Dimensional Emotion Prediction from Spontaneous Head Gestures for Interaction with Sensitive Artificial Listeners

Hatice Gunes and Maja Pantic

Department of Computing, Imperial College London
180 Queen's Gate, London SW7 2AZ, U.K.
{h.gunes,m.pantic}@imperial.ac.uk

Abstract. This paper focuses on dimensional prediction of emotions from spontaneous conversational head gestures. It maps the amount and direction of head motion, and occurrences of head nods and shakes into arousal, expectation, intensity, power and valence level of the observed subject as there has been virtually no research bearing on this topic. Preliminary experiments show that it is possible to automatically predict emotions in terms of these five dimensions (arousal, expectation, intensity, power and valence) from conversational head gestures. Dimensional and continuous emotion prediction from spontaneous head gestures has been integrated in the SEMAINE project [1] that aims to achieve sustained emotionally-colored interaction between a human user and Sensitive Artificial Listeners.

Keywords: spontaneous head movements, dimensional emotion prediction, virtual character-human interaction.

1 Introduction

Researchers have extensively studied the behavior of the head within social interaction context. They have observed that 'the head movements we make when we speak are not random; these movements mark the structure of the ongoing discourse and are used to regulate interaction' [2], [3]. For instance, side-to-side shakes have been shown to correlate with verbalizations expressing inclusivity, intensification, and uncertainty [3]. Some speaker head nods have been shown to have an interactive function in triggering back-channels to which listeners appear to respond within a second. Therefore, automatic detection of head nods and shakes can be seen as a valuable non-verbal component for achieving natural and affective computer-human and virtual character-human interaction.

When it comes to recognizing human affect and emotion, the mainstream research has mostly focused on facial and vocal expressions and their recognition in terms of seven discrete, basic emotion categories (neutral, happiness, sadness, surprise, fear, anger and disgust). In line with these, most of the past research on automatic emotion sensing and recognition has focused on recognition of facial

J. Allbeck et al. (Eds.): IVA 2010, LNAI 6356, pp. 371–377, 2010.

and vocal expressions in terms of basic emotional states, and then based on data that has been posed on demand or acquired in laboratory settings [4]. However, a number of researchers have shown that a single label (or any small number of discrete classes) may not reflect the complexity of the affective state conveyed by various rich sources of information [5]. These researchers advocate the use of dimensional description of human affect, where an affective state is characterized in terms of a number of latent dimensions (e.g., [5], [6]).

The search for the optimal low-dimensional representation of emotion remains open [7]. Fontaine et al. showed that four dimensions are needed to satisfactorily represent similarities and differences in the meaning of emotion words for three languages: valence (the level of positivity or negativity), activation (the degree of excitement or apathy), power (the sense of control), and expectation (the degree of anticipating or being taken unaware) [7]. Ideally it should be possible to obtain the overall level of intensity from other more specific dimensions. However, to obtain a more complete description of affective coloring, at times intensity is included as the fifth dimension.

This paper focuses on automatic and dimensional prediction of emotions from head gestures. More specifically, we focus on the mapping between the amount and direction of head motion, and occurrences of head nods and shakes and the arousal, expectation, intensity, power and valence level of the observed subject as there has been virtually no research bearing on this topic.

2 Related Work

Various techniques have been developed for the detection and recognition of head nods and shakes. [8] observed the spatial evolution of the 'between eyes' circle that serves as a template for tracking and a basis for head movement estimation. The system was tested on 450 frames collected from three people who moved their heads up and down and left and right. An overall recognition accuracy of about 80% was reported. [9] recognizes head nods and head shakes based on two Hidden Markov Models (HMMs) trained and tested using 2D coordinate results from an eye gaze tracker. A total number of 110 samples from 10 subjects answering a number of factual questions (with a head nod or a head shake) were collected, and a recognition rate of 78.46% was reported. [10] calculates eye coordinate using an AdaBoost-based classifier along with two HMMs (one for nods and one for shakes). Data (80 samples for training and 110 samples for testing) were collected by asking the participants a number of factual questions, and a recognition accuracy of 85% was reported. [11] uses a head pose tracker that outputs a head rotation velocity vector. Based on that the head movement is classified by a two-class (nods and shake) SVM. The system was trained with 10 natural head gesture sequences taken from interactions with an embodied agent and 11 on-demand head gesture sequences. When tested on 30 video recordings of 9 subjects interacting with an interactive robot, a true detection rate of 75% for nods and 84% for shakes was obtained. Incorporating speech context further improved detection.

When it comes to automatic and dimensional emotion recognition, the most commonly employed strategy is to reduce the recognition problem to a 4-class problem (classification into the quadrants of 2D arousal-valence space, e.g. [12]), a 3-class valence-related classification problem (positive, neutral, and negative emotion classification), or a 2-class problem (positive vs. negative and active vs. passive classification). Systems that target automatic dimensional emotion recognition, considering that the emotions are represented along a continuum, generally tend to quantize the continuous range into certain levels. Representative works include quantization into low, medium and high [13] and excited-negative, excited-positive and calm-neutral [14]. More recently, works focusing on continuous prediction of arousal and valence from the audio modality have also emerged (e.g. [15]). See [16] for details.

As is reflected by the summary of related work, virtually no work has focused on dimensional prediction of emotions in terms of arousal, expectation, intensity, power and valence dimensions from spontaneous conversational head gestures.

3 Methodology

3.1 Data and Annotations

The Sensitive Artificial Listener database (SAL-DB) [17] consists of emotionally colored conversations. The SEMAINE Database (SEMAINE-DB) [18] builds upon and extends the concept of SAL-DB, and has been created as part of the SEMAINE project [1]. In both cases, spontaneous (induced) data was collected by recording conversations between a participant and an operator undertaking the role of a Sensitive Artificial Listener (SAL) with four personalities: Poppy (happy), Obadiah (gloomy), Spike (angry) and Prudence (pragmatic). The SAL characters are virtual dialogue partners, based on audiovisual analysis and synthesis [19]. Despite their very limited verbal understanding, they intend to engage the user in a conversation by paying attention to the user's emotions and non-verbal expressions.

Automatic detection of head nods and shakes from visual cues requires the existence of annotated nod and shake videos. To this aim, training data was obtained by visually inspecting the SAL-DB and manually cutting 100 head nod and 100 head shake clips of variable length.

The SEMAINE-DB contains 20 participants, a total of 100 character conversational and 50 non-conversational recordings of approximately 5 minutes each. Recordings have been annotated by multiple coders who provided continuous annotations with respect to five affect dimensions (arousal, expectation, intensity, power and valence) using the Feeltrace annotation tool. Feeltrace allows coders to watch the affective behavior recordings and move their cursor within each emotional dimension separately, i.e. within the value range of [-1,+1], to rate their impression about the emotional state of the subject (see [16] for details).

For our preliminary experiments, we chose a maximum number of sessions that have been annotated by the highest number of coders (not all coders annotated

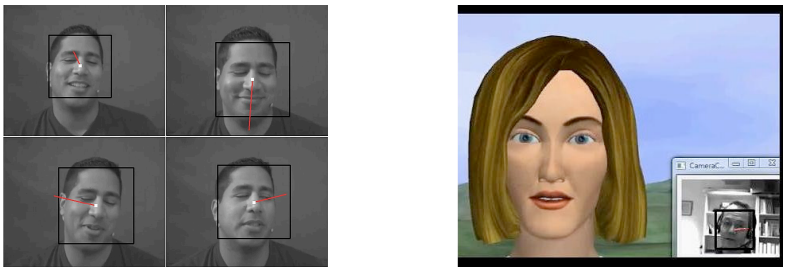
all sessions). This provided us with data from 7 subjects (27 sessions and 351, 510 video frames) annotated by 3 different coders.

3.2 Feature Extraction

Automatic detection of head nods and shakes is based on the 2-dimensional (2D) global head motion estimation. The facial region is detected using the well known technique of [20]. In order to determine the magnitude and the direction of the 2D head motion, optical flow is computed between two consecutive frames. It is applied to a refined region (i.e., resized and smoothed) within the detected facial area to exclude irrelevant background information. After preliminary analysis, the angle feature has been considered as the distinguishing feature in order to represent nods and shakes. The angle measure has then been discretized by representing it with directional codewords. The directional codeword is obtained by quantizing the direction into four codes for head movements (for rightward, upward, leftward and downward motion, respectively) and one for no movement. Figure 1(a) illustrates results from the feature extraction process.

3.3 Nod and Shake Detection

The directional codewords generated by the visual feature extraction module were fed into an HMM for training a *nodHMM* and a *shakeHMM*. However, to be able to distinguish other head movements from the actual head nods/shakes, we (i) threshold the magnitude of the head motion, (ii) build an *otherHMM* to be able to recognize any movement but nods/shakes, and (iii) statically analyze the likelihoods outputted by the nod/shake/other HMM (maximum likelihood vs. training classifiers on the outputted likelihoods). In order to analyze the visual data continuously, we empirically chose a window size of 0.6 secs that allows the detection of both brief and longer instances of head nods/shakes (similarly to other related work [9], [10]).



(a) (b)

Fig. 1. (a) Illustration of the 2D global head motion and angle estimation; (b) illustration of the SEMAINE system where a human subject is conversing with Poppy

3.4 Dimensional Emotion Prediction

From the global head motion features extracted and the head movements (nod or shake) detected, we created a window-based feature set that consists of total duration of the head movement in terms of codewords, average of the magnitude and angle values (within the window), standard deviation of the magnitude and angle values, loglikelihoods outputted by nodHMM, shakeHMM and otherHMM, the results of the maximum likelihood classification and the classifier trained on the outputted likelihoods. The ground-truth for the window at hand consists of the dimensional annotations averaged over that window, for each coder separately. Such a representation allows us to consider each feature vector independently of the others using the so-called static (frame-based) regressors. We considered the Support Vector Machines for Regression (SVR) [21] to the aim of dimensional emotion prediction from head gestures as they are among the most widely used regressors in the field [16]. The final feature set was scaled in the range of $[-1, +1]$. The parameters of SVR, for each coder-dimension combination, were optimized using 10-fold cross-validation on a subset of the data at hand.

We employ subject-dependent cross-validation evaluation as most of the works in the field report only on subject-dependent dimensional emotion recognition when number of subjects and data are limited (e.g.,[15]). For evaluation we used all data from 7 subjects and 27 sessions: $11,717$ instances obtained by processing $351,510$ video frames. Evaluation then has been done by adopting 10-fold cross-validation over the full data set. The measure of performance is the mean squared error (MSE) that measures the average of the square of the errors. MSE is reported for each coder-dimension combination separately.

4 Preliminary Experiments and Results

The aim of the dimensional emotion prediction experiments is to demonstrate how well the trained SVR models are able to predict the arousal, expectation, intensity, power and valence level of the user compared to the human coders (i.e., the ground truth). The MSE for each coder's annotation was estimated by constructing vectors of coder annotation pairs that correspond to each video session, and averaging the results over all estimations (that the coder has contributed to).

Our findings, presented in Table 1, confirm that it is possible to obtain dimensional emotion prediction from conversational head gestures, and overall, the trained SVR predictors provide an MSE level comparable to human coders.

To the best of our knowledge, it is not possible to directly compare our results to other state-of-the art systems in the field due to lack of works reporting on automatic dimensional emotion prediction from spontaneous head movements. The work presented by Wollmer et al. [15] that extracts audio cues to obtain automatic dimensional emotion prediction in arousal and valence space is the most similar one to the work reported in this paper in terms of context and data used. They conducted subject-dependent hold-out evaluation using spontaneous

Table 1. Comparison of the MSE values obtained: human coders' annotation (C1-C3) vs. prediction of the trained SVRs (P1-P3) for five emotional dimensions

coder	arousal	expectation	intensity	power	valence	average
C1	0.137	0.044	0.091	0.121	0.068	0.092
P1	0.128	0.128	0.099	0.143	0.145	0.129
C2	0.160	0.064	0.160	0.149	0.089	0.124
P2	0.069	0.114	0.049	0.123	0.060	0.083
C3	0.191	0.055	0.143	0.141	0.078	0.122
P3	0.064	0.093	0.092	0.119	0.103	0.094

emotional data (i.e., the SAL database), and reported a mean squared error MSE=0.18 using SVR, and MSE=0.08 using Long Short Term Memory Neural Networks as the best results. Although different data sets and experimental conditions have been used, our results can be seen as comparable to these.

Note that the coders have annotated the expectation dimension using only the positive hemisphere [0,+1]. In order to make the MSE values obtained during evaluation comparable to the other dimensions, during experimentation we normalized the expectation ground truth into the range of [-1,+1]. This somewhat explains the difference observed in Table 1 between the coders' MSE and the predictors' MSE for the expectation dimension. Additionally, the table also indicates that modeling the annotations provided by *coder 1* (C1) is somewhat more challenging compared to other coders. This might be due to the fact that non-verbal cues other than the head movements (possibly) influenced the annotations of C1.

5 Conclusions

This paper focused on dimensional and continuous emotion prediction from spontaneous head movements occurring in a conversational setting. The preliminary experimental results suggest that it is possible to automatically predict the arousal, expectation, intensity, power and valence dimensions from conversational head movements and gestures.

Dimensional emotion prediction from spontaneous head gestures has been integrated in the SEMAINE project [1] that aims to support sustained emotionally-colored computer-human interaction using non-verbal expressions. One of the key ways the SEMAINE system creates a sense of interaction is having SAL characters producing head movements, and responding to the user's head movements. Fig. 1(b) illustrates a human subject conversing with Poppy while his spontaneous conversational head movement is analyzed. This is utilized to simultaneously predict the arousal, expectation, intensity, power and valence level of the user. The prediction result is then used by the SEMAINE framework to choose appropriate back-channels and sustain an ongoing interaction between the virtual character and the user.

Acknowledgment. The work of H. Gunes is funded by the EC's 7th Framework Programme [FP7/2007-2013] under grant agreement no 211486 (SEMAINE). The work of M. Pantic is funded in part by the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB).

References

1. The SEMAINE project, `http://www.semaine-project.eu/`
2. Kendon, A.: Facial Expression of Emotion. In: Some functions of the face in a kissing round, pp. 117–152. Cambridge University Press, Cambridge (1990)
3. McClave, E.Z.: Linguistic functions of head movements in the context of speech. J. of Pragmatics 32, 855–878 (2000)
4. Zeng, Z., et al.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Tran. on PAMI 31, 39–58 (2009)
5. Russell, J.A.: A circumplex model of affect. J. of Personality and Social Psychology 39, 1161–1178 (1980)
6. Scherer, K.: Psychological models of emotion. In: The Neuropsychology of Emotion, pp. 137–162. Oxford University Press, Oxford (2000)
7. Fontaine, J.R., et al.: The world of emotion is not two-dimensional. Psychological Science 18, 1050–1057 (2007)
8. Kawato, S., Ohya, J.: Real-time detection of nodding and head-shaking by directly detecting and tracking the between-eyes. In: IEEE FGR, pp. 40–45 (2000)
9. Kapoor, A., Picard, R.W.: A real-time head nod and shake detector. In: Workshop on Perceptive User Interfaces (2001)
10. Tan, W., Rong, G.: A real-time head nod and shake detector using hmms. Expert Systems with Applications 25(3), 461–466 (2003)
11. Morency, L.–P., et al.: Contextual recognition of head gestures. In: ICMI, pp. 18–24 (2005)
12. Glowinski, D., et al.: Technique for automatic emotion recognition by body gesture analysis. In: CVPR Workshops, pp. 1–6 (2008)
13. Kulic, D., Croft, E.A.: Affective state estimation for human-robot interaction. IEEE Tran. on Robotics 23(5), 991–1000 (2007)
14. Chanel, G., et al.: Valence-arousal evaluation using physiological signals in an emotion recall paradigm. In: IEEE SMC, pp. 2662–2667 (2007)
15. Wollmer, M., et al.: Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In: Interspeech, pp. 597–600 (2008)
16. Gunes, H., Pantic, M.: Automatic, dimensional and continuous emotion recognition. Int. Journal of Synthetic Emotions 1(1), 68–99 (2010)
17. Douglas-Cowie, E., et al.: The Humaine database: addressing the needs of the affective computing community. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 488–500. Springer, Heidelberg (2007)
18. The SEMAINE database, `http://semaine-db.eu/`
19. Schroder, M., et al.: A demonstration of audiovisual sensitive artificial listeners. In: ACII, pp. 263–264 (2009)
20. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE CVPR, pp. 511–518 (2001)
21. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001), `http://www.csie.ntu.edu.tw/~cjlin/libsvm`