Predicting mood from punctual emotion annotations on videos

C. Katsimerou, I. Heynderickx, and J.A. Redi

Abstract — A smart environment designed to adapt to a user’s affective state should be able to decipher unobtrusively that user's 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 propose a model for estimating mood from a known sequence of punctual emotions. To validate our model experimentally, we rely on the human annotations of two well-established databases: the VAM and the HUMAINE. We perform two analyses: the first serves as a proof of concept and tests whether punctual emotions cluster around the mood in the emotion space. The results indicate that emotion annotations, continuous in time and value, facilitate mood estimation, as opposed to discrete emotion annotations scattered randomly within the video timespan. The second analysis explores factors that account for the mood recognition from emotions, by examining how individual human coders perceive the underlying mood of a person. A moving average function with exponential discount of the past emotions achieves mood prediction accuracy above 60%, which is higher than the chance level and higher than mutual human agreement.

Index Terms—Emotion recognition, automatic mood recognition, affective computing, pervasive technology

1 INTRODUCTION

Affect-adaptive systems are commonly embedded into so-called smart ambiances, ready to read our needs and address them, and as such providing a higher quality of life. Technology endowed with emotional intelligence can, among other things, drive or maintain the user to a positive affective state, for instance by adapting the lighting system in a care center room to comfort the inhabitants [1]. In this example, the adaptive system needs to be equipped with empathic technology, which is able to identify (unobtrusively) the user’s affective state, at least as well as a caretaker would do.

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 [2]. A large body of work has been devoted to mapping visual representations of facial expressions [3], and body postures [4] into short-term affective states, commonly referred to as emotions. Emotions are short-term, intense and highly volatile affective states. Their automated recognition from video input is, therefore, often performed frame-by-frame. In certain affect-adaptive systems, however, this choice may be suboptimal: adapting the system behavior to the dynamics of instantaneous emotions may be redundant, if not counterproductive. Take the case of a lighting system that adapts its configuration (in terms of intensity and color temperature, for example) to improve the occupant’s affective state: it is neither necessary nor desirable that the light changes at the speed of instantaneous emotional fluctuations. The recognition of longer-term, underlying affective states would be more appropriate for such a system.

In psychological literature, long-term diffused affective states are often referred to as ‘mood’. Mood is distinguished from emotions typically based on its duration [5] and the intensity of its expression, although these differences are hardly quantified in an operational way. Little is known, for example, about the time spanned by either emotional or mood episodes. Thus, 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 a certain time window.

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 explicitly perform automatic mood recognition from visual input, except for some remarkable yet preliminary attempts [6] [7], discussed in more detail in section 2. On the other hand, psychological literature recognizes a relationship between underlying mood and expressed emotions [8] [9]. Thus, it may be possible to infer mood from a sequence of emotions, recognized from
e.g. video. 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 propose a methodology for inferring the mood of a person portrayed in a video, similar to what an external human observer would do. In the applicative context of a care-home, the mood perceived by the experienced care-takers can be considered a reliable ground truth, especially when it concerns fragile groups (e.g., elderly, patients), who are often not aware or cannot report properly their mood. We therefore refer to the target of our model as ‘perceived mood’. Such model then should infer perceived mood as a function of the instantaneous emotions expressed by the person in the video. As a proof of concept, we validate this functional relationship by examining how human annotators, while watching an emotionally colored video of a person, relate the emotions they perceive over time with the overall mood they perceive that person to be in. Perceived affect serves as ground truth for training any automatic affect recognition system, and thus, we consider it a required bed-set for proving the validity of our idea.

In summary, this paper aims at answering the following research questions:

1) To what extent can we identify (perceived) mood from punctually perceived emotions?
2) How do humans account for the (perceived) displayed emotions when judging the overall mood of another person?

Answering these questions brings us closer to retrieving a model of human affective intelligence that can then be replicated for machine-based mood perception, possibly re-using existing emotion recognition systems. In answering our research questions, we contribute to the field by a) bringing out automatic mood recognition as different (yet related) problem from emotion recognition, b) formulating a model where mood is the outcome of a function with punctual emotions as arguments, c) indicating an experimental setup for validating this model, d) defining the cases where this model is applicable, e) specifying the best fitting mood function out of a number of heuristics, and f) optimizing the parameters (e.g. minimum frequency for emotion recognition) of such function in terms of accuracy and computational complexity.

The remainder of the paper evolves as follows: section 2 presents the related background in psychological literature, for distinguishing and connecting emotions to mood, and in affective computing literature for automatic mood recognition. Section 3 defines the affective domains in which emotions and mood are expressed, and formulates the mood model within these domains. Section 4 presents the databases we used to validate our model. In section 5 we introduce a simple experimental setup as a proof of concept for the mood model. In section 6 we zoom into the model and quantify its salient parameters. Finally, we discuss in the last section the insights gained by the experiments, along with limitations of our analysis and future improvements.

2 RELATED WORK

The necessary background for our work stems from two directions: the psychological literature and the affective computing literature. Psychology provides the theoretical insights on how emotions and mood are different concepts, and leads to our claim that automatic mood recognition diverges from the problem of automatic emotion recognition (section 2.1). A second notion we retrieve from psychology is how the expressed emotions and the underlying mood are related, so that we can use the emotions to infer the mood. Subsection 2.2 lists the state-of-the-art efforts from the affective computing perspective to automatically recognize mood, which in essence boils down to emotion recognition.

2.1 Mood and emotions in psychology

A lot of psychologists who study affect in its nature, tend to consider emotion and mood as affective states very much associated, but at the same time distinguish them in terms of duration, intensity, stability, clarity and direction (awareness of cause). Jenkins [10] considers emotions, mood and temperament as affective states with increasing duration respectively, although without quantifying the difference in duration. Lane and Terry [11] define mood as “a set of feelings, ephemeral in nature, varying in intensity and duration, and usually involving more than one emotion”. Russell [12] describes mood as being prolonged core affect without object or with no specific object. Timing was proposed as a distinctive factor between emotions and mood also in [13], where emotions are said to evolve under clear dynamic phases (onset-apex-offset), whereas mood “lingers somewhere in the background of consciousness”. Finally, in [18], the authors conducted a so-called ‘folk psychological study’, in which they asked ordinary people to describe how they experience the difference between emotion and mood. A qualitative and quantitative analysis on the responses indicated cause and duration as the two most significant distinctive features between the two concepts.

Even though emotions and mood are well established as different constructs, literature agrees also on the consequence of the two, to the point that the terms are used interchangeably and a one-to-one mapping between the two is often assumed [14]. Ekman [15], 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 behavior that matches joy. Likewise, stress as an emotion would imply an anxious mood or even further an anxious personality. Mood can be instantiated due to an emotion [16], even though temporally remote from it [8] and in a dissipated intensity [9]. On the other hand, when eliciting emotions, the underlying mood of a person works potentially as a bias: it fortifies the emotions related to it and alleviates the not-relevant ones [9].

To summarize, emotions and mood are different concepts, yet tightly linked. As such, in this work we aim at capturing a functional relationship between these constructs, yet without attempting to determine causes and effects regulating this relationship.

2.2 Prediction of mood from audio-visual signals

From the affective computing point of view, automating the process of mood recognition (prediction of the perceived mood) entails linking data collected from sensors monitoring the user (e.g. cameras, microphones, physiological sensors) to a quantifiable representation of the (perceived) mood. In the case of visual input, very scarce results are retrievable in the 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 of 2011 [17] and 2012 [18]. 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 [6] 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 their expressions perhaps less genuine. Another reference to mood comes from [7], where the authors inferred again the bipolar happiness-sadness mood axis from data of 3D pose tracking and motion capturing. Finally, the authors of [19] were the first to briefly tap in the concept of summarizing puntual 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 of the affective state, without taking into account their temporal distribution.

In this study, we significantly extend 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, i.e., after the punctual (on frame level) emotion recognition module (see Fig. 1).

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 are represented. In affective computing there are two main trends for affect representation: the discrete [20] and the dimensional one [21] [22]. 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 point-vectors. Analogously, mood can be represented in a Euclidean (mood) space as a pair of valence and arousal values. In this work, we opt to discretize the mood space into four partitions, corresponding to the four quadrants defined by the valence and arousal axes. As a result, we define mood as a discrete variable belonging to one of the following mood classes (shown in Fig. 2):

1. positive valence - high arousal (PH), including moods such as happiness and excitement,
2. negative valence - high arousal (NH), including anxiety and anger,
3. negative valence - low arousal (NL), including sadness and sombreness,
4. positive valence - low arousal (PL), including calmness and serenity.

This 4-class mood discretization gives a trade-off between ambiguity and simplicity. It is assumed to be able to capture all possible moods sufficiently distinguishable, yet to eliminate redundancies. It should be noted, however, that this space partition is based on the psychological theoretical model, and thus, may be suboptimal; we demand the identification of more appropriate space configuration strategies to further research.

3.2 Problem formulation

Suppose we have a video $i$ representing a person in a mood, which would be perceived by an external human observer as mood $m_i$. The person in the video portrays over time different emotions, which are perceived by a
human (or perceived by a system) as emotions $e_i$. Emotion recognition is performed at certain time-steps, for example for every video frame $k$ during the entire video $i$.

For every independent video $i$ consisting of $n_i$ frames, the punctual emotion vector $e_i$ corresponding to the perceived emotion at frame $k$ is expressed in the VA space as:

$$e_i(k) = (v_i(k), a_i(k)), k = 1, 2, ..., n_i,$$

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

$$E_i = (e_i(1), e_i(2), ..., e_i(n_i)).$$

is known, we want to express the mood vector $m_i = (m_i^v, m_i^a)$ as

$$m_i = F(E_i),$$

where $F$ is the function mapping the emotion sequence into the mood vector, whose components $m_i^v$ and $m_i^a$ represent the mood valence and arousal respectively. We finally retrieve from the mood vector the quantized mood $M_i$ with

$$M_i = Q(m_i^v, m_i^a),$$

where the function $Q: \mathbb{R}^2 \rightarrow \{PH, NH, NL, PL\}$ maps the continuous mood vector in the quantized mood space defined in Sec. 3.1, and is defined as:

$$Q(x, y) = \begin{cases} PH & \text{if } \text{sgn}(x) \geq 0 \text{ and } \text{sgn}(y) \geq 0 \\ NH & \text{if } \text{sgn}(x) < 0 \text{ and } \text{sgn}(y) > 0 \\ NL & \text{if } \text{sgn}(x) < 0 \text{ and } \text{sgn}(y) < 0 \\ PL & \text{if } \text{sgn}(x) > 0 \text{ and } \text{sgn}(y) < 0 \end{cases}.$$ (5)

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

### 4 Data

#### 4.1 Databases

The first step to unveil experimentally the link between punctual emotions and mood was to retrieve ground truth that would illustrate this relationship. Therefore, we searched the literature for affective databases which would include videos portraying affective behavior, and for which both punctual emotion annotations (over time) and global annotations (one for the whole affective episode) were reported. We eventually retrieved the following two: a) the Vera am Mittag (VAM) [23] and b) the HUMAINE exemplar dataset [24]. Both of them are audio-visual. Snapshots of both databases are presented in Fig. 3.

VAM consists of videos extracted from a German reality show (Vera am Mittag), where the “protagonists” are plain people, narrating personal emotionally complicated situations. The creators/authors of the VAM database argue that these videos are indicative of naturalistic behavior, but the presence of cameras makes this argument doubtful [25]. The database is mainly speech oriented, but we chose to focus on the subset VAM-faces. For the latter, frames were extracted for a subset of 20 characters, whose face was captured by the cameras under a frontal or quasi-frontal view. A pool of annotators was asked by the authors of [25] to indicate, for each face and snapshot, the emotion expression in terms of valence and arousal, by means of a 5-point scale represented by the Self-Assessment Manikins (SAM) [26]. These 5 ordinal values were then mapped in the range [-1,1], distributed at equal intervals. Global affective annotations of each episode were provided in the form of adjectives with an affective connotation (e.g., sad). These global annotations corresponded to a general impression of the affective state of the person over the whole video. In the following section, we equate these global annotations to mood annotations.

The HUMAINE Database is a collection of videos capturing a large spectrum of the expression of emotion in everyday life. It includes naturalistic, induced and acted videos, providing diversity in the naturalness of the affective state expression. This makes the context of the whole database vary from social interaction to passive situations, from full-blown to suppressed emotions, from extreme to mild expressions. In total, HUMAINE consists of 46 videos, each of which revolves around a person, either acting or talking or both. Also in HUMAINE, a pool of annotators was asked to label the emotions and mood of the people portrayed in the video. In particular, emotions were annotated continuously in time and value (valence and arousal), using Anvil [27] as annotation tool. At the end of the video, the annotators were instructed to report the underlying mood they perceived the person in the video to be in. According to the instructions to the annotators, the mood had to represent the core affect of the person throughout that video, without focusing on a particular event.

![Fig. 3 Examples of the VAM Faces dataset (top two rows) and HUMAINE database (last row).](image)
Table 1 summarizes the main characteristics of the two databases used. An initial comparison shows that VAM consists of significantly longer videos. However, the annotations of emotions are generally available at random and sparse moments (due to the fact that only snapshots of frontal faces were annotated), leaving the largest part of the video unlabeled. On the other hand, HUMAINE contains in general short videos, which are annotated in a very dense and regular way time-wise, allowing to conduct analyses related to the effect of the emotion annotation frequency. We should take into account that for VAM on average 13.9 annotators (between 8 and 34) were recruited and the evaluated frames with low agreement (<30%) were discarded [28]. In HUMAINE each video was annotated by 4-6 annotators, characterized as experts, whereas no information on the level of expertise is specified for the annotators of VAM.

### 4.2 Preprocessing of the data

Before we could use the data in our analysis, we had to deem the annotations compatible. Our intention was to intervene as least as possible on the data, to avoid inserting additional noise. The emotion annotations were in both databases conveniently given as valence-arousal numerical values, so we did not process them further. This, however, was not the case for the global mood annotations. In HUMAINE, mood values were expressed on a discrete 7-point scale (in the interval [-3,3]) for valence and [0,6] for arousal, which we translated directly to [-3,3]). From these scores, we could determine the corresponding mood class by means of eq. (5). In VAM the global annotations were instead descriptive (e.g. grumpy, angry, vigorous, factual, etc.), and therefore, a preliminary mapping of these words into a numerical representation in the VA space was necessary. The transformation was based on the Whissell’s Dictionary of Affective language [29], in which each word is mapped on a tuple of valence and arousal values, according to its affective connotation, e.g., sad = (-0.625, -0.5714). Thus, all descriptive annotations were first converted into valence and arousal values through Whissell’s Dictionary, and then eq. (5) was applied to map the obtained values into the corresponding mood class, e.g., $Q(-0.625, -0.5714) = NL$.

Finally, we had to manually assign time stamps to the emotion frames of VAM, since this information was not provided. This was done by comparing annotated emotional frames (in image format) to the actual video frames, and assigning to the annotated image (and related emotion) the timestamp of the most similar video frame.

#### 4.2.1 Videos portraying more than one mood

In both databases, we came across videos in which either a shift in mood (i.e., the person in the video was perceived in one mood first and in another later, during the same video) or co-existing moods (i.e., the person in the video was perceived to be simultaneously in two or more different moods) were annotated. In both cases, this meant that, for the same video, more than one mood annotation was available. In HUMAINE this information was explicitly provided per video by the authors. In VAM we had to look for temporal relationships in the mood description. So, when a description would include the words ‘then’ or ‘first’ (e.g. “sad, then angry”), it would be translated into a ‘shift’ in the mood (from sad to angry). If that was not the case (e.g., “sad, angry”), the annotated moods were considered to be ‘co-existing’, i.e. blended in time, without being able to discriminate a temporal order.

#### 4.3 Average-coder annotation

When creating ground truth for multimedia system training, it is common practice to merge the different judgments given by multiple annotators to the same content (image or video), into a single, aggregated judgment (e.g. the Mean Opinion Score [30]). When dealing with affective databases, though, there are a few challenges to be taken into consideration. The existence of differences in the empathic skills of annotators is a problem stated frequently in the literature [19][31]. Those are manifested on both an absolute and relative scale. The first one refers to the assignment of different scale values to the same valence and arousal expression, and is attributed to the different level of ‘sensitivity’ of each annotator. A way to compensate for this is to normalize the valence and arousal values rated by each annotator per video to their zero mean value [32]. However, this type of normalization assumes normal distribution of the data, which was not the case with the emotion annotations of each annotator across all videos.

The relative scale in continuous annotation over time represents the ability to identify the current emotion in relation to the previous ones; in other words, being able to follow the general trend of the emotion sequence. High correlation between the continuous annotations of two different coders of the same video reflects good relative agreement, even if the judgments differ a lot in absolute terms. Incongruences could be attributed to either subjective perceptual differences in perceiving changes of emotions, or to the reaction lag [33], namely the delay be-

<table>
<thead>
<tr>
<th>TABLE 1</th>
<th>Summary of the main features of HUMAINE and VAM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HUMAINE</strong></td>
<td><strong>VAM</strong></td>
</tr>
<tr>
<td>Emotion Annotations</td>
<td>Continuous values of VA</td>
</tr>
<tr>
<td>Mood annotations</td>
<td>Discrete 7-point scale of VA</td>
</tr>
<tr>
<td>Avg. Video duration (range)</td>
<td>47 sec. (5-sec-3 min)</td>
</tr>
<tr>
<td>Frequency of emotion annotations</td>
<td>Quasi-continuous (at frame rate)</td>
</tr>
<tr>
<td>Modalities</td>
<td>Face-Gestures-Speech</td>
</tr>
<tr>
<td>Context</td>
<td>Daily situations, interactive &amp; non-interactive</td>
</tr>
<tr>
<td>Naturalness</td>
<td>Naturalistic Induced Posed</td>
</tr>
</tbody>
</table>
tween the moment the annotator perceives an emotion and the mouse click on the annotation tool (see Fig. 4, left figure). This lag is annotator-dependent and we need to take it into account when aggregating continuous individual annotations into one, to ensure as much as possible that all the aggregated values correspond to the same moment.

Various techniques to cope with annotator-specific lags in obtaining one continuous ground truth from multiple noisy annotations have been proposed. For example, it is possible to shift the continuous annotations with a certain time delay until their cross-correlation is maximized, such as in [34], or to use more sophisticated methods, which consider that the lag can change during the annotation, such as in [31]. Inspired by the latter, we aligned the emotion sequences by means of Dynamic Time Warping (DTW) [35]. DTW is a non-linear transformation applied on two given (usually time-dependent) signals, shifting them with non-linear time warping to minimize the Euclidean distance between them. In this work, we consider the emotion annotation sequences as signals evolving over time, and we apply DTW to align them. Taking into account that the responses of two annotators on the same video stimulus cannot lie infinitely away from each other, we added a locality constraint by demanding the algorithm to not shift any point of the signal further than 5 seconds. Since DTW is applied on two signals and we have up to 6 emotion sequences per video, as a first step, we pick a reference emotion sequence. This is chosen among all emotion annotation sequences available for a video, as the sequence that is higher correlated with the rest of them, after computing all possible pairwise correlations among them. Then, we align with DTW each of the emotion annotation sequences with the reference one, in order to minimize their distance. Having aligned the sequences, we can finally aggregate them into one “average-annotator” emotion sequence, by taking the mean of all warped annotations per time-step, separately for valence and arousal. It should be noted that the temporal alignment was not necessary for the VAM emotion annotations, since those were performed statically on predefined snapshots of video frames, identical for all annotators. Therefore, we calculated the averaged annotation per frame, as a simple arithmetic mean across all annotators.

5 EXPERIMENT 1: PRELIMINARY ANALYSIS

The first question we wanted to answer was whether our assumption that mood can be inferred from a sequence of recognized emotions had some grounding when using real data. Therefore, we set up an experiment to perform a proof of concept of our method. Specifically, we started from the assumption that the temporal distribution of the punctual emotions in each cluster had significantly different medians, and therefore, these clusters indicated two affective states sequential in time. Videos for which the null-hypothesis was rejected were classified as shifts; otherwise, they were labeled as videos containing co-existing moods. Table 2 reports the confusion matrices for the shift detection on the two databases separately. For HUMAINE all shift cases were mood (degenerating to a simple mean of the emotions), this was not the case for videos annotated with mood shifts or coexisting moods. Thus, as a first step, we also investigated whether by clustering emotions in the VA space, coexisting or shifted moods could be identified.

5.1 Detection of co-existing and shifted moods

For each of the two databases we considered, we first excluded for simplicity the cases annotated as neutral mood. In total, we ended up with 28 videos from HUMAINE and 16 from VAM. Of these, 11 from HUMAINE and 14 from VAM included multiple moods (i.e., shifts or co-existing moods).

For each video $i$, we used (a) the average-coder emotion sequence $E_i$ (retrieved in the way described in 4.3), expressed in the VA space (eq. (2)), (b) a sequence of timestamps $T_S^i$ containing the timestamps of the emotions in $E_i$ and (c) a mood value $M_i$ encoded as one of the 4 classes of eq. (5). In this first analysis, we aimed at discriminating between videos with co-existing and shifted moods. So, for all videos presenting more than one mood, we clustered the emotions in vector $E_i$ using the k-means algorithm [36]. The clustering was performed in a supervised way, limiting the number of clusters to the (known) number of annotated moods. The clustering algorithm returned the centroids $c_1,\ldots,c_k$ of each cluster in the VA space (see also Fig. 5), with $k$ being the number of moods for video $i$. In addition, the algorithm provided the punctual emotions $e_i$ belonging to each cluster.

We then checked whether the clusters were temporally ordered by evaluating the timestamps of the emotion annotations belonging to each cluster. To do so, we considered the clusters as independent variable and the distribution of timestamps within each cluster as the dependent variable, and we tested whether these distributions were significantly different with a Mann-Whitney U test. Rejection of the null-hypothesis implied that the temporal distribution of the punctual emotions in each cluster had significantly different medians, and therefore, these clusters indicated two affective states sequential in time. Videos for which the null-hypothesis was rejected were classified as shifts; otherwise, they were labeled as videos containing co-existing moods. Table 2 reports the confusion matrices for the shift detection on the two databases separately. For HUMAINE all shift cases were
recognized correctly, but the co-existing moods were not recognized as such, resulting to a reasonable accuracy (73%). This is not the case for VAM, where the decision “shift” versus “co-existing” seems to be random (50%). We discuss the reason for this discrepancy later in section 5.3. For the time being, we observe that this simple procedure is not sufficiently accurate for identifying mood shifts and we demand its improvement to further work. In the rest of the analysis, we consider only videos annotated with a single mood or for which the shift or co-existence of moods was identified correctly.

5.2 Mood Prediction

For the videos with correctly identified mood shifts and co-existing moods, and the single mood videos we checked whether the centroids of the clusters of punctual emotions $e_i$ would predict the annotated mood class. For single mood videos, we just applied eq. (5) to the coordinates of the cluster centroids, to map them into one of the 4 mood classes defined in Section 3.1. For shifted moods, we first ordered the centroids temporally based on the average of the timestamps of the annotations in each cluster, and then we checked whether they would correspond to the annotated moods, in sequence. Finally, for the case of co-existing moods, the temporal order did not matter and we simply allowed all pairwise comparisons between the centroids and the annotations, until we found a match.

To measure the accuracy of this simple method, we computed the ratio of the number of correct mood predictions over the (number of videos) x (number of mood annotations in each video). Viewing the mood prediction as a 4-classification problem, we can easily compute the 4x4 confusion matrix $CM$. In essence, the accuracy is the ratio of the trace of the confusion matrix to the total number of elements:

$$\text{accuracy} = \frac{\sum_{\text{row}=1}^{4} \sum_{\text{col}=1}^{4} CM(\text{row},\text{col})}{\sum_{\text{row}=1}^{4} \sum_{\text{col}=1}^{4} CM(\text{row},\text{col})},$$

where row and col stand for the rows and columns of $CM$, respectively.

5.3 Results and discussion

Tables 3 and 4 present the confusion matrices ($CM$) obtained from the comparison of the position of the centroids $e_i$ with the annotated moods for both databases, after eliminating the incorrectly predicted shifted and co-existing moods. We present the actual number of mood annotations, since the classes are not equally sized. The rows represent the annotations ("actual") and the columns summarize the outcome of the prediction. The diagonal elements indicate the accuracy per mood class. In both cases, mood class NH (negative valence-high arousal) is predicted better than the others. Mood class PL (positive valence-low arousal) never occurred in the dataset, but is predicted in some cases. Mood classes PH and NL are not predicted as well as NH, suggesting perhaps that not all moods are related in the same way with emotions.

From an overall perspective, clustering the punctual emotions can estimate mood with an overall accuracy of 53% in HUMAINE and of 44% in VAM (note that in VAM we were left with only 9 moods to be predicted). The results are still above randomness (i.e., 25%, since our classifier predicts mood without prior knowledge of the asymmetry of each class). In HUMAINE all diagonal elements are higher than or equal to the off-diagonal elements (except of the mood class PL). In VAM this is also

![Image](image-url)
the case for the mood classes NH and NL.

A possible reason for the poorer performance of predicting mood from clustered punctual emotions on VAM is that this dataset is very sparsely and irregularly annotated, and at the same time the videos are much longer than in HUMAINE. The average ratio \((\text{number of annotated frames})/(\text{total number of frames})\) is 0.11 for VAM, whereas it is 1 for HUMAINE. To counterbalance this difference, we subsampled the emotion annotations in HUMAINE to reach the ratio of 0.11, and repeated the clustering on these less frequently annotated videos. Interestingly, this did not compromise prediction accuracy. This suggests that the problem with VAM may stem from the irregularity of the annotations. Sampling very irregularly (i.e., very dense annotations at moments in between no annotations for very long parts) captures a small part of the emotions expressed and ignores completely emotions and emotional fluctuations occurring in the rest of the video. In a rather large video of about 15 minutes, a random sampling of a few seconds may not at all be representative for the whole emotional summary. This may also explain why we were reasonably able to detect mood shifts in HUMAINE, but not in VAM: the global annotations were provided after watching the entire video, but the annotated punctual emotions may have been captured only from a subset of the video, possibly clustered on one of the two moods before or after the shift.

To gain a clearer notion on the performance of clustering punctual emotions to predict perceived mood, we separated the single mood videos from the multiple mood ones, and we repeated the analysis described in section 5.2 for both databases. For HUMAINE this approach predicted the mood correctly in 5 out of 13 mood annotations from the multiple mood videos (38%) and in 11 out of 17 mood annotations from the single mood videos (65%). In VAM, we had only 2 single mood videos, and they were both predicted correctly (class NH), whereas for the multiple mood videos, only 2 out of 7 mood annotations were correctly predicted, resulting in a 29% accuracy. The main conclusion is that a simple clustering of punctual emotions in the VA space, is better in predicting single moods, but lacks accuracy when multiple moods are present in a video.

When comparing mood prediction on each database, it is essential to bear the substantial differences between the two databases in mind: in VAM we map emotion annotations of visual modalities only (i.e., facial expression and upper body) to a global annotation on the complete audio-visual material, whereas for HUMAINE the annotators were provided with both visual and audio modality when annotating both punctual emotions and mood. Another difference is the annotation task: VAM’s emotion annotations are discrete time-wise, while HUMAINE’s are quasi-continuous. The latter differs from the first in the way that the annotators have the whole contextual information of present and past moments when judging for the current emotion. However, their response has to be almost instantaneous as the frames flee in front of the eye, and there is no room for meta-thinking, while judging emotion from images statically can be done in a more careful way, paying attention on features from face and body. The different tools used for emotion annotation in turn can add up to the differences, but it is uncertain whether SAM is more user-friendly than ANVIL. Finally, the VAM annotations at the end of the video are not explicitly mood annotations, but rather summaries of emotions, which may to some extent differ from mood.

Despite all the above disclaimers, this preliminary experiment allows us to (1) confirm that our hypothesis that mood can be inferred from emotion sequences has some grounding, (2) grasp the notion of the temporal parameters for emotion annotation: too irregular emotion annotations are not sufficient for safely predicting mood, and (3) identify a core problem in mood detection, which is that of detecting shifts in mood and identifying these moods. With respect to the latter, we demand the solution of such problem to future research. The first two results allow us instead to proceed with our investigation by building a more sophisticated function mapping emotions into mood, and by focusing on single mood videos densely annotated, as those included in the HUMAINE dataset.

### 6 EXPERIMENT 2: MOOD PREDICTION MODEL

The second experiment aims at retrieving a more accurate function for estimating mood from a sequence of emotions. More specifically, with this analysis we answer questions like whether and to what extent factors like intensity, occurrence frequency, temporal order and granularity of the punctual emotion annotations contribute to the overall perceived mood.

Based on the results of 5.1 and 5.2, we decided to focus only on the HUMAINE database for this analysis, be-

<table>
<thead>
<tr>
<th>predicted</th>
<th>PH</th>
<th>NH</th>
<th>NL</th>
<th>PL</th>
<th>Overall Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PH</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>53%</td>
</tr>
<tr>
<td>NH</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>NL</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>PL</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>predicted</th>
<th>PH</th>
<th>NH</th>
<th>NL</th>
<th>PL</th>
<th>Overall Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PH</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>44%</td>
</tr>
<tr>
<td>NH</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>NL</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>PL</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
cause it is annotated in a continuous way, allowing us to
tune factors like the frequency of annotated emotions.
This also allowed us to make a different choice in the
experimental setup. HUMAINE provides the emotional
and mood data separately per annotator. Because of the
subjectivity issues in obtaining one ground truth from
multiple annotators remarked in the previous section, we
decided to focus on mapping emotions to mood per an-
notator separately, without aggregating the annotations into
mean valence and arousal values. This can be seen as a
form of adding annotator information, so that the mood
model can predict more precisely the individual mood
labels. Due to the bad prediction of shifted and co-
existing moods from the former analysis, we also exclud-
ed from our analysis the videos with reported shifted or
coexisting moods. We consider that these cases require
separate attention, and we demand their analysis to fur-
ther work. As a result, in this experiment we analyzed 168
single-mood annotations in total: 38 of class PH, 34 of class
NH, 51 of class NL, and 46 of class PL. Due to our analy-
sis per annotator, different mood annotations can corre-
spond to the same video.

We first formulate simple hypotheses about how mood
and emotions are related, based on the literature,
proposed computational models and common sense. We
formulate these hypotheses as functions that map emo-
tion sequences to mood and test how well these functions
predict the mood annotations of each annotator of HU-
MAINE. Then, we subsample the continuous (i.e., at
frame rate, so 25 fps) emotion sequence with iteratively
decreasing frequency in the emotion annotation, in order
to define the most economical sampling for an automatic
system. Lastly, we investigate which temporal portion of
the emotion sequence is more meaningful for predicting
mood.

6.1 Modeling mood from a sequence of punctual
emotions

6.1.1 Basic mapping functions

We propose several possible formulations of the function
F in eq. (3), which maps punctual measurements of emo-
tion into a representative value for the overall mood. This
value is then used in eq. (5) to predict the mood class,
unless stated otherwise.

**Predictor 1: The mean emotion (mean).** Probably the easi-
est assumption is that mood is formed by the equal con-
tribution of all the emotions within a given timespan [19].
The average of the punctual emotions will then be the
“station” mood, which acts as a gravitational force on
them [37]. More formally, the mean of an emotion se-
quence over a particular time window is the predictor of
the overall mood for this time window:

\[ M_i = Q(\bar{F}(E) = \frac{1}{n} \sum_{k=1}^{n} e(k) \) /n \]

\[ = Q \left( \frac{1}{n} \sum_{k=1}^{n} v(k) / n, \sum_{k=1}^{n} a(k) / n \right) \]  

**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

\[ \|e_k(k)\| = \sqrt{(v_k(k))^2 + (a_k(k))^2}, k = 1, 2,.., n \]  

(8)

Then the mood occurs from the emotion vector that max-
izes the intensity over the sequence \( E_k \) or

\[ M_i = Q(F(E_k) = \argmax \{Q(e_k(k), k = 1, 2,.., n)\} = Q(e_q) \]  

(9)

**Predictor 3: The longer emotion (long).** Another hypo-
thesis is that the emotions that occur more within a given
timespan sustain the associated mood [38]. Thus, we may
map individual emotion vectors into mood vectors direct-
ly 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 four disjoint subsets of \( E_q \) defined as

\[ E_q = \{e(k) | Q(e_k(k)) = q, k = 1, 2,.., n \} \]  

each with cardinality \( C(q) = |E_q| \), then the mood corre-
sponding to the longer emotion is

\[ M_i = Q(F(E) = \argmax \{C(q) \} \]  

(11)

**Predictor 4: First emotion (FE).** A reasonable property of
the mood function is memory [39], in the sense that mood
recognition involves the assessment of not only the cur-
rent 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 ep-
isode, resonating the impact of the first points to the cur-
rent moment. Therefore, we may hypothesize that the
first of a sequence of emotions over a certain time win-
dow is a predictor for the overall mood for this time win-
dow, or

\[ M_i = Q(F(e_1)) \]  

(12)

**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 last recognized emo-
tion defines the overall mood, that is:

\[ M_i = Q(F(e_n)) \]  

(13)

6.1.2 Complex predictors

The simple models proposed in section 6.1.1 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 emo-
tional episode. We can formulate this as
where $w$ is the size of the memory window. In fact, eq. (14) 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 [40], that is, it should gradually neglect the past, as it moves along the emotion sequence. This can be modelled through a discount function $D_w$ of the previous frames, either linear $LD$ (eq. 15), as seen in [38], or exponential $ED$ (eq. 16).

$$D_w(k) = \begin{cases} \frac{k-(n_i-w)}{n_i-w}, k = n_i-w,...,n_i, \\ \exp\left(\frac{k-(n_i-w)}{n_i-w}\right), k = n_i-w,...,n_i \end{cases}$$

Then the mood is the weighted average of the last $w$ emotions:

$$M_i = Q(F(E_i)) = Q\left(\sum_{k=n_i-w}^{n_i} w_i(k) : D_w(k) : \sum_{k=n_i-w}^{n_i} D_w(k)\right).$$

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

### 6.2 Results and discussion

The prediction accuracy (given by eq. (6)), obtained by the basic mood functions of section 6.1.1, is presented in Fig. 6, per annotator and in Fig. 7 for the average annotator. The random benchmark marks the lowest bound of randomly assigning moods to one of the 4 classes (i.e., 25%). A second benchmark we chose to use is the annotator agreement, namely how well a human annotator would predict the mood perceived by a fellow human annotator, to check whether our model can do just as well. To estimate the agreement of one annotator $j$ with a fellow annotator $c$, we counted the number of matched mood annotations (quantized according to eq. (5), i.e., the target of our proposed model), normalized by the number of videos they annotated in common $N_{vc}$ (not all annotators annotated all the videos). We repeated this across all fellow annotators and took the average value over the number of fellow annotators $C$ (C=6 in HUMAINE), to obtain the level of agreement of annotator $j$ with all the other annotators throughout the database:

$$A_j = \frac{1}{C-1} \sum_{c \neq j} C \sum_{i=1}^{N_{vc}} \delta(M_{ij}, M_{ic}),$$

where $A_j$ stands for the agreement of annotator $j$ with the fellow annotators, $M_{ij}$ the mood assessment of the $c$-th annotator on the $i$-th video, and $\delta$ the difference function, defined as

$$\delta(x, y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise} \end{cases}$$

The annotator agreement is presented per annotator with the solid orange column in Fig. 6. Then, the average annotator agreement across the database or human agreement $HA$, would be computed as

$$HA = \frac{1}{C} \sum_{i=1}^{C} A_j,$$

and resulted in 44% (marked with the dashed line in Fig. 7). This number represents a rather low inter-coder agreement in terms of mood, underpinning the subjectivity of the annotated (perceived) mood.

From Fig. 6, it can be derived that the mood estimation process is also subjective, even though mean and LE seem to represent quite well the majority of the individual mapping functions. Mean is the most popular predictor, predicting best the mood annotations of annotators 1, 3 and 6. For annotators 4 and 5 the most accurate model is LE, being also a good predictor according to annotator 3. For annotator 2 max outperforms the other models. Overall (Fig. 7), mean predicts mood most accurately (60%), with a similar accuracy for LE (59%), indicating the importance of previously recognized emotions as well as current emotions in mood. The max is in general a worse predictor than the long, which implies that duration is more important than intensity in mood prediction. FE is the worst predictor. Except for FE, the predictors are significantly above random choice ($p<0.01$). We also performed paired t-tests between the annotator agreement $A_j$

**Fig. 6.** Accuracy of mood prediction from emotions for the 5 simple models per annotator.
The annotations of annotator $j$. Only the functions mean and LE could predict the mood annotations systematically more accurately than the fellow annotators ($p<0.035$).

Further on, we continue the analysis only for the average annotator, whose best predictor represents the overall best predictor.

It is interesting at this point to check how properties of the video, i.e., its length, the granularity of the emotion sequence and the temporal position of the emotions in the sequence influence the mood prediction. For this analysis, we investigate these effects on the mood model based on the mean function given by eq. (7), as this was the best predictor obtained.

**Length of the video.** 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 moods. 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 (at least for videos up to 3 minutes) the duration of the emotional episode did not influence the correctness of the mood prediction.

**Granularity of emotion sequence.** 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 rates $f_i = \left\lfloor \frac{(N-1) / (n_i-1)} \right\rfloor$, with $N$ the number of samples per sub-sampled sequence $E_i^{ss}$. Then we applied $F$ on $E_i^{ss}$ with $N$ as parameter. The hypothesis is that

$$F(E_i) = F(E_i^{ss}) = \frac{1}{F\left(\frac{(i-1)}{f_i+1} + 1, \ldots, \frac{i \cdot (N-1)}{f_i+1} + 1\right)}$$

(21)

where $F$ is estimated from eq. (7). For $N=2$ only the first and the last point of the sequence are taken into account. In our experiments, we investigate the effect of subsampling the emotional sequence from 2 up to 11 points. Fig. 8a illustrates the prediction accuracy of mood (given by eq. (6)) 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. (21) is essentially satisfied.

In real applications, the duration of the video is not known a priori, thus, 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 iteration those videos 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. 8b.

![Fig. 7. Accuracy of mood prediction from emotions for the 5 simple models for the average annotator.](image)

![Fig. 8. Mood prediction accuracy of the mean predictor, as a function of a) number of samples in the sequence, b) the frequency of the annotated emotions expressed in seconds, c) temporal position of the emotions in the sequence. d) Overall valence, arousal and mood prediction accuracy of the average annotator with the two best basic (mean, LE) and the three complex (MA, LD, ED) models.](image)
computed how well they predict annotations of mood valence (positive/negative) and arousal (high/low), respectively. As shown in Fig. 8d, overall valence is mainly better predicted than overall arousal, and they are both better predicted than mood, as expected, since mood stems from the combination of these two noisy predictions. Mean predicts valence with an accuracy higher than 80%, whereas the overall arousal is predicted better by 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 (Table 5). The misclassifications occur in their vast majority between classes that share the valence or arousal axis, namely PH and NH, NH and NL, NL and PL, and PL and PH. Only in 6 cases misclassifications occur between opposite moods, i.e., between classes PH and NL (5 videos), and classes NH and PL (1 video). 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 (NL or PL) are more easily confused with the neighboring low arousal mood (PL or NL), rather than with the neighboring mood of the same valence (NH or PH). 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. Sadness/somberness (i.e., class NL) is the best predicted from punctual emotions, whereas calmness/serenity (i.e., class PL), is the least, which may imply that it is more complicated to judge calmness from an emotion sequence with the proposed model.

7 CONCLUSIONS AND OUTLOOK

We introduced a novel framework for predicting mood, as perceived by other humans, from the emotional expressions of a person. To the best of our knowledge, this is the first systematic effort to automatically predict perceived mood from a sequence of recognized emotions. Another novelty of this work consists in introducing an experimental setup for validating the proposed computational models, using emotion and mood annotations of humans on affective databases. On this experimental platform, we tested simple computational models relating the perceived emotions to perceived moods and we explored their properties.

Our experimental results showed that clustering emotions in the VA space can predict single moods much better than multiple moods perceived within the same video. Additionally, a model that has memory of the last recognized emotions, yet exponentially discounts their importance in the overall mood prediction is able to predict perceived mood with 62% accuracy, which is well above random, as well as above human agreement. However, this score implies that the emotions-mood relation-

<table>
<thead>
<tr>
<th>Actual</th>
<th>PH</th>
<th>NH</th>
<th>NL</th>
<th>PL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PH</td>
<td>26</td>
<td>2</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>NH</td>
<td>1</td>
<td>23</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>NL</td>
<td>2</td>
<td>6</td>
<td>36</td>
<td>5</td>
</tr>
<tr>
<td>PL</td>
<td>11</td>
<td>1</td>
<td>17</td>
<td>18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Valence</th>
<th>Arousal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pos.</td>
<td>Neg.</td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Pos.</td>
<td>62</td>
</tr>
<tr>
<td>Neg.</td>
<td>8</td>
</tr>
<tr>
<td>High</td>
<td>40</td>
</tr>
<tr>
<td>Low</td>
<td>20</td>
</tr>
</tbody>
</table>

We can deduce that a safe rate for sampling emotions (which still capture the global picture without discarding essential emotional information) is every 3 seconds (98% of the accuracy achieved without subsampling). This finds us in perfect agreement with the economical coding scheme of the SEMAINE database [41]. However, sampling up to every 20 seconds would still grant 90% of the accuracy of the original sequence.

Temporal position of the emotions. We investigated to what extent the temporal position of a punctual emotion in the sequence contributes to the final mood prediction. For this, we segmented each sequence in 10 equal non-overlapping temporal sub-windows, and predicted the mood, as if we only had one of these portions of the sequence available. The average accuracy (again given by eq. (6)) is depicted in Fig. 8c. 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. 14), LD (eq. 15) and ED (eq. 16) as 30% of the video length n. The average accuracy of these complex models for the average coder is summarized in Fig. 8a, along with the accuracy for the mean and LE models as reference. The performance of the complex models is very similar, although the moving average with exponentially discounted precedent values scores the best. 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 exploration of a mood function. 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

![Table 5](image_url)
ship is not as straightforward as considered in the literature, and a simple mean over the recognized emotions is insufficiently adequate to predict mood.

In our experiments, we also found that sparsely recognized emotions, distributed irregularly within the video timespan, do not reveal sufficient information for the perceived mood (in VAM the mood prediction was hardly above random). However, when the annotations are regular, a sequence of emotions recognized as sparsely as every 3 seconds predicts mood as well as a sequence of emotions recognized at video frame rate, which is relevant information for decreasing the computational complexity involved in designing empathic systems working in real time.

Lastly, we showed that valence and arousal values are predicted quite satisfactorily from our models. Yet, the prediction accuracy could be improved. One possible limitation of our method is that we predefined the boundaries of the mood classes; our mood space partitioning was based on a psychological model of affect, and may be suboptimal for the utilized dataset. Additionally, the short duration of the videos is not necessarily adequate to unveil the perceived mood; the actual mood 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.

The subjectivity of the mood annotations represented by the low human agreement might raise questions regarding the subjectivity of the proposed mood model. Nevertheless, even though the mood annotations are subjective, the function mapping perceived emotions to perceived mood can still be universal, and therefore objective. The proposed model of the average annotator is a generally objective model, that can be applied directly on machine recognized emotions.

Another 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). For instance, one annotator commented on a video he rated: “He (the person speaking in the video) is talking with pride about his family, but I can see he is deeply sad”. In this case, the positive, high arousing emotions scored continuously do not seem to reveal anything about the overall mood (class NL), advocating the importance of contextual information in the mood estimation process.

An additional worthwhile future extension of this work would be to prove the accuracy of the proposed models when applied to a sequence of machine-recognized rather than human-recognized emotions. However, as already stated, we modeled automatic mood prediction as a two-stage process. The first stage, namely automatic emotion recognition, has been well studied during the last two decades. We believe that the second stage, i.e., mood estimation from recognized emotions, is still at its infancy and requires separate attention.

REFERENCES


