A computational model for mood recognition

Christina Katsimerou[1], Judith A. Redi[1], and Ingrid Heynderickx[2]

[1] Interactive Intelligence Group, Technical University Delft, The Netherlands
{C.Katsimerou,J.A.Redi}@tudelft.nl
[2] Human-Technology Interaction Group, Technical University Eindhoven, The Netherlands
{I.E.J.Heynderickx@tue.nl}

Abstract In an ambience designed to adapt to the user's affective state, pervasive technology should be able to decipher unobtrusively his underlying mood. Great effort has been devoted to automatic punctual emotion recognition from visual input. Conversely, little has been done to recognize longer-lasting affective states, such as mood. Taking for granted the effectiveness of emotion recognition algorithms, we go one step further and propose a model for estimating the mood of an affective episode from a known sequence of punctual emotions. To validate our model experimentally, we rely on the human annotations of the well-established HUMAINE database. Our analysis indicates that we can approximate fairly accurately the human process of summarizing the emotional content of a video in a mood estimation. A moving average function with exponential discount of the past emotions achieves mood prediction accuracy above 60%.

Keywords. Emotion recognition, mood estimation, affective computing, pervasive technology

1 Introduction

An indispensable feature for emotionally intelligent systems and affect-adaptive ambiences is recognizing (unobtrusively) the user's affect [1]. Technology endowed with this skill can, among others, drive or maintain the user to a positive affective state, for instance adapting the lighting system in a care centre room to comfort the inhabitants [2], or engage the user in natural interaction within virtual scenarios [3] [4].

Automatic affect recognition is often based on visual input (images and videos), due to the unobtrusiveness of visual sensors and the fact that people convey important affective information via their facial and bodily expressions [5]. A large body of work has been devoted to mapping visual representations of facial expressions [6], and body postures [7] into short-term affective states, commonly referred to as *emotions*. However, in certain applications based on affective monitoring, adapting the system behaviour to the dynamics of instantaneous emotions may be redundant, if not counterproductive. Take the case of a lighting system that adopts the optimal configuration to improve the occupant's affective state: it is neither necessary nor desirable that the light changes at the speed of instantaneous emotional fluctuations. A sys-

tem that retains and summarizes the affective information over a certain time window, and adapts smoothly over time would be more appropriate.

It is useful, at this point, to make a distinction between two types of affective states: *emotion* and *mood*. In psychological literature these terms are typically distinguished based on the duration [8] and the intensity of their expression [9]. Unfortunately, these differences are hardly delineated in a quantifiable way, as little is known e.g. about the time spanned by either emotional or mood episodes. To cope with this vagueness and make as few assumptions as possible, in the rest of the paper we will use the term *emotion* to refer to a punctual (i.e. instantaneous) affective state and the term *mood* as an affective state attributed to the whole affective episode, regardless of the duration of this episode.

From an engineering perspective, mood is typically assumed to be synonym to emotion and the two terms are often used interchangeably. Very little research, indeed, has tried to perform explicitly automatic mood recognition from visual input, except for some remarkable yet preliminary attempts [10] [11], discussed in more detail in section 2. Nevertheless, psychological literature recognizes a relationship between underlying mood and expressed emotions [12] [13]. Thus, it may be possible to infer mood from a series of recognized emotion expressions. This would entail the existence of a model that maps, for a given lapse of time, punctual emotion expressions into an overall mood prediction (Fig. 1).

In this paper, we describe an experimental setup for gaining basic insights in how humans relate mood to recognized punctual emotions, when they annotate emotionally coloured video material. The research questions that we aim to answer are a) to what extent we can estimate (recognized) mood from punctual emotions and b) how a person accounts for the (recognized) displayed emotions when judging the overall mood of another person. Answering these questions will bring us closer to retrieving a model of human affective intelligence that can then be replicated for machine-based mood recognition.

As such, the main contributions of this work are: a) we formulate a model where the mood we want to unveil is a function with punctual emotions as arguments, b) we indicate an experimental setup for validating this model, c) we specify the best fitting mood function out of a number of possible ones, and d) we optimize its parameters in terms of accurate prediction and computational complexity.

2 Related work

In psychology, emotion and mood are considered to be highly associated affective states, yet differing in terms of duration [8], intensity and stability [9], dynamics [14] and attribution (awareness of cause) [15]. Even though most literature agrees that there is a distinction between emotion and mood, in practice the terms are often used interchangeably and a one-to-one mapping between the two is typically assumed. Ekman [16], for example, claims that we infer mood from the signals of the emotions we associate with the mood, at least in part: we might deduce that someone is in a cheerful mood because we observe joy; likewise, stress as an emotion may imply an anxious mood. In the literature we find only one empirical study trying to identify the distinction between emotion and mood [12]. The authors conducted a so-called folk

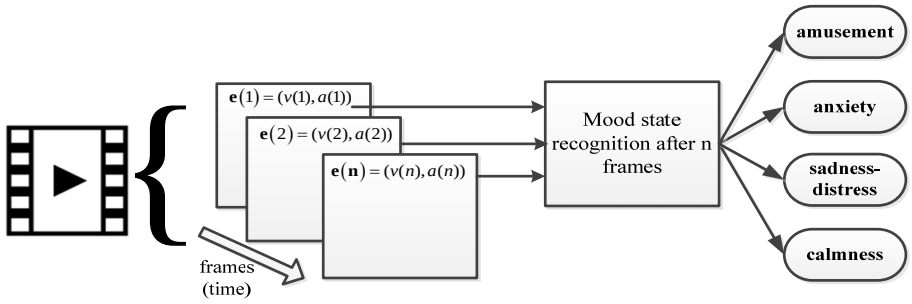

Fig. 1. Framework of automatic mood recognition module from a sequence of emotions.

psychological study, in which they asked ordinary people to describe how they experience the difference between the two terms. A qualitative analysis on the responses indicated *cause* and *duration* as the two most significant distinctive features between the two concepts; nevertheless, their difference was not quantified.

Automating the process of mood recognition entails linking data collected from sensors monitoring the user (e.g. cameras, microphones, physiological sensors) to a quantifiable representation of the (felt) mood. In the case of visual input very scarce results are retrievable in literature. In fact, the latest studies in the field have been geared towards recognizing continuously the emotions along videos rich in emotions and emotional fluctuations, e.g. as requested by the AVEC challenge [17]. However, typically a decision on the affective state is made on frame-level, i.e., for punctual emotions, whereas no summarization into a mood prediction is attempted.

In [10] we find explicit reference to mood recognition from upper body posture. The authors induced mood with music in subjects in a lab, and recorded eight-minute videos focusing on their upper body after the induction. They analyzed the contribution of postural features in the mood expression and found that only the head position predicted (induced) mood with an accuracy of 81%. However, they considered only happy versus sad mood and the whole experiment was very controlled, in the sense that it took place in a lab and the subjects knew what they were intended to feel, making the genuine expression doubtful. Another reference to mood comes from [11], where the authors inferred again the bipolar happiness-sadness mood axis from data of 3D pose tracking and motion capturing. Finally, the authors of [18] were the first to briefly tap in the concept of summarizing punctual annotations of affective signals to an overall judgment. However, they only considered the mean or the percentiles of the values of valence and arousal as global predictors, without taking into account their temporal position.

In this study, we will extend significantly the latter work, by constructing systematically a complex mood model from simple functions, analyzing its temporal properties and proposing it as an intermediate module in automatic mood recognition from video, after the punctual (on frame level) emotion recognition module (Fig. 1).

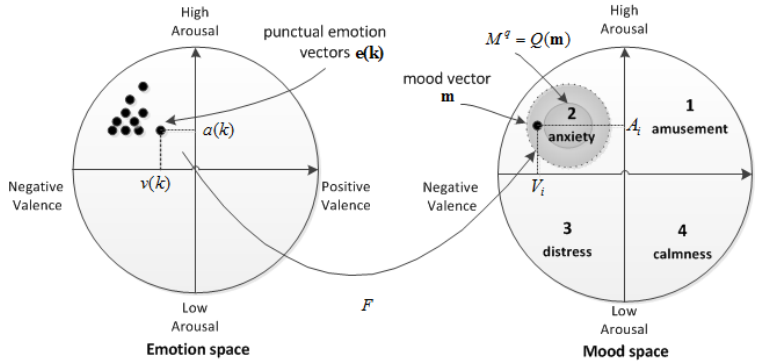

Fig. 2. Model of emotion and mood space. Each emotion is a point in the emotion space and the trajectory of points in the emotion space is mapped as a point in the mood space.

3 Problem setup and methodology

3.1 Domains of emotion and mood

To define a model that maps punctual emotion estimations into a single mood, it is necessary to first define the domains in which emotion and mood will be expressed. In affective computing there are two main trends for affect representation: the discrete [19] and the dimensional one [20] [21]. The latter most commonly identifies two dimensions, i.e., valence and arousal, accounting for most of the variance of affect. It allows continuous representation of emotion values, capturing in this way a wider set of emotions. This is why we resort to it in our work.

In this study we assume the valence and arousal dimensions to span a Euclidean space (hereafter referred to as the VA space), where emotions are represented as points-vectors. Analogously, mood can be represented in a Euclidean (mood) space as a tuple of valence and arousal values. We quantize the mood space in four partitions, corresponding to the four quadrants defined by the valence and arousal axes, namely (1) positive V- high A, (2) negative V- high A, (3) negative V- low A, and (4) positive V- low A[1] (shown in Fig.2). This 4-class mood discretization gives a trade-off between ambiguity and simplicity, being able to capture different possible moods expressed sufficiently, yet eliminating redundancies and diminishing the problem dimensionality.

3.2 Problem formulation

Suppose we have a video clip *i* representing an affective episode of a user with a total number of frames n_i. Punctual emotions can be estimated from each video frame (static images) and the overall (quantized) mood M_i characterizes the whole clip *i*. In this study both emotions and mood refer to the affective state of the active person of the clip, as perceived by human annotators.

[1] The class numbering 1-4 serves for notation and not ranking.

For every independent clip i the punctual emotion vector $\boldsymbol{e}_i$ corresponding to the recognized emotion at frame k is expressed in the VA space as

$$\mathbf{e_i}(k) = \left(v_i(k), a_i(k)\right),\ k=1,2,..,n_i\ , \tag{1}$$

where $v_i(k)$ and $a_i(k)$ are recognized valence and arousal values of the emotion expressed at frame $k \le n_i$ of the clip i. Assuming that the sequence of punctual emotion vectors for clip i

$$\mathrm{E_i} = \left(\mathbf{e_i}(1), \mathbf{e_i}(2), .., \mathbf{e_i}(n_i)\right), \tag{2}$$

is known, and we intend to estimate the overall mood, we want to express the mood vector $\mathbf{m}_i = (V_i, A_i)$ as

$$\mathbf{m_i} = F(\mathrm{E_i})\ , \tag{3}$$

where F is the function mapping the emotion sequence to the mood vector. We finally retrieve from the mood vector the quantized mood M_i through the function Q, defined as:

$$M_i = Q(\mathbf{m_i}) = Q(V_i, A_i) = \begin{cases} 2 - sgn(V_i) & if \quad sgn(V_i \cdot A_i) > 0 \\ 3 + sgn(V_i) & if \quad sgn(V_i \cdot A_i) < 0 \cdot \\ 0 & otherwise \end{cases} \tag{4}$$

In this study, we set M_i as the ultimate target of our discrete prediction model and F the function to be modeled.

3.3 Modeling mood from a sequence of punctual emotions

3.3.1 Basic mapping functions

We propose several possible formulations of the function F in eq. (3), which map punctual measurements of emotion into a representative value for the overall affective episode. This value is then used in eq. (4) to predict the mood class, unless stated otherwise.

Predictor 1: The mean emotion (mean). Probably the easiest assumption is that mood is formed by the equal contribution of all the emotions within a given timespan [18]. The average of the emotion points will then be the "station" mood point, which acts a gravitational force on them [22]. More formally, the mean of an emotion sequence over a particular time window is predictor of the overall mood for this time window:

$$M_i = Q\left(F(\mathrm{E_i})\right) = Q\left(\sum_{k=1}^{n_i} \mathbf{e_i}(\mathrm{k}) / n_i\right) = Q\left(\left(\sum_{k=1}^{n_i} v_i(k)/n_i, \sum_{k=1}^{n_i} a_i(k)/n_i\right)\right). \tag{5}$$

Predictor 2: The maximum emotion (max). Intuitively, the emotion with the highest intensity is expected to have a high impact on the overall mood within a given timespan. Thus, we may hypothesize it to be a predictor for the overall mood for the given time window. As a measure for the intensity of the emotion we use the Euclidean norm of the emotion vector, defined as

$$\left\|\mathbf{e_i}(\mathrm{k})\right\| = \sqrt{\left(v_i(k)\right)^2 + \left(a_i(k)\right)^2}, k = 1, 2, .., n_i\ . \tag{6}$$

Then the mood occurs from the emotion vector that maximizes the intensity over the sequence $\mathrm{E_i}$, or

$$M_i = Q\left(F\left(\mathrm{E_i}\right)\right) = Q\left(\underset{\mathbf{e_i}(\mathrm{k})\in \mathrm{E_i}}{\arg\max}\left(\left\|\mathbf{e_i}(\mathrm{k})\right\|, k = 1,2,..,n_i\right)\right) = Q\left(\mathbf{e}_{\max}\right). \quad (7)$$

Predictor 3: The longest emotion (long). Another hypothesis is that the emotions that occur more within a given timespan will sustain the associated mood [23]. Thus, we may map individual emotion vectors into mood vectors directly and take the quadrant of the mood space containing the majority of them (see Fig.2); this quadrant may then be a predictor of the recognized mood. More formally, if we consider 4 disjoint subsets of *Ei* defined as

$$\mathrm{E_i^q} = \left\{\mathbf{e_i(k)} \,|\, Q\left(\mathbf{e(k)}\right) = q, k = 1,2,..,n_i\right\}, q = 1,2,3,4, \quad (8)$$

each with cardinality $C(q) = \left|\mathrm{E_i^q}\right|$, then the mood corresponding to the longest emotion is

$$M_i = F\left(\mathrm{E_i}\right) = \underset{q\in\{1,2,3,4\}}{\arg\max}\left(C(q)\right). \quad (9)$$

Predictor 4: First emotion (FE). A reasonable property of the mood function is memory [24], in the sense that mood recognition involves the assessment of not only the current emotion, but also of the previously recognized ones. In the extreme case, the time span of the memory window may extend back to the beginning of the emotional episode, resonating the impact of the first points to the current moment. Therefore, we may hypothesize that the first of a sequence of emotions over a certain time window is a predictor for the overall mood for this time window, or

$$M_i = Q\left(\mathbf{e}(1)\right). \quad (10)$$

Predictor 5: Last emotion (LE). Contrary to the previous hypothesis, we may assume that the significance of the previously recognized emotions in the mood estimation decreases as time lapses and only the latest recognized emotion defines the overall mood, that is:

$$M_i = Q\left(\mathbf{e}(n_i)\right). \quad (11)$$

3.3.2 A more complex model of mood recognition from emotions

The simple models proposed in section 3.3.1 may further be combined into a more complex one, occurring from a moving average of emotions over time, with memory retention expanding back to the preceding recognized emotions for a given portion of the timespan of the emotional episode. We can formulate this as

$$M_i = Q\left(F\left(\mathrm{E_i}\right)\right) = Q\left(\sum_{k=n_i-w}^{n_i}\left(\mathbf{e_i}(\mathrm{k}) / w\right)\right), \quad (12)$$

where w is the size of the memory window. In fact, (12) is a moving average (*MA*) over w frames. In this formulation we consider a hard limit function to retain only the last w recognized emotions, disregarding the rest. In reality, a desirable property of mood assessment is smoothness over time [25], that is, it should gradually neglect the past, as it moves along the emotion sequence. This can be modeled through a discount function D_w of the previous frames, either linear *LD* (Eq. (13)), as seen in [23], or exponential *ED* (Eq. (14)).

$$D_w(k) = \frac{k-(n_i-w)}{n_i-w}, k = n_i - w,..,n_i, \quad (13)$$

$$D_w(k) = e^{\frac{k-(n_i-w)}{n_i-w}}, k = n_i - w,..,n_i \,. \tag{14}$$

Then the mood will be the weighted average of the last *w* emotions:

$$M_i = Q\left(F\left(\mathrm{E_i}\right)\right) = Q\left(\sum_{k=n_i-w}^{n_i} \mathbf{e_i}(\mathrm{k}) \cdot \left(D_w(k) / \sum_{k=n_i-w}^{n_i} D_w(k)\right)\right). \tag{15}$$

We expect these refined models to be able to properly capture the processes that regulate the relationship between recognized emotions and mood.

4 Experimental validation

4.1 Data and Overview

To check whether any of the models proposed in section 3 properly captures the relationship between punctual emotions and mood, we searched the literature for an affective database which includes videos portraying affective episodes, and for which both punctual emotion annotations (over time) and global mood annotations (for the whole clip) were reported. The HUMAINE [26] audio-visual affective dataset proved appropriate for the purpose, including natural, induced and acted mood videos, portraying a wide range of expressions. It consisted of 46 relatively short video clips (5seconds-3 minutes), each annotated continuously by 6 experts in the VA domain, using ANVIL [27] as annotation tool. The continuous annotations were encoded into the emotion sequence of Eq. (2). Per video, also global mood annotations were given on a 7-point scale, for valence and arousal separately. From the latter, we determined the quantized mood we targeted by applying Eq.(4) to each VA mood annotation.

To overcome subjectivity issues in obtaining one ground truth per clip [18], we decided to focus on the mapping from emotions to mood *per coder* separately, which meant that the same video annotated by *x* coders would produce *x* different instances in our experimental set. This choice allowed us to study the interpersonal differences in the process of mood estimation. For simplicity, we excluded from our analysis the clips with multiple annotated moods, i.e. shifted (consecutive) moods or co-existing (simultaneous) moods. We consider that these cases require separate attention, and we demand their analysis to further work. As a result, in this experiment we analyzed 168 *single-mood* video-instances in total: 38 of class 1, 34 of class 2, 51 of class 3 and 46 of class 4.

In the following sections, we first test the simple models proposed in section 3.3.1. Then, we explore the temporal properties of the mood function. Based on the outcomes of this analysis, we set the parameter *w* of the more complex mood models proposed in section 3.3.2. All models are evaluated based on their accuracy, which is calculated as the ratio of the correct mood predictions over the number of instances.

4.2 Experimental results

The prediction accuracy obtained by the basic mood functions of section 3.3.1 is presented in Fig.3, per coder and the “average” coder. The random benchmark marks the

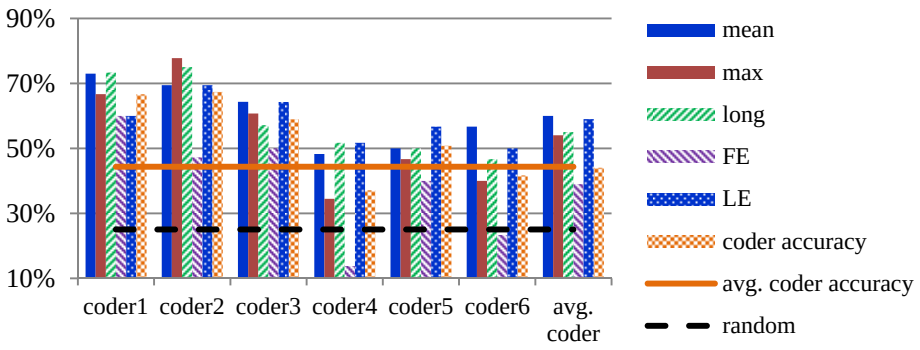

Fig. 3. Accuracy of mood prediction from emotions for the 5 simple models per coder and the average coder.

lowest bound of randomly assigning moods to one of the 4 classes[2] (i.e., 25%). A second benchmark is the coder accuracy, namely how well human coders agree on the mood of a video (note that these annotations of mood use the full video and not a sequence of emotions - as our model does). We estimated the coder accuracy as the average rate of pairwise agreement in recognized mood per video between one coder and the rest. The average agreement across all possible pairs of coders results in the average coder accuracy (44%), marked with the solid line in Fig. 3.

For coders 1,3 and 6 the model predicting their mood annotations best is mean. For coders 4 and 5 the most accurate model is LE, being also a good predictor according to coder 3. For coder 2 max outperforms the other models. Overall mean predicts more accurately the mood (60%), similar accuracy with LE (59%), indicating an importance of previously recognized emotions as well as current emotions in mood recognition. The maximum emotion is in general a worse predictor than the longest emotion, which implies that duration is more important than intensity in mood prediction. FE is the significantly worst predictor (all pairwise t-tests between FE and each of the other predictors across all coders, df=5, resulted in $p<0.05$). Except for the FE, the predictors are significantly above the random choice ($p<0.01$). We also performed paired t-tests between the coder accuracy and the accuracy of each predictor. Only mean and LE could predict the mood annotations of a coder more accurately than the rest of the coders at the 5% significance level ($p<0.035$).

It is interesting at this point to check how properties of the video, i.e., its length, the density of the emotion sampling and the temporal position of the recognized emotions, influence the mood prediction. For this analysis, we investigate these effects on the *mean* mood model given by eq. (5) only, as this was the best predictor obtained from the previous test**.**

Length of the affective episode. We ran a Mann-Whitney U-test between the lengths of the videos for which the mood was correctly predicted and those of the misclassified videos. The test returned no significant difference between the median of the length of the correctly and misclassified videos ($p=0.31$, $h=0$, $z=1.01$), indicating that

[2] Our classifier handles each video independently, without prior knowledge of the number of videos in each class.

(at least for videos up to 3 minutes) the duration of the emotional episode does not influence the correctness of the mood prediction.

Sampling rate. A typical emotion sequence produced by continuous annotations is very rich in points, considering that they are sampled according to the video frame rate (e.g. 25 frames/second). Hence, a sparser sampling may be sufficient, as well as desirable in affect-adaptive applications, reducing computational requirements and allowing the system to remain idle in cases of uncertainty. To determine the lowest possible sampling rate, we sub-sampled the original E_i with sampling rate $f_s = \lfloor (n_i - 1)/(N-1) \rfloor$, with N the number of samples per sequence. Then we applied F on the sub-sampled sequence E_i^{sub} with f_s as parameter. The hypothesis is that

$$F(E_i^{sub}) = F(\langle \mathbf{e_i}(1), \mathbf{e_i}(f_s+1), \mathbf{e_i}(2\cdot f_s+1)..,\mathbf{e_i}((N-1)\cdot f_s+1)\rangle) = F(E_i), \quad (16)$$

where F is estimated from eq. (5). For N=2 only the first and last point of the sequence are taken into account. In our experiments, we investigate the effect of sub-sampling the emotional sequence from 2 up to 11 points. Fig.4a illustrates the prediction accuracy of mood in relation to N. For N>3, namely sampling every temporal third of the sequence, the accuracy stabilizes above 99% of its value when calculated on the whole sequence, so eq. (16) is essentially satisfied. In real applications, where the length of the affective episode is not known a priori, it would be useful to determine a bound for the sampling rate in seconds rather than number of samples. Thus, we sampled the emotion sequences iteratively, increasing the sampling rate up to 1 minute with steps of 0.5 seconds. Based on the previous results, we excluded at each

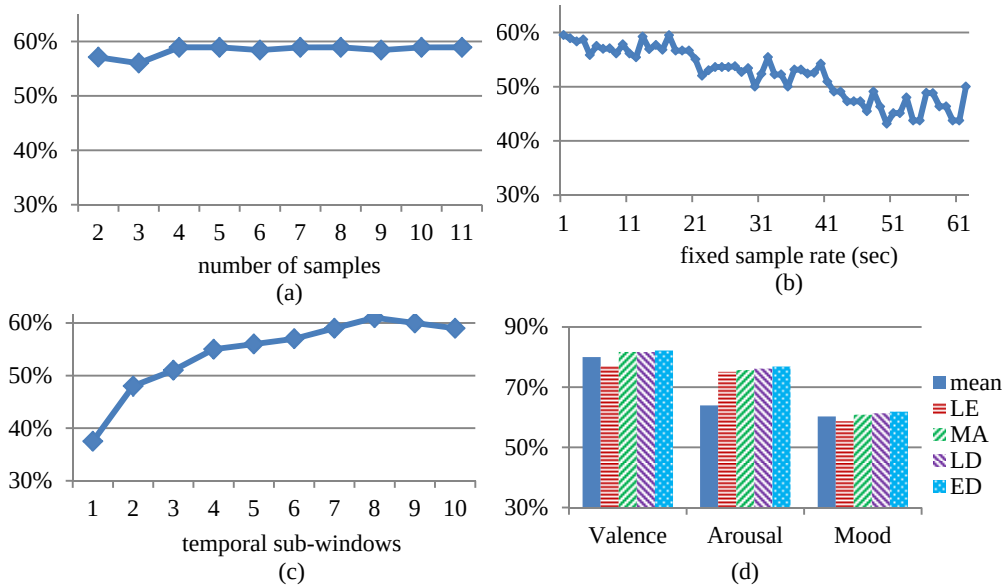

Fig. 4. Mood prediction accuracy of the mean predictor, as a function of a) number of samples in the sequence, b) sampling rates expressed in seconds, c) temporal position of sub-windows in the sequence. d) Overall valence, arousal and mood prediction accuracy of the average coder with the two best basic (mean, LE) and the three complex models (MA,LD,ED).

Table 1**.** Confusion matrix of the predictor ED for mood, valence and arousal.

Mood				
Actual \ Pred.	1	2	3	4
1	26	2	3	7
2	1	23	10	0
3	2	6	36	5
4	11	1	17	18

Valence		
	Pos.	Neg.
Pos.	62	23
Neg.	8	75

Arousal		
	High	Low
High	40	20
Low	20	76

iteration those clips with length lower than double the sampling rate, for which we would have sampled only the first and last points of the sequence. The accuracy of the mood prediction from the subsampled sequences obtained is illustrated in Fig. 4b. We can deduce that a safe sampling that still captures the global picture without discarding essential emotional information is every 3 seconds (98% of the accuracy achieved without subsampling). However, sampling up to every 20 seconds would still grant 90% of the accuracy of the original sequence.

Temporal portions contribution. We investigated to what extent the temporal position of an emotion point in the sequence contributes to the final mood decision. For this, we segmented each sequence in 10 equal non-overlapping temporal subwindows, and predicted the mood as if we only had one of these portions of the sequence at our disposal. The average accuracy is depicted in Fig.4c. Mood is predicted from the last 30% of the sequence as accurately as if we used the whole sequence.

Based on the results above, we defined the length of the window w for the complex models *MA* (eq. 12), *LD* (eq. 13) and *ED* (eq. 14) as *30%* of the clip length n_i. The average accuracy of these models is summarized in Fig.4d, along with the accuracy for the mean and LE models as reference. The complex models perform in a similar way, although the ED scores the best, significantly better than LE ($p=0.048$), but not than the rest. Thus, a model where mood is influenced constantly by the past emotions, but the current state plays a more definitive role, may be a good template for further extensions of this work. Also, the performance of the discount factor seems to depend on the speed of the attenuation. The exponential discount is steeper than the linear, which seems to be beneficial to the model accuracy.

To gain more insights on the performance of the models, we also applied F on each of the sequences of punctual annotations of valence and arousal independently and computed how well they predict the global annotations of valence and arousal, respectively. As shown in Fig. 4d, overall valence is better predicted than overall arousal. Mean predicts valence with an accuracy higher than 80%, whereas the overall arousal is predicted better by the LE. ED predicts more accurately both the overall valence and arousal.

To go even more in detail, we report the confusion matrices of mood, valence and arousal prediction for the best performing ED model. The misclassifications occur in their vast majority between classes that share the valence or arousal axis, namely 1 vs.2, 2 vs.3, 3 vs.4 and 4 vs.1. Only in 6 cases misclassifications occur between opposite moods, i.e. between classes 1 and 3 (5 clips), and classes 2 and 4 (1 clip). This is rather intuitive, since the four mood classes occurred from a partition of the

mood continuum, in which neighboring classes lie closer perceptually and, therefore, should be closer also computationally. It can also be noticed that low arousal moods (3 or 4) are more easily confused with the neighboring low arousal mood (4 or 3), rather than with the neighboring mood of the same valence (2 or 1). In fact, valence is predicted more accurately (82%) than arousal (77%), and in most cases the misclassification is due to an underestimation of valence. This is not true for arousal, for which the false negatives and false positives are in equal number. Finally, mood classes are not predicted equally well. Class 3 is the best predicted from punctual emotions, whereas class 4 is the least, which may imply that it is more complicated to judge moods of positive V – low A from an emotion sequence.

5 Conclusion and outlook

We proposed a set of computational models that infer the long-term affective state of a person from a series of recognized emotion expressions. We showed that a model that has memory of the last recognized emotions, yet exponentially discounts their importance in the overall mood prediction, is able to recognize mood with 62% accuracy, which is well above random, as well as human agreement. In our experiments, we also found that a sequence of emotions recognized as sparsely as every 3 seconds predicts mood as well as a series of emotion recognized at video frame rate, which is relevant information for decreasing the computational complexity involved in designing empathic systems working in real time. Finally, we showed that valence and arousal values are predicted quite satisfactorily from our models, yet their combination into mood is less precise.

The latter could be due to the fact that we approached the mood prediction problem as classification, whereas in reality the mood space is continuous; our mood space partitioning was done arbitrarily, yet intuitively, and therefore may be suboptimal. Additionally, the short duration of the videos is not necessarily representative of the actual mood of people, which could last for an indeterminate lapse of time beyond the emotional episode represented in the video. Thus, the deployment and setup of the discount function may need to be optimized for longer videos.

A remark is that an emotional signal can reveal information about the mood, but presumably it is not the only factor people take into account. To fit the intricate mood estimation process we may need more complex models of emotion dynamics, also taking into account other factors (e.g., situational context, scene semantics, personal prejudices). Finally, the robustness of these models when applied on a series of machine-recognized rather than human-recognized emotions, still has to be proven.

Bibliography

1. R. W. Picard, "Affective computing," in *MIT press*, 2000.
2. A. Kuijsters et. al., "Improving the mood of elderly with coloured lighting," in *Ami*, 2011.
3. K. Porayska-Pomsta et. al., "Modelling Users' Affect in Job Interviews: Technological Demo," in *UMAP*, Rome, 2013.
4. C. Conati and H. Maclaren, "Empirically building and evaluating a probabilistic model of user affect," in *User Modeling and User-Adapted Interaction*, pp. 267-303, 2009.
5. C. Darwin, "The expression of the emotions in man and animals," in *Oxford University*

Press, 1998.

6. F. De la Torre and J. F. Cohn, "Facial expression analysis," in *Visual Analysis of Humans*, pp. 377-409, 2011.
7. A. Kleinsmith and N. Bianchi-Berthouze, "Affective Body Expression Perception and Recognition : A Survey," in *Transactions on Affective Computing*, pp. 15–33, 2013.
8. J.Jenkins et al.,"Human emotions: A reader," Malden, MA:Blackwell, 1998.
9. A. M. Lane and P. C. Terry, "The nature of mood: Development of a conceptual model with a focus on depression," in *Journal of Applied Sport Psychology 12.1,* pp. 16-33, 2000.
10. M. Thrasher, et.al., "Mood recognition based on upper body posture and movement features," *ACII,* 2011.
11. M. Livne, et. al., "Human attributes from 3d pose tracking," in *ECCV*, 2010.
12. C. Beedie, P. Terry and A. Lane, "Distinctions between emotion and mood," in *Cognition & Emotion*, vol. 19, no. 6, pp. 847-878, 2005.
13. A. Mehrabian, "Pleasure-Arousal-Dominance: A General Framework for Describing and Measuring Individual Differences in Temperament," in *Current Psychology*, pp. 261-292, 1996.
14. B. Parkinson, et.al., "Changing moods: The psychology of mood and mood regulation", Harlow, UK.: Addison Wesley Longman, 1996.
15. J. A. Russell, "Core affect and the psychological construction of emotion," in *Psychological review 110.1:* 145, 2003.
16. P. Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*, pp. 45-60, 1999.
17. B. Schuller, M. Valstar, F. Eyben and R. Cowie, "AVEC 2012-The Continuous Audio / Visual Emotion Challenge," 2012.
18. A. Metallinou and S. Narayanan, "Annotation and Processing of Continuous Emotional Attributes : Challenges and Opportunities," in *EmoSPACE Workshop*, Shangai, 2013.
19. P. Ekman, "An argument for basic emotions," in *Cognition & Emotion* , Vol. 6.3-4, pp. 169-200, 1992.
20. J. Rusell, "A circumplex model of affect," in *Personality and Social Psychology*, pp. 1161-1178, 1980.
21. R. E. Thayer, "The origin of everyday moods: Managing energy, tension, and stress," in *Oxford University Press*, 1996.
22. R. Dietz and A. Lang, "Affective agents: Effects of agent affect on arousal, attention, liking and learning," in *3rd ICTC*, San Francisco, 1999.
23. P. Gebhard, "ALMA – A Layered Model of Affect," in *4rth International Conference of AAMAS*, 2005.
24. M. M. Bradley, "Emotional Memory: A dimensional analysis," in *"Emotions: Essays on Emotion Theory*, pp. 97-134,1994.
25. A. Hanjalic and X. Li-Qun, "Affective video content representation and modeling," in *Multimedia, IEEE Transactions*, pp. 143-154, 2005.
26. E. Douglas-Cowie and e. al., "The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data," in *ACII 2007*, pp. 488-500.
27. M. Kipp, ""Anvil": The video annotation research tool", 2007.