Fusion of Facial expressions and EEG for implicit affective tagging

Sander Koelstra[a,*], Ioannis Patras[a]

[a]School of Computer Science and Electronic Engineering, Queen Mary University of London (QMUL), Mile end road, E1 4NS London

Abstract

The explosion of user-generated, untagged multimedia data in recent years, generates a strong need for efficient search and retrieval of this data. The predominant method for content-based tagging is through slow, labour-intensive manual annotation. Consequently, automatic tagging is currently a subject of intensive research. However, it is clear that the process will not be fully automated in the foreseeable future. We propose to involve the user and investigate methods for implicit tagging, wherein users' responses to the interaction with the multimedia content are analysed in order to generate descriptive tags.

Here, we present a multi-modal approach that analyses both facial expressions and Electroencephalography(EEG) signals for the generation of affective tags. We perform classification and regression in the valence-arousal space and present results for both feature-level and decision-level fusion. We demonstrate improvement in the results when using both modalities, suggesting the modalities contain complementary information.

Keywords: Emotion classification, EEG, Facial expressions, Signal processing, Pattern classification, Affective computing.

1. Introduction

In this paper, we utilize methods for facial expression and EEG signal analysis (introduced in Section 2.2) to investigate the possibilities for multi-modal fusion in affect recognition and implicit tagging. We use a dataset with recordings of participants watching video clips designed to elicit emotional responses. These responses, in the form of detected facial expressions and EEG signals, are classified into arousal, valence, and control classes. Both feature-level fusion and decision-level fusion methods are explored and shown to improve upon the single modality results. In addition, we show how this method can be effectively used for the implicit tagging of videos by aggregating affect estimates from multiple participants. The main novelty in this work is the combination of the facial expression and EEG modalities for affect recognition and implicit tagging of videos. To the best of our knowledge, this is the first work to attempt this.

The remainder of this paper is organized as follows: We first give an overview of related works in Section 2. The used dataset is described in Section 3. Next, we detail the methodology used for feature extraction in each modality and the methods used for fusion in Section 4. We present the obtained classification results from single modalities and from both feature-level and decision-level fusion, as well as the results for implicit affective video tagging in Section 5. Results for regression of the single modalities and fusion are given in Section 6. Section 7 concludes the paper.

2. Related work

2.1. Implicit tagging

Implicit tagging concerns the automated annotation of multimedia data by analysis of users' behaviour. The advantage over explicit tagging is that it is done passively and requires little or no active or conscious involvement from users (other than their usual interaction with the multimedia data). For example, a video could be implicitly tagged as humorous if a user is filmed smiling or laughing in response to it. Then, other users searching for humourous videos can benefit from these tags. Where explicit tagging can be a slow and labour-intensive process, implicit tagging is done in the background, can be performed each time a multimedia item is viewed and can deliver a wealth of annotation that can be used in search and indexing of multimedia data and can be combined with any tags derived explicitly (e.g. through crowd-sourcing of tags). In addition, implicit tagging can be used to assess the validity of existing tags, as well as for user profiling (storing particular preferences of a user based on his reactions to content)[1].

So far, tagging is mostly done explicitly and manually by humans, or automatically using computer vision algorithms. Both types of existing tags suffer from various drawbacks. Manual tagging is greatly increasing with the rise of social media websites, allowing users to attach tags

*Corresponding author

Email addresses: sander.koelstra@eecs.qmul.ac.uk (Sander Koelstra), i.patras@eecs.qmul.ac.uk (Ioannis Patras)

to uploaded multimedia items. However, as mentioned in [1], users do not typically tag with the intent of enriching the data for automated search and indexing. Instead users tag based on their own personal and social needs, bringing the value of these tags into question. Machine-based tagging suffers from the issue of the semantic gap, where tags may include recognized objects/locations/faces, the detected amount of motion, shot transition speed, etc. but tagging algorithms still have great difficulty assigning semantic meaning to multimedia items, such as detecting plot keywords or affective content. Implicit tagging may be able to alleviate some of these shortcomings.

Broadly speaking, there exist two approaches to the problem of implicit tagging: game-based tagging and observation-based tagging. In the former, the tagging of the data is a by-product of playing a game. In game-based tagging, users do actively participate and contribute tags, but are not necessarily consciously aware of it. The tagging itself is not the user's goal or intention; they are merely playing a game. In this sense, game-based tagging can be viewed as a form of implicit tagging, but rather than deriving tags from observation of the user, users are 'tricked' into producing the tags themselves. The most well-known of these approaches is probably the ESP game[2]. In the game, two users are paired and are given the task of assigning tags to an image. Points are awarded when the same tag is given by both users. The users do not know each other and can not communicate, so their only way to score points is to assign straightforward tags to the image in the hope that their counterparts will assign the same ones. Since the publication of this work, several authors have expanded and refined this concept for a diverse set of tag types such as object localization[3], music metadata[4] and moods[5].

In observation-based tagging, users' responses as they view the media are recorded and analysed in order to extract tags describing the media. User responses can be recorded from a variety of modalities. In this work, we focus on the modalities of facial expressions and EEG signals. Of the possible passive observation modalities, facial expressions probably are the most informative, while EEG signals may reveal some otherwise unobservable affective states and may well complement the former modality.

2.2. EEG signal analysis for tagging

Electroencephalography (EEG) is a non-invasive technique for measuring a participant's brainwave patterns, by recording electrical activity via electrodes placed on the scalp. A cap is placed on the participant's head, electrode gel is applied to ensure good conductivity and then electrodes are attached to the cap. An international standard known as the 10-20 system determines the location of the electrodes on the scalp (see Fig. 1).

The use of EEG in annotating multimedia data is a very new research direction and so far only a few works have investigated this area.

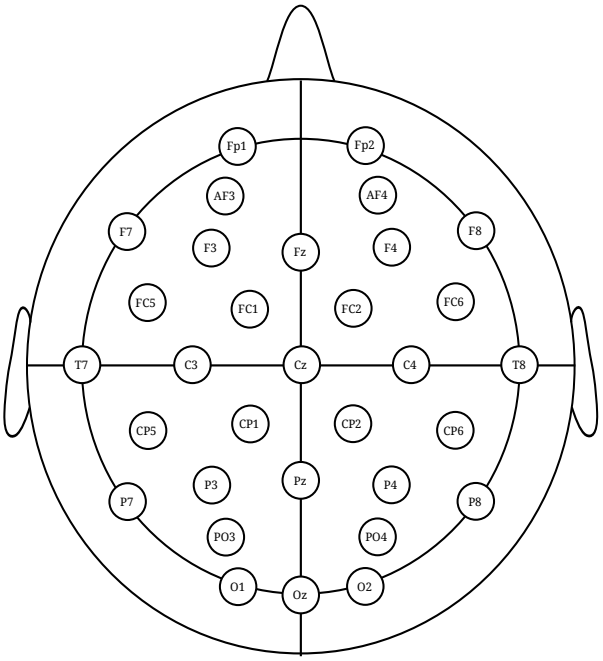

Figure 1: Electrode placement in the 10-20 system.

In [6], an oddball paradigm is used in which images of a forest environment were shown to participants for 100 ms each. The goal was to detect a small subset of target images that contained pedestrians. The target images elicit a P300 event-related potential(ERP) which is then classified using Fisher linear discriminant analysis. Another test is run without the EEG modality, where participants press a button upon seeing the target images. The results show no significant differences in target image detection accuracy between the use of the EEG modality and the use of buttons.

In [7, 8], categories of images are classified based on EEG measurements recorded as the images were presented. The used categories are faces, animals and inanimate objects. This is based on the notion that the human visual system responds very differently to these categories of images. The authors propose a vision-based algorithm that uses pyramid match kernels to initially classify the images. The EEG data is then combined with the vision-based features using a kernel-alignment method. The combination of the two modalities outperforms the individual methods.

In [9], the RAPID system is proposed. The authors use ERP analysis in combination with eye tracking to assist intelligence analysts in rapidly reviewing and categorizing satellite imagery. The analyst is assigned a target category to look for in the images. When participants see an image in the target category, an ERP occurs in the EEG data which is then classified. Eye tracking is used to determine points of interest within the images.

In [10], a method is proposed for validation of tags displayed in conjunction with videos using EEG signals. Videos are displayed with either valid or invalid tags. It has

been shown that in cases of two semantically mismatching categories an N400 ERP occurs at around 400 ms after the second stimulus is presented. The authors demonstrate significant differences in the EEG signals recorded during the display of valid versus invalid tags.

Several works have attempted recognition of emotions from EEG signals. In [11], participants are asked to remember an episode in their life that corresponds to positive/excited and one that corresponds to negative/excited emotions. A third emotional state called calm/neutral is elicited by asking the participants to stay calm and relax. For these three classes, a classification accuracy of 63% is reported using the short-time Fourier transform for feature extraction and a linear SVM for classification.

In [12], participants watch a series of music videos selected to elicit emotions. The participants then rate the felt emotions in terms of valence, arousal and like/dislike. In performing a binary classification, accuracies of up to 62% are attained based on EEG bandpower features and a Gaussian Naïve Bayes classifier. Regression results for the same experiment are reported in [13].

In [14], 5 different emotions (joy, anger, sadness, fear, relaxation) are elicited by using video stimuli in 12 participants. Using a one-vs-all SVM classifier, a classification rate of 41.7% is reported.

Besides these works, much research has been done in psychology into ERP analysis and correlations with emotion (e.g. [15, 16, 17]). These works show clear associations between ERP activity and valence/arousal. However, they mostly have in common that they work with time-locked stimuli (such as pictures), and average the ERP signal over several trials to increase the signal-to-noise ratio. In the implicit tagging paradigm, the focus is mainly on making judgments on single trials, since the main goal is to derive tags without active participation from users, rendering it nearly impossible to obtain multiple recordings of responses to the same stimulus. In addition, we can not rely on time-locked stimuli, as in this paradigm it is generally not known in advance what the stimulus will be (the user should be able to watch anything he/she likes). Without time-locked stimuli and using only single trials, performing a proper ERP analysis as in these works is virtually impossible.

2.3. Facial expression analysis for tagging

Two main streams in the current research on automatic analysis of facial expressions consider facial affect (emotion) detection and facial muscle action (action unit) detection[18, 19, 20]. The most commonly used facial expression descriptors in facial affect detection approaches are the six basic emotions (fear, sadness, happiness, anger, disgust, surprise), proposed by Ekman and discrete emotion theorists, who suggest that these emotions are universally displayed and recognized from facial expressions. The most commonly used facial muscle action descriptors are the Action Units (AUs) defined in the Facial Action Coding System (FACS; [21]). The basic emotion categories form only a subset of the total range of possible facial displays and categorization of facial expressions can therefore be forced and unnatural. Boredom and interest, for instance, do not seem to fit well in any of the basic emotion categories. Moreover, in everyday life, these prototypic expressions occur relatively rarely; usually, emotions are displayed more subtly. To detect such subtle expression a model of atomic facial signals, such as FACS, is needed. FACS classifies atomic facial signals into Action Units, considered to be the smallest visually discernible facial movements, according to the facial muscles that cause them. It defines 9 upper face AUs and 18 lower face AUs.

Figure 2: Apex phases of 8 AUs of the FACS system.

However, Action Units do not encode the semantic or affective meaning of the expression, which, for many applications, is the main focus. Thus, methods are needed to map the occurrence of AUs to the presence of higher-level affective states. For discrete emotions, the EMFACS [21] method (for basic emotions) and the FACSAID[1] (for various affective states) methods provide rules to map AU (co-)occurrences to discrete emotions. FACS is also used for determination of other complex psychological states such as depression[22] or pain[23], or other higher-level states [24].

When it comes to associations between AUs and dimensional models of emotion, such as the valence-arousal model, little research is available[25]. Russell's valence-arousal scale, widely used in research on affect, is used to quantitatively describe emotions. In this scale, each emotional state can be placed on a two-dimensional plane with arousal and valence as the horizontal and vertical axes. While arousal and valence explain most of the variation in emotional states, a third dimension of control is often included in the model [26]. Arousal can range from inactive (e.g. uninterested, bored) to active (e.g. alert, excited), whereas valence ranges from unpleasant (e.g. sad, stressed) to pleasant (e.g. happy, elated). Control ranges from a helpless and weak feeling (without control) to an empowered feeling (in control of everything). To the best of our knowledge, only one work tries to estimate emotions in a dimensional model by first detecting AUs[27]. They first detect the AUs present in the video and subsequently classify the valence value, but do not list associations found of particular AUs with valence. The original work propos-

[1] http://www.face-and-emotion.com/dataface/facsaid/description.jsp

ing the valence-arousal model[26] does however give information on the relation of valence and arousal to discrete emotions. Thus, one could map the AUs to a discrete emotion using for instance FACSAID, and then find the approximate location in valence-arousal space from [26].

The use of facial expressions for multimedia tagging is also a new concept and consequently only few works are available. In [28], facial expressions are utilised for implicit feedback to determine the relevance of search results. An excerpt of the result document is shown to the user and the result is then classified as relevant or irrelevant for the query based on the user's facial expressions. A similar methodology is applied in [29].

Jiao and Pantic[30] performed an experiment on implicit tag validation using facial expressions. They use geometric features obtained from a particle filter tracker with a Hidden Markov Model to assess the correctness of tags displayed alongside images from the participants' facial expressions. On average, 54% of the trials are correctly classified.

2.4. Multi-modal approaches

In general, approaches for modality fusion can be classified into two broad categories, namely, feature fusion (or early integration) and decision fusion (or late integration) [31].Some works have also attempted a combination of both methods in hybrid fusion[32]. In feature-level fusion, the features extracted from signals of different modalities are concatenated to form a composite feature vector and then inputted to a recognizer. In decision fusion, on the other hand, each modality is processed independently by the corresponding classifier and the outputs of the classifiers are combined to yield the final result. Various methods can be used for decision-level fusion, such as simple rule-based methods (for instance taking the sum or product of class probabilities) to classifier-based approaches, where a meta-classifier is trained taking the decisions from individual classifiers as its features.

Each approach has its own advantages. For example, implementing a feature fusion-based system is straightforward, while a decision fusion-based system can be constructed by using existing unimodal classification systems. Moreover, feature fusion can consider synchronous characteristics of the involved modalities, whereas decision fusion allows us to model asynchronous characteristics of the modalities flexibly. In addition, in feature-level fusion, correlations between feature sets can be exploited by the classifier, while these are lost in decision-level fusion.

The most common modalities to be merged are combinations of audio, video and text modalities[32]. In facial expression analysis, the main modalities to be used are aural and visual features[20]. The EEG modality is most often fused with fMRI , EMG, MEG or PET in clinical settings. For multimedia applications, EEG has been fused with peripheral physiological signals and gaze information in [33], with peripheral physiological signals and audio-visual features in [13, 12] and with audio-visual features in [7, 8]. To the best of our knowledge, the combination of EEG and facial expressions has not been attempted previously. For a more thorough review of previous methods for multi-modal fusion, the reader is referred to the recent surveys by Sebe et al.[34], Zeng et al.[20] and Atrey et al.[32].

3. Dataset

Figure 3: A participant in the MAHNOB HCI experiment.

We use the MAHNOB HCI[33] dataset in this experiment, which contains EEG, video, audio, gaze and peripheral physiological recordings of 30 participants. Each participant watched 20 clips extracted from hollywood movies and video websites such as YouTube.com and blip.tv. The stimuli were selected in order to elicit 5 emotions (disgust, amusement, joy, fear and sadness). In addition, various weather reports were included as neutral stimuli. The stimuli videos range in duration from 35 to 117 seconds. After watching each stimulus, the participants used Self-Assessment Manikins (SAM)[35, 36] to rate their felt arousal, valence and control on a discrete scale of 1 to 9.

Facial video was recorded by 6 cameras at 60 frames per second from different angles. Only the frontal camera is used in this study. 32 channel EEG, placed according to the 10-20 system (see Figure 1) was recorded at 256Hz using a BioSemi ActiveTwo system. Figure 3 shows a participant during the experiment.

For 6 of the 30 participants, various problems such as technical failures occurred during the experiment. Here, only the 24 participants for which all data is available were used.

4. Affect recognition using EEG and Facial expressions

4.1. EEG features

The EEG signal is down-sampled to 128Hz and a 4-45Hz bandpass filter is used to reduce artefacts. As EOG

Preselected electrodes[12]	
Arousal	**Valence**
CP6*, Cz*, FC2*	Oz**, CP1**, T7**, C4**, FC6**, PO4**, Cz*, CP6*, CP2*, T8*, F8*

Table 1: Preselected features based on earlier work[12] (*=$p < .01$, **=$p < .001$).

was not recorded, eye movement artefacts are not suppressed. The average of the 15-second baseline signal before each trial is subtracted from the trial data and it is referenced to the common average (CAR).

Power spectral density (PSD) in the Theta (4 - 8Hz), slow alpha (8 - 10Hz), alpha (8 - 12Hz), beta (12 - 30Hz) and gamma (30 - 45Hz) bands is computed, as well as the lateralization for 14 left-right pairs. This gives a total of (5 bands $\times$ (32 channels + 14 asymmetry pairs)) = 230 features.

PSD features have been previously used for EEG signal analysis, for example in [12]. In that work, significant correlations with arousal and valence are found for a subset of the 32 electrodes. Table 1 gives an overview of these electrodes. For the control dimension, we have not found a list of important electrodes in literature. However, the results presented here with the reduced electrode set, improve upon results for the full set for all three dimensions.

We also include electrodes Fp1 and Fp2 as they often respond to eyebrow motion and/or forehead wrinkling. While this is usually considered noise in the EEG literature as it is not caused by neurophysiological activity, we consider this a valid feature as eyebrow action may wel correlate with our affective targets. In this work, we limit the EEG feature vector to contain these 14 electrodes (and their 9 associated left-right pairs), giving a final 115 features.

4.2. Facial expression features

The method for facial expression analysis used here is based on the work on facial Action Unit (AU) detection described in [37], which performs frame-by-frame recognition of the activation of Action Units. Fig. 4 gives an overview of this system. In the preprocessing phase, the face is located in the first frame of an input video and head motion is suppressed by an affine rigid face registration. Next, non-rigid motion is estimated between consecutive frames by the use of Non-rigid Registration using Free-form Deformations (FFDs). For each AU, a quadtree decomposition is defined to identify face regions related to that AU. In these regions, orientation histogram feature descriptors are extracted. Two GentleBoost classifiers are trained per AU, one to detect onset segments and one to detect offset segments. Finally, a Hidden Markov Model (HMM) is used to classify the input video in terms of AUs and their temporal segments, based on the output of the GentleBoost classifiers. For more details on this method, the reader is referred to [37].

In this work, the final classification target is not the AUs themselves, but rather the arousal, valence and control ratings. Here, we use a two-step approach to accomplish this. First, we use the system proposed in [37] to detect frame-by-frame AU activations. Second, a set of meta-features is extracted from the output of this system, which are subsequently used for classification and regression in terms of arousal, valence and control.

We note that no AU ground truth annotation is available for the MAHNOB HCI dataset. Manually annotating the dataset of in total 480 videos in terms of AUs was not considered feasible. Therefore, we can not train the AU detection system on this dataset. Instead, we use the system as trained on the (posed) MMI dataset, which showed success in classifying AUs in the posed Cohn-Kanade (CK) and the spontaneous SAL datasets [37]. The same 18 AUs are detected as in the CK dataset in [37] (AUs 1, 2, 4, 5, 6, 7, 9, 10, 11, 12, 14, 15, 17, 20, 24, 25, 27 and 45).

Unfortunately, little work is available on the associations between AUs and dimensional models of emotion. For discrete emotions there is however some earlier work available. EMFACS[21] is an addition to FACS, that allows annotators to only annotate those AUs and AU combinations associated with the 6 basic emotions. Table2 lists these AUs. As these basic emotions can be mapped with some accuracy in dimensional models[26, 38], we can hypothesize that these AUs may also correlate with arousal, valence, control, etc.

Therefore we limit the AUs used in our system to the intersection of Table2 and those AUs we can detect. For each AU (combination), we extract three features: the number of onset detections in the video, the number of offset detections in the video, and the difference in mean output of onset- and offset-classifiers (a possible indication of the strength of the AU activation). This gives a total of $28 \times 3 = 84$ features.

Figure 5 shows several screenshots of face videos in the MAHNOB HCI dataset. This dataset is quite challenging for facial expression analysis, partly due to the spontaneous nature of the expressions. In addition, unlike in some other datasets such as SAL, there is no (virtual) conversation partner or in fact anyone else in the room. This means the number of expressions is quite sparse and expressions that are shown are generally quite subtle.

4.3. Classification

We perform binary classification on the arousal, valence and control ratings, which are thresholded into high (rating 6-9) and low (rating 1-5) classes. Those videos that are rated 5 (in the middle of the 9-point scale) are grouped with the low class, since this gives the most balanced class distribution.

We use recursive feature elimination (RFE) to select features for classification. This is done by iteratively calculating the feature weights for a linear SVM classifier and

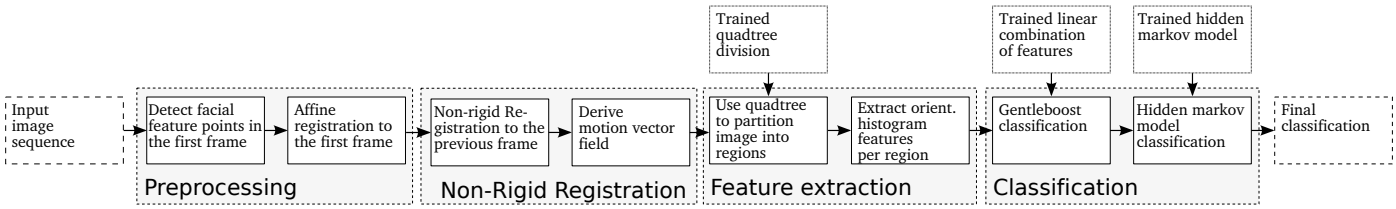

Figure 4: Outline of the proposed method.

Happiness	**Fear**	**Sadness**	**Surprise**	**Anger**	**Disgust**
12	1+2+4	1+4	1+2+5+**26**|27	4+5+7+10+**22**+**23**+25|**26**	9|10+17
6+12	1+2+4+5	1+4+11|15	1+2+5	4+5+7+10+**23**+25|**26**	9|10+**16**+25|**26**
	1+2+4+5+20+25|**26**|27	1+4+15+17	1+2+**26**|27	4+5+7+17+**23**|24	9
	1+2+4+5+25|**26**|27	6+15	5+**26**|27	4+5+7+**23**|24	10
	1+2+5+25|**26**|27	11+17	4+5|7		
	5+20+25|**26**|27	1	17+24		
	5+20				
	20				

Table 2: AU's and AU combinations associated with the 6 basic emotions, according to EMFACS. "|" indicates OR. AUs in boldface are not detected by our system.

then removing the 10% of features with lowest weights. This continues until the target number of features remains. The number of selected features is optimized by a 10-fold inner cross-validation on the training set.

We compare the performance of RFE to feature reduction through Independent component analysis (ICA). Independent component Analysis [39] is an extension of principal component analysis (PCA) and a form of blind source separation (BSS). In principal component analysis, the objective is to separate the signal into orthogonal components of decreasing variance. In independent component analysis however, we try to find statistically independent components which may not be orthogonal. ICA decomposes the source signal into a linear combination of maximally independent components. The assumption is that measured electrical activity is a linear mixture of underlying sources in the brain. Here, we use an implementation of the FastICA algorithm, originally developed by Hyvärinen and Oja[40]. The number of components to use is again determined by 10-fold inner cross-validation on the training set.

For classification we use a Gaussian Naïve Bayes (GNB) classifier as implemented in the scikits-learn toolkit[2]. The naïve Bayes classifier G assumes independence of the features and is given by:

$$G(f_1, .., f_n) = \underset{c}{\operatorname{argmax}}\, p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c) \quad (1)$$

where F is the set of features and C the classes. $p(F_i = f_i | C = c)$ is estimated by assuming Gaussian distributions of the features and modelling these from the training set.

[2]http://scikit-learn.sourceforge.net/

This is a simple and generalizable classifier which is able to deal with unbalanced classes in small training sets and has the added advantage of providing probabilistic outputs, which can be used for decision-level fusion.

We perform two types of cross-validation. First, we perform 10-fold cross-validation on the entire set of 480 trials. Second, we perform a per-subject leave-one-trial-out cross-validation, where the classifier is trained on 19 trials from the same subject and tested on the 20th.

4.4. Regression

For regression, we first scale the arousal, valence and control ratings to the $[0, 1]$ range for convenience. We use again RFE and ICA for feature selection, but in this case we use the cross-correlation between the features/components and ratings as the feature/component weights. The model we use for regression is Bayesian Ridge Regression as implemented in the scikits-learn toolkit[2]. This model is similar to ordinary least squares regression, but attempts to avoid overfitting by penalizing large values of its coefficients (see [41] for more information).

The same cross-validation approach is taken as in classification.

4.5. Classification Fusion

We investigate both feature-level and decision-level fusion in this work. In feature-level fusion, the feature vectors from both modalities are stacked together and the total number of features becomes 199. Next, classification proceeds the same as for the single modalities.

In decision-level classification fusion, we first classify the modalities individually as described above and then combine the classifier outputs in a linear fashion. Figure 6 depicts the two different approaches graphically. We

Figure 5: Example screenshots of face video in the MAHNOB HCI dataset.

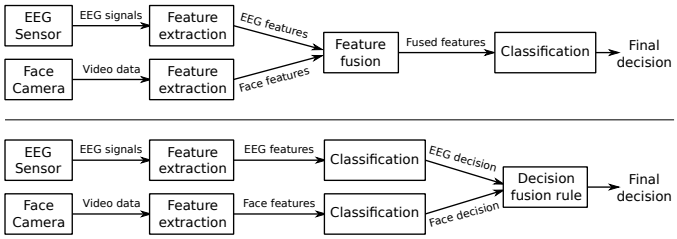

Figure 6: Fusion approaches depicted graphically. Top: feature-level fusion, bottom: decision-level fusion.

distinguish two different types of approaches to decision-level fusion.

First, we present methods for estimating a per-sample weighting α for the different modalities where the final decision is a weighted sum of the outputs from classification of the individual modalities. For each trial, let p_e^x, p_f^x and $p_o^x \in [0,1]$ denote the classifier probability for class $x \in \{1,2\}$ for EEG, facial expressions and fusion respectively. Then the output class probabilities are given by

$$p_o^x = \alpha p_e^x + (1-\alpha) p_f^x. \quad (2)$$

Besides the probability p of each class as given by the GNB classifier, we additionally take into account the classifier's training set F1-score for each trial. The training set performance is determined by a 10-fold inner cross-validation on the training set. Let $F_e, F_f \in [0,1]$ denote the training set F1-score for EEG and facial expression classification respectively. We first normalize these scores to ensure the fusion probabilities for all classes sum up to 1. The normalized training set performance t_e and t_f are given by

$$t_e = \frac{F_e}{\alpha F_e + (1-\alpha)F_f}, t_f = \frac{F_f}{\alpha F_e + (1-\alpha)F_f}. \quad (3)$$

Then, the output class probabilities are given by

$$p_o^x = \alpha(p_e^x t_e) + (1-\alpha)(p_f^x t_f). \quad (4)$$

It can then be shown that

$$p_o^1 + p_o^2 = \alpha t_e + (1-\alpha) t_f = \frac{\alpha F_e + (1-\alpha)F_f}{\alpha F_e + (1-\alpha)F_f} = 1. \quad (5)$$

Second, we also present a meta-classification fusion method, where a meta-classifier is trained on the outputs from classification of the individual modalities. The different methods investigated for decision-level classification fusion are each briefly explained below.

*Equal weights fusion(***W-EQ** *and* **W-EQ**^{T}*)*

W-EQ is the most straightforward method, where the output probabilities for each class are an equal weighting of the class probabilities from each single modality ($\alpha = 0.5$). That is,

$$p_o^x = 0.5 p_e^x + 0.5 p_f^x. \quad (6)$$

For **W-EQ**^{T}, the training performance (t_e and t_f) is also considered. That is,

$$p_o^x = 0.5(p_e^x t_e) + 0.5(p_f^x t_f). \quad (7)$$

*Estimated weights fusion(***W-EST** *and* **W-EST**^{T}*)*

For a given sample, we numerically approximate the optimal decision weight α for the set of training samples. This is done by varying α between 0 and 1 in steps of 0.01 and choosing the value which gives the highest F1-score on the training samples. The estimated weight is then applied to the current sample as in Equation 2. For **W-EST**^{T}, the training performance is also used as in Equation 4.

*Regression-estimated weights fusion(***W-REG** *and* **W-REG**^{T}*)*

Here we first train a linear support vector regressor (SVR) on the training samples with the class probabilities from the individual modalities as features and the optimal weight, determined as above, as the target. The SVR is then used to predict the optimal weight for the current sample and fusion class probabilities are computed as in Equation 2. In **W-REG**^{T}, the training set performance F_e and F_f are also included as features and fusion class probabilities are computed as in Equation 4.

*Meta-classification of class label(***M-CLASS** *and* **M-CLASS**^{T}*)*

Here we train a linear SVM classifier on the probabilistic outputs in order to directly predict the class of the current sample. In **M-CLASS**^{T}, the training set F1-scores F_e and F_f are included in the set of features.

4.6. Regression Fusion

For regression feature-level fusion, the same combined 199-feature vector is used as for classification. For decision-level fusion, we proceed in a similar fashion as for classification. Instead of class probabilities, here we use the ratings predicted by the single modality regressors directly in Equation 4. In addition, for the training set performance we replace F_e, F_f (F1-scores) by the reciprocal of the mean squared error (MSE) to indicate the quality of regression on the training set. This is then normalized as in Equation 3. **W-EQ**, **W-EQ**^{T}, **W-EST**, **W-EST**^{T}, **W-REG** and **W-REG**^{T} can then be used for fusion of the output of the single modality regressors. Rather than using meta-classification as in **M-CLASS**, here we use meta-regression (**M-REG**). In **M-REG**, the aforementioned Bayesian Ridge Regression model is used with the predicted values by the single modality regressors as features (and the training set performance in the case of **M-REG**^{T}).

5. Classification Results

5.1. Single modalities and feature fusion

		Arousal		Valence		Control	
	Modality	**CR**	**F1**	**CR**	**F1**	**CR**	**F1**
ICA	**EEG**	66.0	64.7	71.5	70.9	67.5	**67.4**
	Face	65.0	63.8	64.5	62.8	64.5	64.3
	Fusion	68.0	**66.2**	72.5	**71.3**	67.5	**67.4**
RFE	**EEG**	67.5	66.1	70.0	69.3	63.5	63.5
	Face	67.5	66.3	64.0	63.3	62.0	62.0
	Fusion	68.5	**67.1**	73.0	**71.5**	68.5	**68.4**†
	Random	50.0	48.1	50.0	48.7	50.0	48.2
	Majority	62.0	37.6	62.6	38.3	65.6	39.4
	Class ratio	56.7	50.0	54.6	50.0	56.7	50.0

Table 3: Average classification rates (CR) and F1-scores (average of F1-score for each class) for the single modalities and feature fusion for the two different feature sets. †indicates that the fusion F1-score distribution is significantly higher than the score distribution of the best performing single modality according to a related two-sided t-test ($p < .05$). As a baseline, expected results are given for classification based on random voting, voting according to the majority class and voting with the ratio of the classes. The highest score per feature set and target is shown in bold.

Table 3 gives the classification results for each single modality and rating target. As a baseline, we also give the expected values (analytically determined) of classifying randomly, classifying according to the majority class in the training data, and classifying by choosing a class with the probability of its occurrence in the training data. For determining the expected values of majority voting and class ratio classification, we used the class ratio of each participant's ratings during the experiment. All classification F1-scores are significantly better than class ratio voting according to an independent one-sample t-test ($p < .01$).

The result for feature-level fusion for both ICA and RFE features is also given here. In all cases, fusion equals or slightly outperforms the single modalities. However, the difference is only statistically significant for the control target with RFE features.

Figure 7 further investigates the balance between number of components or features, and performance for EEG features, face features and feature-level fusion. The top plot depicts the relationship between number of RFE-selected features and F1-score. In most cases, the performance tapers off as the number of features increases. For face features, the performance peaks early, and adding more features degrades performance. Possibly the number of meaningful face features is limited, an additional features just add noise and complexity. For EEG and feature-level fusion, the result is more constant. While peak performance of feature-level fusion is highest in most of the plot, there are also areas where feature-level fusion performs worse than the single modalities. Also shown is the actual number of selected features by the inner cross-validation loop mentioned earlier. We can see the estimate is quite effective, although in some cases a more optimal number of features exists (e.g. for EEG in valence and control).

The second plot shows the number of ICA components versus F1-score, while the third plot shows the percentage of variance explained by the ICA components. Again we see the face features as performing least and showing the largest drop in performance as the number of components increases. We can also see that the explained variance increases fastest for face features, which may again indicate less information existing in these features. The inner cross-validation selects the optimal number of components for all cases except for feature-level fusion/control.

Feature-level fusion does not consistently improve the results, especially when the disparity in performance between EEG and face features is large.

All results are significantly above the class ratio baseline level. The RFE feature selection method slightly outperforms ICA feature reduction, although the difference is not significant.

In general, EEG seems to outperform the facial expression analysis. It may be possible to improve facial expression analysis results by training the AU detector on the MAHNOB HCI dataset, rather than the currently used MMI dataset. However, this would require the manual annotation of the entire set for AUs. Another possible improvement may be possible by directly classifying arousal and valence, rather than first classifying the AUs as an intermediary step. We leave these improvements as future work.

Table 4 lists the 10 most selected feature categories in Recursive feature elimination. Somewhat surprisingly, valence turns out to be the hardest target to classify in the case of facial expressions. One might expect AU 12 (smile) to be a clear indicator of valence, but it is rarely even

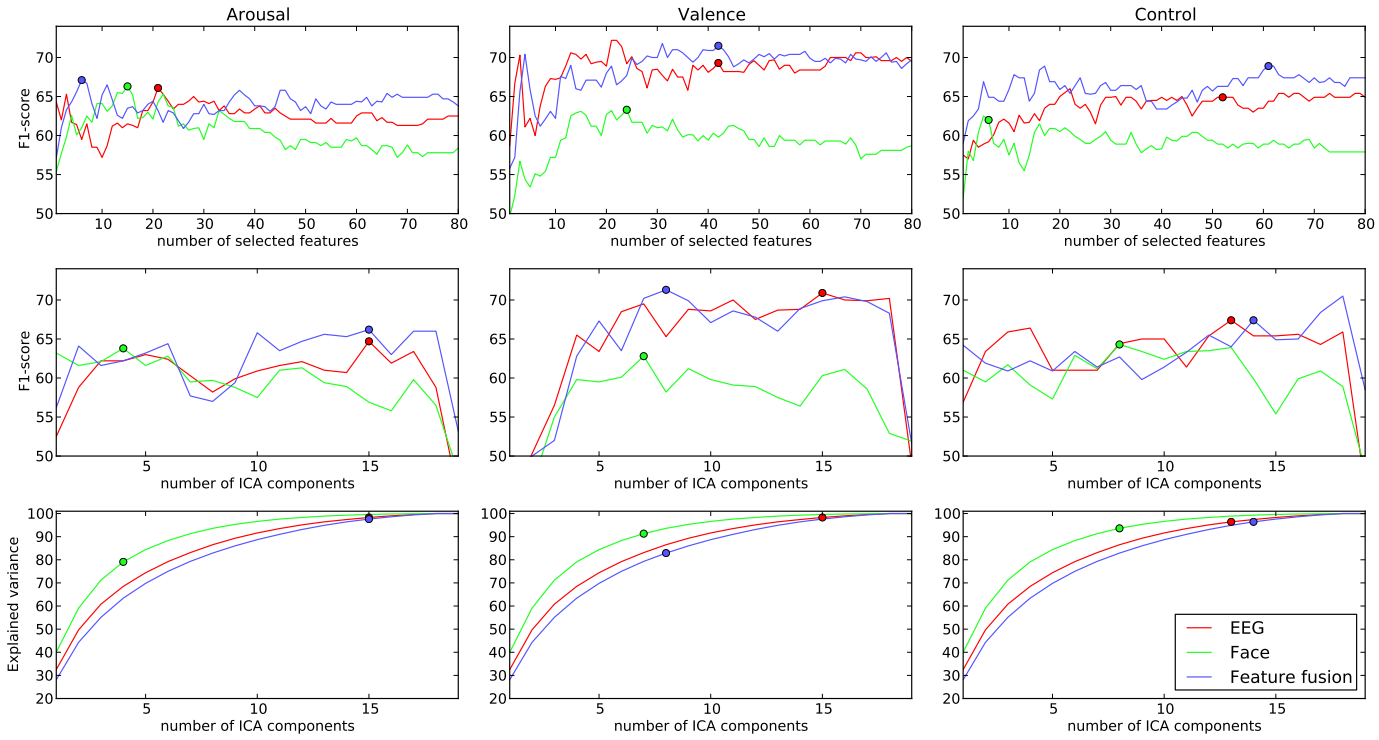

Figure 7: The relationship between number of components/features and performance (averaged over participants). a) Number of RFE-selected features plotted against F1-scores. b) Number of ICA components plotted against F1-scores. c) Number of ICA components plotted against explained variance. In each plot, the circle denotes the number of actual selected features/components in the paper (selected by an inner cross-validation loop).

Top 10 RFE selected AUs			Top 10 RFE selected electrodes			Top selected frequency bands		
Arousal	**Valence**	**Control**	**Arousal**	**Valence**	**Control**	**Arousal**	**Valence**	**Control**
20	4+7	9	FC5 - FC6	FC5/FC6	FC5/FC6	4-8Hz	30-45Hz	4-8Hz
1+4	1	20	F7 - F8	Fp1/Fp2	F7/F8	8-10Hz	4-8Hz	12-30Hz
9	5+20+25	9+17	PO3 - PO4	PO3/PO4	CP5/CP6	12-30Hz	8-10Hz	8-10Hz
17+24	9+17	4+7	Cz	CP1/CP2	Fp1/Fp2	8-12Hz	12-30Hz	8-12Hz
10+17	9+25	17+24	CP5 - CP6	FC1/FC2	FC1/FC2	30-45Hz	8-12Hz	30-45Hz
4+5	20	5+20+25	T7 - T8	CP5/CP6	CP1/CP2			
9+17	17+24	1+2+4+5+25	FC2	F7/F8	T7/T8			
1+2+4+5	10+17	1+2+5	T7	Fp2	PO3 - PO4			
1+2+4+5+25	1+2+4	1	C3 - C4	PO4	Oz			
1	1+2+4+5	10+25	CP1 - CP2	FC2	FC2			

Table 4: AU combinations, electrode channels and frequence bands most often selected by Recursive feature elimination (listed in decreasing order). Note that items with a / denote a lateralization pair.

selected by RFE. Upon visual inspection of the data, it was found that the displayed smiles are firstly very subtle (and thus hard to detect), and do not only occur in high valence cases (e.g. awkward smiles can occur during disgusting videos). The same in general holds for all archetypical universal emotions, which rarely occur clearly throughout the dataset.

Table 4 also shows the EEG electrodes most often selected by RFE. The selected electrodes for valence and arousal align reasonably well with those shown in Table 1, which were shown to correlate with these dimensions in [12]. The two most often selected features for each target are frontal electrodes, indicating a likely effect due to muscle activations, rather than neural activity.

The table also shows the most selected frequency bands. For arousal, the most selected bands are theta and alpha, which matches our earlier work [12, 42], and a relationship between alpha power and arousal has been reported elsewhere too [43]. For valence, the most selected band is the gamma band, which also showed the strongest correlation in [12]. Earlier work has also shown such correlations [44]. For control, the most features were selected from the theta and beta bands.

5.2. Decision-level fusion

		Arousal		Valence		Control	
	Modality	**CR**	**F1**	**CR**	**F1**	**CR**	**F1**
ICA	**EEG**	66.0	64.7	71.5	70.9	67.5	67.4
	Face	65.0	63.8	64.5	62.8	64.5	64.3
	W-EQ	67.0	**65.5**	72.5	**71.0**	71.5	**71.5**
	W-EQ^T	66.5	**65.0**	72.5	**71.0**	71.0	**70.9**
	W-EST	68.0	**66.7**	72.5	**71.6**	70.0	**70.0**
	W-EST^T	69.5	**68.3**	71.5	70.2	72.5	**72.4**†
	W-REG	69.0	**67.7**	73.0	**72.2**	71.0	**71.0**
	W-REG^T	70.0	**68.8**†	74.0	**73.0**	73.0	**73.0**†
	M-CLASS	63.5	63.2	71.5	70.7	69.0	**69.0**
	M-CLASS^T	66.5	**65.8**	71.5	70.7	67.5	67.4
RFE	**EEG**	67.5	66.1	70.0	69.3	63.5	63.5
	Face	67.5	66.3	64.0	63.3	62.0	62.0
	W-EQ	71.0	**69.7**	71.0	**70.3**	67.5	**67.4**
	W-EQ^T	71.5	**70.0**	71.5	**70.7**	65.5	**65.4**
	W-EST	72.0	**70.4**†	72.5	**71.3**	60.0	59.9
	W-EST^T	71.0	**69.4**	67.5	66.2	64.0	**63.9**
	W-REG	72.5	**70.9**†	73.0	**71.8**	67.0	**67.0**
	W-REG^T	71.5	**70.0**	71.5	**70.4**	67.0	**66.9**
	M-CLASS	69.0	**68.4**	71.5	**70.5**	67.0	**66.7**
	M-CLASS^T	70.5	**70.0**	70.5	**69.7**	66.5	**66.4**

Table 5: Average classification rates (CR) and F1-scores (average of F1-score for each class) for decision-level fusion. For comparison the results of single modalities are also shown. F1-scores higher than the best single-modality result are bold. †indicates that the fusion F1-score distribution is significantly higher than the score distribution of the best performing single modality according to an independent two-sample t-test ($p < .05$).

For decision-level fusion, we investigated the use of several different methods as described in Section 4.5. Results are shown in Table 5. While in most of the performed tests the fusion does outperform the single modalities, in only five cases the difference is statistically significant.

Significant results were attained only for the **W-REG** and **W-EST** methods (and their counterparts with training set performance included). This suggests that trying to predict or approximate a weighting between the modalities generally works better than trying to directly classify the samples using the individual modalities outputs as features. In addition, using the training set performance does not seem to lead to consistent improvements.

Figure 8 gives a graphical overview of the results for single modalities, feature fusion, and decision level fusion.

5.3. Implicit Video tagging

Table 6 shows the results for implicit tagging of the videos based on the classifier outputs of all participants. This is challenging since the classes here are not objectively defined. Arousal, valence, and control are subjective measures and participants frequently disagree on the affective content of videos. Nevertheless, in most cases there is a reasonable agreement between participants, which can be seen in the table. Here, we assign a binary class to each video clip (e.g. low/high arousal/valence/control) based on the majority opinion of the participants. Video clips with less than 55% agreement for one of the modalities (4,5,6,9 and 12) were excluded from this test. Next, we estimate the class based on the outputs of the EEG, facial expression, and feature-level fusion classifiers. This is based on the per-subject leave-one-video-out cross-validation. As can be seen from the table, the feature-level fusion performs much better on this task, frequently correcting the errors of one or both modalities. For valence and control, a classification rate of 80% and 86.7% are achieved. This is a strong indication that the gathering and analysis of implicit responses in large quantities can provide effective and reliable emotional tags.

Figure 9 shows the classification accuracy for different numbers of participants. This is the average classification accuracy over videos for the given number of participants, determined by averaging over all possible combinations of the 24 participants in the experiment. It should be noted that the number of combinations is not the same over the x-axis, e.g. the result for 12 participants is the average over 2704156 combinations, while the results for 1 and 23 subjects is the average of only 24 combinations. This explains the somewhat erratic behaviour at the left and right edges of the plots. Thus this plot should be interpreted with care. Nevertheless, a clear advantage can be seen in aggregating the results from multiple participants.

Surprisingly, it seems the aggregation works better for the control dimension, perhaps due to the modalities being more complementary for this target (as is consistent with gains for this dimension in decision-level fusion). Naturally it is impossible to obtain a perfect classification as the ground truth is itself subjective and the human annotators themselves do not always agree on the appropriate class label. In most of the plots above, we see the performance levelling off, for instance for EEG prediction of valence, it seems that increasing the number of participants will not increase the performance. For others, the performance plateau does not seem to have been reached yet (e.g. face/arousal), and using more participants may yet increase the results, although it seems highly likely a plateau exists and a 100% classification rate is not expected to be achievable, especially given the limited human agreement.

6. Regression results

Table 7 gives the results for regression based on single modalities and feature-level fusion. We report the mean squared error (MSE), the mean absolute error (MAE) and it standard deviation as well as the coefficient of determination (R^2). R^2 is defined as

$$R^2 \equiv 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{8}$$

where y_i is the ground truth rating for sample i, $\bar{y}$ is the mean ground truth rating over all samples and $\hat{y}_i$ is the

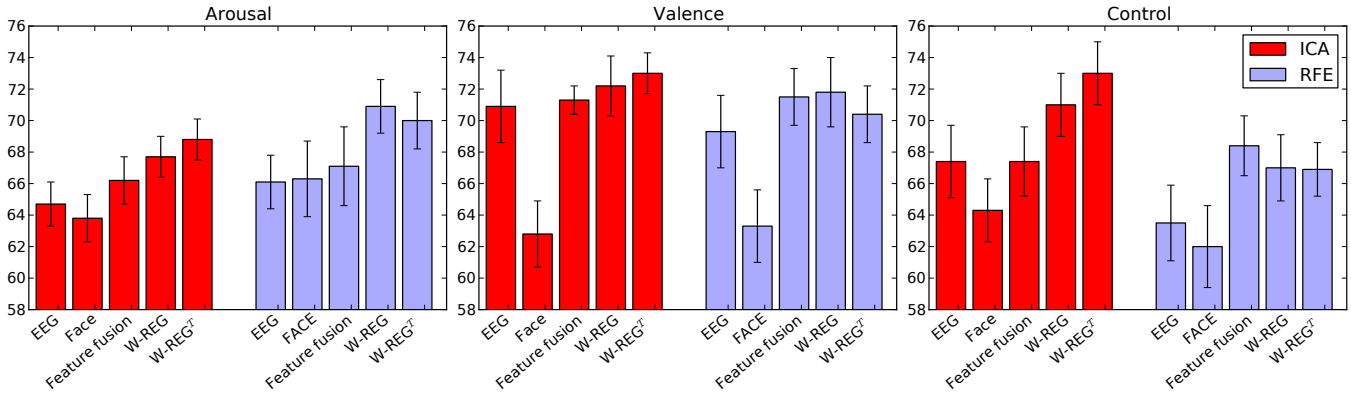

Figure 8: Results for classification on single modalities, feature fusion, and the two best decision fusion methods. Error bars shown correspond to the standard error of the mean

	Arousal					Valence					Control				
Video	**HA**	**Class**	**EEG**	**Face**	**FlF**	**HA**	**Class**	**EEG**	**Face**	**FlF**	**HA**	**Class**	**EEG**	**Face**	**FlF**
1	62%	+	−	+	+	100%	−	−	−	−	83%	−	−	−	−
2	70%	+	−	−	−	100%	−	−	−	−	91%	−	+	−	−
3	83%	−	−	−	−	91%	+	+	−	+	79%	+	+	+	+
7	83%	−	−	−	−	83%	+	+	−	−	70%	+	+	+	+
8	75%	−	−	−	−	87%	+	−	−	−	87%	+	−	−	−
10	70%	−	−	−	−	79%	+	+	−	+	75%	+	+	+	+
11	70%	+	+	+	+	95%	−	−	−	−	100%	−	−	+	−
13	70%	+	+	−	+	95%	−	−	−	−	70%	−	−	+	+
14	66%	−	−	−	−	100%	−	−	−	−	87%	−	−	−	−
15	95%	−	−	−	−	79%	−	−	−	−	66%	+	+	−	+
16	70%	+	−	+	−	100%	−	−	−	−	87%	−	−	+	−
17	100%	−	−	−	−	83%	−	+	−	−	70%	+	+	+	+
18	70%	−	+	−	−	95%	+	−	+	−	79%	+	+	+	+
19	91%	−	−	−	−	83%	−	−	−	−	58%	+	+	−	+
20	58%	+	−	−	−	70%	+	+	+	+	66%	+	−	+	+

Mean human agreement and modality classification rates

Target	**Human**	**EEG**	**Face**	**Fusion**
Arousal	75.5%	66.7%	80.0%	80.0%
Valence	89.3%	80.0%	73.3%	80.0%
Control	77.9%	80.0%	60.0%	86.7%

Table 6: Video class labelling based on human annotation. The assigned class (**C**) of each video is determined by a majority vote among the 24 human raters, where + indicates the positive and − the negative class. Predicted classes are also given by + and −, where green and red coloring means correct and incorrect predictions.**HA** stands for human agreement with the class label **C**. **FlF** stands for feature-level fusion. Also given is the mean human agreement with the class labels and the classification rate for each modality. Trials with less than 55% human agreement were left out of this study.

rating as estimated by the regressor. An R^2-score of 1 indicates a perfect agreement between the regression output and the ground truth and an R^2-score of 0 means the regressor performs the same as taking the sample mean of the ground truth as the estimate.

As can be seen from the results, except for regression of valence based on facial expression features, all results are better than taking the sample mean. In general, as was the case for classification, the EEG features perform better than the facial expression features. Unlike in the classification case, here the feature-level fusion only outperforms the single modalities in 3 of the 6 tests. This may be due to the large discrepancy in results between the EEG and face modalities. In decision-level fusion, similar to the classification case, **W-EQ**, **W-EST** and **W-REG** seem to perform the best and the inclusion of training set performance does not consistently improve the results. Compared to the classification case, decision-level fusion performs better for regression, in most cases also better than feature-level fusion.

7. Conclusions

In this paper, we explore several methods for the fusion of EEG and facial expression modalities for implicit, affective tagging. A large dataset containing recordings of 24 subjects each watching 20 video clips is utilized for evaluation of these methods. In a binary classification of arousal,

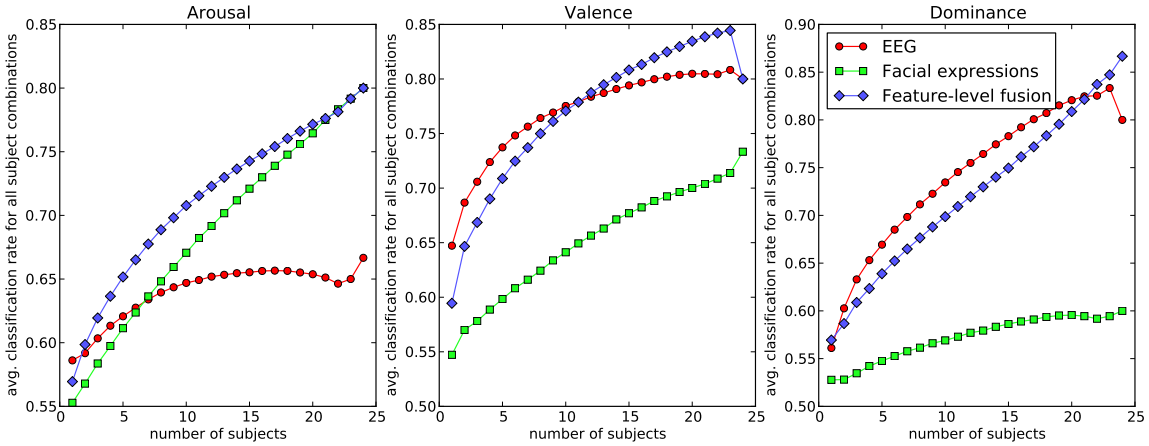

Figure 9: The number of participants plotted against the average classification rate over all possible combinations of participants.

			Arousal			Valence			Control		
	Modality	**Method**	$\mathbf{R^2}$	**MSE**	**MAE(std)**	$\mathbf{R^2}$	**MSE**	**MAE(std)**	$\mathbf{R^2}$	**MSE**	**MAE(std)**
ICA	**EEG**		.067	.076	.236(.144)	.239	.073	.221(.157)	.131	.100	.263(.175)
	Face		.041	.078	.236(.151)	.063	.090	.243(.176)	.067	.107	.286(.160)
	Feature fusion		**.086**	.075	.230(.147)	.230	.074	.223(.156)	**.186**	.094	.260(.162)
	Decision fusion	**W-EQ**	**.083**	.075	.234(.142)	**.267**	.071	.223(.144)	**.166**	.096	.266(.158)
		W-EQ^T	**.082**	.075	.234(.143)	**.273**	.070	.221(.145)	**.154**	.097	.268(.159)
		W-EST	**.074**	.076	.236(.142)	**.276**	.070	.220(.146)	**.157**	.097	.266(.162)
		W-EST^T	**.077**	.075	.235(.141)	**.271**	.070	.219(.149)	**.134**	.100	.269(.165)
		W-REG	**.083**	.075	.234(.141)	**.284**	.069	.219(.146)	**.168**	.096	.264(.161)
		W-REG^T	**.086**	.075	.234(.141)	**.279**	.069	.218(.149)	**.148**	.098	.267(.164)
		M-REG	.064	.076	.239(.140)	**.265**	.071	.222(.146)	**.144**	.098	.271(.158)
		M-REG^T	.047	.078	.241(.141)	**.252**	.072	.224(.147)	.125	.101	.274(.159)
RFE	**EEG**		.178	.067	.216(.143)	.229	.074	.222(.159)	.188	.093	.242(.187)
	Face		.011	.081	.234(.161)	-.020	.098	.261(.174)	.078	.106	.271(.180)
	Feature fusion		.134	.071	.219(.151)	.187	.078	.225(.167)	**.223**	.089	.237(.182)
	Decision fusion	**W-EQ**	**.217**	.064	.215(.134)	**.248**	.072	.224(.149)	**.268**	.084	.244(.156)
		W-EQ^T	**.235**	.063	.212(.132)	**.262**	.071	.225(.143)	**.277**	.083	.241(.158)
		W-EST	**.212**	.064	.213(.137)	**.261**	.071	.220(.151)	**.260**	.085	.244(.160)
		W-EST^T	**.224**	.063	.212(.136)	**.269**	.070	.220(.148)	**.259**	.085	.240(.166)
		W-REG	**.219**	.064	.212(.137)	**.268**	.070	.219(.151)	**.269**	.084	.242(.159)
		W-REG^T	**.231**	.063	.211(.135)	**.276**	.070	.219(.148)	**.270**	.084	.239(.165)
		M-REG	**.202**	.065	.218(.132)	**.252**	.072	.223(.149)	**.249**	.086	.249(.156)
		M-REG^T	**.190**	.066	.220(.133)	**.240**	.073	.225(.151)	**.243**	.087	.251(.156)

Table 7: R^2-score, MSE and MAE for the single modalities, feature-level fusion and decision-level fusion. R^2-scores higher than the best single-modality result are shown bold.

valence, and control affect dimensions, significant results are attained for both single modalities. A feature-level fusion approach is demonstrated to improve upon these single modality results in most cases. In addition, several methods are investigated for decision-level fusion, resulting in some improvement. Results are also reported for regression. Here, feature-level fusion did not consistently improve upon the single-modality results, though for decision-level fusion, improvements are more convincing.

While in most cases fusion improves upon single modality results and sometimes significantly so, the differences are small and the number of samples too limited to provide a definite answer on the benefits of fusion. Unfortunately, when using EEG, it is often difficult to obtain much larger sample sizes, due to the limited time participants can use the equipment before fatigue sets in and effectivity of the electrode gel becomes an issue. In addition, having participants take part in multiple sessions can degrade performance as significant differences in brain activity can occur between sessions, due to mood changes, slightly different positioning of electrodes, or even the time of day. At the

same time, it is not desirable to shorten the video duration too much, as in many cases some time is needed to set a certain mood. Nevertheless, it is advisable to try and obtain as many samples as possible, and it seems more than 20 (as in the used MAHNOB dataset), should be attainable.

In addition, it is advisable to implement a rigorous stimuli pre-selection methodology, to ensure maximum effect of the used videos. In most datasets currently available, stimuli are selected merely at the whim of the researchers compiling the dataset, and not pre-screened in any way. For the MAHNOB dataset used here, a pre-screening was performed, with participants rating the videos online. Nevertheless, the candidates to select were drawn from a relatively small pool of 21 movies. It would be a significant step forward if a large stimuli set of affective videos was available, including affective ratings, in a similar fashion to the International Affective Picture System (IAPS) [45] for photo's. Another way to improve responses to stimuli may be to show them to several subjects at the same time, as it is common for people to exhibit stronger facial responses when in a group than alone.

We have also attempted to show the potential of this method for implicit tagging when aggregating the classification results of multiple participants. For arousal, valence, and control, video tag classification rates of 80.0%, 80% and 86.7% are attained respectively when aggregating across all 24 participants. Interestingly, when aggregating results over participants, large increases in tag classification accuracy are shown, in some cases up to 30% higher. This is a compelling validation of the implicit tagging approach, where even relatively weak individual results can generate valuable and reasonably reliable tags. It would be very interesting to examine this behaviour on a larger scale, with an expanded, perhaps more specific, set of tags. Another step forward can be to try and achieve such results with equipment that can be used outside the laboratory, such as low-cost EEG headsets and webcams, and in a spontaneous setting, in order to further validate the implicit tagging paradigm. Finally, it would be interesting to investigate whether the resulting tags are truly perceived as valuable by users.

In relation to this work in particular, an interesting way forward may be to consider more closely the relation between AUs and EEG activity at specific timepoints in the videos. For instance, one could possibly weight eeg features lower when certain AUs are detected as the associated muscular acitivity introduces strong noise in the EEG. Conversely, one could utilize certain EEG artefacts to add in AU detection, for instance brow motions will likely correlate with EMG signals from frontal electrodes. Another interesting direction is to combine this method with more modalities, for instance the gaze and physiological features already present in the MAHNOB dataset. Other modalities of interest can be audio and video features retrieved from the stimuli videos and any already present tags for the stimuli video. Fusion results may well prove more convincing when using more modalities.

References

[1] A. Vinciarelli, N. Suditu, M. Pantic, Implicit Human-centered Tagging, IEEE Int'l Conf. Multimedia and Expo (2009) 1428–1431.

[2] L. von Ahn, L. Dabbish, Labeling images with a computer game, in: Proc. Conf. Human factors in computing systems, 2004, pp. 319–326.

[3] L. von Ahn, R. Liu, M. Blum, Peekaboom, in: Proc. Conf. Human factors in computing systems, 2006, pp. 55–64.

[4] M. Mandel, D. Ellis, A web-based game for collecting music metadata, Journal of New Music Research 37 (2) (2008) 151–165.

[5] Y. Kim, E. Schmidt, L. Emelle, Moodswings: A collaborative game for music mood label collection, in: Proc. Int'l Conf. Music Information Retrieval, 2008, pp. 231–236.

[6] A. D. Gerson, L. C. Parra, P. Sajda, Cortically coupled computer vision for rapid image search., IEEE Trans. neural systems and rehabilitation engineering 14 (2) (2006) 174–179.

[7] A. Kapoor, P. Shenoy, D. Tan, Combining Brain Computer Interfaces with Vision for Object Categorization, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[8] A. Kapoor, D. Tan, P. Shenoy, E. Horvitz, Complementary computing for visual tasks: Meshing computer vision with human visual processing, Proc. IEEE Conf. Face and Gesture Recognition 18 (4) (2008) 1–7.

[9] A. J. Cowell, K. Hale, C. Berka, S. Fuchs, A. Baskin, D. Jones, G. Davis, R. Johnson, R. Patch, E. Marshall, Brainwave-Based Imagery Analysis, Digital Human Modeling: Trends in Human Algorithms, Lecture Notes in Computer Science 4650 (2008) 17–27.

[10] S. Koelstra, C. Mühl, I. Patras, EEG analysis for implicit tagging of video data, in: Proc. Workshop on Affective Brain-Computer Interfaces, 2009, pp. 27–32.

[11] G. Chanel, J. J. M. Kierkels, M. Soleymani, T. Pun, Short-term emotion assessment in a recall paradigm, Int'l Journal of Human-Computer Studies 67 (8) (2009) 607–627.

[12] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, DEAP: A Database for Emotion Analysis Using Physiological Signals, IEEE Trans. Affective Computing 3 (1) (2012) 18–31.

[13] M. Soleymani, S. Koelstra, I. Patras, T. Pun, Continuous Emotion Detection in Response to Music Videos, in: IEEE Int'l Conf. Automatic Face & Gesture Recognition and Workshops, 2011, pp. 803–808.

[14] K. Takahashi, A. Tsukaguchi, Remarks on emotion recognition from multi-modal bio-potential signals, in: IEEE Int'l Conf. Systems, Man and Cybernetics, Vol. 2, 2003, pp. 1654–1659.

[15] J. K. Olofsson, S. Nordin, H. Sequeira, J. Polich, Affective picture processing: an integrative review of ERP findings., Biological psychology 77 (3) (2008) 247–65.

[16] B. Cuthbert, H. Schupp, M. Bradley, N. Birbaumer, P. Lang, Brain potentials in affective picture processing: covariation with autonomic arousal and affective report, Biological psychology 52 (2) (2000) 95–111.

[17] J. Polich, Updating P300: an integrative theory of P3a and P3b., Clinical neurophysiology 118 (10) (2007) 2128–48.

[18] M. Pantic, L. J. M. Rothkrantz, Automatic analysis of facial expressions - the state of the art, IEEE Trans. Pattern Analysis and Machine Intelligence 22 (12) (2000) 1424–1445.

[19] M. Pantic, M. S. Bartlett, Machine Analysis of Facial Expressions, I-Tech Education and Publishing, Vienna, Austria, 2007.

[20] Z. Zeng, M. Pantic, G. I. Roisman, T. S. Huang, A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions, IEEE Trans. Pattern Analysis and Machine Intelligence 31 (1) (2009) 39–58.

[21] P. Ekman, W. V. Friesen, J. C. Hager, The Facial Action Coding System: A Technique for the Measurement of Facial Movement, A Human Face, Salt Lake City, UT, 2002.

[22] P. Ekman, E. L. Rosenberg, What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS), Oxford University Press, 2005.
[23] A. C. Williams, Facial expression of pain: An evolutionary account, Behavioral and Brain Sciences 25 (4) (2002) 439–488.
[24] R Kaliouby, Real-time inference of complex mental states from facial expressions and head gestures, in: B. Kisačanin, V. Pavlović, T. Huang (Eds.), Real-time vision for human-computer interaction, Springer US, 2005, pp. 181–200.
[25] H. Gunes, B. Schuller, M. Pantic, R. Cowie, Emotion representation, analysis and synthesis in continuous space: A survey, Proc. IEEE Conf. Face and Gesture Recognition (2011) 827–834.
[26] J. Russell, A circumplex model of affect, Journal of personality and social psychology 39 (6) (1980) 1161–1178.
[27] D. McDuff, R. E. Kaliouby, Affect valence inference from facial action unit spectrograms, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2010, pp. 17–24.
[28] M. Hussein, T. Elsayed, Studying Facial Expressions as an Implicit Feedback in Information Retrieval Systems (2008).
[29] I. Arapakis, I. Konstas, J. M. Jose, Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance, in: Proc. ACM Int'l Conf. Multimedia, 2009, pp. 461–470.
[30] J. Jiao, M. Pantic, Implicit image tagging via facial information, in: Proc. Int'l Workshop on Social Signal Processing, 2010, pp. 59–64.
[31] J.-S. Lee, C. H. Park, Robust audio-visual speech recognition based on late integration, IEEE Trans. Multimedia 10 (5) (2008) 767–779.
[32] P. K. Atrey, M. A. Hossain, A. El Saddik, M. S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, Multimedia Systems 16 (6) (2010) 345–379.
[33] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, A multimodal Affective Database for Affect Recognition and Implicit Tagging, IEEE Trans. Affective Computing 3 (1) (2012) 42–55.
[34] N. Sebe, I. Cohen, T. Huang, Multimodal emotion recognition, Handbook of Pattern Recognition and Computer Vision (2005) 981–256.
[35] M. M. Bradley, P. J. Lang, Measuring emotion: the Self-Assessment Manikin and the Semantic Differential., Journal of Behavior Therapy and Experimental Psychiatry 25 (1) (1994) 49–59.
[36] J. D. Morris, Observations: SAM: The Self-Assessment Manikin; An Efficient Cross-Cultural Measurement of Emotional Response, Journal of Advertising Research 35 (8) (1995) 38–63.
[37] S. Koelstra, M. Pantic, I. Patras, A dynamic texture-based approach to recognition of facial actions and their temporal models., IEEE Trans. Pattern Analysis and Machine Intelligence 32 (11) (2010) 1940–54.
[38] K. Scherer, What are emotions? And how can they be measured?, Social science information 44 (4) (2005) 695–729.
[39] A. Bell, T. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural computation 7 (6) (1995) 1129–1159.
[40] A. Hyvärinen, Fast and Robust Fixed-Point Algorithms for Independent Component Analysis, IEEE Trans. Neural Networks 10 (3) (1999) 626–634.
[41] D. Mackay, Bayesian Interpolation, Neural computation 4 (3) (1992) 415–447.
[42] S. Koelstra, A. Yazdani, M. Soleymani, C. Muehl, J. Lee, A. Nijholt, T. Pun, T. Ebrahimi, I. Patras, Single Trial Classification of EEG and Peripheral Physiological Signals for Recognition of Emotions Induced by Music Videos, Proc. Brain Informatics (2010) 89–100.
[43] R. J. Barry, A. R. Clarke, S. J. Johnstone, C. A. Magee, J. A. Rushby, EEG differences between eyes-closed and eyes-open resting conditions, Clinical Neurophysiology 118 (12) (2007) 2765–2773.
[44] J. Onton, S. Makeig, High-frequency Broadband Modulations of Electroencephalographic Spectra, Frontiers in Human Neuroscience 3 (61) (2009) 1–18.
[45] P. Lang, M. Bradley, B. Cuthbert, International affective picture system (IAPS): Affective ratings of pictures and instruction manual, Tech. Rep. A-8, University of Florida, USA (2008).