frames.

Combining E_i with A_i we have a 9-component feature vector V_i which describes both activity and visual information on the user at time *i* (given by the i-th frame of the video sequence). This vector is used by the system in order to estimate the user attention at time *i*.

Before doing this, however, we need an (off-line) training phase in which the system compares examples of feature vector values with video sequence frames manually labeled (the ground truth) and learns the statistical distribution of each class (being a class an attention level). In other words, we have conducted experiments in which real users have been analyzed during their teleconferencing activity. By monitoring each single user activity we have a set of feature vectors (one for each frame of each video sequence). Each frame is labeled with a human attention level, ranging in three levels: low, medium and high level. These levels have been obtained by interviewing the users after the experiments. By comparing each frame label with the corresponding frame vector, we trained off line the statistical classifier. We chose a Naïve Bayes classifier with Gaussian class-conditional probability density function of each feature [6]. The main reason for choosing this classifier is its ability to cope with uncertainties in the input data. Once the parameters of the statistical model have been learned by the system, the statistical classifier is used (on-line) to automatically estimate the posterior probability of each class (i.e., the user attention level) given a vector V_i representing the user combined emotional and activity state at time *i*.

Our experiments involved five different persons. We have collected videos of voluntary students having a background in different disciplines (computer science, psychology and statistics). All the five persons have participated to two video-chat sessions. The topics of the two sessions have been pre-chosen, using two different subjects for the two different sessions. During the first one the volunteers were invited to talk about a recent proposal for an Italian law about unmarried couples; the second one was on their summer holiday plans.

During the chat, each person could see all the other participants on her/his screen and could communicate with them either by sending text or by using the webcam positioned on her/his monitor. No audio information has been used. This was done for two reasons. First, the lack of speech simplifies the facial expression analysis: since the users have not spoken during the session, facial muscle movements have been influenced only by emotional state changes. Second, since audio information is not used by our system, we forced people to communicate by means of only text which, in turn, is analyzed by the Morae system [7].

All the involved persons were not aware of the actual aim of the experiment. What they knew was that the aim of the experiment was to test a particular video-chatting tool. Thus, they did not know that we were interested in "measuring" their activity and facial expressions as well as their attention level. This has been done in order to assure genuine facial expressions and emotional states. We declared the actual aim of the experiment just after its completion. Finally, we separately asked to each volunteer to:

A) Watch their own video sequence and mark prefixed-size subsequences with an attention label ranging in low, medium and high.

B) To provide a personal interest-based ranking of the subjects they dealt with during the video-chat.

By comparing data from A) and B) we associated each video sequence frame to an attention level L_i. The set of pairs feature vector-attention level, (V_i, L_i), is used as ground truth in the training stage.

4 Attention visualization to the user

We define an attention measure based on activities and facial expressions. In this section, we present an example for the corresponding visualization. During a remote communication, producing a sampled video stream we

selected a particular frame and we extracted information about the activities: number of mouse clicks and keystrokes. This frame is also analyzed to recognize facial expressions: we obtain seven components, one for each emotional state.

Table 1 contains information about activities. The total number of activities is 77, mouse clicks are 2 and keystrokes are 75. These numbers mean that in the time interval considered (corresponding to 80 video frames of the whole video chat session) the user has done 75 keystrokes and 2 mouse clicks.

Each frame represents an interval of about twenty seconds. So 77 is the total number of activities performed by the user during the last interval.

Frame	Mouse Clicks	Keystrokes	Total Events
0:02:32.33	2	75	77

Table 1 – Example of Activities.

Figure 4 contains the representation of facial expressions: for each emotional state we fill in a triangle to illustrate the relative values.

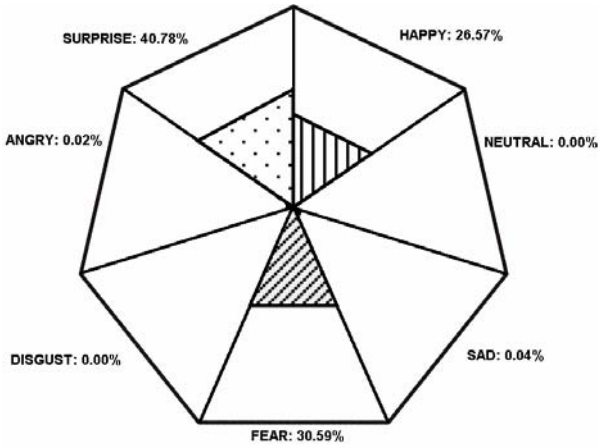

Figure 4 – Representation of Facial Expressions.

Finally, Figure 5 shows the representation of the attention level, in this case medium, obtained from the user during the training of the system.

We believe that combining the two modalities (i.e., facial expressions and activity) we will not only be able to achieve a better estimate of the attention displayed by a user during the remote communication but we will be also capable of coping with situations in which one modality is not visible or is not accurate enough (e.g., not frontal face, no mouse or keyboard activity, etc.).

Figure 5 – Representation of Attention Level.

5 Conclusions

We presented a way to detect a particular feature of awareness: attention. With a preliminary experiment, we constructed a Naïve Bayes model that can output an attention value. Inputs to this model are: 1) number of activities (2) and 2) a set of facial expressions (7). We have shown how activities and emotional states can help to capture the users' awareness during a remote communication. Our purpose is to cover the gap between face-to-face and remote communication. This work is still in an initial phase, yet, with future experiments, our model can be improved and used on-line to compute the attention level for each frame of a video stream. In this way everybody participating in a remote communication activity can then know the attention level of the other participants.

Acknowledgements

Work supported by project MODUWEB (TIN2006-09678) of the Spanish Ministry of Education and Science

References

[1] J. M. Carroll, M. B. Rosson, G. Convertino, C. H. Ganoe. Awareness and Teamwork in Computer-Supported Collaborations. *Interacting with Computers* 18(1): 21-46, 2006.

[2] P. L. Isenhour, J. M. Carroll, D. C. Neale, M. B. Rosson, D. R. Dunlap. The Virtual School: An integrated collaborative environment for the classroom. *Educational Technology and Society* 3:1436-4522, 2000.

[3] I. Cohen, N. Sebe, A. Garg, L.S., Chen, L. S., T.S. Huang, Facial expression recognition from video sequences: temporal and static modeling. *Comput. Vis. Image Underst.* 91(1-2):160-187, 2003.

[4] N. Sebe, M.S. Lew, I. Cohen, Y. Sun, T. Gevers, T.S. Huang, Authentic Facial Expression Analysis, *Proc. International Conference on Automatic Face and Gesture Recognition*, 517- 522, 2004.

[5] www.afn.org/~gestalt/. *The Gestalt Center* of Gainesville.

[6] R. O. Duda and P. E. Hart and D. G. Strorck, *Pattern classification (2nd ed.)*, Wiley Interscience, 2000.

[7] Morae Usability Testing for Software and Web Sites, © 1995-2007, TechSmith Corporation

[8] J. Grudin, Computer-supported Cooperative Work: Its History and Participation, *IEEE Computer*, pp. 19-26, May 1994.

[9] P. Ekman, Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique, *Psychology Bulletin* 115(2): 268–287, 1994.

[10] R. W. Picard. Affective Computing. MIT Press. September 1997. ISBN-10: 0-262-16170-2

[11] Kaptelinin, V. and Nardi, B. A. 2006 *Acting with Technology: Activity Theory and Interaction Design (Acting with Technology)*. The MIT Press.