User modeling via gesture and head pose expressivity features

George Caridakis
Image, Video and Multimedia Systems Lab
National Technical University of Athens
gcari@image.ntua.gr

Stylianos Asteriadis
Image, Video and Multimedia Systems Lab
National Technical University of Athens
stiast@image.ntua.gr

Kostas Karpouzis
Image, Video and Multimedia Systems Lab
National Technical University of Athens
kkarpou@image.ntua.gr

Abstract

Current work focuses on user modeling in terms of affective analysis that could in turn be used in intelligent personalized interfaces and systems, dynamic profiling and context-aware multimedia applications. The analysis performed within this work comprises of statistical processing and classification of automatically extracted gestural and head pose expressivity features. Computational formulation of qualitative expressive cues of body and head motion is performed and the resulting features are processed statistically, their correlation is studied and finally an emotion recognition attempt is presented based on these features. Significant emotion specific patterns and expressivity features interrelations are derived while the emotion recognition results indicate that the gestural and head pose expressivity features could supplement and enhance a multimodal affective analysis system incorporating an additional modality to be fused with other commonly used modalities such as facial expressions, prosodic and lexical acoustic features and physiological measurements.

1 Introduction

Intelligent personalized systems often ignore the affective aspect of human behavior and focus more on tactile cues of the user activity. A complete user modeling though should also incorporate cues such as facial expressions, speech prosody and gesture or body posture expressivity features in order to dynamically profile the user fusing all available modalities since these qualitative affective cues contain significant information about the user's non verbal behavior and communication. Additionally, since user's affective status is part of the interaction's semantic context, context aware multimedia applications should cater for semantic context modeling and extraction of related parameters. Towards this direction this work focuses on automatic extraction of gestural and head pose expressivity features and related statistical processing and affective classification.

An abundance of research within the fields of psychology and cognitive science related with the non verbal behavior and communication stress out the importance of qualitative expressive characteristics and cues of body motion, posture, gestures and in general human action during an interaction session [10]. Nevertheless, it is hard to identify specific characteristics of body language that could help us assess a user's emotional state. First of all, there is no clear mapping from gestures to emotional states. Secondly, the use of gestures differs from person to person and from situation to situation. Although such research work study primarily and mainly context of human to human interaction such approach can be extended to human computer interaction. Some work has incorporated gesture expressivity in HCI context but the vast majority concentrates on the expressively enhanced synthesis of gestures by virtual agents and ECAs [9]. Currently, research on the automatic analysis of gesture expressivity is still immature and this fold of human action analysis is asymmetrically studied with reference to the synthesis counterpart.

The rest of the article is organized as follows: Section 2 describes the corpus design and recording process while sections 3.1 and 3.2 discuss the feature extraction process for gestures and head pose respectively. Section 4.1 presents the computational formulation of the expressivity features of the gestures (section 4.1.1) and the head pose (section 4.1.2), while sections 4.2 and 4.3 present a statistical study on the expressivity features and experimental results of emotion classification. Finally, section 5 concludes the presented work and presents future research directions.

2 Corpus construction

The recorded corpus [4] features modalities such as speech and facial expressions but the focus is on hand gesture expressivity. Thus, this is the primary modality and is recorded using three methods: bare hands, Nintendo Wii remote controls and datagloves. The recordings took place in three European countries, namely Greece, Germany and Italy. Present work deals solely with the Greek subset of the corpus and only with the facial and bare hands modalities.

Emotion induction and recording procedure The adopted emotion elicitation method was inspired by the Velten mood induction technique [11] where people had to read aloud a number of sentences that put them in particular emotional state. The users were encouraged to use their own words as long as they helped them feel a particular emotion. The sentences were shown in three coherent blocks with first positive, then neutral and finally negative sentences in order to put the users gradually into the desired mood. We selected a total of 120 sentences (40 for each target class). The order of the emotions (positive-neutral-negative) was chosen in such a way so as not to switch directly between the two emotional extremes. Furthermore, users usually feel less motivated towards the end of the experiment and it would be harder to put them into a positive emotional state. During the first 20 sentences subjects are wearing a data glove by HumanWare. The next 10 sentences the glove is exchanged by two Wii remote controls, which the users hold in their hands. Finally, the remaining 10 sentences were performed with free hands.

Hardware setup During the recordings the user stands in front of a neutral background. The stimuli, i. e. the Velten sentences, is projected on a screen in front of him. Below the projection, in a distance of two meters and approximately at the height of the user's face, two high-quality cameras (720x576 pixels, 25 fps, 24 bit colour depth) are placed in order to capture the user's complete body and the user's face. To avoid occlusions in the videos a stand is used to locate the microphone on top of the user's head. Present work focuses on the last interaction mode which is freehand.

Participants Regarding the participants (see Table 1), in Greece 11 subjects (6 male and 5 female) between 23 and 40 years old took part in the experiment, while in Germany 21 subjects (11 male, 10 female) were following our scenario. Their age varied between 20 and 28 years old, while in Italy 19 subjects (11 males and 8 females) took part in the experiment, between 24 and 48 years old. The feature extraction and affective analysis presented here refers to the Greek subjects.

Table 1. Experiment users' demographics

	Greece	Germany	Italy
Male	6	11	11
Female	5	10	8
Age variation	23–40	20–50	24–48

3 Feature extraction

3.1 Gestures

Regarding the hand and head detection and tracking problem which is a required step for extracting expressivity features from a gesture, several approaches have been reviewed. Amongst them only video based methods were considered since motion capture or other intrusive techniques would interfere with the person's emotional state which is a crucial issue in this kind of analysis. The major factors taken under consideration are computational cost and robustness, resulting in an accurate near real-time skin detection and tracking module, as can be illustrated in Figure 1.

The overall process is described in detail in [3] and briefly includes creation of moving skin masks and tracking the centroid of these skin masks among the subsequent frames of the video depicting a gesture. Real time color model of the human skin is constructed by sampling the upper area of the box containing the head which corresponds to the forehead of the user, as provided by the Viola-Jones head detection module [12]. Such an approach tackles illumination issues which often impede the process of modeling and detecting human skin. Additionally enhances the robustness since the head detection module rarely outputs false positives. Object correspondence between two frames is performed by a heuristic algorithm based on skin region size, distance with reference to the previous classified position of the region, flow alignment and spatial constraints. In the case of occlusions (hand object merging and splitting), we establish a new matching of the left-most candidate object to the user's right hand and the right-most object to the left hand.

The described algorithm is lightweight, allowing real time implementation (see Figure 2). The object correspondence heuristic makes it possible to individually track the hand segments correctly, at least during usual meaningful gesture sequences. In addition, the fusion of color and motion information eliminates any background noise or artifacts, thus reinforcing the robustness of the proposed approach.

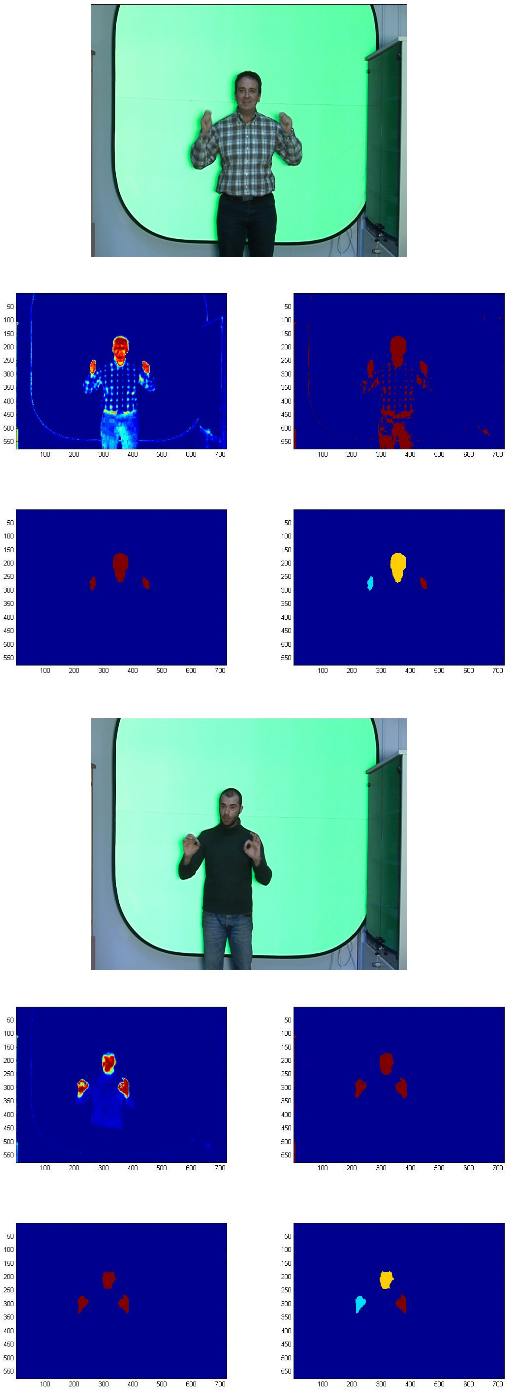
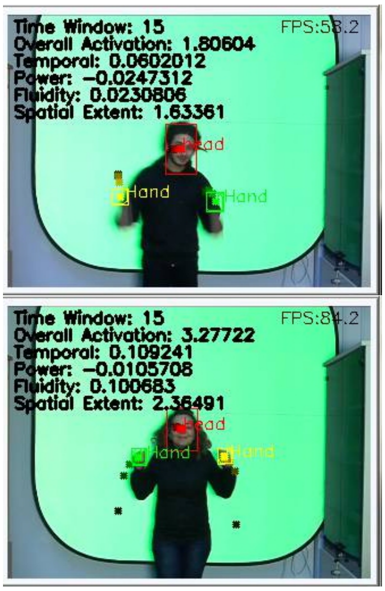


Figure 1. Image processing intermediate steps and results



```
Time Window: 15                     FPS:58.2
Overall Activation: 1.80604
Temporal: 0.0602012
Power: −0.0247312
Fluidity: 0.0230806
Spatial Extent: 1.63361
```

```
Time Window: 15                     FPS:84.2
Overall Activation: 3.27722
Temporal: 0.109241
Power: −0.0105708
Fluidity: 0.100683
Spatial Extent: 2.36491
```

Figure 2. Real time implementation of the described algorithm

3.2 Head pose

Head Pose is estimated based on the position of the eyes midpoint with regards to its position when the user is facing the camera frontally. For this reason, a series of rules have also been employed, for discriminating between frontal and rotated views of the head, and for restarting the system based on expected geometrical relations among facial points and natural human motion criteria. The eyes midpoint distortions are normalized with the inter-ocular distance, as calculated every time the user is facing the camera frontally. In this way, the system is scale independent and can give reliable results for various distances of the user with regards to the camera, whenever re-initialization occurs. For tracking facial features coordinates, a three-pyramid Lucas-Kanade tracker [2] is used, in order to handle large and sudden point movements. The above system can perform real time and the rules employed for re-initialization render it robust to sudden changes in lighting and spontaneous. Further details of the system are given in [1].

4 Affective analysis

4.1 Expressivity features extraction

4.1.1 Gestures

Features and cues of non verbal behavior are an integral part of the communication process since they provide information on the current emotional state and the personality of the interlocutor [8]. Common classification schemes include binary categories such as slow/fast, restricted/wide, weak/strong, etc. Our gesture expressivity modeling is close to these schemes in the sense they provide formulation and quantitative measurement of the respective aspects of the gesture. Adopting a subset of the gesture synthesis expressivity modeling parameters [6] we define five expressivity features: Overall activation, Spatial extent, Temporal, Fluidity and Power.

A computational formulation of these parameters is described in detail in [3] and in order to provide a more strict definition let us consider a gesture G as a sequence, of T frames, consisting of coordinates of the left and right hand, (x_{li}^G, y_{li}^G) and (x_{ri}^G, y_{ri}^G), respectively and $i \in [1, T]$. The coordinates of hands are relative to the position of the head which is defined as the center of the bounding box of the region of head as provided by the head detection module and normalized with reference to the diagonal of this box which is considered indicative of the size of the head. These transformations are required in order to ensure that the coordinates are invariant to the position and the distance of the user in front of the camera, parameters that are not known a priori. Thus a gesture is formally defined as:

$$\begin{aligned} G = [((x_{l1}^G, y_{l1}^G), (x_{r1}^G, y_{r1}^G)), ((x_{l2}^G, y_{l2}^G), \\ (x_{r2}^G, y_{r2}^G)), \cdots, ((x_{lT}^G, y_{lT}^G), (x_{rT}^G, y_{rT}^G))] \end{aligned} \tag{1}$$

For simplicity reasons (x_{li}^G, y_{li}^G) and (x_{ri}^G, y_{ri}^G) will be referred to L_i^G and R_i^G respectively from this point forward. Additionally, the quantity of motion D_i for one hand during the time period between frame i and frame $i+1$ is defined as the norm of the vector defined by points (x_i, y_i) and (x_{i+1}, y_{i+1}):

$$D_i = \left|\overrightarrow{(x_i, y_i)(x_{i+1}, y_{i+1})}\right| \tag{2}$$

Overall activation is considered as the quantity of movement during a dialogic discourse and is formally defined as the sum instantaneous quantities of motion:

$$OA_G = \sum_{i=1}^{T-1} D_{li}^G + D_{ri}^G \tag{3}$$

Spatial extent is expressed with the expansion or the condensation of the used space in front of the user (gesturing space). In order to provide a strict definition of this expressivity feature spatial extent is considered as the maximum value of the instantaneous spatial extent during a gesture. Let e_i be the norm of the vector defined by the points (x_{li}, y_{li}) and (x_{ri}, y_{ri}) during time i. Thus, the spatial extent expressivity parameter corresponds to the maximum value of this instantaneous spatial extent e_i during the stroke phase of the gesture:

$$SE_G = \max e_i, i \in [1, T], e_i = \left|\overrightarrow{(x_{ri}, y_{ri})(x_{li}, y_{li})}\right| \tag{4}$$

The temporal expressivity parameter denotes the speed of hand movement during a gesture and dissociates fast from slow gestures. Given that the quantity D_i denotes instantaneous hand speed during time i the temporal expressivity parameter is defined as the as the arithmetic mean of this quantity and since OA_G, as defined earlier corresponds to the discrete integral:

$$TE_G = \frac{OA_G}{T} \tag{5}$$

On the other hand, the energy expressivity parameter refers to the movement of the hands at during the stroke phase of the gesture. Gestures are constituted by three phases: preparation, stroke and withdrawal. The message is primarily conveyed during the stroke phase, while the phases of preparation and withdrawal occur while the hands move from and to their neutral position respectively. The formalization of the energy expressivity feature according to this definition however is far from trivial since the automatic detection of the gesture phases is a quite challenging task. Alternatively we opted to associate this parameter qualitatively with the first derivative of the norm of D which refers to the acceleration of hands during a gesture:

$$PO_G = |D|' \tag{6}$$

Fluidity differentiates smooth / elegant from the sudden / abrupt gestures. This concept attempts to denote the continuity between hand movements and is suitable for modeling modifications in the acceleration of the upper limbs. Under this prism, we formally define as the gesture's fluidity the variation of the energy expressivity parameter as described in the previous paragraph:

$$FL_G = var(PO_G) \tag{7}$$

The reader is prompted to note that the quantity FL_G corresponds to is reversely proportional to the notion of fluidity. Thus, a gesture with high value of the FL_G expressive parameter demonstrates low fluidity and consequently is categorized as a sudden/abrupt gesture. Inverting the definition of fluidity is not a trivial process since the upper and lower bound of the measure are not a priori known.

4.1.2 Head pose

The formulation of the expressivity parameters for the user's hand gestures has been adapted appropriately for head pose features. Let H be a sequence of head pose cues for the corresponding temporal segment, consisting of T frames, for the gesture G (see Equation 1):

$$H = [((y_1^H, p_1^H), (y_2^H, p_2^H), \dots (y_T^H, p_T^H),)] \qquad (8)$$

where y_i^H, p_i^H are the yaw and pitch angles respectively, as described in section 3.2. Equivalently the head pose expressivity features for sequence H are

$$OA_H = \sum_1^T dYaw + dPitch$$

$$TE_H = mean(dYaw) + mean(dPitch)$$

$$PO_H = \sum_1^T \frac{\vartheta dYaw}{\vartheta y} + \frac{\vartheta dPitch}{\vartheta p}$$

$$FL_H = \frac{var(dYaw) + var(dPitch)}{2}$$

$$SE_H = \sqrt{(max(y) - min(y))^2 + (max(p) - min(p))^2}$$

where $dYaw = \frac{\vartheta H}{\vartheta y}$ and $dPitch = \frac{\vartheta H}{\vartheta p}$.

4.2 Expressivity features study

Initially, the means of all expressivity features were calculated and grouped per emotion and the results are shown in Figure 3. The emotion specific, user independent expressivity features are plotted in the same order as performed by the users and more specifically positive (blue bar), neutral (green bar) and negative (red bar). It should be noted that the mean values for the expressivity features are not user dependent and include collected values for all participants of the recordings of the Greek subset of the overall corpus. The order of the expressivity features is OA_H, OA_G, TE_H, TE_G, PO_H, PO_G, FL_H, FL_G, SE_H, SE_G, where the respective measurements are defined in sections 3.1 and 3.2 for the gesture and head pose expressivity features respectively.

Commenting on the mean values of the expressivity features there are two points worth noticing concerning the relation of the same expressivity features for different emotions (e.g. OA_H) and the correlation of pairs of the same expressivity feature for gestures and head pose (e.g. OA_H and OA_G, TE_H and TE_G). One could readily distinguish that for example OA_H has high mean values for positive emotions, average mean values for neutral emotions and low mean values for negative emotions. Similar conclusions can be drawn for the rest of the expressivity features. One could also note that positive emotions display high expressivity parameter values for all the features where as neutral and negative emotions demonstrate similar values for almost all the affective categories. This lead us to perform a positive against all others classification, presented in section 4.3.

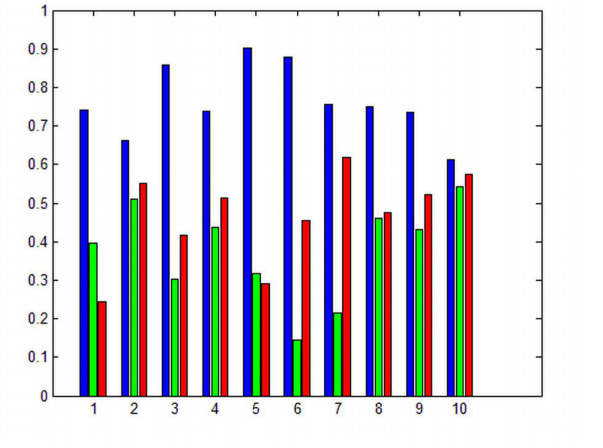


Figure 3. Means of expressivity features for the three affective categories

Additionally, similar patterns can be found for pairs of the same expressivity feature for gestures and head pose. An illustrative example would be that of the TE_H and TE_G pair (columns 3 and 4 of figure 3) where positive emotions display high, neutral display low and negative display intermediate temporal values. This correlation and the fact that overall activation (both gestural and derived from head pose cues) is highly correlated with the rest of the expressivity parameters is also depicted in figure 4.

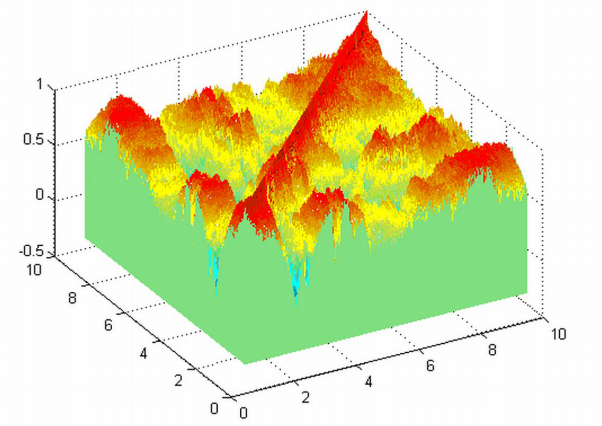


Figure 4. Correlation of expressivity features of head pose and gestures

4.3 Emotion classification

For discriminating between positive and non positive states, we employed a neural-fuzzy classifier. We used expressivity features coming from both modalities for training. Prior to training, our data were clustered using the subcluster algorithm described in [5]. This algorithm, instead of using a grid partition of the data, clusters them and, thus, leads to fuzzy systems deprived of the curse of dimensionality. The number of clusters created by the algorithm determines the optimum number of the fuzzy rules. After defining the fuzzy inference system architecture, its parameters (membership function centers and widths), were acquired by applying a least squares and back-propagation gradient descent method [7]. In our case, the fuzzy inference system gave a set of three rules, while for data clustering we used a radius equal to half of the maximum absolute value of each expressivity feature and output. Setting any output value equal to 1 when larger than 0.5 (positive) and 0 when smaller than 0.5 (negative/neutral) allowed classification between the two classes. For training, we followed a leave-one-out protocol, creating a fuzzy system for each user using as training data expressivity features from the rest of the users.

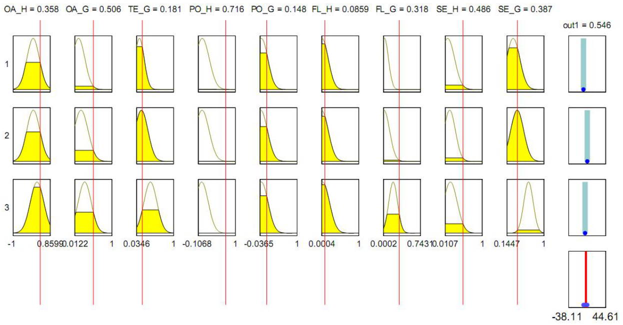


Figure 5. Fuzzy Inference System firing strengths

Using the Fuzzy Inference System as a classifier for positive and non-positive instances as can be shown in Figure 5 which is an example where the output is 0.546 (larger than 0.5, classified as positive) with the values of the expressivity features seen in the premise part of the rules. The total precision obtained was 60% and the total recall 67.33%. We also calculated the average f-measure for all users, which resulted 0.63. It should be noted that almost all expressivity features contributed positively to the final results.

5 Conclusions and future work

The work presented here includes the computational formulation of qualitative expressive cues of body motion and head pose, statistical and correlation study of the extracted expressivity features and a preliminary emotion classification attempt. The focus of this work is the definition of gestural and head pose expressive cues, how they are associated and what could be their role in a multimodal affective analysis system.

Future work includes application of the algorithms on the complete corpus described in section 2. Such an expansion would allow us to perform multimodal experimentation on fusion of modalities such as facial expressions, prosodical and lexical acoustic features as well as expand the range of investigation on gesture expressivity on other recording techniques such as Nintendo Wii remote controls and datagloves. A complete corpus affective analysis, which is considered ongoing and future work, related cross cultural and interaction obtrusiveness issues will be studied and significant conclusions are expected to be drawn both on these issues as well as the importance of gesture and head pose expressivity cues and their correlation amongst themselves and with other modalities.

References

[1] S. Asteriadis, P. Tzouveli, K. Karpouzis, and S. Kollias. Estimation of behavioral user state based on eye gaze and head poseapplication in an e-learning environment. *Multimedia Tools and Applications*, 41(3):469–493, 2009.

[2] J. Bouguet et al. Pyramidal implementation of the lucas kanade feature tracker description of the algorithm. *Intel Corporation, Microprocessor Research Labs*, 2000.

[3] G. Caridakis, A. Raouzaiou, E. Bevacqua, M. Mancini, K. Karpouzis, L. Malatesta, and C. Pelachaud. Virtual agent multimodal mimicry of humans. *Language Resources and Evaluation 41 (3-4), Special issue on Multimodal Corpora, pp. 367-388, Springer*, 2007.

[4] G. Caridakis, J. Wagner, A. Raouzaiou, Z. Curto, E. Andre, and K. Karpouzis. A multimodal corpus for gesture expressivity analysis. In *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, LREC, Malta, May 17-23, 2010*, 2010.

[5] S. Chiu. Fuzzy model identification based on cluster estimation. *Journal of intelligent and Fuzzy systems*, 2(3):267–278, 1994.

[6] B. Hartmann, M. Mancini, S. Buisine, and C. Pelachaud. Design and evaluation of expressive gesture synthesis for embodied conversational agents. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, page 1096. ACM, 2005.

[7] J. Jang. ANFIS: Adaptive-network-based fuzzy inference system. *IEEE transactions on systems, man, and cybernetics*, 1993.

[8] A. Mehrabian. *Nonverbal communication*. Aldine, 2007.

[9] C. Pelachaud. Studies on gesture expressivity for a virtual agent. *Speech Communication*, 51(7):630–639, 2009.

[10] S. Trenholm and A. Jensen. *Interpersonal communication*. Oxford University Press, USA, 2007.

[11] E. Velten. A laboratory task for induction of mood states. *Behaviour Research and Therapy*, 35:72–82, 1998.

[12] P. Viola and M. Jones. Rapid Object Detection using a Boosted Cascade of Simple. In *Proc. IEEE CVPR 2001*, 2001.