Emotion Tracking for Remote Conferencing Applications using Neural Networks.

Emotion Tracking for Remote Conferencing Applications Using Neural Networks

Paul Smith and Sam Redfern

Discipline of Information Technology, College of Engineering & Informatics, National University of Ireland, Galway, Ireland.
{p.smith2, sam.redfern}@nuigalway.ie

Abstract. In face-to-face work, discussion and negotiation relies strongly on non-verbal feedback, which provides important clues to negotiation states such as agreement/disagreement and understanding/confusion, as well as indicating the emotional states and reactions of those around us. With the continued rise of virtual teams, collaborative work increasingly requires tools to manage the reality of distributed and remote work, which is often hampered by a lack of social cohesion and such phenomena as participants multi-tasking rather than paying full attention. This paper discusses the use of a neural network-based emotion recognition system and describes its application to the monitoring of presence and emotional states of participants in virtual meetings. Experimental analysis shows our Emotion Tracking Agent (ETA) to have marginally better accuracy at recognising universal emotions than human subjects presented with the same data.

Keywords: Neural Networks; Emotion Tracking; Virtual Teams; Multitasking; CSCW;

1 Introduction and Motivation

In face-to-face work, discussion and negotiation relies strongly on non-verbal feedback, which provides important cues related to conversational negotiation states such as agreement/disagreement and understanding/confusion, as well as indicating the emotional states and reactions of those around us. This level of virtual presence is typically missing in remote collaboration tools. With the continued rise of virtual teams, collaborative work increasingly requires tools to manage the reality of distributed and remote work, which is often hampered by a lack of social cohesion and such phenomena as participants multi-tasking rather than paying full attention to the collaborative work.

In this paper we describe the development and experimental assessment of neural network based software agents for monitoring the presence and emotional states of co-participants in virtual meetings. The Emotion Tracking Agent (ETA) tracks participants' facial movements in real-time and uses a feed forward neural network to estimate the emotions being portrayed. More detail on our ETA's intended application is given in [1].

In order to assess the accuracy of the technique, we initially apply it to the six 'universal' emotions (e.g. fear, surprise) published in the works of Paul Ekman [2] in the psychology literature. The deployed ETA, however, operates on the more ambiguous 'non-universal emotions' (e.g. agreement, confusion), which are more useful in the context of meetings. The non-universal emotions are recorded by meeting participants and taught to the ETA's neural network on a per-user basis.

2 Application Background

In modern work environments, teleconferencing and other forms of virtual meetings have become increasingly important due to the number of companies employing distributed workers. Many companies such as Intel (70%), IBM (40%) and Sun Micro Systems (nearly 50%) already have high percentages of virtual and distributed workers [3]. Some of these workers are based at home or abroad or travel frequently and for these individuals, virtual conferencing is the most convenient and economical means of communication with colleagues and clients.

Some companies opt not to have a physical base of operations at all, choosing instead to work exclusively in virtual teams. The company Accenture, for example, reports that it has no physical workspace but chooses instead to meet virtually or while crossing paths while travelling from client to client across the globe [4].

It is predicted that by 2012, 40% of the USA'S workforce will qualify as distributed and the rest of the world will soon follow [5]. The improvement of software for the support of virtual teamwork is clearly of great significance to the future of knowledge work.

An important current topic in the literature relates to the advantages and disadvantages of multitasking in both traditional and virtual work environments [6]. Successfully managing worker multitasking and providing solutions to specific multitasking disadvantages is of key importance to future work practices.

Receiving much attention in this area is the effect of multitasking on a user's attention and performance in a virtual meeting [7]. The problem here lies in the fact that during virtual meetings, far more than in face to face meetings, participants often juggle multiple tasks: for example checking emails or browsing the internet, performing other work or non-work activities. In a recent survey [5] of 385 respondents it is reported that 90% of all teleconference participants engage in multitasking during meetings. Some of the activities named by these respondents were: unrelated work (70%), email or instant messaging (50%), eating (35%), muting for side conversations (35%), surfing the web (25%) and even driving (12%).

Attention and awareness are important elements in communication and collaborative work, in order to fully understand and follow the reactions and sentiments of co-workers. However, multitasking during conference calls causes certain information points and important non-verbal information to be missed [8], due largely to the importance of non-verbal communication (NVC) [8]. Facial expressions are particularly important, since this channel is the most immediate indicator of the emotional state of a person [2].

The importance of NVC in everyday collaboration has been well covered in the literature, and it is reported in a number of studies that the quantity of information portrayed through NVC is significantly larger than verbal communication alone [10][11]. The receiver in a communicative interaction is naturally predisposed to base the sender's intentions on the non-verbal cues received. If this communication channel is ignored during virtual meetings, whether due to reasons of multitasking, lapse of attention or otherwise, it is clear that a large amount of the information available to a user in a normal face-to-face interaction is missing, which will hinder communication effectiveness as well as efficiency for everyone involved in the interaction.

Emotion gives communication life. With the absence of emotion, the messages we perceive from what a person says can become ambiguous. Paul Ekman developed a list of six emotions whose facial displays are universally recognized and transcend culture and ethnicity [2]. These emotions are: Anger, Disgust, Sadness, Fear, Happiness and Surprise, and are referred to as primary or 'universal' emotions.

3 Automated Emotion Recognition

3.1 Background

Much of the research conducted into the synthesis and recognition of emotions through non-verbal signals has been aimed at creating natural human-machine interfaces. Other application domains for automated analysis of facial expressions include its use in behavioral science and medicine [12][13].

Research into the automatic recognition of facial expressions focuses on issues related to the representation and classification of the static or dynamic characteristics of the deformations of facial components, and their special relations which form the basis of facial expressions [14]. Systems designed for automatic recognition of facial expressions are commonly structured as a sequence of processing blocks built around a conventional pattern recognition model. These blocks include image acquisition, pre-processing, feature extraction and classification [14].

The pre-processing stage is one of the most important due to the extent of distortions caused by head rotation, distance from camera and in the varying head and facial feature proportions. The methods used to solve these problems commonly involve some form of standardization or normalization of the input image before the data is sent to the recognition system itself: typically, this involves geometric correction in order to ensure adherence to a standard image which exhibits an ideal facial pattern. These standardization techniques ensure that all affine transformations such as skewing, rotation or scaling do not affect the data being tested.

In the face recognition literature, face orientation has received particular attention. For simplicity, most research projects in this area assume that the orientation of the face is limited to only in-plane movement [15], or that out-of-plane movement is negligible [16][17].

Typically, the process of expression categorisation is achieved by the use of a classifier, commonly consisting of a pattern recognition model paired with some form of decision process for calculating the final output. Many different methods have been used for creating automatic facial expression classifiers consisting of both parametric and non-parametric techniques [18]. In the majority of these classifiers, the final output decision is given as a description using facial action units (FAUs) [19], or as one of the six universal facial expressions defined by Ekman [20]. Some approaches use a technique of grouping specific facial muscles and describing the movement of these groups in combination with each other [21].

3.2 Design of the Emotion Tracking Agent (ETA)

For the purposes of the current project, we choose to focus primarily on the classification stage of the problem, and on the use of the classifier in a proof-of-concept remote collaboration tool. In order to simplify the image acquisition and pre-processing stages, we therefore use a specialist optical motion capture camera, the Optitrack FLEX: C120 developed by the company Naturalpoint [22] (see Fig. 1). This camera is an integrated image capture and processing unit which uses a B&W CMOS imager to captures 120 frames of video per second and an onboard image processer which transfers marker data over standard USB to a computer for display and post processing. The camera pre-processes the image to remove most light, preventing anything from being displayed in the image unless it is from a highly reflective surface. The camera contains a ring of 12 infrared LEDs which are used to illuminate the tracking markers which are made of a highly reflective material. The Software Development Kit (SDK) provided with the camera allows developers to track these markers and write their own tracking applications.

Fig. 1. Front and side views of our chosen tracking solution, the Optitrack FLEX: C120 developed by Naturalpoint.

The ETA is written in C++, and incorporates a Multi-Layer Perceptron (MLP), trained using back propagation. Data is received from the Optitrack camera via the SDK. The MLP with backpropagation was chosen due to its prevalence in the facial

tracking literature, and due to the flexibility it would afford in allowing users to record their own specific emotions and use them as training for the network [23].

The main considerations in the design of our ETA were the representation of network inputs, designing appropriate pre-processing and feature extraction methods, and developing a suitable output interpretation model for the decision phase of the recognition process. The basic structure of our completed network model took the form of an MLP consisting of three separate layers: the input layer, a single hidden layer where most of the classification is performed and a final output layer in which a single emotion is identified as the dominant among Ekman's universal emotions (see Fig. 2).

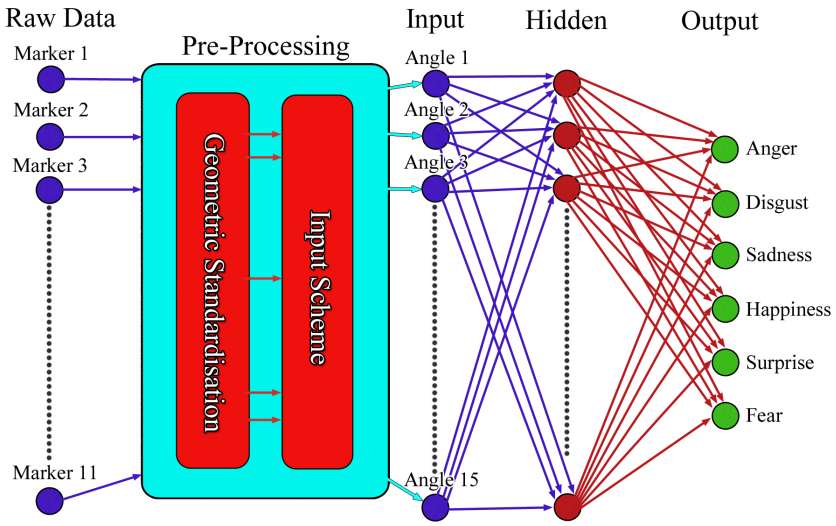

Fig. 2. The basic structure of the ETA. Raw x and y coordinates of the 11 markers are sent to the pre-processor, where they are geometrically corrected and transformed into an appropriate scheme (consisting of angles between specific markers) for the neural network. The hidden layer contains a single node for each node in the input layer, and the output layer contains nodes for each of Ekman's universal emotions.

The inputs to our ETA take the form of the x and y coordinates of 9 tracked markers, which are placed in positions on the user's face that have previously been shown to have the most significant movement during facial expression display [24] (see Fig. 3). The recorded facial data is pre-processed into appropriate form using our network input conversion scheme, which is described below and illustrated in Fig. 5. This input schema converts the coordinate data into a form which efficiently captures the shape of the facial expression pattern, and allows the data to be served to the hidden layer of our neural network for output calculation and analysis.

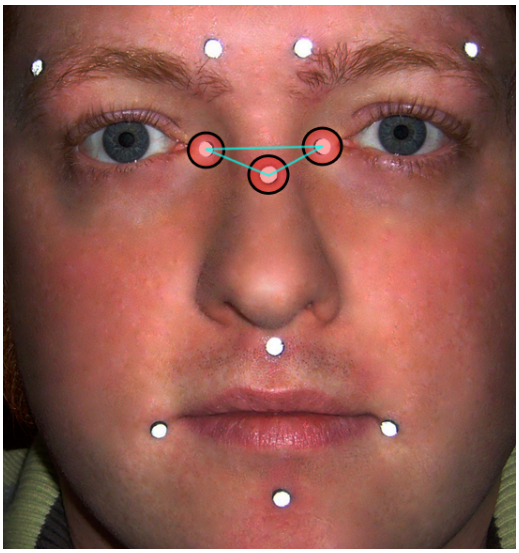

Fig. 3. Positions of tracking markers placed on a user's face. The 3 highlighted markers are those used in pre-processing to calculate and remove affine transformations distorting the input data.

Our standardization method consists of the calculation of rotation and fluctuation in scale of the user's head from its ideal position, and re-translating the marker coordinates to remove these affine transformations, in order to prevent distortion to the input data given to the neural network. This was achieved by adding two extra markers to the user's face placed in the medial canthus of both eyes (see Fig. 3 above). These markers and the marker placed on the upper bridge of the nose were chosen since they are placed in positions on a person's face which have no significant movement during the display of facial expressions [24]. The angles of the triangle formed by these points are computed along with the length of each side. These calculations are then compared with our ideal lengths and angles to compute any transformations or scaling of the head. The distance and angle between these points and the camera viewpoint origin (0, 0) is also calculated to compute the rotation of the head. The markers are all rotated about the centre of the rectangle formed by the boundary x and y dimensions of the 9 markers (see Fig. 4).

The final pre-processing stage consists of the calculation of specific angles between the vectors formed by the marker coordinates. These angles are displayed in Fig. 5 and are labelled a1 to a15. This fifteen-angle scheme was chosen as the best of among a number of candidate schemes that we experimentally assessed (see discussion below).

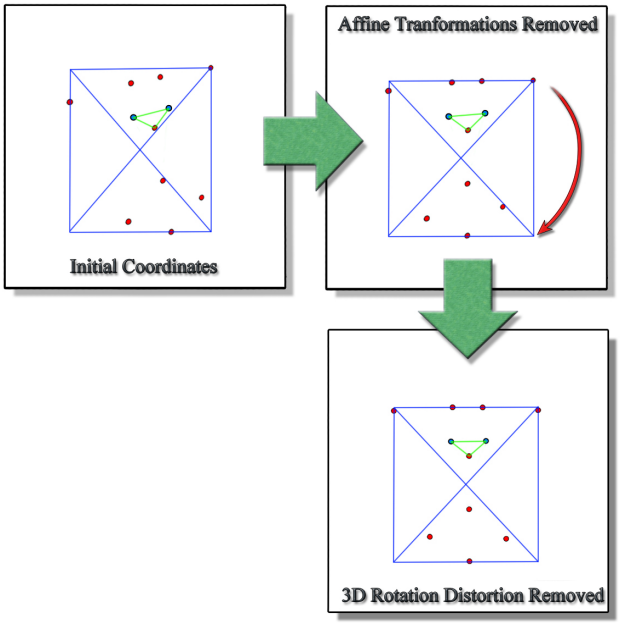

Fig. 4. Removal of affine transformations from recorded marker coordinates is achieved via three stationary markers (those that form the small triangle) before input to the neural network.

The output layer of the MLP consists of 6 nodes representing each of Ekman's prototypic emotions, each of which holds a value ranging from zero to one representing the probability of the node's corresponding emotion matching the tested input pattern. Once these values have been computed an output interpretation model is used to make the final emotion classification decision, and this information is transmitted to all clients logged into the online interaction. Alerts, warnings and other feedback is then broadcasted to the participants depending on the emotions being exhibited and the current state of each participant.

3.3 Input Representation Scheme

In order to arrive at an appropriate input representation for our marker coordinates, a number of experiments were carried out to evaluate candidate input schema. There were 3 main candidates used in the initial evaluation experiment. The first scheme computed the y distance between the markers, the second used both the x and y coordinates of each marker to compute their 2-dimensional distance, and the last used the angles between various markers in an attempt to describe the pattern formed by the facial expression as accurately as possible.

Data was recorded from a set of facial expressions depicting universal emotions, and the data was transformed according to the candidate schema before being submitted to a cluster analysis. During this experiment, 30 instances of each of the six

universal emotions were recorded from 3 separate recording sessions. The cluster analyses indicated that the angle representation scheme provided a better clustering of similar emotions than did the other schema.

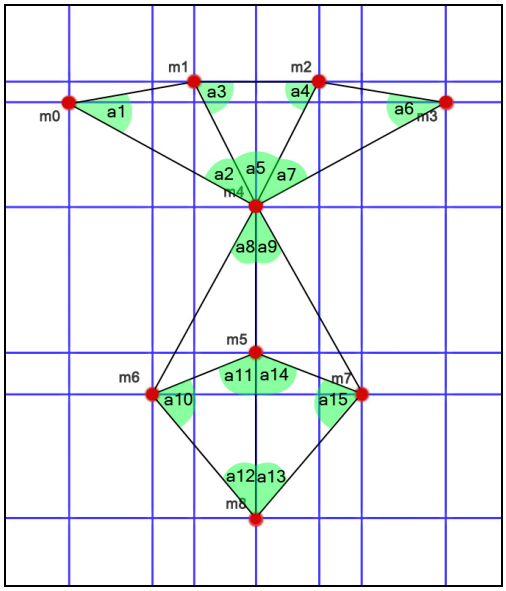

Fig. 5. The angles which are calculated to represent the facial expressions pattern using our chosen input scheme.

Further tests were then carried out to deduce the best combination of angles to use in order to optimize network accuracy, from among a number of candidate sub-schema involving 7, 9, 11, 13 and 15 angles. Fig. 5. illustrates the final input representation scheme adopted.

3.4 Assessment of the ETA

Our MLP was initially trained on 40 sets of recorded marker coordinates for each of the six universal emotions. Although these are not, realistically, the emotions that are highly relevant to efficient communication in normal work circumstances, they allow us to directly assess the accuracy of the system in comparison to human subjects presented with video data captured at the same time. The ambiguity of the non-universal emotions would make this assessment impossible.

Each set of data was recorded from a face imitating the descriptions given in Paul Ekman's research for the emotion it represented. The training algorithm used is back-propagation, and we applied an early-stopping approach in order to prevent over-training, through the application of unseen data.

The accuracy of the MLP was tested against a dispersed group of 15 people. These experiment participants were shown videos of the same facial expressions used in the

previous experiments. These videos were recorded by a digital video camera at the same time as the Optitrack FLEX:C120 camera captured the testing data for the ETA. This ensured the validity of the experiment, in that the human subjects were presented with identical footage, in video form, as that received by the ETA.

18 of the 180 recorded emotions were chosen for the test (3 for each primary emotion) emotion, and the 18 corresponding sets of marker coordinate data were used to test the accuracy of the neural network. This allowed a direct comparison to be conducted between both the recognition rate of the human and neural network tests. The results of these experiments are shown in Tables 1 and 2. During the human test we observed a 78% accurate recognition rate, while our neural network using our best output interpretation model resulted in 96%.

Table. 1. The recognition results from our human test showing the summed number of correct, incorrect, partially correct (ie a similar emotion commonly mistaken for said emotion eg. Anger and Disgust) and entirely incorrect answers

Correct	Incorrect	Partially Correct	Incorrect Entirely
211	59	41	18

Table. 2. The recognition results from our neural network using our best output interpretation algorithm, which labels the emotion output with the highest value as the identified emotion

Correct	Incorrect	Partially Correct	Incorrect Entirely
259	11	8	3

4 Conclusions

The ETA software described in this paper provides an additional level of communication and awareness for users collaborating within a virtual meeting environment. It provides a means of tracking all participants' facial expressions, and of prompting co-participants with helpful details related to emotional states, in order to reduce the negative effects of multitasking and lack of attention. It attempts to raise the user's awareness of other people's reactions during a meeting.

We are currently undertaking a series of experiments and a case study in which the ETA is assessed in its deployed state in a multi-user collaborative environment.

References

1. Smith, P., Redfern, S.: Facial Expression Tracking for Remote Conferencing Applications: An Approach to Tackling the Disadvantages of Remote Worker Multitasking. In: Proceedings of the Second International Conference on Games and Virtual Worlds for Serious Applications, pp. 91--92 (2010)
2. Ekman, P., Friesen, W.V.: Unmasking the Face. Prentice-Hall Inc, New Jersey (1975)
3. Barczak, G.: Have advice will Travel: Lacking Permanent Offices, Accenture's Executives Run'Virtual' Company on the Fly. Wall street Journal (2006)

4. Conlin, M.: The Easiest Commute of All, Business Week, (2005) http://www.businessweek.com/magazine/content/05_50/b3963137.htm
5. Gilbert, A.: Can't Focus on the Teleconference? Join the Club. C-Net, (2004) http://news.cnet.com/Cant-focus-on-the-teleconference-Join-the-club/2100-1022_3-5494304.html
6. Appelbaum, S.H., Marchionni A., Fernandez, A.: The multitasking paradox: perceptions, problems and strategies. Management Decision 46(9), 1313--1325 (2008)
7. Lojeski, K.S., Reilly R., Dominick, P.: Multitasking and Innovation in Virtual Teams. In: Proceedings of the 40th Hawaii International Conference on System Sciences, pp. 44b (2007)
8. Vlaar, P. W. L., van Fenema, P. C., Tiwari, V.: Cocreating understanding and value in distributed work: how members of onsite and offshore vendor teams give, make, demand, and break sense. MIS Quaterly, 32(2), 227 -- 255, ISSN 0276-7783 (2008)
9. Knapp, M. L.: Nonverbal Communication in Human Interaction (2nd Edition). Holt, Rinehart and Winston Inc.,New York (1978)
10. Argyle, M., Salter, V., Nicholson, H., Williams, M., Burgess, P.: The communication of inferior and superior attitudes by verbal and non-verbal signals. British journal of social and clinical psychology vol. 9, pp. 222--231 (1970)
11. Pantic, M., Rothkrantz, L.J.M.: Automatic Analysis of Facial Expressions: the State of the Art. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1424--1445 (2000)
12. Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P., Sejnowski, T.J.: Classifying Facial Actions. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 10, pp. 974--989 (1999)
13. Essa, I.A., Pentland, A.P.: Coding, Analysis, Interpretation, and Recognition of Facial Expressions. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 757--763 (1997)
14. Chibelushi, C.C., Bourel, F.: Facial Expression Recognition: A Brief Tutorial Overview. In: CVonline: On-Line. Compendium of Computer Vision (2003)
15. Bartlett, M.S., Hager, J.C., Ekman, P., Sejnowski, T.J.: Measuring facial expressions by computer image analysis. Psychophysiology, vol. 36, pp. 253--264 (1999)
16. Lien, J.J.J., Kanade, T., Cohn, J.F., Li, C.C.: Detection, tracking, and classification of subtle changes in facial expression. Journal of Robotics and Autonomous Systems, in press, 31, 131--146 (2000)
17. Lien, J.J.J., Kanade, T., Cohn, J.F., Li, C.C.: Automated facial expression recognition. In: Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 390--395 (1998)
18. Pantic, M., Rothkrantz, L.J.M.: Automatic Analysis of Facial Expressions: the State of the Art. In: IEEE Transations on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1424--1445 (2000)
19. Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P., Sejnowski, T.J.: Classifying Facial Actions. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 10, pp. 974--989 (1999)
20. Ekman, P.: Emotion in the Human Face. Cambridge University Press, New York (1982)
21. Ekman, P., Friesen, W.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press (1978)
22. NaturalPoint, Optitrack Motion capture solution, http://www.naturalpoint.com/optitrack/
23. Pantic, M., Rothkrantz, L.J.M.: An Expert System for Multiple Emotional Classification of Facial Expressions. In: Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence, pp. 113--120 (1999)
24. Zhang, Z.: Feature-based facial expression recognition: Sensitivity analysis and experiments with a multilayer perceptron. International Journal of Pattern Recognition and Artificial Intelligence, 13(6), 893--911 (1999)