Happy and Agreeable? Multi-Label Classification of Impressions in Social Video

Gilberto Chávez-Martínez
Idiap Research Institute
Switzerland
gchavez@idiap.ch

Salvador Ruiz-Correa
Instituto Potosino de
Investigación Científica y
Tecnológica
Mexico
src@cmls.pw

Daniel Gatica-Perez
Idiap Research Institute
École Polytechnique Fédérale
de Lausanne
Switzerland
gatica@idiap.ch

ABSTRACT

The mobile and ubiquitous nature of conversational social video has placed video blogs among the most popular forms of online video. For this reason, there has been an increasing interest in conducting studies of human behavior from video blogs in affective and social computing. In this context, we consider the problem of mood and personality trait impression inference using verbal and nonverbal audio-visual features. Under a multi-label classification framework, we show that for both mood and personality trait binary label sets, not only the simultaneous inference of multiple labels is feasible, but also that classification accuracy increases moderately for several labels, compared to a single-label approach. The multi-label method we consider naturally exploits label correlations, which motivate our approach, and our results are consistent with models proposed in psychology to define human emotional states and personality. Our approach points to the automatic specification of co-occurring emotional states and personality, by inferring several labels at once, compared to single-label approaches. We also propose a new set of facial features, based on emotion valence from facial expressions, and analyze their suitability in the multi-label framework.

Author Keywords

Mood; Personality; Classification; Social Video; YouTube.

INTRODUCTION

In recent years, mobile devices have enabled social video to become one of the major trends in the social media landscape. This trend has also been propelled by the popularity of mobile platforms and applications that allow for ubiquitous access and sharing of video content, whether it is for social interaction and communication, or for entertainment. Mobile video services are sprawling, from the rise of Snapchat, an application widely used for video messaging among young users [11], to Periscope, a platform acquired by Twitter early this year that enables users to share and watch live video broadcasts from mobile devices [14]. In addition, there are long well-known services that are some of the most important nowadays. Such is the case of the 10-year-old YouTube platform, where, according to recent statistics,[1] the number of video views ascends to billions everyday (50% of which are generated on mobile devices), and every minute 300 hours of video are uploaded. The platform's ubiquity is confirmed by the fact that it is available in 61 languages, and approximately "60% of a creator's views come from outside their home country."

Today, the ubiquitous and mobile nature of video content has placed video blogs as one of the most popular video kinds, among a variety of video content available in social media. This has generated increasing interest in conducting studies from several perspectives; for instance, human behavior in video blogs has been studied from the computational view, for a variety of potential applications in affective and social computing. More concretely, there have been recent studies such as mood [23] and personality trait [4] inference from the perspective of *impressions*, i.e. judgments of mood or traits given by video viewers, as opposed to self-reported judgments. In fact, "the tendency to make a correspondent inference about personality based on behavior is both automatic and ubiquitous" [10].

To computationally model human traits, the usual practice has been to count on psychology constructions that attempt to define them and that are generally supported by observations. Human behavior is displayed through verbal and nonverbal audio-visual information, and from the computational perspective, this type of information has been used in inference tasks, e.g. [26, 23, 4].

Regarding human personality, the Big-Five (or Five-Factor) model is one of the most consistent and used models in the psychology field, and characterizes personality along five independent dimensions [19]. These dimensions have been obtained through factor-analysis of questionnaire data, and correspond to the *Extraversion*, *Agreeableness*, *Conscientiousness*, *Openness*, and *Emotional Stability* traits.

On the other hand, moods are defined as temporary emotional states, and their classification has also relied on dimen-

MUM '15, November 30-December 02, 2015, Linz, Austria
© 2015 ACM. ISBN 978-1-4503-3605-5/15/11…$15.00
DOI: http://dx.doi.org/10.1145/2836041.2836051

[1] Retrieved from https://www.youtube.com/yt/press/statistics.html, accessed August 2015.

sional models. The Circumplex model [22] is one of the most widespread. Supported by factor analysis evidence, the model suggests that emotional states can be represented in a two-dimensional space, embedded inside a circumference. The axes defining the space are *Activation*, or *Arousal*, and *Evaluation*, or *Valence* (also called *Pleasantness*). As an example, both *Sadness* and *Anger* have negative valence, but they have low and high arousal, respectively, whereas *Happiness* and *Sadness* have opposed valence.

The problem we consider here is simultaneous mood and personality trait impression classification, using audio-visual features extracted from video blog content. Research in psychology has shown that there are correlations between personality traits and emotional states, and other work has related the two in a common framework [30]. Other research works have studied the influence of facial expressions of emotion and other nonverbal expressive cues on the human inference of personality traits [10, 13].

To perform simultaneous inference of mood and trait categories, we rely on a multi-label classification scheme, which generalizes the usual classification problem, by allowing each observation to have more than one label assigned. Our problem can be stated as follows. We are given a training set of audio-visual features and labels, $\mathscr{D} = \{\{\mathbf{x}_1, \mathbf{y}_1\}, \ldots, \{\mathbf{x}_N, \mathbf{y}_N\}\}$, where $\mathbf{x}_n \in \mathbb{R}^p$ are the audio-visual feature vectors, and $\mathbf{y}_n \in \{0,1\}^q$ are the binary labels. Each pair $\{\mathbf{x}_n, \mathbf{y}_n\} \in \mathscr{D}$, is extracted from a specific video of the training set. Our goal is to use the data in $\mathscr{D}$ to train a multi-label classifier which takes advantage of the correlations between feature and label vectors, to predict the label vector $\mathbf{y}_*$ of a new video in a test set, on the basis of its features $\mathbf{x}_*$. We study three different cases: 1) $q = 8$ mood categories, 2) $q = 5$ personality trait categories, and 3) a particular combination of $q = 6$ mood and trait categories.

The audio-visual features include linguistic, acoustic, visual and facial cues, and the mood categories are *Overall Mood*, *Happy*, *Excited*, *Angry*, *Disappointed*, *Sad*, *Relaxed*, and *Bored*; whereas the personality categories come from the Big-Five model.

To the best of our knowledge, no previous work exists on simultaneous inference of mood and trait categories by multi-label classification in video blogs. In the context of the psychology works motivating our work, we argue that this modeling is beneficial as, under the Circumplex model, some moods are correlated and can co-occur, while others are mutually exclusive, and thus label correlations can convey important information. The multi-label framework can exploit this to provide a more integral specification of the video blogger's emotional state, by predicting all mood categories at once. Analogously, since human personality can be specified by the Big-Five model dimensions, the simultaneous inference of the five personality trait labels enables the inference of a more complete specification of the personality, rather than only particular traits. Finally, the simultaneous inference of mood and traits is also justified by works in social psychology [30, 13, 10].

As part of our work, we also propose a set of facial features aimed at capturing emotion valence, which are suitable for the multi-label setting. This is motivated by findings in psychology stating that compound facial expressions are more common than simple ones, and that these compound expressions are formed by single expressions having the same valence [15]. Our contributions are summarized as follows:

■ We motivate our study by performing a correlation analysis of mood and trait scores, and illustrate the consistency of the results with the models proposed in psychology to model human emotional states and personality.

■ We use a recently proposed, state-of-the-art multi-label classification method to classify mood and trait binary labels, and demonstrate that simultaneous inference can be achieved with statistically significant performance for both. We compare our approach with a single-label one [23], concluding that the accuracy can be moderately increased for some categories using the multi-label framework. We also evaluate the multi-label method itself using proper measures, and evaluate the predictive power of several feature combinations.

■ Under the same setting, we study the problem of inferring a combination of mood and trait labels, also showing improvement in the performance, in comparison with the mood-only and trait-only multi-label experiments.

■ We propose a set of facial features based on the valence of facial expressions, which in combination with the other audio-visual feature sets, shows one of the most accurate results in the multi-label framework.

We will first discuss related work in the next section, followed by a general overview of our approach. The next two sections describe the dataset and the features, respectively, the latter including our proposed facial features. After that we present the formulation of the multi-label classification method we used. This is followed by the section describing our experiments and the obtained results, which includes a comparison with the baseline methods, described in the same section. At the end of the paper we present our conclusions.

RELATED WORK

Emotional States and Personality

From the psychology perspective, research works have considered relationships between emotional states and personality, as well as between facial expressions and personality inference. For example, "nonverbal expressive cues of emotion influence the perception of personality traits" [10], particularly for negative emotions and strong expressive cues; there is a correlation between affective cues and the possession of personality traits [10]. In [13], the hypothesis that facial expressions of emotion, in addition to conveying emotional information, also affect the inference of interpersonal traits (such as dominance and affiliation), was successfully tested. For example, sadness and fear can convey low dominance. In [15], the authors study the effect of perceiving blended facial expressions. They conclude that these compound expressions are more familiar and spontaneous, and thus easier to recognize, especially if the expressions have a similar valence ("hedonic tone"). Regarding the direct relationships between emotional states and personality, there have been works, e.g. [30], in which correlations are found. The authors note that affect can be predicted from personality, and they study the links between these human aspects by making use of the Circumplex model.

Mood and Personality Inference

Recent work on automatic personality trait prediction has addressed, among others, the case of video data, such as group meetings using audio-visual features. In [1] it is found that the *Extraversion* trait can be predicted with promising accuracy for regression and classification tasks, and the authors study the effect of considering thin video slices or whole meetings. In[16], the personality prediction task is formulated as a regression problem on scores given by the meeting participants for two traits. The specific case of personality inference in video blog data has been addressed recently, using several approaches to predict scores given by annotators. In [2], a step-wise linear regression procedure on the personality scores showed significant results only for the Extraversion trait, using acoustic and visual (motion) features. In [4], acoustic and visual features were proposed and used in a regression task on each personality trait score, obtaining significant results for *Extraversion* and other traits. The use of facial features, computed from basic facial expression cues, was studied in [27], obtaining significant performance only for *Extraversion*. The work in [5] studied the predictive performance of verbal content features. Multivariate regression was used to model the five traits simultaneously in [9], obtaining promising results, although not better than the single-target regression approach. This could be due to the regression methods that were considered, which do not fully exploit targets correlations.

With respect to emotional states and mood inference, previous work has addressed the task of recognizing various emotional states and moods from several sources such as audio [25] and text [7]. The task of analyzing tweets was explored in [7], where, based on word analysis, over 200 terms that referred to a mood were found, and embedded into the Circumplex model space with consistency. Regarding the case of video data, the task of sentiment polarity analysis using text, visual, and audio features, has been addressed for YouTube movie [29] and product [20] review videos, concluding that the multimodal approach is more convenient than using single modal features. Mood binary classification was also addressed for the case of video blogs, using multimodal features [23]. In that work, a classification task was defined for each mood category, comparing SVM regression and random forests regression. The best accuracies were obtained using the latter, being the highest for *Excited* (68.3%) and *Overall Mood* (68.9%). They concluded that a multimodal approach was the most suitable, although the best combination of multimodal features varied among mood categories. This study was later extended in [24] to include a supervised mood ranking procedure, obtaining promising results for the task of video retrieval based on mood.

Multi-Label Classification (MLC)

In recent years, the multi-label classification framework has received significant attention [32]. It relates to the problem of multi-task learning, and commonly mentioned applications include text, video, and image categorization. A problem somehow related to ours is the assignment of multiple emotion labels to music, which has become a common benchmark in the evaluation of multi-label methods. This has been formally addressed in [28]. The basic idea behind all these methods is to encode label correlations, whether it is label pair-wise or of a higher order. The algorithms can be grouped into 2 broad categories: *problem transformation* and *algorithm adaptation* [32]. We give a succinct description of a couple of methods representative of each category, as presented in [32], which are among the most popular.

The *binary relevance* algorithm [6] is an example of problem transformation. It decomposes the problem into independent binary classification problems by considering the *relevance* of each observation to each label [32], and then constructing a binary training set for each particular label (this is called *cross-training*). *Classifier chains* [21] is another example of problem transformation methods. The idea is to transform the problem into a hierarchical chain of binary classification problems, where each one is built upon the previous prediction.

On the other hand, the algorithm adaptation category includes the *multi-label k-nearest neighbor* [31], which works by performing nearest-neighbor classification using a measure of neighbors for each label, then computing a MAP estimate for the posterior probability of an observation having the correct label, for all possible labels. The *multi-label decision tree* [8] is also in this category. It constructs a decision tree recursively based on a multi-label entropy measure [32].

A technique that is worth mentioning is utilizing the *label power set*. It consists of computing the power set of the labels, and performing multi-class classification on it, assigning one class to each element of the power set [32].

The method we consider in our experiments is based on error-correcting output codes [33]. It works by maximizing the correlation between the feature and the label vectors, constructing a codeword from these, and then it learns a model to predict the codeword for a new observation. The prediction of the new label vector is made in a codeword-decoding stage. We give further details later on.

OVERVIEW

In Figure 1 we illustrate our approach for the case of mood and trait classification. Given a set of video blogs and the corresponding mood and personality trait binary labels, we are interested in inferring these labels from the blogger's verbal and nonverbal behavior. We use a set of features that encode both types of information. Regarding the verbal content, we use a feature vector that has been extracted from the video's text transcript, and consists of word counts, categorized into linguistic categories by specific software. As to the nonverbal information, it includes acoustic (or audio) cues extracted from the video blog audio track, such as pitch and speaking rate. Additionally we use visual information, such as measurements of body motion, head turns, and gaze activity estimations. Finally, facial expression features are considered. They consist of several measures extracted from binary segmentations of the video, which indicate the presence or absence of basic

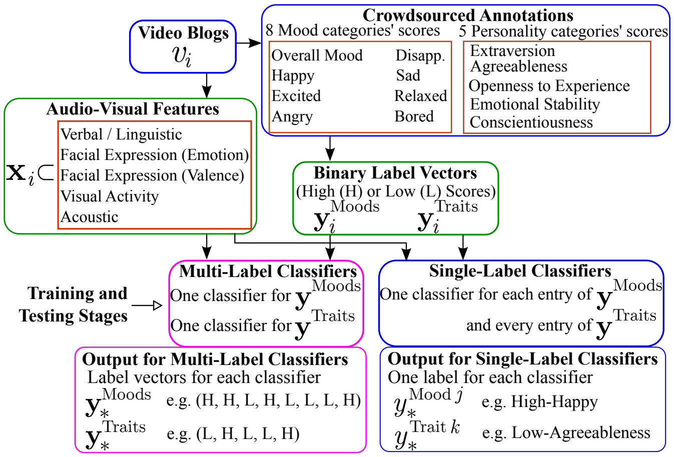
Video Blogs
v_i
Crowdsourced Annotations
8 Mood categories' scores
Overall Mood
Happy
Excited
Angry
Disapp.
Sad
Relaxed
Bored
5 Personality categories' scores
Extraversion
Agreeableness
Openness to Experience
Emotional Stability
Conscientiousness
Audio-Visual Features
$\mathbf{x}_i \subset$
Verbal / Linguistic
Facial Expression (Emotion)
Facial Expression (Valence)
Visual Activity
Acoustic
Binary Label Vectors
(High (H) or Low (L) Scores)
$\mathbf{y}_i^{\text{Moods}}$ $\mathbf{y}_i^{\text{Traits}}$
Training and Testing Stages
Multi-Label Classifiers
One classifier for $\mathbf{y}^{\text{Moods}}$
One classifier for $\mathbf{y}^{\text{Traits}}$
Single-Label Classifiers
One classifier for each entry of $\mathbf{y}^{\text{Moods}}$
and every entry of $\mathbf{y}^{\text{Traits}}$
Output for Multi-Label Classifiers
Label vectors for each classifier
$\mathbf{y}_*^{\text{Moods}}$ e.g. (H, H, L, H, L, L, L, H)
$\mathbf{y}_*^{\text{Traits}}$ e.g. (L, H, L, L, H)
Output for Single-Label Classifiers
One label for each classifier
$y_*^{\text{Mood } j}$ e.g. High-Happy
$y_*^{\text{Trait } k}$ e.g. Low-Agreeableness

Figure 1. Overview of our approach.

facial expressions of emotion, such as joy, surprise, or anger. Here we incorporate our proposed facial features, which can be seen as facial expression valence cues. We compute 3 binary signals for each video, indicating the presence or absence of *positive*, *negative*, or *neutral* valence (pleasantness) in the facial expression for each frame. We will give a detailed description of these features in the corresponding section.

We use all these features, as well as the video labels, to perform multi-label classification, which predicts more than one label for each observation. We train a classifier for the set of mood labels, as well as one for the set of trait labels, and a third one using a combination of mood and trait labels (not shown in Figure 1), to compare the accuracy rates with respect to the mood-only and trait-only cases. We use 10-fold cross-validation to obtain several performance measurements. This is done for each possible feature type combination (among linguistic, visual, facial, and acoustic). We also replicate the experiments from [23], to establish a comparison between this particular approach with ours, for the mood and trait label cases (although in [23] only the case of moods is studied). Besides, we evaluate the multi-label classification method itself, using performance measures which are adequate for the multi-label setting, and we perform a statistical test to assess the extent to which the simultaneous inference of the labels is possible. The details about the experiments and their discussion will be given in more depth.

DATASET

Raw Data

We used the dataset from [23], which includes 264 one-minute video blogs downloaded from YouTube using search keywords, each one depicting a unique person speaking while facing the camera. There is one video per blogger and, aside from the spoken language being English, there is no restriction regarding the content or recording settings of the videos. Approximately 70% of the video bloggers have been categorized as being below 24 years old [24], and 80% as being Caucasian. The gender distribution is 53% males and 47% females. The dataset also includes manual text transcriptions for each video, made from the audio channel.

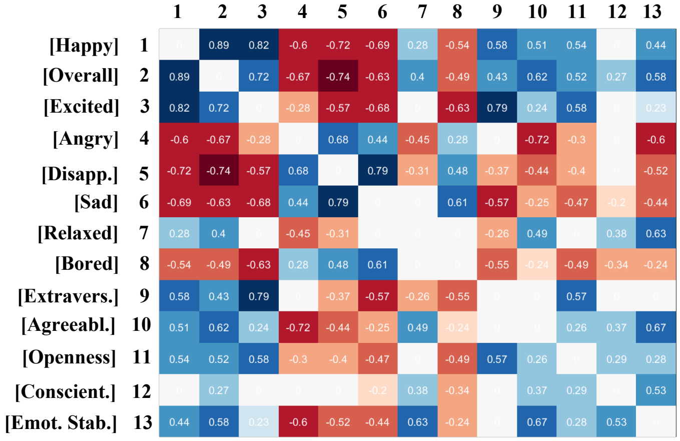

Figure 2. Correlation heatmap of mood and trait scores. Nonempty entries have $p < 10^{-3}$.

Each video has a set of 11 mood and 5 personality trait impression scores. These were obtained by averaging ordinal scores assigned by five annotators in a crowdsourcing platform, using a scale from 1 to 7, for each mood and trait. To obtain binary labels, we thresholded the scores using each category's median value; this represents a *high* or *low* score for each category. We use this binary label setting for several reasons. First, it provides reliable scores (they represent binary ordinal scores), since the annotators agreement can vary for some instances (see next subsection). In addition, while simple, the binary setting allows to potentially contrast our approach with sentiment analysis techniques widely used e.g. in text sources (where sentiment polarity is often used). Finally, it allows to generate basic comparisons across a variety of methods and categories. Note that it is not uncommon (e.g. in computer vision) to use other approaches (such as regression on continuous labels).

Impression Annotation Reliability

The reliability of the annotations has been studied in [3], using the intraclass correlation coefficient (ICC), which measures the inter-annotator agreement. The mood categories, ordered by their ICC, are *Happy* (0.76), *Overall Mood* (0.75), *Excited* (0.74), *Angry* (0.67), *Disappointed* (0.61), *Sad* (0.58), *Relaxed* (0.54), *Bored* (0.52), *Stressed* (0.50), *Surprised* (0.48), and *Nervous* (0.25); whereas the personality traits ordered by their ICC are *Extraversion* (0.77), *Agreeableness* (0.65), *Openness* (0.47), *Conscientiousness* (0.45), and *Emotional Stability* (0.42). As was previously noted in [23], high-arousal moods, such as *Happy* or *Angry*, have higher ICC, likely because they are manifested in a more obvious manner by bloggers. Similarly, the *Extraversion* trait is reported to be easier to judge at first sight in many scenarios [3].

Correlation Analysis

An important question regarding these categories is whether some of them have significant correlations, which is to be expected. We computed the Pearson correlation coefficient of the annotation scores, illustrated in Figure 2, where categories are ordered by decreasing ICC (entries 1–8 for moods, and 9–13 for traits). Looking first at the mood categories block (rows and columns 1–8), we can observe that there is important correlation information among several categories. For instance, *Happy* is positively correlated with *Excited* (0.82, with $p < 10^{-3}$), and negatively correlated with *Disappointed* (-0.72) and *Sad* (-0.69). In fact, in the context of the Circumplex model, another important observation can be made: moods with the same valence are positively correlated, whereas moods with opposed valence are negatively correlated. This suggests that indeed it is likely to see some moods co-occur, while others are mutually exclusive.

In the case of personality trait impressions (rows and columns 9–13), there are also some positive correlations among them, but no negative ones; this agrees with the formulation of the Big-Five model (although theoretically, personality traits should be uncorrelated [19], but in practice this is not always the case). There is a high correlation between *Extraversion* and *Openness* (.56), and also *Emotional Stability* is correlated with *Agreeableness* (.67) and *Conscientiousness* (.53).

Finally, we analyze the block corresponding to mood (entries 1–8) and trait (entries 9–13) categories. Interestingly, there are high positive and negative correlations between them. The most notorious are the following: *Happy* and *Overall Mood* are positively correlated with 4 of the 5 traits; *Extraversion* is positively correlated with *Excited* and negatively correlated with *Sad*; and finally *Angry* is negatively correlated with *Agreeableness* and *Emotional Stability*. We also observe that *Conscientiousness* does not present high correlations with moods. Research in psychology has investigated the relationships between emotional states and personality, e.g. in [30].

As a concluding remark, the co-occurrence of some categories, as well as the mutual exclusivity of some others, is the motivation to consider an inference methodology that can exploit these relationships. We hypothesize that incorporating this information –about the relationships between categories– into the mood and personality trait inference tasks, could improve the performance. We investigate this in the experiments' section.

FEATURES

In this section we briefly describe the features used in our experiments. We first describe the basic audio-visual features from [23, 24], and then introduce the second set of facial features we propose.

Basic Features

Mood and personality are displayed through verbal and nonverbal cues; the basic feature set includes audio-visual features extracted from the videos that encode these types of cues, and they comprise visual activity, facial expressions, and prosody information. These basic features have been shared by the authors of [23, 24]; they comprise the following:

Linguistic features. The Linguistic Inquiry and Word Count software[2] was used to process the video text transcriptions. Each word in the transcript was assigned to one or more linguistic categories, and the final count for all the categories is taken as the 81-dimensional linguistic feature vector of the video. It corresponds to the video blogger's verbal content.

Acoustic features. The set of acoustic (or audio) features comprises statistics about pitch, intensity, speaking rate, and formants and their bandwidth. First, the audio channel was processed using the Praat software[3], which gives frame-by-frame audio signals. Then, statistics such as the mean, variance, median, maximum, minimum, and entropy were computed, encoding the audio information for the whole video, resulting in a 98-dimensional vector.

Visual Activity features. Using a frontal face detector, looking activity and pose cues were extracted, under the assumption that face detections occur when the person is looking directly to the camera [4]. These cues include looking time, number of looking turns, proximity to the camera, and blogger's position relative to the frame's center. Additionally, body activity was encoded using weighted motion energy images, where the accumulated motion for each video is represented globally as a grayscale image, which can be seen as a motion "heatmap". From these images, the features correspond to measures such as mean, entropy and center of mass. Finally, the looking-while-speaking and looking-while-not-speaking times and their ratio were incorporated as features; these were computed from the video's looking and speaking segments. The final visual activity feature vector is 48-dimensional.

Basic facial features. Facial features were computed with the approach proposed in [27], which we summarize here. Using the Computer Expression Recognition Toolbox (CERT) [17], 9 continuous signals were extracted for each video, corresponding to the per-frame probabilities of 8 facial expressions of emotion (*Anger*, *Contempt*, *Disgust*, *Fear*, *Joy*, *Neutral*, *Sadness*, and *Surprise*), plus *Smile*. This basic emotion recognizer is trained on image features that measure the activation of facial action units (AUs, also computed by CERT) [17], so, for example, a nose wrinkle would be characteristic of disgust.

CERT signals encode the strength of the facial expressions, with estimates generated at each frame. To obtain the facial features for each video, the signals were processed to obtain binary segments (active/inactive regions) using a 2-state HMM, from which 36 features were extracted, including the proportion of time of active segments, their rate, their average duration, and the proportion of time of short active segments, for each facial expression.

It is important to mention that, despite the correspondence between some mood categories and CERT emotion categories, we cannot establish a single mapping between the two. In addition to the different nature of the entities being categorized (a person's predominant mood and a spontaneous facial expression), emotions and moods are not equivalent in psychological terms.

Proposed Facial Features

Compared to the previous facial features, the facial features we propose encode information about the *valence* of the emotions (rather than the emotions themselves), which corresponds to the evaluation (pleasantness) axis under the Circumplex model. Hence, facial expressions can have *positive* or *negative* valence. From the set of basic emotions, the ones with positive

[2] http://www.liwc.net/

[3] http://www.praat.org/

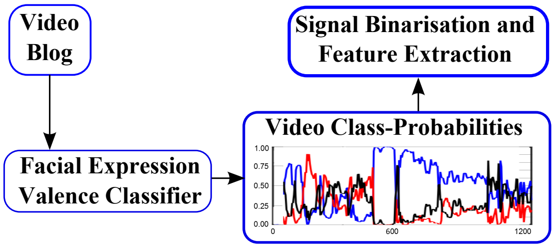
Video Blog
Signal Binarisation and Feature Extraction
Video Class-Probabilities
Facial Expression Valence Classifier
1.00
0.75
0.50
0.25
0.00
0
600
1200

Figure 3. Overview of the computation of facial valence features.

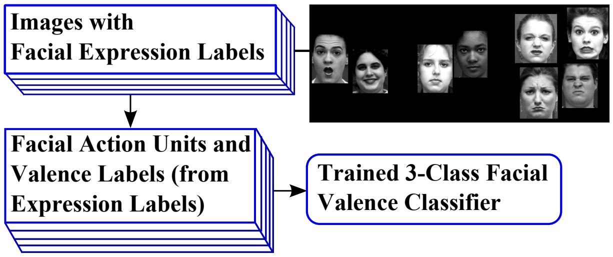
Images with Facial Expression Labels
Facial Action Units and Valence Labels (from Expression Labels)
Trained 3-Class Facial Valence Classifier

Figure 4. Construction of the facial expression valence classifier.

valence are *Joy* and *Surprise*, whereas the ones with negative valence are *Anger*, *Contempt*, *Disgust*, *Fear* and *Sadness* [22, 15]. Psychology research has shown that compound facial expressions of emotion are more frequent, and generally the basic expressions forming the compound ones have the same valence [15].

The overview of the construction of our features is illustrated in Figure 3. First, we construct a *facial valence* image classifier, which gives the 3 probability outputs for each valence class. Then, we compute the continuous probability signals for each video by applying the classifier on it. Finally, we obtain the features analogously to the case of basic emotion facial features, by segmenting the signals and extracting the measures. The details are given below.

Facial Expression Aggregation and Classification. We used a subset of the image dataset presented in [27], consisting of 521 video blog frames containing a blogger's face. Each image has the CERT estimate of 28 AUs, and also a unique label corresponding to the basic facial expression, obtained through crowdsourcing. These 521 images were chosen from an initial 1400-image dataset as having the majority of the annotators' high scores for a particular facial expression label. We augmented this dataset with the images from the Cohn-Kanade dataset [18], that consists of labeled facial expressions. To obtain the 28 AU measurements for this dataset, we processed the images with CERT. This completed a classification dataset, consisting of 1157 labeled facial expression images and their corresponding 28-dimensional AU feature vector.

Grouping the examples by their valence (see first paragraph at the beginning of this subsection), we obtained 355 examples from the *Positive* valence class, 222 examples from the *Negative* valence class, and 580 from the *Neutral* valence class. Classification experiments were performed using an implementation of the Robust Multi-Class Gaussian Process Classifier [12], which is robust to mislabeled examples and naturally provides class probability outputs. The 10-fold cross-validation evaluation for the balanced 3-class problem gave a mean classification rate of 87.9% (90.1% for *Positive*, 86.8% for *Negative*, and 87.6% for *Neutral*). An illustration of the construction of the classifier is shown in Figure 4.

High-Level Facial Feature Extraction. For each frame in a video, we computed the valence-class probabilities using the trained classifier, generating 3 continuous facial valence signals per video. From here, we performed the same procedure to obtain the features as in the case of basic emotion signals: we binarized the 3 valence signals using a 2-state HMM, to give active/inactive regions, and from these we computed, for the whole video, the proportion of time of active segments, their rate, their average duration, and the proportion of time of short active segments, for each valence binary signal.

MULTI-LABEL CLASSIFICATION

We now describe succinctly the multi-label learning method we used [33]. It is based on error-correcting output codes and canonical correlation analysis (CCA), and comprises two main stages: *encoding* and *decoding*. In the encoding phase, each observation of feature and label vectors $(\mathbf{x},\mathbf{y})$ is transformed into a codeword $\mathbf{z}$ that contains the label vector $\mathbf{y}$ as well as the projection vector $\mathbf{v} = (\mathbf{v}_1^{\mathrm{T}}\mathbf{y},\ldots,\mathbf{v}_d^{\mathrm{T}}\mathbf{y})$ from CCA, where each $\mathbf{v}_i$ is obtained by maximizing the correlation between projections of $\mathbf{x}$ and $\mathbf{y}$. Using a training set, models can be learned to predict new codewords. At the second stage, the class label is obtained by decoding the predicted codeword.

Let $\mathbf{x} \in \mathbb{R}^p$ be a feature vector, and $\mathbf{y} \in \{0,1\}^q$ the corresponding label vector. In canonical correlation analysis, the correlation between the projected *canonical variates* $\mathbf{u}^T\mathbf{x}$ and $\mathbf{v}^T\mathbf{y}$ is maximized, with $\mathbf{u} \in \mathbb{R}^p$ and $\mathbf{v} \in \mathbb{R}^q$. Under the general formulation of the problem, we have a $n \times p$ matrix $\mathbf{X}$ of features and a $n \times q$ matrix $\mathbf{Y}$ of label vectors, and vectors $\mathbf{u}$ and $\mathbf{v}$ to optimize. The correlation maximization problem is stated as a constrained optimization problem, which is reduced to the eigenproblem

$$\begin{aligned}\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X}\mathbf{u} &= \lambda\mathbf{X}^T\mathbf{X}\mathbf{u},\\ \mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{v} &= \lambda\mathbf{Y}^T\mathbf{Y}\mathbf{v}.\end{aligned} \tag{1}$$

The first d pairs $\{(\mathbf{u}_k,\mathbf{v}_k)\}_{k=1}^d$ of projection vectors that maximize the correlation can be computed.

We can now define a codeword in terms of the projection vectors $\mathbf{u}$, $\mathbf{v}$. In analogy to message transmission using codewords, the codeword $\mathbf{z}$ must have some kind of redundancy in order to better recover the original message $\mathbf{y}$, such that from the codeword prediction given $\mathbf{x}$, the label vector $\mathbf{y}$ can be recovered. For label vector $\mathbf{y} = (y_1,\ldots,y_q)^T$ and its canonical variates $\{\mathbf{v}_k^T\mathbf{y}\}_{k=1}^d$, the codeword is specified as

$$\mathbf{z} = (y_1,\,\ldots,\,y_q,\,\mathbf{v}_1^T\mathbf{y},\,\ldots,\,\mathbf{v}_d^T\mathbf{y})^T. \tag{2}$$

For codeword prediction, a set of q classifiers $\{\hat{c}_1,\ldots,\hat{c}_q\}$, and d regression models $\{\hat{r}_1,\ldots,\hat{r}_q\}$, based on random forests, are learned from training examples, to predict the q labels and the d variates, respectively. Then, at the decoding stage, for a test case $\mathbf{x}_*$, the classification and regression models provide a predictive joint distribution for the label vector, through a Bernoulli PDF, for each label y_j,

$$\phi_j(y_j) = \hat{c}_j(\mathbf{x}_*)^{y_j}(1-\hat{c}_j(\mathbf{x}_*))^{1-y_j}, \tag{3}$$

and a Gaussian PDF, for each canonical variate $\mathbf{v}_k^T\mathbf{y}$,

$$\psi_k(\mathbf{y}) \propto \exp - \frac{(\mathbf{v}_k^T\mathbf{y} - \hat{r}_k(\mathbf{x}_*))^2}{2\hat{\sigma}_k^2}. \tag{4}$$

The joint log probability for the label vector $\mathbf{y}$ given $\mathbf{x}_*$ can then be written as

$$\log P(\mathbf{y}|\mathbf{x}_*) = -\log Z + \sum_{k=1}^{d} \log \psi_k(\mathbf{y}) + \sum_{j=1}^{q} \log \phi_j(y_j), \tag{5}$$

where Z is the normalizing constant. Exact inference has an inconvenient time complexity, so the implementation uses a mean-field approximation $P(\mathbf{y}) \approx Q(\mathbf{y}) = \prod_{j=1}^{q} Q_j(y_j)$, which is found by minimizing the Kullback-Leibler divergence between Q and P, using a particular fixed-point equation for each factor Q_j.

Since the output of the algorithm is given in terms of probabilities, in a formal setting these could be used to get a more quantitative prediction than using binary labels. This is of importance for the inference problem we are considering.

EXPERIMENTS AND DISCUSSION

We conducted three multi-label classification (***MLC***) experiments. The first one corresponds to mood inference, so we incorporated the 8 mood labels (the ones with ICC greater than 0.5) into this experiment. The second one corresponds to personality trait inference; therefore, we used all 5 trait labels. In the final experiment, we combined 4 mood and 2 trait labels into the multi-label problem. In addition, we conducted single-label regression and binary classification experiments, in order to compare our multi-label classification results with those obtained with these approaches. In our experiments, we tested all possible combinations of feature types, among linguistic, acoustic, visual, and facial, for both the standard features and the proposed ones. We begin by summarizing the baseline methods and the performance evaluation procedures in the first two subsections. Next, we will present and discuss the results for mood, trait, and joint mood-trait experiments.

Baseline Methods

Random forest regression. For the regression task, we replicated the experiments setup from previous work [23], to be able to make a direct comparison between their approach and ours. Random forest regression (***RFR***) was reported to give the best rates for mood labels in [23], so we used this method as the first baseline method. We performed a regression task on every target score independently, and, as in [23], we assigned a binary label to each prediction, based on whether the predicted score was above or below the targets median, which is the same procedure we used to obtain the binary labels of the categories.

Random forest classification. Additionally, we performed binary random forest classification (***RFC***), where the binary labels were assigned as mentioned. This allows to establish a second, classification-based baseline method. Also in this case, a model was trained and used for prediction, for each category independently.

Evaluation measure. We used the *accuracy per label*, which is the usual classification accuracy, to compare the baseline methods mentioned above against the multi-label classification approach, that is, the evaluation measure is computed for each label category regardless of the number of labels that were simultaneously predicted with the multi-label method.

MLC Performance Evaluation

We performed multi-label classification (MLC) using the method described previously. In this case, more than one category can be learned and predicted simultaneously, so proper evaluation measures are needed (in addition to evaluating each category independently using the evaluation measure mentioned above); we used two of such evaluation measures. We also performed a statistical test on each label rate, using the majority class percentages.

Macro-average. The average of the label-wise rates can be used as a multi-label evaluation measure; it is referred to as the *macro-average* label-based metric in the multi-label classification paradigm [32]. As an example, let us consider a set of 8 labels predicted by the multi-label approach for all test cases. First, we compute the usual accuracy per label, which gives us 8 rates. Then, averaging these 8 rates, we can obtain the macro-average measure.

Exact-accuracy. The *exact-accuracy* (or *exact-rate*) measures the proportion of inferred label vectors that are equal to the ground truth label vectors. Following the previous example, it is the proportion of test cases where, for each example, all 8 predicted labels are equal to the ground truth label vector. Under this measure, high rates are difficult to achieve in general, especially if the number of labels is large [33, 32].

Majority class baseline and statistical testing. To statistically assess the predictive power of a feature combination with the multi-label method, and to evaluate the extent to which the feature set can infer all labels, we performed a two-tailed binomial test for each label rate. This is used to test whether the proportion of correct predictions per label is (statistically) significantly higher than the proportion given by the majority class percentage for that label (which is roughly 50%). To perform this test we compare the accuracy per label against the majority class percentage.

To train the classifiers for our experiments and evaluate their performance we used 10-fold cross-validation. Since we tested all possible feature combinations, we computed the evaluation measures for each method and each feature combination.

Mood Classification

We commence by describing the experiments regarding the classification of mood categories. We first compare the results between the single-label and the multi-label methods, and then we give the details about the evaluation of the multi-label method.

	Overall Mood		*Happy*		*Excited*		*Angry*		*Disappointed*		*Sad*		*Relaxed*		*Bored*	
RFR	67.1	*LF*	63.5	*LFV*	**68.5**	*FAV*	63.9	*LF*	66.0	*LPV*	62.8	*LF*	61.0	*LPAV*	**63.0**	*FA*
RFC	67.0	*LF*	62.6	*LFV*	66.2	*FAV*	65.5	*LF*	64.0	*LPV*	61.1	*LF*	65.4	*LPA*	61.9	*LFA*
MLC	**67.8**	*LFA*	**66.0**	*LFV*	68.0	*LFAV*	**69.6**	*LF*	**67.3**	*LV*	**65.4**	*LFV*	**66.4**	*LPA*	62.5	*FV*

	Extraversion		*Agreeableness*		*Openness*		*Conscient.*		*Emotional Stab.*	
RFR	**72.6**	*FAV*	63.6	*LF*	**64.4**	*LFV*	**66.0**	*LPA*	65.5	*LF*
RFC	71.3	*FA*	69.7	*LF*	61.0	*LF*	63.2	*LF*	66.0	*L*
MLC	71.7	*FA*	**71.0**	*LP*	63.4	*LFV*	63.4	*LF*	**66.3**	*LPV*

Table 1. Best classification results per label (accuracy per label). Top: Mood labels. Bottom: Trait labels. Best in bold. Possible feature sets are *linguistic L*, *basic facial emotion F*, *proposed facial valence P*, *acoustic A*, and *visual V*.

Comparison of Results Per Label

To perform the comparison between our multi-label approach with the single-label baseline methods, we utilize the usual accuracy per label as mentioned in the previous subsection. As we mentioned, we computed the classification rates for all feature type combinations, among *linguistic* ***L***, *basic facial emotion* ***F*** or *proposed facial valence* ***P***, *acoustic* ***A***, and *visual* ***V***. We report the feature set that gave the best accuracy per label in Table 1 (Top), including for the multi-label method. This means that a particular combination of features in the multi-label case did not necessarily improve the accuracy for all the other labels, although it did for some of them.

We observe that the multi-label rates surpass the single-label ones for 6 of the 8 mood labels. The improvement is greater for *Angry* (increment of 4.1% over the second best), *Sad* (2.6%), *Happy* (2.5%), and *Disappointed* (1.5%). These moods show some of the strongest correlations in Figure 2, and, under the Circumplex model, have opposed pleasantness (e.g. *Happy* and *Sad* or *Anger*) and opposed arousal (e.g. *Angry* and *Sad*). The results suggest that, looking at the results per label individually, label correlations can moderately boost the accuracy for some labels under the multi-label approach. Below we investigate whether a particular feature combination can give significant rates for *all* labels at once.

The results in Table 1 (Top) also provide information about the suitability of the features for each mood label, since there is a recurrent presence of some features types for several moods. As we can see, not all feature types are needed to achieve the best accuracy for each mood, a fact that was already stated in previous work [23].

Multi-Label Evaluation

We now turn to the evaluation of the multi-label method itself, with the *macro-average* and *exact-accuracy* as evaluation measures. The best macro-average was 64.5%, achieved using our proposed features combined with linguistic and acoustic ones, *LPA*. Furthermore, the best exact-rate, 19.5%, was achieved using the same features *LPA* (we recall the fact that the exact-rate is a difficult measure). This is of particular interest, since, looking at the results *label-wise* (as seen in Table 1 (Top)), our proposed facial valence features *P* were not better in general than the basic ones *F*, however, under the multi-label evaluation measures, they perform better than the basic facial features. This suggests that our features could be encoding global facial expression information, allowing them to be suitable for the inference of several mood categories. In Table 2 (Top) we summarize some of the best results for both multi-label evaluation measures (note that the results from Table 1 are obtained using a different evaluation measure, so no comparison can be made between results from the two tables).

Macro-Average	**64.5**	64.3	64.3	63.3
Exact-Accuracy	**19.5**	16.5	15.1	17.9
Features	*LPA**	*LFA*	*LFV*	*LA**

Macro-Average	**65.8**	65.4	65.3	65.2
Exact-Accuracy	16.3	14.9	16.6	**17.8**
Features	*LF*	*LPV**	*LFV**	*LFA*

Table 2. Best multi-label classification results. Top: Mood labels. Bottom: Trait labels. * = Simultaneous stat. significance for all labels w.r.t. majority class classifier rates (significance level of 0.05). Best in bold.

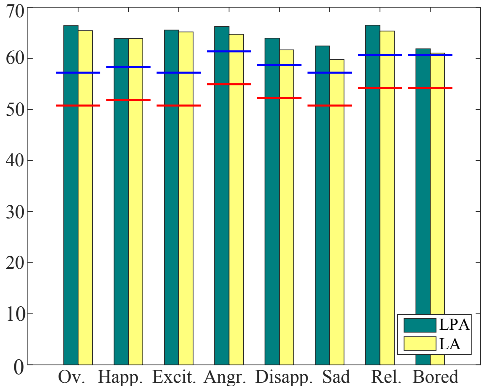

Figure 5. Multi-label mood rates for two feature sets that achieved statistical significance (at level 0.05) for all labels. Red line: majority class percentage; blue line: statistical significance threshold. Possible feature sets are *linguistic L*, *basic facial emotion F*, *proposed facial valence P*, *acoustic A*, and *visual V*.

As we mentioned, we also performed a two-tailed binomial test, to examine whether the proportion of correct predictions per label category is significantly higher than the majority class classifier rates for that category (which are around 50%); we seek to obtain a feature set that has significant rates for all labels. We found that, at a 0.05 significance level, the feature set *LPA* (which in fact performed best under both multi-label performance measures) achieved statistically significant performance simultaneously for all labels. The respective rates per label are shown in Figure 5, where we mark in red the majority class percentage, and in blue the significance threshold value. The categories for which the rate was higher than their single-label counterparts are *Happy*, *Angry*, and *Relaxed* (3 out of 8). The other combination of features that accomplished this was *LA*, with rates per label shown also in Figure 5. These results show that the multi-label approach indeed enables the simultaneous inference of the mood categories with significant results, for particular feature combinations.

We argue that the good performance of our facial-valence features might be the result of encoding global emotion information through valence (which, besides arousal, is one of the fundamental components of mood under the Circumplex model), and thus can be used to predict jointly a wider range of mood categories, rather than individually.

Personality Trait Classification

We replicated the experiments and evaluation from the previous subsection, using all 5 personality trait labels. As before, we first discuss the comparison between the single-label and multi-label results, and then we discuss the evaluation of the multi-label method.

Comparison of Results Per Label

In Table 1 (Bottom) we show the best results per label and feature combinations (*linguistic* ***L***, *basic facial emotion* ***F*** or *proposed facial valence* ***P***, *acoustic* ***A***, and *visual* ***V***), for regression (***RFR***), single-label (***RFC***) and multi-label (***MLC***) classification. First, we will discuss briefly the single-label results. We observe that the *Extraversion* and *Agreeableness* traits are the most predictable; this is consistent with other works using regression [4, 5, 9] and classification [1]. With respect to feature combinations, the linguistic features *L* combined provide the best results, except for *Extraversion*, which improves with the use of acoustic and facial *FA*. For each category independently, we can obtain significant accuracy, using a suitable feature combination, as in the case of mood labels.

Regarding the multi-label results label-wise, in contrast to the moods case, only 2 of 5 categories (*Agreeableness* and *Emotional Stability*) had a slight improvement, in comparison with the single-label methods. Similar behavior was also noted in recent work using multivariate regression [9] on personality scores. This is in spite of the presence of correlation between some traits, as shown in Figure 2 (although all of these correlations are positive, thus they could be encoding the dependencies differently).

The results, however, are in agreement with the definition of the Big-Five model, in which trait dimensions are uncorrelated. This means that in principle we might not expect one trait to provide very significant information about another, as opposed to mood categories and the Circumplex model.

Multi-Label Evaluation

With respect to the multi-label evaluation using the macro-average and exact-rate, we summarize the best results in Table 2 (Bottom). The best exact-rate was 17.8%, achieved using *LFA*, and the best macro-average was 65.8%, using *LF*. We can observe that, although the number of mood labels we considered in the previous experiment is higher than the number of trait labels, the exact-rate in the moods' case was slightly higher (as we stated, a larger number of labels represents a more difficult problem). This could be explained by the richer dependencies present among mood categories, in contrast with the trait categories; although in essence these are two different inference problems.

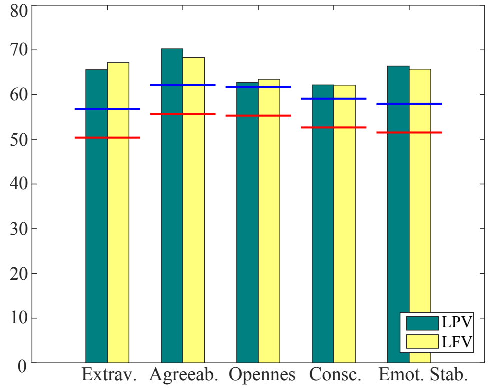

Figure 6. Multi-label trait rates for two feature sets that achieved statistical significance (at level 0.05) for all labels. Red line: majority class percentage; blue line: statistical significance threshold.

In Table 2 (Bottom) we also present the second-best and third-best macro-averages, achieved, respectively, using *LPV* (which contains the proposed features *P*) and *LFV*. We also performed a statistical significance test to assess the classification rates, in comparison with the majority class percentage value. These two feature sets (*LPV* and *LFV*) also achieved simultaneous statistically significant performance for all labels; we show their rates in Figure 6. The rates that were higher than the corresponding single-label ones were *Agreeableness* and *Emotional Stability*, for features *LPV*. The overall results suggest that the simultaneous inference of all trait labels is feasible under the multi-label setting, with promising rates for particular combinations of features.

As in the case of mood labels, we observe that, although our proposed features do not improve greatly the classification accuracy label-wise, they manage to obtain good results globally, under the multi-label approach. We recall, however, the fact that some trait categories have low ICC, which can affect the overall results presented in this subsection.

Combining Trait and Mood Labels

The task of finding correlations between emotional states and personality traits has been studied in the psychology field [30, 13, 10]. For example, *Agreeableness* correlates positively with positive affective states. In this context, an interesting question arises: would mood and trait impressions influence each other in the multi-label classification framework? We considered an experiment using 4 moods that have high correlation, *Happy*, *Excited*, *Angry*, and *Sad*, and also 2 traits, *Extraversion* and *Agreeableness*, the ones with highest ICC. For this experiment, we make a comparison with the mood-only and trait-only multi-label results from the previous subsections.

The best results per label are shown in Table 3. Compared to the respective multi-label results in Table 1 (mood-only and trait-only experiments), there is an observable improvement for *Excited* and *Angry*, as well as for *Extraversion* and *Agreeableness* (all these in bold), which suggests that inference benefits from both mood and trait labels' information. These underlying dependencies concur with the findings in psychology we have mentioned, about the existing relationships between personality traits and emotion or well-being [30].

Happy		*Excited*		*Angry*		*Sad*		*Extraversion*		*Agreeableness*	
63.7	*LFV*	**69.3**	*FAV*	**69.8**	*LF*	63.4	*LPV*	**72.3**	*FAV*	**71.8**	*LF*

Table 3. Best multi-label classification results per label for some moods and traits. Better than results from Table 1 (mood-only and trait-only) in bold.

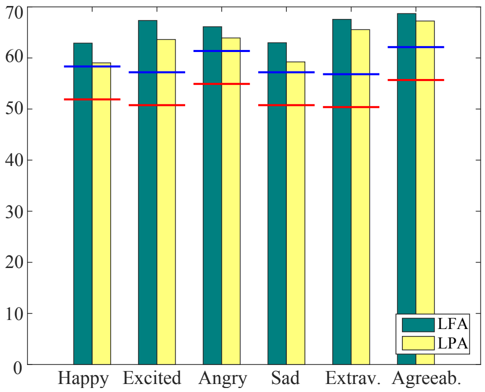

Figure 7. Multi-label rates for 4 mood and 2 trait labels, for two feature sets that achieved statistical significance (at level 0.05) for all labels. Red line: majority class percentage; blue line: statistical significance threshold.

With respect to the multi-label evaluation measures, the highest macro-average 65.9% and exact-rate 23.7% were both obtained using the feature set *LFA*. The label performances in this case were all statistically significant as well; we show them in Figure 7.

The second-best exact-accuracy (22.7%) was achieved using the feature set *LPA*, which contains our proposed features. Furthermore, it had statistically significant rates for all labels as well (using the same statistical test as before); we include its label rates in Figure 7 for comparison. For this experiment there were several feature combinations that accomplished this. Hence, also in this case our proposed features show some of the top multi-label results. Additionally, in comparison with the exact-rates from previous subsections, we observe an increase in this evaluation measure for this experiment. All these results overall point to the benefit of simultaneously inferring personality traits and emotional states for our problem, although optimal label combinations must be investigated in more depth.

Limitations

Experiments show that there is no feature set that gives the best classification rate for all categories at once. Some categories require a particular combination of features that is not necessarily the best for other categories. This can be inconvenient since one of our purposes has been to infer sets of labels simultaneously with acceptable accuracy. Perhaps a feature selection step can be used to alleviate this. On the other hand, we have observed that the classification rates can vary depending on the categories considered in the multi-label learning procedure. Investigating optimal label combinations has not been addressed in depth. Although not formally evaluated, the computational time of the selected multi-label algorithm can represent a disadvantage. A comparison with other recent multi-label learning methods is necessary, including accuracy and computational times of these and the baseline single-label methods. Finally, at the data level, the binarization of the categories' scores can signify loss of information; more complex label spaces can be considered in future work, including appropriate methods to handle them (for example, regression).

CONCLUSIONS

In the context of ubiquitous social video, we have considered the problem of simultaneous inference of both mood and personality impressions using video blog data and multimodal features, in a multi-label classification framework. We showed that the classification accuracy per label increased slightly for 6 out of 8 mood categories, in comparison with a recently proposed single-label approach, and that both mood and trait labels can be jointly predicted with statistical significance. We performed a proper evaluation of the multi-label classification method, to be able to establish its accuracy and to assess the suitability of several feature combinations. We also showed that a combination of some mood and trait categories can be inferred, showing an additional increase in the accuracy. Finally, we proposed a set of facial features based on emotion valence that, under the multi-label framework, gave some of the top results for both mood and trait labels, despite the fact that label-wise they did not show considerable improvement, in comparison with facial features proposed in recent works.

Future work will study the effects of certain trade-off measures arising in our setting, such as precision and recall in mood or trait retrieval tasks. Additional research includes considering label probabilities, as an alternative to performing regression tasks on continuous target scores. Also, future research directions include a deeper study of mood and trait category relationships in the multi-label setting, and studying the suitability of other multi-target learning methods that rely on exploiting label correlations, including regression for continuous, or even ordinal, targets. We chose to conduct our experiments using multi-label classification based on canonical correlation analysis on feature and label vectors; however, alternative state-of-the-art tools such as structural SVMs or transfer learning algorithms could also be considered to achieve our goals. A comparative study among these techniques is the topic of a future paper.

Acknowledgements: This work was supported by the SNSF UBImpressed project and the NISHA project. We thank Dairazalia Sanchez-Cortes (Idiap) and Lucía Teijeiro-Mosquera (U. Vigo) for sharing data, and Kazuhiro Otsuka and Shiro Kumano (NTT) for discussions.

REFERENCES

1. Oya Aran and Daniel Gatica-Perez. 2013. One of a Kind: Inferring Personality Impressions in Group Meetings. In *Proc. ACM Int. Conf. on Multimodal Interaction (ICMI)*.

2. Joan-Isaac Biel, Oya Aran, and Daniel Gatica-Perez. 2011. You Are Known by How You Vlog: Personality Impressions and Nonverbal Behaviour in YouTube. In

Proc. AAAI Int. Conf. on Weblogs and Social Media (ICWSM).

3. Joan-Isaac Biel and Daniel Gatica-Perez. 2012. The Good, the Bad, and the Angry: Analyzing Crowdsourced Impressions of Vloggers. In *Proc. AAAI Int. Conf. on Weblogs and Social Media (ICWSM)*.
4. Joan-Isaac Biel and Daniel Gatica-Perez. 2013. The YouTube Lens: Crowdsourced Personality Impressions and Audiovisual Analysis of Vlogs. *IEEE Trans. on Multimedia* 15, 1 (January 2013), 41–55.
5. Joan-Isaac Biel, Vagia Tsiminaki, John Dines, and Daniel Gatica-Perez. 2013. Hi YouTube! Personality Impressions and Verbal Content in Social Video. In *Proc. ACM Int. Conf. on Multimodal Interaction (ICMI)*.
6. Matthew R. Botuell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. 2004. Learning Multi-Label Scene Classification. In *Pattern Recognition*, Vol. 37. 1757–1771.
7. Munmun De Choudhury, Scott Counts, and Michael Gamon. 2012. Not All Moods are Created Equal! Exploring Human Emotional States in Social Media. In *Proc. AAAI Int. Conf. on Weblogs and Social Media (ICWSM)*.
8. Amanda Clare and Ross D. King. 2001. Knowledge Discovery in Multi-Label Phenotype Data. In *Lecture Notes in Computer Science*, L. De Raedt and A. Siebes (Eds.). Springer.
9. Golnoosh Farnadi, Shanu Sushmita, Geetha Sitaraman, Nhat Ton, Martine De Cock, and Sergio Davalos. 2014. A Multivariate Regression Approach to Personality Impression Recognition of Vloggers. In *Proc. ACM Int. Conf. on Multimedia (ACM-MM), Workshop on Computational Personality Recognition (WCPR)*.
10. Judith A. Hall, Sarah D. Gunnery, and Susan A. Andrzejewski. 2011. Nonverbal Emotion Displays, Communication Modality, and the Judgment of Personality. *Journal of Research in Personality* 45, 1 (2011), 77–83.
11. Ellis Hamburger. 2014. Real Talk: The New Snapchat Brilliantly Mixes Video and Texting. *The Verge* (May 2014). http://www.theverge.com/2014/5/1/5670260/real-talk-the-new-snapchat-makes-texting-fun-again-video-calls, consulted August 2015.
12. Daniel Hernández-Lobato, José Miguel Hernández-Lobato, and Pierre Dupont. 2011. Robust Multi-Class Gaussian Process Classification. In *Proc. Neural Information Processing Systems (NIPS)*.
13. Brian Knutson. 1996. Facial Expressions of Emotion Influence Interpersonal Trait Inferences. *Journal of Nonverbal Behavior* 20 (1996), 165–182.
14. Yoree Koh and Evelyn M. Rusli. 2015. Twitter Acquires Live-Video Streaming Startup Periscope. *The Wall Street Journal* (March 2015). http://www.wsj.com/articles/twitter-acquires-live-video-streaming-startup-periscope-1425938498, consulted August 2015.
15. Debi LaPlante and Nalini Ambady. 2000. Multiple Messages: Facial Recognition Advantage for Compound Expressions. *Journal of Nonverbal Behaviour* (2000).
16. Bruno Lepri, Nadia Mana, Alessandro Cappelletti, Fabio Pianesi, and Massimo Zancanaro. 2009. Modeling the Personality of Participants During Group Interactions. In *Proc. Int. Conf. on User Modeling, Adaptation and Personalization (UMAP)*.
17. Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian Fasel, Mark Frank, Javier Movellan, and Marian Bartlett. 2011. The Computer Expression Recognition Toolbox. In *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*. 298–305.
18. Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression. In *Proc. IEEE Workshop on CVPR for Human Communicative Behavior Analysis*.
19. Robert R. McCrae and Oliver P. John. 1992. An Introduction to the Five-Factor Model and its Applications. *Journal of Personality* (1992).
20. Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web. In *Proc. ACM Int. Conf. on Multimodal Interaction (ICMI)*. 169–176.
21. Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2009. Classifier Chains for Multi-Label Classification. In *Lecture Notes in Artificial Intelligence*, W. Buntine, M. Grobelnik, and J. Shawe-Taylor (Eds.). Springer.
22. James A. Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* (1980).
23. Dairazalia Sanchez-Cortes, Joan-Isaac Biel, Shiro Kumano, Junji Yamato, Kazuhiro Otsuka, and Daniel Gatica-Perez. 2013. Inferring Mood in Ubiquitous Conversational Video. In *Proc. Int. Conf. on Mobile and Ubiquitous Multimedia (MUM)*.
24. Dairazalia Sanchez-Cortes, Shiro Kumano, Kazuhiro Otsuka, and Daniel Gatica-Perez. 2015. In the Mood for Vlog: Multimodal Inference in Conversational Social Video. *ACM Trans. on Interactive Intelligent Systems, Special Issue on Behavior Understanding for Arts and Entertainment* 5, 2 (July 2015), 1–24.
25. Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge. *Speech Communication* 53, 9 (2011), 1062–1087.

26. Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S. Huang. 2006. Emotion Recognition Based on Joint Visual and Audio Cues. In *Proc. Int. Conf. on Pattern Recognition (ICPR)*.

27. Lucía Teijeiro-Mosquera, Joan-Isaac Biel, José Luis Alba-Castro, and Daniel Gatica-Perez. 2015. What Your Face Vlogs About: Expressions of Emotion and Big-Five Traits Impressions in YouTube. *IEEE Trans. on Affective Computing* 6, 2 (April–June 2015), 193–205.

28. Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis Vlahavas. 2008. Multilabel Classification of Music into Emotions. In *Proc. Conf. of the International Society for Music Information Retrieval (ISMIR)*.

29. Martin Wollmer, Felix Weninger, Tobias Knaup, Bjorn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. YouTube Movie Reviews: Sentiment Analysis in an Audio-visual Context. *IEEE Trans. on Intelligent Systems, Special Issue on Concept-Level Opinion and Sentiment Analysis* (March 2013).

30. Michelle S. M. Yik, James A. Russell, Chang-Kyu Ahn, José Miguel Fernández Dols, and Naoto Suzuki. 2002. Relating the Five-Factor Model of Personality to a Circumplex Model of Affect. In *The Five-Factor Model of Personality Across Cultures*.

31. Min-Ling Zhang and Zhi-Hua Zhou. 2007. ML-kNN: A Lazy Learning Approach to Multi-Label Learning. *Pattern Recognition* 40, 7 (2007), 2038–2048.

32. Min-Ling Zhang and Zhi-Hua Zhou. 2014. A Review On Multi-Label Learning Algorithms. *IEEE Trans. on Knowledge and Data Engineering* (2014).

33. Yi Zhang and Jeff Schneider. 2011. Multi-label Output Codes using Canonical Correlation Analysis. In *Proc. Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*.