Inferring Mood in Ubiquitous Conversational Video

Dairazalia Sanchez-Cortes[1], Joan-Isaac Biel[1], Shiro Kumano[3], Junji Yamato[3],
Kazuhiro Otsuka[3] and Daniel Gatica-Perez[1,2]
[1] Idiap Research Institute, Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[3] NTT Communication Science Laboratories, Japan
(dscortes,jibiel,gatica)@idiap.ch,jy@acm.org,(kumano.shiro,otsuka.kazuhiro)@lab.ntt.co.jp

ABSTRACT

Conversational social video is becoming a worldwide trend. Video communication allows a more natural interaction, when aiming to share personal news, ideas, and opinions, by transmitting both verbal content and nonverbal behavior. However, the automatic analysis of natural mood is challenging, since it is displayed in parallel via voice, face, and body. This paper presents an automatic approach to infer 11 natural mood categories in conversational social video using single and multimodal nonverbal cues extracted from video blogs (vlogs) from YouTube. The mood labels used in our work were collected via crowdsourcing. Our approach is promising for several of the studied mood categories. Our study demonstrates that although multimodal features perform better than single channel features, not always all the available channels are needed to accurately discriminate mood in videos.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Human Factors

Keywords

Nonverbal behavior, Verbal content, Moods, Sentiment analysis

1. INTRODUCTION

Thanks to social media and mobile computing, conversational video is truly becoming ubiquitous. Many applications are allowing people to talk via video. This includes the now "traditional" forms of video blogging on sites like YouTube, two-way communication via FaceTime, or multiparty calls on Skype or Google hangouts, but also a new generation of mobile applications like Vine, SnapChat, and MixBit, that allow to share short video snippets with friends.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...$10.00.

The naturality and ubiquitousness of conversational video, where people share personal news, ideas, opinions, and experiences, opens the possibility to study human mood in real-life settings. Mood is defined as "a conscious state of mind or predominant emotion" [2] or "a temporary state of mind or feeling" [4]. Automatic mood inference on social video could be used to search for mood effects with respect to political concerns, services, etc., as it is currently done on textual sources like blogs and tweets [18, 35, 20]. Furthermore, it could be used to recommend video playlists based on similar mood to the video creators themselves or to a general audience.

The research progress in recognizing mood in text is significant [18]. Automatic mood labeling in written text logs has received much attention, with the aim of understanding associated moods to products, as well as political opinions and population habits [33, 34, 25, 21, 38]. This said, video brings with it the rich nonverbal dimension: people often express their mood by exhibiting measurable behaviors in their speech, facial expressions, and gestures, emphasizing and modulating what is said [16]. A substantial amount of research has also been done in audio processing and computer vision as individual modalities for automated mood inference tasks using posed and naturalistic data [14, 26, 44]. In addition, an increasing amount of work is attempting to understand mood from multimodal cues in both scripted and realistic situations [42], and is exploring other individual variables like personality traits [7]. In our view, much of the existing knowledge can be transferred to the social video setting, including techniques to reliably label human mood, computational modules for text and perceptual processing, as well as previous experiences about the many challenges associated to these tasks.

In this paper, we present a novel, systematic study on automated inference of human mood in conversational social video. We study the feasibility of inferring a broad set of 11 mood categories (happiness, excitement, relax, sadness, boredom, disappointment, surprise, nervousness, stress, anger, and overall mood) on YouTube vlogs from a rich set of automatically extracted nonverbal and verbal cues. Our contributions are as follows:

- We present a social video dataset of 264 vlogs downloaded from YouTube (3 minutes in average per video), where the 11 annotated moods were manually produced via crowdsourcing, and manual speech transcriptions have also been generated. Vlogs are an excellent example to study social conversational video given their wide availability, and might be used as a first

testbed as new video sharing applications emerge. Furthermore, the set of mood categories was designed to study recognition tasks beyond the positive/negative mood polarity task.

- We use state-of-the-art methods to automatically extract nonverbal features in ubiquitous video that contains variations in quality, background, lighting, etc. Our feature set, while not new in terms of new extraction techniques, is comprehensive and includes speaking activity, prosody, visual activity, facial expressions, and linguistic and paralinguistic categories that have been validated in psychometric terms. This allows for a systematic evaluation.
- Using two supervised learning methods, we conduct a study of the effect of single and combined modalities (verbal and nonverbal) on mood inference performance for each of the mood categories. We also examine the effect of inter-annotator agreement on mood inference performance. The study shows that even though mood inference is a challenging task, we can recognize several categories in a binary classification setting, with promising results for Overall mood and Excited (69% and 68%), both statistically better than a majority class baseline. Moreover, although multimodal features perform better than single channel features, not always all the available channels are needed to discriminate mood levels.

The paper is organized as follows. We discuss related work in Section 2. Our approach is summarized in Section 3. In Section 4 we describe our corpus and annotations. Section 5 describes the nonverbal and verbal cues and the machine learning framework used in the study. We present and discuss results in Section 6. We conclude in Section 7.

2. RELATED WORK

In this section, we discuss previous work that has examined mood inference from textual and perceptual data.

Mood inference from text. Studies in psychology have revealed strong connections between the words we use to express ourselves in written and spoken forms with personal traits and emotional states [40, 30]. It is thus not surprising that text analysis techniques have exploded to examine these dimensions in text blogs, product reviews, and social media, under the umbrella term of sentiment analysis [18]. One of the first mood classification approaches using written blogs was presented in [33], using the LiveJournal dataset (815k blogs, 200 words per blog on average.) A set of mood labels were provided by the blogger themselves (from a list of available moods along with an option to add a new labels) when submitting the blog entries. The method used n-grams and other features capturing basic statistics from the text, and SVMs. Another early approach proposed to classify moods in blogs used term frequency/inverse document frequency (tf/idf) and the 5,000 most frequent English words, also using LiveJournal blogs (168 words per blog on average) [25]. A separate method on LiveJournal proposed for mood categorization used orientation scores from positive and negative words, verbs, and adjectives in addition to Bag-of-Words (BoW) and text statistics [21]. This work showed that sentiment orientation improved mood classification performance up to 63.5%, as compared to only using text statistics (40%). More recently, Nguyen et al. [37] presented also mood classification from blogs using tf/idf, BoW and Affective Norms of English (ANEW) words, using two blog datasets in the analysis, and reporting performance of up to 77.6% classification accuracy based on different feature selection schemes per mood.

There has also been an explosion of work to characterize sentiment in social media short text sources (tweets, comments, tips, etc.). The characterization of polarity of tweets is challenging due to the brevity of text (140 characters in a tweet vs. 200 words in a blog entry) and the use of idiosyncratic jargons. Well-known examples of mood analysis include [35] who presented visualizations of mood fluctuations over time and space in the US context, and [20], which examined daily and seasonal fluctuations of mood worldwide according to a number of contextual factors. A recent work that examined the potential of crowdsourcing-based labeling of tweet mood is [13] based on the circumplex model (that describes valence and activation dimensions). While our work also uses crowdsourcing to obtain mood labels, in contrast to all the above literature, our study integrates the video and audio modalities to text, and so brings in the possibility of complementing sentiment analysis techniques.

Mood inference from audio and video. Many psychological studies have demonstrated the relationship between affective states, including mood and emotions, and expressive human behavior. A significant body of work has also studied mood inference from audio and video but without specifically addressing social video. Regarding audio, mood in non-written forms has been explored using acoustic features. As one early example, the work in [24] used acoustic features to distinguish between negative and non-negative emotions from females and males, using labeled utterances from a call center application. The study found consistent improvement of performance using combined acoustic and language features (emotionally salient words), for both females and males. Several other affective states related to emotion have been studied in the speech community for several years, with some comparison initiatives (e.g. [14]), but not using social video as we do in this paper.

Regarding visual processing, facial expressions reveal internal states [16], and numerous efforts have been made to develop video-based automatic recognition systems of facial expressions, e.g. [44]. As a result, advanced facial expression analyzers are now publicly/commercially available, e.g. [26] and [3]. Based on these techniques, the automatic analysis of spontaneous facial expressions in the wild is one key topic in affective computing. The target affective states include prototypical emotions [44], emotional dimensions such as valence and arousal [32], empathy [23], pain [29, 27], and depression [19]. Some of them have focused on the observers' impressions about the target person [32, 23], like the present study. One recent study classified viewers' preferences for video advertisements from their smiles produced during video watching [31]. A fundamental difference between that work and ours is that, instead of analyzing the passive behavior of observers, we are interested in modeling the mood of active speakers in social video.

Finally, the combination of audio and video cues for recognizing affective states has been studied in the past. A well known study in a laboratory setting reported classification of 11 emotional states using prosodic features from sub-

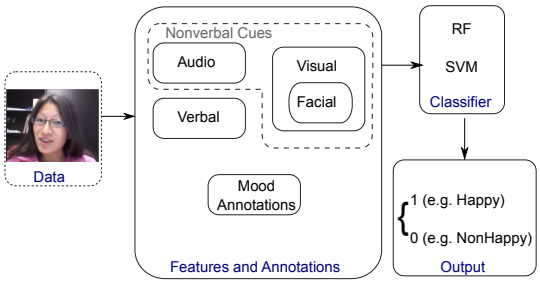

Figure 1: Our approach. We address the automatic inference of the perceived mood in conversational social video from single and multimodal features.

jects displaying requested emotions [42]. The single acoustic modality performed relatively poorly, although combined modalities (motion facial units and prosodic features) improved significantly the classification performance. To our knowledge, the closest work to ours is [36], which used 47 videos from YouTube where people expressed product reviews. Each video was normalized at 30-second duration and manually labeled as negative, neutral or positive. While single modalities showed low performance of up to 41.9% (F-measure), additional experiments using multimodal features (gaze, smile, words polarity, pause, and pitch) showed a considerable increase of performance up to 55.3%. In contrast to this work, we study a significantly richer set of moods (11 categories, one of which corresponds to overall mood) on a significantly larger (5.6 times as many videos and users) and more diverse dataset, containing a variety of topics and user intentions, i.e., not only product reviews.

3. OVERVIEW OF OUR APPROACH

Figure 1 presents our approach. First we automatically extract a large number of nonverbal and verbal cues per vlog that allow multimodal analysis. Then we use a classification framework to infer binary mood labels.

The nonverbal features include **audio cues**, i.e., acoustic features (e.g., pitch, energy, speaking rate, formants and bandwidths) computed from the audio channel; **visual features**, that capture looking activity, pose cues and visual activity, and additionally, we also compute Facial expressions cues. For the **verbal cues**, we computed word categories using Linguistic Inquiry Word Count (LIWC) from manual transcriptions from the vlogs. We describe the feature extraction process in Sections 5.1 and 5.2.

Regarding classification, to have balanced classes and to avoid overfitting, we divided the samples per mood using the median value from the mood labels, and applied 10-fold cross-validation, where train and test sets are disjoint sets. The features are normalized using z-normalization z=N(0,1) and they are passed to the binary classifier, in this case Support Vector Machine (SVM) and Random Forest (RF). For analysis, we perform inferences using features from single modality cues, and then we perform feature fusion. The outputs of the classifier are then thresholded and assigned to their respective mood class (e.g., Happy and Non-Happy).

4. DATASET

We used a dataset of YouTube conversational social video shared by Biel and Gatica-Perez [7]. This data includes 264 vlogs, each one featuring one single vlogger talking in front of the webcam. The spoken language is English. The collection had no restriction in terms of the topics addressed by the vloggers or the recording setting, so the dataset is quite diverse with respect to the content and the audio and visual quality of the videos. The typical vlog is recorded indoors with a commercial webcam, lasts about 3min, and features the head and shoulders of the vlogger.

The dataset also includes annotations of mood and demographic impressions that were collected from people watching vlogs in Mechanical Turk. The reason to use non-experts in the annotations is supported by the findings reported in [15, 43], which affirm that untrained observers can accurately judge spontaneous and natural emotions. Moreover, one of the advantages of labeling mood via crowdsourcing is that the annotators watch the video in ecologically valid conditions, i.e., watching them directly from YouTube. Concerning the demographics of the dataset, approximately 70% of the vloggers were labeled as below 24 years old (majority between 18-24 years), and around 80% of the population was reported as Caucasian. With respect to the gender, it is mostly balanced: 53% Female and 47% Males. More details on these annotations can be found in [6]. Clearly our sample is not a fair sample of the world population, but reflects the statistics of the YouTube, English speaking video blogger community.

For each vlogger, five MTurk workers annotated ten items that cover ten affective states (one item per state), as well as one overall judgment of mood (positive or negative). The use of these number of workers for the annotation task, is supported by the findings of Snow et al. [43]: "For an affect recognition task we find that we require an average of 4 non-expert labels per item in order to emulate expert-level quality". Note however that the task in [43] and ours are not the same. The study of this issue in more detail is part of future work.

As measure of reliability for the annotations, we use the Intraclass Correlation Coefficient measure $ICC(1,k)$, which is a standard measure used in psychology. These judgments are averaged across annotators (used as ground-truth in our paper), and are reliable with the following intra-class correlations ($ICC(1,k)$, $k=5$): Overall mood (.75), Happy (.76), Excited (.74), Angry (.67), Disappointed (.61), Sad (.58), Relaxed (.54), Bored (.52), Stressed (.50), Surprised (.48), Nervous (.25). These ICC values show that high arousal moods such as Excited, Happy, or Angry are easier to judge by annotators, a result that may in part be explained by this moods manifesting more explicitly in vloggers' behavior. As reported in [6], the overall judgments of mood can be explained mainly as a combination of Happiness, Excitement, Relax, and Anger judgments.

We complemented this dataset with the manual transcriptions of vlogs, which was performed by a professional company. The transcriptions have in average 625.9 words per vlog.

5. AUTOMATIC MOOD INFERENCE

We integrated several audio processing, computer vision and text analysis technologies to characterize vloggers' nonverbal and verbal behavior. In Section 5.1, we describe the methods used to compute nonverbal cues from audio and video, while in Section 5.2 we explain the analysis technique used to characterize verbal content. In Section 5.3 we give details about the classifiers.

5.1 Nonverbal Cues

In this work, we investigate three different nonverbal behavioral sources that have been documented by research in nonverbal communication as conveying emotional information [22]: vocal cues, visual activity, and facial expressions.

5.1.1 Audio nonverbal cues

Voice is a primary channel for expressing emotion in humans [22]. Research has shown that emotion perception depends on changes in pitch, volume and speaking rate [41], and has repeatedly showed that automatically extracted prosodic cues are useful to capture personal and emotional information [24, 42].

We extracted **prosodic cues** that estimate the pitch, energy and speaking rate of vloggers. First, we processed the audio channel of vlogs using PRAAT [11] to generate frame-by-frame estimates of these and other related signals (e.g. the second and third formants and their bandwidth). Second, we aggregated features across the whole video duration by computing the mean, median, mean-scaled standard deviation, maximum, minimum, and entropy. In total, we computed 98 prosodic cues.

5.1.2 Visual activity nonverbal cues

Gesture, gaze, posture, and movement are rich sources of personal and affective states available though the visual channel. The extraction of these nonverbal cues in social video is challenging due to the variety of content available, but has nevertheless been successfully addressed to build computational models of vlogger personality [7].

We extracted three different types of visual nonverbal cues. First, we extracted **looking activity cues** (cues related to gaze) obtained from looking-non-looking segmentations including the time looking at the camera, the average duration of looking segments, and the number of looking turns. These looking activity segmentations were produced following a method based on a frontal face detector and that has been shown useful to capture looking behaviors in vlogs at scale, i.e., without manual intervention [8]. Second, we used the position and size of facial detections to compute **pose cues** such as the proximity to the camera and the horizontal and vertical framing of the vlogger (i.e., the position of the vlogger with respect to the center of the frame). Finally, we characterized the **visual activity** of vloggers through the computation of weighted motion energy images (wMEI). wMEIs are gray scale images that measure the accumulated motion through the whole video (one single image is generated per video, where brighter pixels correspond to regions with higher motion). From this, we computed several features such as the entropy, mean, median, and the vertical and horizontal center of mass.

In addition to the visual only activity features, we also extracted a few **multimodal cues** generated from looking-not-looking and speech-non-speech segmentations. In particular, we computed the looking-while-speaking time (L&S), the time looking-while-not-speaking (L&NS), and the multimodal ratio (L&S/L&NS), which capture joint patterns of speech and gaze. The total number visual and multimodal cues sums up to 31.

5.1.3 Facial expressions

Facial expressions are very important cues in human perception [22], accounting for personality traits [5], as well as cognitive and psychological states [16]. Today, real-time facial analysis can be addressed with tools such as the Computer Expression Recognition Toolbox (CERT) [26]. Though these technologies have been developed for videos without speech, research has also shown that applied to conversational social video, automatic facial expression cues derived from CERT can be used to predict vlogger personality [9].

We followed the approach used in [9] to aggregate the frame-by-frame outputs of CERT which include seven expressions of emotion, a neutral expression, and one smile signal. First, we converted frame-by-frame estimates to a binary segmentation that divides each expression signal into active-inactive regions, and then, we computed features such as the active time, the duration of active time and number of active turns. Active-inactive segmentations were obtained with two approaches: based on thresholding the raw output (THR) as well as using a a two-state (active/inactive) Hidden Markov model to detect the active states. Each method generates 27 facial expression cues.

5.2 Verbal Cues

Social psychology research has shown that the words people use in their daily interactions reflect information about people's psychological constructs and concerns [39]. Compared to videos, text is easier to process and can be automatically analyzed using tools such as the Linguistic Inquiry Word Count (LIWC) which categorizes words into 77 linguistic and paralinguistic categories that have been validated in psychometric terms. This tool has been previously applied to analyze essays and text blogs [1].

We explore the use of verbal content to predict mood through the analysis of manual transcriptions of vlogs. Each transcript was processed with LIWC to breakdown word category usage based on relative word occurrences (note that in LIWC, words can be assigned to more than one category at a time). Each LIWC category count is treated as verbal cue. This results in feature vectors of 81 dimensions.

5.3 Inferring Mood

To infer the mood in vlogs, we use SVM (Regression) with Gaussian kernel ($k(x, x') = exp(-\gamma||x - x'||^2)$), given that it manages data with many attributes, finds the optimal solution, and it has shown efficiency modelling complex real-world problems [10, 33, 25].

We also use Random Forest Regression given that it does not tend to overfit the data (OOB: out-of-bag samples to estimate the generalization error), it is fast to build (grows trees in parallel), it is robust to outliers, it can handle data from mixed types, and often performs automatic selection of features [12].

We train the supervised learners, one per mood (k={happy, excited, ...}) using single and multimodal cues, where the input vector contains the respective set of features (f). In the test phase, the outputs from the learners are thresholded (using the median value) to perform two-class classification per mood.

$$Mood_k^f(vlog_i) = \begin{cases} 1 & \text{if } y(vlog_i) \geq Median_k; \\ 0 & \text{if } y(vlog_i) < Median_k. \end{cases}$$

Where $mood_k^f$ means the label assigned to the $vlog_i$, tested with the mood classifier k (k={happy, excited, ...}) given the features f. The output of the classifier $y(vlog_i)$ is then

thresholded using the median value of the mood k. Later on, we estimate the significance of the accuracy (at 95% confidence level) using a two-tailed standard binomial significance test with $z = N(0, 1)$, i.e., mean=0 and standard deviation=1 [28] with respect to the baseline. The baseline per mood corresponds to majority class performance.

6. RESULTS AND DISCUSSION

In this section, we present the results for the automatic mood inference. We first present the results organized per cue modality, followed by a discussion about the best results obtained for each mood. Figures 2 and 3 summarize the performance per mood, using SVM and RF. In the figures, the blue line represents the majority class baseline performance (note that this is around but not exactly 50% due to several vlogs having the same median mood value); and the red line corresponds to performance that is statistically better than the baseline at 95% confidence interval.

6.1 Audio Nonverbal Cues (A)

For Audio features as single modaltity, the performance for 9 moods is not statistically better than the baseline. In Figure 3 (RF), we only observe significant performance improvement for Excited and Bored at 61.9% and 60.6% respectively.

6.2 Visual Nonverbal Cues (V)

The single visual channel includes gaze, posture, motion and multimodal (gaze and speaking) patterns, described in Section 5.1. From Figures 2 and 3, for Excited mood we observe that these cues perform significantly better than the baseline (63.3% and 65.3% for SVM and RF respectively). Which could be explained by the fact that highly excited vloggers exhibit high motion in their videos.

Moreover, we can observe from Figure 3 that Visual cues and RF can infer 3 additional moods including Disappointed (62.6%), Sad (59.6%) and Bored (61.0%). On one hand, higher changes in posture and motion capture properly the Disappointed mood, on the other hand, slow motion and slow pace of gaze and speaking patterns make accurate differences between Sad/Non-Sad and Bored/Non-Bored.

Facial expressions as single cue can infer Happy and Excited moods; both learners (SVM and RF) perform statistically better than the baseline (See Figures 2 and 3). For Happiness, we obtain 61% (SVM) and 62.4% (RF); perhaps the explained by the accurate detection of smiles from frontal faces in the video. For Excitement, we obtain 61.1% (SVM) and 60.3% (RF), possibly due to the accurate detection of basic expressions of joy and smiles. We also can observe significantly accuracy for Overall mood (58.4%) and Bored (61.4%) from Figure 3 (RF). Finally, Sad (60% SVM) in Figure 2, also performs statistically better than the baseline, reasonably explained by accurate detection of sad expressions (eyebrows and lip corner depressor).

6.3 Verbal Cues (L)

The word categories derived from the Verbal content, show significantly better performance than the baseline for the Overall mood (60.1% for SVM and 64.5% for RF, from Figures 2 and 3 respectively).

From Figure 3 we observe significantly accurate performance for Happy (61.3%), Disappointed (59.1%) and Sad (59.1%). Perhaps, is not surprising that Sad mood can be accurately detected if the verbal content reveal high percentages on word categories like *death* (e.g., die, alive, war), *sad* (e.g., unsuccessful, tragic, sad) and *quantifiers* (e.g., a lot, anymore, less).

From Figure 2, we also observe statistically significant performance for Angry (up to 64.6%). Intuitively, Angry mood could be strongly associated with spoken word categories like *Swear*, *Anger* and *Negative Emotions*.

6.4 Multimodal Cues

For Overall mood, Happy and Angry, although all features perform statistically significant, the best multimodal combination is with Verbal and Facial Expression Cues (**L+F**). As we can observe in Figure 3 (RF), Overall mood and Happy (68.98% and 64.0% respectively), and Angry mood (65.1% with SVM) in Figure 2.

From figures 2 and 3, we observe that the best multimodal combination using Audio, Visual and Facial Cues (**AVF**, i.e., only nonverbal cues), performs the best for Excited (67.2% with SVM and 68.3% with RF).

For Disappointed and Surprised, the best multimodal combination is using Verbal and Visual Cues (**L+V**). From Figure 3, RF preforms at 65.96% for Disappointed. For Surprised, only SVM performs significantly better than the baseline (64.0% from Figure 2).

For Sad, Relaxed and Bored, **All** the features (Audio, Visual, Facial and Verbal Cues) are needed to reach the best performance. As we can observe in Figure 2 (SVM) for Sad and Relaxed (64.9% and 66.0% respectively) and Figure 3 (RF) for Bored (64.1%).

6.5 Discussion of Overall Results

Table 1 shows the summary of best accuracy achieved per mood using any possible combination of modalities. Each mood is illustrated by a snapshot of a vlog that has one of the top scores for the specific mood. Furthermore, moods whose ICC > 0.5 and ≤ 0.5 are separated by a line. Moods are ordered according to their ICC reliability. Note that we only include results for which the performance is statistically better than the baseline. This shows that 2 moods could not be classified better than majority class (Stressed and Surprised, see empty entries in Table 1). For the other 9 mood labels, we see that SVM and RF split the number of times they perform the best (4 for SVM, 5 for RF), although the performance differences are not statistically significant. The overall mood task resulted in the highest performance (69%). Two important observations are the following. First, for all moods it was a combination of features (although not necessarily the same ones) what produced the best performance. Second, we do not observe any clear pattern between performance and reliability for moods with ICC > 0.5. This means that the reliable moods tend to produce similar performance than less reliable ones (which correspond to noisier tasks). That said, the results for the least reliable moods (Stressed, Surprised, Nervous, ICC ≤ 0.5) are largely not statistically significant.

In addition, we also present the computed Receiver operating characteristics (ROC), area under the curve (AUC). The AUC in binary classification, is equivalent to: "the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance" [17]. The greater the area, the better the performance of the classifier. In Table 1, we observe that the

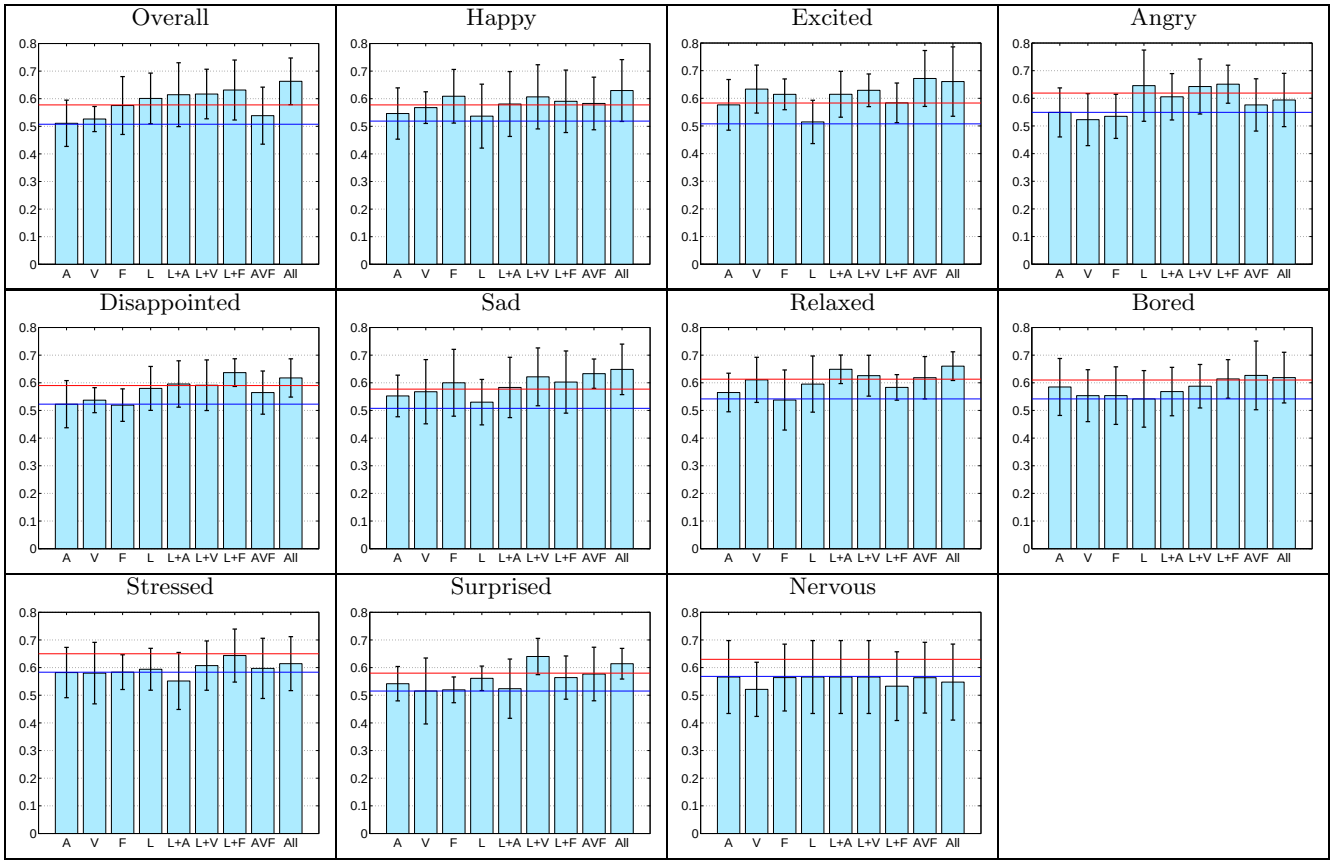

Figure 2: Mood Classification Accuracy comparison using SVM. Moods are ordered according to their ICC reliability value (see Section 4). A: Audio, V: Visual, F: Facial, L: Verbal, L+A: Verbal and Audio, L+V: Verbal and Visual, L+F: Verbal and Facial, AVF: Audio, Visual and Facial, All: All features. Blue line: Baseline method (Majority class), Red line: Significantly better than baseline at 95% confidence interval.

Overall Mood label also results in highest AUC (0.75), and that the most reliable moods seem to obtain only slightly higher AUC (0.69-0.75) compared to the rest (0.65-0.7).

As an example, Figure 4 shows the ROC curves from the AUC values using RF, the ROC curves are computed merging the 10 folds. We can observe that RF is promising classifier for Happy with AUC=0.70 (confidence interval (c.i.)=[0.63,0.76]), Excited with AUC=0.74 (c.i.=[0.68,0.80]), Disappointed with AUC=0.70 (c.i.=[0.63,0.76]) and Angry with AUC=0.69 (c.i.=[0.63,0.75]).

We conclude this section by discussing our findings in comparison with previous work:

- Although no direct comparison with text blogs is possible because of different number of moods and different data sources, we can point out the best overall performance, 63.5% using word sentiment orientation, verbs, adjectives, BoW and text statistics in [21].
- With respect to video, no direct comparison to previous work is possible for the same reasons, nevertheless we can cite the work by Morency et al., [36] with reported performance of up to 55.3% (F-measure) using multimodal features to discriminate between negative, neutral or positive product reviews. In our case, for the overall mood we obtain 69% accuracy on a binary task using a more diverse (in topics) and larger dataset.

7. CONCLUSIONS

In this paper we presented a systematic study of ubiquitous social video from verbal and nonverbal cues. Based on a set of crowdsourced mood categories, our work showed that, while the task is challenging, several of these categories can be recognized in the simplest binary classification setting with performance that is statistically better than a majority class baseline. The best performance was obtained for Overall mood and Excited (69% and 68% accuracy), which are categories that can be of great value in social video applications (e.g. related to sentiment analysis). We also found that mood categories that have low reliability in terms of annotation agreement result in non-significant performance improvement.

We showed that although multimodal features perform better than single channel features, not always all the available channels are needed to accurately discriminate mood in videos. We observed that the verbal content augmented the nonverbal information for many of the moods. Our findings revealed that to model mood is important to know the spoken categories appearing in a video, including categories related to *health* (e.g., hungover, pain), *swearing words*, *anger* (e.g., hate, annoyed), *anxiety* (e.g., worried, nervous), etc., in addition to the *positive* and *negative* emotion categories, that have shown improvement in mood inference from text blogs [33].

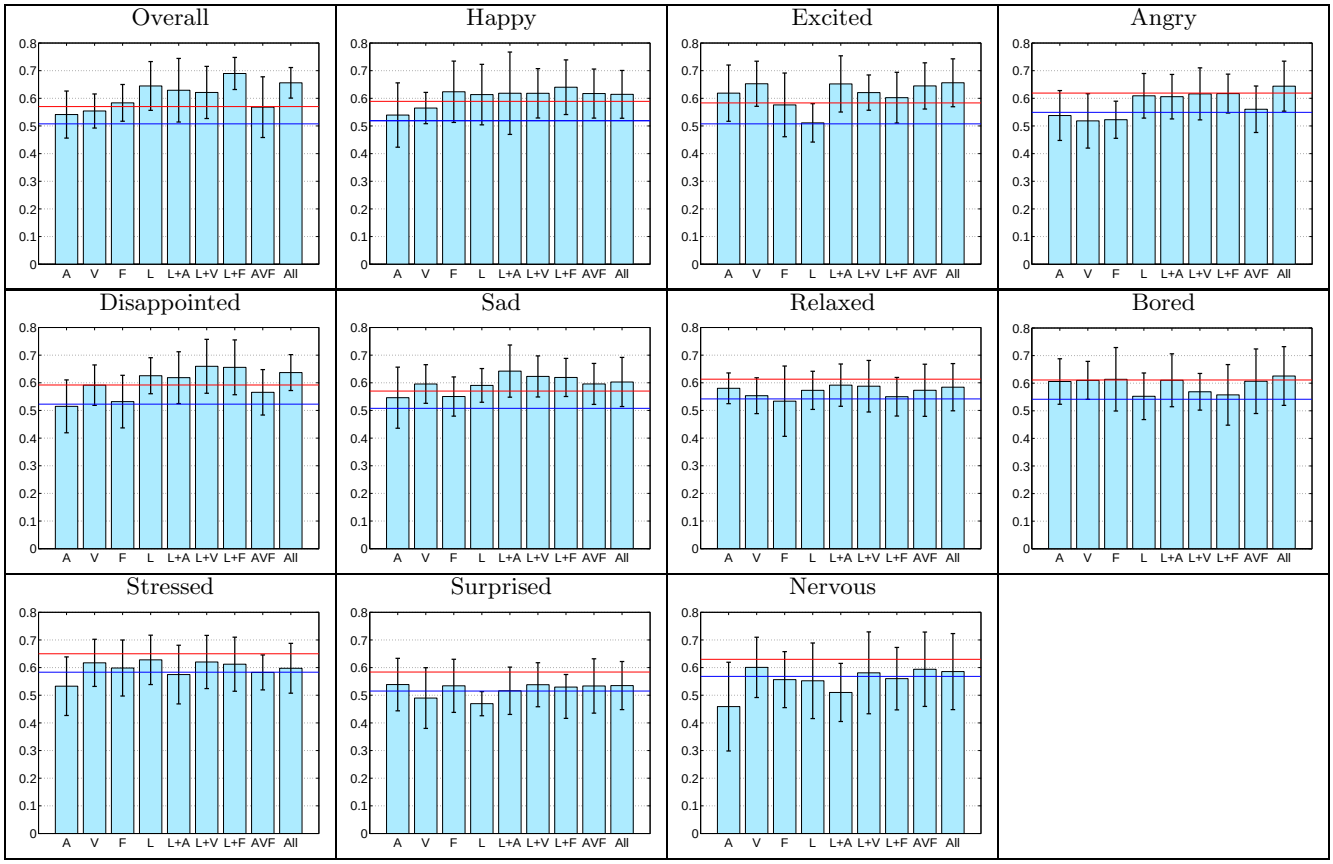

Figure 3: Mood Classification Accuracy comparison using RF. Moods are ordered according to their ICC reliability value (see Section 4). A: Audio, V: Visual, F: Facial, L: Verbal, L+A: Verbal and Audio, L+V: Verbal and Visual, L+F: Verbal and Facial, AVF: Audio, Visual and Facial, All: All features. Blue line: Baseline method (Majority class), Red line: Significantly better than baseline at 95% confidence interval.

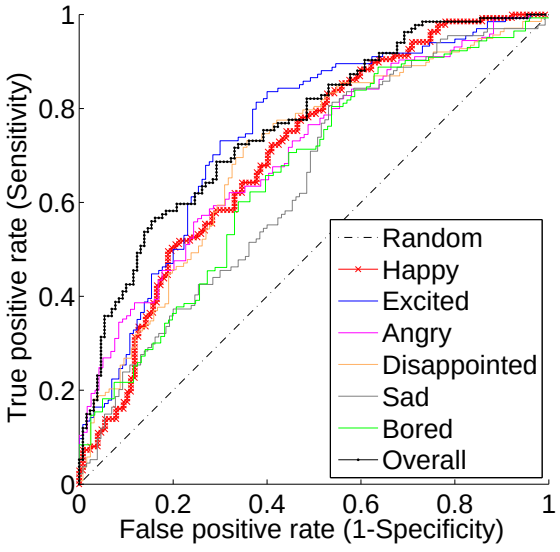

Figure 4: AUC from best performed moods using RF.

As part of future work, first we consider that a larger dataset would improve the training phase and thus might provide better inferences. Secondly, we plan to apply automatic speech recognition and replicate our experiments using partial verbal content (only most accurate recognized keywords), to explore the feasibility of a fully automatic system. Third, since in practice there could be several "active" moods per vlog, we could aggregate the outputs of each binary mood classifier and use posterior probabilities to keep the top 2-4 moods to characterize each vlog, following the evaluation of posterior probabilities of empathy proposed in [23]. Finally, the automatic inference of overall mood (positive or negative) could be embedded in a social video recommendation system prototype for a ubiquitous video service, or to infer the affective state of patients in online support groups.

8. ACKNOWLEDGMENTS.

This work was supported by the NISHA NTT-Idiap project and the SNSF National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM)2.

9. REFERENCES

[1] Liwc inc. http://www.liwc.net/index.php.

Mood	Baseline	SVM			RF		
		Features	Accuracy	AUC	Features	Accuracy	AUC
Overall	50.8	All	66.3	0.69	Verb + Facial	**69.0**	0.75
Happy	51.9	All	63.0	0.67	Verb + Facial	**64.0**	0.70
Excited	50.8	AVF	67.2	0.71	AVF	**68.3**	0.74
Angry	54.9	Verb + Facial	**65.1**	0.69	All	64.4	0.69
Disappointed	52.3	Verb + Facial	63.7	0.63	Verb + Visual	**66.0**	0.70
Sad	50.8	All	**64.9**	0.69	Verb + Audio	64.2	0.62
Relaxed	54.2	All	**66.0**	0.70	-	-	-
Bored	54.2	AVF	62.7	0.66	AVF	**64.1**	0.65
Stressed	58.3	-	-	-	-	-	-
Surprised	51.5	Verb + Visual	**64.0**	0.65	-	-	-
Nervous	56.8	-	-	-	-	-	-

Table 1: Best classification results per mood. Moods are ordered according to their ICC reliability. All non-empty entries are Statistically better than baseline at 95% confidence interval. In bold, we show the best classifier (SVM or RF). The horizontal line separates mood categories whose ICC above or below 0.5

[2] Merriam-webster online dictionary. http://www.merriam-webster.com/dictionary/mood.

[3] Okao vision - omron tech. http://www.omron.com.

[4] Oxford online dictionary. http://oxforddictionaries.com/definition/english/mood.

[5] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences. a meta-analysis. *Psychological Bulletin*, 111:256–274, 1992.

[6] J.-I. Biel and D. Gatica-Perez. The good, the bad, and the angry: Analyzing crowdsourced impressions of vloggers. In *Proc. of ICWSM*, 2012.

[7] J.-I. Biel and D. Gatica-Perez. The YouTube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 15(1):41–55, 2013.

[8] J.-I. Biel and G. Gatica-Perez. Vlogsense: Conversational behavior and social attention in youtube. *ACM Transactions on Multimedia Computing, Communications*, 7(1):33:1–33:21, 2011.

[9] J.-I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez. Facetube: Predicting personality from facial expressions of emotion in online conversational video. In *Proc. ICMI*, 2012.

[10] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[11] P. Boersma. Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345, 2002.

[12] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[13] M. De Choudhury, S. Counts, and M. Gamon. Not all moods are created equal! exploring human emotional states in social media. In *AAAI ICWSM*, 2012.

[14] L. Devillers and et al. In *Proc. LREC Int. Workshop on Emotion: Corpora for Research on Emotion and Affect*, 2010.

[15] P. Ekman and W. V. Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, 2003.

[16] P. Ekman and E. L. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[17] T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[18] R. Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[19] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, and D. P. Rosenwald. Social risk and depression: Evidence from manual and automatic facial expression analysis. In *FG*, 2013.

[20] S. A. Golder and M. W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.

[21] F. Keshtkar and D. Inkpen. Using sentiment orientation features for mood classification in blogs. In *NLP-KE*, 2009.

[22] M. L. Knapp and J. A. Hall. *Nonverbal Communication in Human Interaction*. Wadsworth, Cengage Learning, 2008.

[23] S. Kumano, K. Otsuka, D. Mikami, M. Matsuda, and J. Yamato. Understanding communicative emotions from collective external observations. In *Proc. Extended abstracts, CHI*, pages 2201–2206, 2012.

[24] C. M. Lee and S. S. Narayanan. Toward detecting emotions in spoken dialogs. *Speech and Audio Processing, IEEE Transactions on*, 13(2):293–303, 2005.

[25] G. Leshed and J. J. Kaye. Understanding how bloggers feel: recognizing affect in blog posts. In *Extended Abstracts, CHI'06*, 2006.

[26] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (cert). In *Proc. FG 2011*, 2011.

[27] G. C. Littlewort, M. S. Bartlett, and K. Lee. Faces of pain: automated measurement of spontaneousallfacial expressions of genuine and posed pain. In *Proc ICMI*, 2007.

[28] R. Lowry. *Concepts and applications of inferential statistics*. R. Lowry, 1998.

[29] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, S. Chew, and I. Matthews. Painful monitoring: Automatic pain monitoring using the unbc-mcmaster shoulder pain expression archive database. *Image and Vision Computing*, 30(3):197–205, 2012.

[30] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–501, 2007.

[31] D. McDuff, R. el Kaliouby, D. Demirdjian, and R. Picard. Predicting online media effectiveness based on smile responses gathered over the internet. In *FG*, 2013.

[32] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic. The semaine corpus of emotionally coloured character interactions. In *Proc. ICME*, 2010.

[33] G. Mishne. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*, page 19, 2005.

[34] G. Mishne and M. de Rijke. Capturing global mood levels using blog posts. In *AAAI 2006 Spring symposium on*

computational approaches to analysing weblogs, pages 145–152, 2006.

[35] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. Pulse of the nation: Us mood throughout the day inferred from twitter. http://www.ccs.neu.edu/home/amislove/twittermood/, 2010.

[36] L.-P. Morency, R. Mihalcea, and P. Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proc. ICMI*, 2011.

[37] T. Nguyen, D. Phung, B. Adams, T. Tran, and S. Venkatesh. Classification and pattern discovery of mood in weblogs. *Advances in Knowledge Discovery and Data Mining*, pages 283–290, 2010.

[38] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, 2008.

[39] J. Pennebaker, M. Francis, and R. Booth. *Linguistic Inquiry and Word Count: LIWC2001*. Mahwah, NJ: Erlbaum Publishers, 2001.

[40] J. Pennebaker and L. King. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312, 1999.

[41] K. R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1):227–256, 2003.

[42] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang. Emotion recognition based on joint visual and audio cues. In *Proc. ICPR*, volume 1, 2006.

[43] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics, 2008.

[44] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *Proc. FG*, 2011.