Implicit Media Tagging and Affect Prediction from RGB-D video of spontaneous facial expressions

Daniel Hadar[13], Talia Tron[2] and Daphna Weinshall[3]
[1] Department of Cognitive Science, Hebrew University of Jerusalem, Israel
[2] Center for Brain Science (ELSC), Hebrew University of Jerusalem, Israel
[3] School of Computer Science and Engineering, Hebrew University of Jerusalem, Israel

***Abstract*— We present a method that automatically evaluates emotional response from spontaneous facial activity. The automatic evaluation of emotional response, or affect, is a fascinating challenge with many applications. Our approach is based on the inferred activity of facial muscles over time, as automatically obtained from an RGB-D video recording of spontaneous facial activity. Our contribution is two-fold: First, we constructed a database of publicly available short video clips, which elicit a strong emotional response in a consistent manner across different individuals. Each video was tagged by its characteristic emotional response along 4 scales: Valence, Arousal, Likability and Rewatch (the desire to watch again). The second contribution is a two-step prediction method, based on learning, which was trained and tested using this database of tagged video clips. Our method was able to successfully predict the aforementioned 4 dimensional representation of affect, achieving high correlation (0.87-0.95) between the predicted scores and the affect tags. As part of the prediction algorithm we identified the period of strongest emotional response in the viewing recordings, in a method that was blind to the video clip being watched, showing high agreement between independent viewers. Finally, inspection of the relative contribution of different feature types to the prediction process revealed that temporal facets contributed more to the prediction of individual affect than to media tags.**

I. INTRODUCTION

The common belief that *"The face is a picture of the mind"* (Cicero c. 63 BC) inspired the computer vision community to find ways to objectively evaluate people's emotional state from their facial expressions. In accordance, in the past two decades we witnessed an ever increasing interest in automatic methods that extract and analyze human facial expressions in order to predict their *emotional state* (e.g. [13]) or provide *media tagging* (e.g. [15]).

Media tagging has become an integral part of the Internet environment in recent years. Many web platforms allow (and even encourage) users to label their content by using keywords or designated scales (e.g. Facebook's reactions, see Fig. 1). As opposed to *explicit* tagging, in which the user is actively involved in the tagging process, *implicit* tagging is done passively and relies only on the normal interaction between the user and the stimulus (e.g. watching a video clip). As such, it requires less effort and is less prone to biases. In parallel, it has been suggested that explicit tagging tends to be rather inaccurate in practice. For example, users tend to tag videos according to their social needs, which yields tagging that could be reputation driven, especially in a setup where the user's friends, colleagues or family may be exposed to the tags [19].

Fig. 1: Facebook's reactions.

Affect computation from facial expressions cannot be readily separated from the underlying theories of emotional modeling, which rely on assumptions made by researchers in the fields of psychology and sociology, and in some cases are still under debate. The relevant theoretical background and previous work are thoroughly reviewed in [8]. Briefly, the Facial Action Coding System (FACS) is often used to analyze facial activity in a quantitative manner. FACS gives a score to the activity of individual facial muscles called Action Units (AUs) based on their intensity level and temporal segments [6] (see Fig. 2).

In our work we took advantage of the emerging technology of depth cameras, which are commonly used for human computer interaction in gaming, in order to achieve better affect recognition and media tagging. Specifically, we used a depth camera (Carmine 1.09) to record participants' spontaneous facial activity when viewing a set of video clips designed to elicit a reliable emotional response. We then developed two types of pertinent prediction models for the automatic quantitative evaluation of affect from facial expressions, including models for the tagging of video clips (i.e. *implicit media tagging*) and models for the prediction

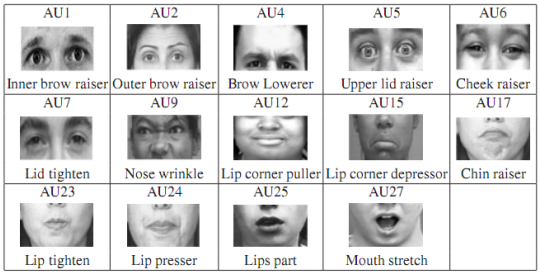


Fig. 2: Examples of Action Units from FACS.

This research was funded by the Intel Collaborative Research Institute for Computational Intelligence (ICRI-CI).

Appeared in: Proc. 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG2017), Washington DC, May 2017.

of viewers emotional state (i.e. *affect prediction*). While focusing on emotion-related indicators based on non-verbal cues, the two types of prediction models differ in their target: *media tagging* aims to predict attributes of *the multimedia stimuli*, while *affect prediction* aims to evaluate a person's *individual emotional response*.

In order to infer descriptive ratings of emotion, we employed the *dimensional approach* [7] to modeling emotion. In accordance, we represented emotion by 2 key scales – *Valence* and *Arousal*, and 2 additional contemporary scales – *Likability* and *Rewatch* (the desire to watch again) – which are more suitable for modern uses in HCI and media tagging.

Our method is based on learning from a tagged database of video clips. In order to train a successful algorithm and be able to test it against some meaningful ground truth, we needed a database of video clips that can invoke strong emotional responses in a consistent manner across individuals. We could not find a reliable database of video clips that suited our needs (see reasoning in [8]), and therefore we constructed a new database from publicly available video clips as described in Section II-B. This database is available to the community[1], including all necessary links to the video clips and the corresponding emotional tags. A similar empirical procedure was used to collect data for the training and evaluation of our method, as described in Section II.

In order to predict emotional tags from RGB-D recordings of spontaneous facial activity, we used the commercial software Faceshift© [1] . Faceshift extracts quantitative measures of over 50 FACS action units. Subsequently we computed a vector representation for each recording based on these measures. The procedure by which we have obtained a concise representation for each video of facial expressions, which involved quantification of both dynamic and spatial features of the AUs activity, is described in Section III-A.

In Section III we further describe the method by which we learned to predict affect from the ensuing representation for each viewer and each viewed clip. In the first step of our method we generated predictions for small segments of the original facial activity video, employing linear regression to predict the 4 quantitative scales which describe affect (namely *Valence*, *Arousal*, *Likability* and *Rewatch*, *a.k.a. VALR*). In the second step we generated a single prediction for each facial expressions video, based on the values predicted in the first step for the set of segments encompassed in the active part of the video. The results are described in Section IV, showing high correlation between the predicted scores and the actual ones.

Prior work

Tagging of media implicitly and predicting viewers affective state based on data obtained from depth cameras is a recent and uncharted territory. A few exceptions include Niese et al. [12] who used 3D information from the raw data obtained by a 2D camera, and Tron et al. [16], [17] who used depth cameras to classify the mental state of schizophrenia patients based on their facial activity.

[1] cs.huji.ac.il/~daphna/Video_tagging/index.html

Implicit Media Tagging: most previous work that used facial expressions for implicit media tagging combined it with additional input, such as content from the video itself or additional physiological measures. For example, Zhao et al. [22] used only facial features to predict categories of movies such as comedy, drama and horror. Bao et al.[2] added the user's acoustic features, motion and interaction with the watching device (tablet or mobile). Wang et al. [20] used head movement as well as facial expressions, combined with "common emotions".

Affect prediction: previous work often used several types of inputs – some physiological (*e.g.* facial expressions, brain activity or acoustic signals) and some derived from the media the participant was exposed to (*e.g.* the video clip's tagging or visual and acoustic features). Moreover, they typically relied on the participant's reported emotional state. Examples include McDuff et al. [11] who used a narrow set of facial features (smiles and their dynamics) to predict a participant's likability response and desire to watch again. Bargal et al. [3] was among the few to propose a model based only on facial expressions. Interestingly, Soleymani et al. [14] reported that results when using facial expressions are superior to results when using EEG.

A more in-depth review of prior work is provided in [8].

II. METHOD

The data collection in our study had two phases: (i) *Database construction:* collect and evaluate a suitable database of video clips which elicit strong and consistent emotional responses in the viewers (Section II-B). (ii) *Emotion prediction and analysis:* record people's spontaneous facial expressions when viewing these clips (Section II-C). The same empirical procedure, described in Section II-A, was used in both phases.

A. Experimental Design

Data collection was carried out in a single room, well lit with natural light, with minimal settings (2 tables, 2 chairs, a white-board, 2 computers and no pictures hanging on the walls). Participants were university students recruited using banners and posters. Thus 52 volunteers between the ages 19-29 ($\mu = 23.4$) with normal vision participated in this study (13 males and 13 females in phase 1 and 14 males and 12 females in phase 2), for which they received a small payment.

Each data collection session consisted of the following stages (see Fig. 3):

1) A fixation cross was presented for 5 seconds, and the participant was asked to gaze at it.
2) A video clip was presented.
3) The participant verbally described his subjective emotion to the experimenter, using two sentences at most.
4) The participant rated her subjective feelings on a preprinted ratings paper (following [5]).

5 discrete scales were used for rating: *Valence, Arousal, Likability*, desire to watch again (*Rewatch*) and *Familiarity*. Specifically, we used SAM Manikins for Valence and

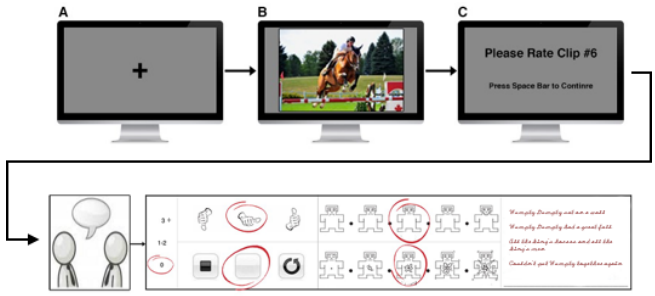


Fig. 3: The Experimental design.

Arousal [10], the liking scale for Likability [9], and self-generated scales for Rewatch and Familiarity. The SAM Manikins method was chosen because it is known to be comprehensive, parsimonious, inexpensive and quick [4]. After every 4 trials, a visual search task was presented (*"Where's Waldo?"*) in order to keep the participants focused, while forcing them to change their head situs, sitting position and focal length. The clips order was randomized and the entire procedure lasted about an hour. We encouraged participants to rate the clips according to their perceived, inner and subjective emotion.

B. Phase 1: Database Construction

We defined the following criteria which should be met by a suitable database of video clips:

1) **Duration.** Clips must be relatively short in order to: (i) avoid eliciting several distinct emotions at different times of the clip, and (ii) make the use of many clips in a single experimental session possible. That being said, the clips should be long enough to elicit a strong clear emotional response.
2) **Diversity.** Clips should be taken from a variety of domains to reduce the effect of individual variability. Clearly the database should not contain only clips of cute cats in action or incredible soccer tricks, but a balanced mixture.
3) **Familiarity.** Clips should be such that uninformed viewers are not likely to be familiar with them while still being publicly available. Moreover, in order to be as universal as possible the clips should not contain any significant dialogue.

Over 110 clips were initially selected from online video sources. We attempted to achieve a diverse set of unfamiliar clips, and therefore focused on lightly viewed ones. We excluded clips that might offend participants by way of pornography or brutal violence (following the YouTube Community Guidelines). Eventually 36 clips were selected. These clips were manually curtailed to remove irrelevant content, scaled to fit a 1440 x 900 resolution, and balanced to achieve identical sound volume.

The familiarity scale indicated that all 36 clips were sufficiently unfamiliar to our subjects ($p < .0001$) irrespective of gender. The remaining 4 scales of *Valence, Arousal, Likability* and *Rewatch* were used to emotionally tag the clips for the next phase of the study. The subjective scores of all raters were averaged for each clip; the average, median, standard deviation and range for each scale across all clips can be found in Fig. 4 and Table I.

TABLE I: Average, median, std and range for the 4 VALR scores over all clips.

	average	median	std	min	max
Valence	3.04	3.21	1.00	1.23	4.42
Arousal	3.08	3.02	0.65	1.73	4.19
Likability	2.02	2.04	0.56	1.12	2.92
Rewatch	1.73	1.75	0.44	1.08	2.54

Rate reliability: As can be seen in Table II, ratings on all scales showed high inter-rater agreement with an average Intra-Correlation Coefficient (ICC) of 0.945. In addition, there were strong correlations between several scales, most notably *Valence-Likability* (Pearson $R = 0.92$), *Valence-Rewatch* ($R = 0.87$) and *Likability-Rewatch* ($R = 0.94$). Interestingly, no significant correlation was found between *Arousal* and *Likability* ($R = -0.23$) or between *Arousal* and *Rewatch* ($R = -0.04$). A small negative correlation was found between *Valence* and *Arousal* ($R = -0.40$), possibly because clips with extremely high V-A values that mostly included pornographic content were excluded. Almost no clips had low ratings for both *Valence* and *Arousal*, probably because melancholic clips tended to elicit high arousal.

C. Phase 2: Collecting data for affect recognition

18 of the 36 clips were selected from the database; aiming for a diverse corpus, we chose clips whose elicited response spanned the spectrum of VALR as uniformly as possible.

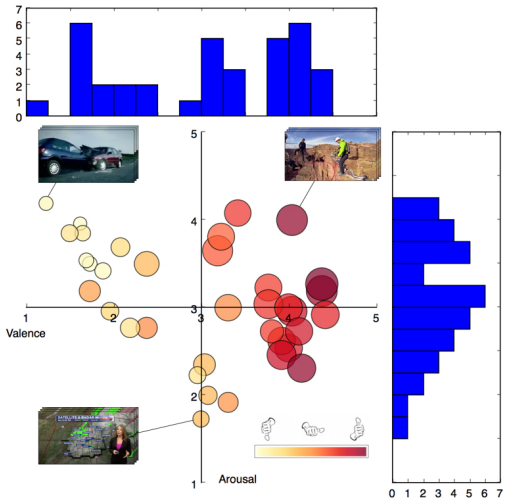


Fig. 4: Distribution of 4 subjective scores over all tested clips, where *Valence* and *Arousal* define the two main axes, also summarized in histogram form above and to the right of the plot. Size and color correspond respectively to the remaining 2 scores of *Rewatch* and *Likability*.

TABLE II: Inter-rater agreement (ICC) on clips ratings.

	Score	*Lower Bound*	*Upper Bound*
Valence	.975	.961	.958
Arousal	.926	.886	.957
Likability	.954	.930	.973
Rewatch	.927	.888	.975

The facial activity of each participant was recorded during the entire procedure, using an $RGB-D$ camera (Carmine 1.09). Participants were informed of being recorded and signed a consent form. Most of the participants (21 of the 26) reported that they were not affected by the recording, and that in fact they had forgotten about being recorded. Moreover, the subjective reports in the second phase on all 4 scales had a very similar distribution to the reports in the first phase (*Valence* ($R = .98,\ p < .0001$), *Arousal* ($R = .90,\ p < .0001$), *Likability* ($R = .97,\ p < .0001$), *Rewatch* ($R = .95,\ p < .0001$)). We therefore believe that the video affect tagging obtained in phase 1 remain a reliable predictor of the emotional response elicited in phase 2 as well.

III. Prediction Models

Below we describe our approach to affect prediction. In Section III-A we describe the features used to represent each RGB-D recording clip. In Section III-B,III-C we describe the prediction algorithm and how it was used to construct the 3 different prediction models.

A. Facial activity extraction

Previous work in this field calculated facial features either by extracting raw movement of the face without relating to specific facial muscles (e.g. [20]), or by extracting the activity level of a single or a few muscles (e.g. [18], [21]). In this work we extracted the *intensity signals* of over 60 AUs and facial gestures; this set was further analyzed manually to evaluate tracking accuracy and noise levels. Eventually 51 AUs were selected to represent each frame in the clip for further analysis and learning, including eyes, brows, lips, jaw and chin movements (see example in Fig. 5).

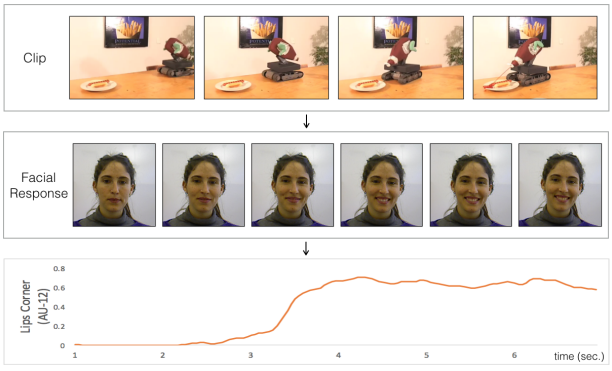
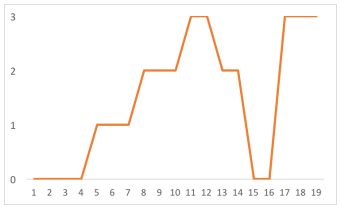


Fig. 5: Facial response (middle row) to video clip (illustrated in the top row), and the time varying intensity of AU12.



(a)

	0	1	2	3
0	4	1	0	1
1	0	2	1	0
2	1	0	3	1
3	0	0	1	3

(b)

Fig. 6: (a) Quantized AU signal ($K = 4$), and (b) its corresponding transition matrix. The number of frames labeled *0,1,2,3* is 6,3,5,5 respectively. It follows that $ActivationRatio = \frac{13}{19}$, $ActivationLength = \frac{7}{19}$ and $ActivationLevel = 1.473$, and $ChangeRatio = \frac{6}{18}$, $SlowChangeRatio = \frac{4}{18}$, $FastChangeRatio = \frac{2}{18}$.

Using this *intensity level* representation, which provided a time series of vectors in $\mathbb{R}^{51}$, we computed higher order features representing the facial expression more concisely. This set of features can be divided into 4 types: *Moments*, *Discrete States*, *Dynamic* and *Miscellaneous*.

- **Moments.** The first 4 moments (mean, variance, skewness and kurtosis) were calculated for each AU in each facial video recording.
- **Discrete State Features.** For each AU separately, the raw intensity signal was quantized over time using *K-Means* ($K = 4$), and the following four facial activity characteristic features were computed (see Fig. 6 for an example):
  - ***Activation Ratio:*** Proportion of frames with any AU activation.
  - ***Activation Length:*** Mean number of frames for which there was continuous AU activation.
  - ***Activation Level:*** Mean intensity of AU activation.
  - ***Activation Average Volume:*** Mean activation level of all AUs, computed once for each facial response video.
- **Dynamic Features.** A transition matrix M was generated, measuring the number of transitions between the different levels described above, and the following three features were calculated for each AU based on M (see Fig. 6 for an example):
  - ***Change Ratio:*** Proportion of transitions.
  - ***Slow Change Ratio:*** Proportion of 1-quantum changes.
  - ***Fast Change Ratio:*** Proportion of 2 or more quantum changes.
- **Miscellaneous Features**, including the number of smiles and blinks in each facial response[2].

B. Highlight Period

In most clips, during most of the viewing time, participants' facial activity showed almost no emotional response.

[2]The number of smiles was calculated by counting peaks in the signals of both lip corners, where *peak* was defined as a local maximum higher by at least 0.75 as compared to its neighboring points. The amount of blinks was calculated in a similar manner, with a threshold of 0.2.

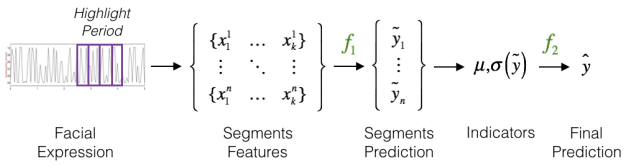


Fig. 7: Illustration of the two-step prediction algorithm.

We therefore computed facial activity features from the AU signals only during the time frame of each facial recording during which facial activity was significant. We implemented a model that localized the *highlight period* solely from the viewer's facial expression, in a technique that was blind to the video clip.

Specifically, for each participant and clip, our model received 51 AU intensity levels for the whole clip's duration. The method isolated the activity of the most informative AUs (namely, smiles, blinks, mouth dimples, lips stretches and mouth frowns), and then localized the 6-seconds window in which these AUs achieved maximal average intensity and variance. This window was fixed as the highlight period; excluding moments, all features were subsequently computed based only on the highlight period. Notice that for some clip C_i, the highlight period can be different for different subjects (although its duration was fixed at 6 seconds).

C. Prediction algorithm (continuous score)

Our algorithm predicts an affective score (in VALR terms) when given as input a facial expression recording, represented by a vector in $\mathbb{R}^d$ feature space ($d = 462$). Since the number of participants in our study was only 26, a clear case of small sample, the full vector representation – if used for model learning – would inevitably lead to overfit and poor prediction power. We therefore started by significantly reducing the dimensionality of the initial representation of each facial activity recording using PCA. Our final method employed a two-step prediction algorithm (see illustration in Fig. 7), as follows:

1) **First step.** After the highlight period of each clip was detected (for each subject), it was divided into n fixed size overlapping segments. A feature vector was calculated for each segment, and a linear regression model (f_1) was trained to predict the 4-dimensional affective scores vector of each segment.
2) **Second step.** Two indicators (mean and standard deviation) of the set of predictions over all segments in the clip were calculated, and another linear regression model (f_2) was trained to predict the 4-dimensional affective scores from these indicators.

Several parameters controlled the final representation of each facial expression clip, including the number of segments, the length of each segment, the overlap between the segments, the final PCA dimension and whether PCA was done over each feature type or over all features combined. The values of these parameters were determined using cross-validation: given a training set of l points, the training and prediction process was repeated l times, each time using $l-1$ points to train the model and predict the value of the left out point. The set of parameters which achieved the best results over these l cross validation repetitions was used to construct the facial expression representation of all data points.

D. Binary prediction

We note that our prediction method provides a continuous score, while the viewer's subjective report is given using a discrete scale. Similarly, the media is tagged using a continuous scale (as these are average ratings). This discrepancy may affect the correlation between the scores, which is how we measure success. In addition, often what is needed in real-life applications is a discrete binary prediction rather than a continuous score, in order to indicate, for example, whether the viewer liked a video clip (or not). We therefore generated an additional binary score from the continuous scores produced by our method, using the mean predicted score as a threshold. To measure success, we binarized the continuous media tag in the same manner, using as threshold the mean media tag.

E. predictive models

We learned three different models that shared the aforementioned mechanism but varied in their prior knowledge and target: the first two models were trained to predict the *clip's affective rating* as stored in the database (IMT), while the third was trained to predict the *viewer's subjective affective state* for each individual (AP).

- **Implicit media tagging of unseen clips (IMT-1).** This model was built using the facial expressions of a single viewer. Given the facial response of this viewer to a *new clip*, the model predicts the *clip's affective score*.
- **Implicit media tagging of unseen clips via multiple viewers (IMT-2).** Given the facial response of a set of familiar viewers to a *new clip*, the model predicts the *clip's affective score*. Generalizing the first model, this second model predicts the new clip's affective score by averaging over the predictions of *all* viewers.
- **Viewer's affect prediction for an unseen clip (AP-1).** This model was built using the facial expressions of a single viewer. Given the facial response of this viewer to a *new clip*, the model predicts the *viewer's subjective affective state*.

IV. RESULTS AND ANALYSIS

To evaluate the predictive power of our models, we divided the set of recordings following a Leave-One-Out procedure. Specifically, we trained each model based on $n-1$ clips ($n = 18$) and tested our prediction on the clip which was left out. The results are described in Section IV-A. In Section IV-B we analyze the relative importance of the different feature types (moments, discrete state features, dynamic features and miscellaneous features). Finally, in Section IV-C we analyze the localization of the highlight period.

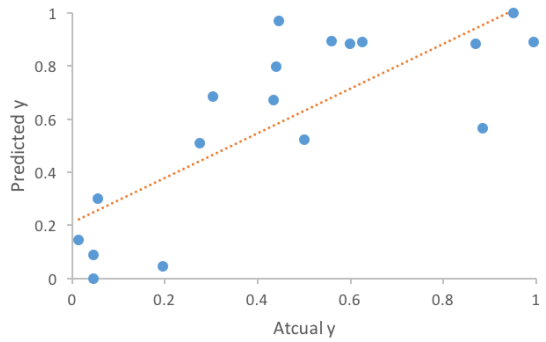


Fig. 8: Correlation example between predicted and actual ratings of a single viewer's valence score (*R=0.791*).

A. Learning Performance

Learning performance was evaluated by *Pearson's R* between the actual VALR scores and the models' predicted ones (see example in Fig. 8). Table III shows the average VALR results over all clips and viewers (all correlations are significant with $p < 0.0001$).

TABLE III: Mean (and std) of Pearson's R between the predicted and actual scores for the 3 prediction models.

	Valence	**Arousal**	**Likability**	**Rewatch**
IMT-1	.752 (.14)	.728 (.07)	.637 (.22)	.661 (.15)
IMT-2	.948 (.22)	.874 (.22)	.951 (.17)	.953 (.19)
AP-1	.661 (.17)	.638 (.19)	.380 (.16)	.574 (.19)

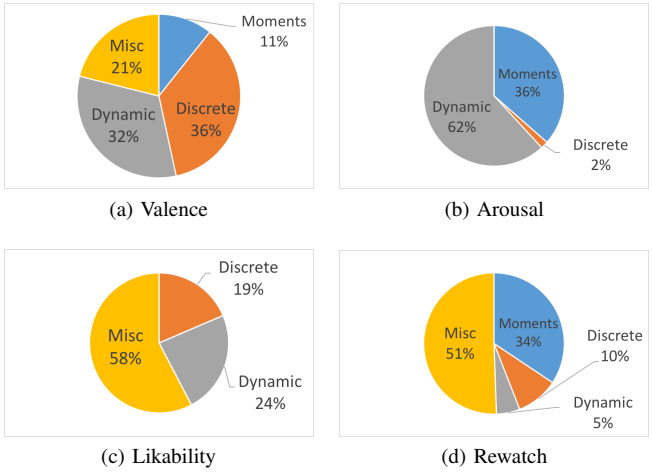
Binary predictions were evaluated by measuring accuracy. Since the scores around the mean are rather ambiguous, we eliminated from further analysis the clips whose original tag was uncertain. This included clips with scores in the range $\mu \pm \sigma$, where μ denotes the average score over all clips and σ its standard deviation, thus eliminating 15% on average of all data points. The results can be found in Table IV.

B. Relative Importance of Features

We analyzed the relative importance of the different facial features for IMT-1, observing that different facial features contributed more or less, depending on the affective scale being predicted. The contribution weight of each set of features was calculated by training the models as described above using only a single type of features at each time, and computing the corresponding success rate. These weights were subsequently normalized, see results in Fig. 9. We see that for IMT-1 the prediction of *Valence* relied on all 4 feature types (including moments, discrete state features, dynamic features and miscellaneous features), while the prediction of *Arousal* didn't use the miscellaneous features at all, but relied heavily on the dynamical aspects of the facial expression. Similarly, the prediction of *Likability* utilized mostly the miscellaneous features, and did not use moments features at all. In isolation, prediction with only miscellaneous features achieved correlation of $R = 0.275$ with the *Likability* score, and $R = 0.287$ with the *rewatch* score.

TABLE IV: Accuracy (std) of the derived binary measures.

	Valence	**Arousal**	**Likability**	**Rewatch**
IMT-1	71% (.21%)	56% (.25%)	68% (.18%)	73% (.18%)
IMT-2	94% (.21%)	90% (.20%)	91% (.23%)	91% (.24%)
AP-1	70% (.24%)	53% (.41%)	49% (.37%)	49% (.55%)

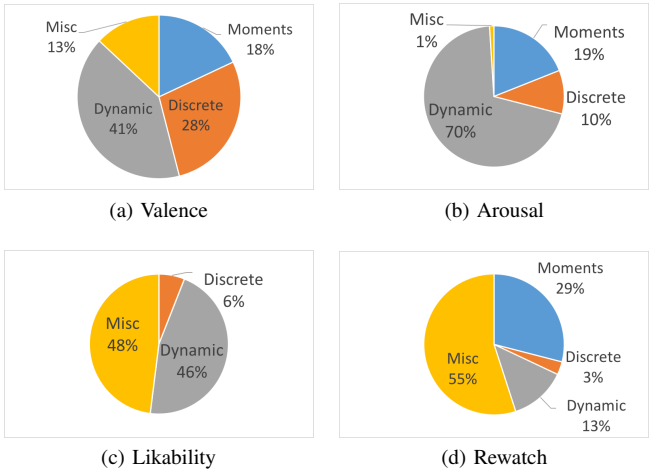


(a) Valence (b) Arousal

(c) Likability (d) Rewatch

Fig. 9: The relative contribution of features to IMT-1.

For comparison, we analyzed the relative importance of the different facial features for the analogous affect prediction algorithm AP-1. As can be seen in Fig. 10, the distribution is similar in both models, except that relative importance of the dynamical features is consistently higher in affect prediction ($+11.75\%$ on average).



(a) Valence (b) Arousal

(c) Likability (d) Rewatch

Fig. 10: The relative contribution of features to AP-1.

C. Localization of Highlight Period

We also analyzed the localization of the response highlight period (HP) within the clip. Although this period was computed bottom-up from the facial recording of each individual viewer and without access to the observed video clip, the correspondence between subjects was notably high ($ICC =$

.941). Not surprisingly, the beginning of the period was usually found a few seconds before the clip's end, and in some clips it lasted after the clip ended (specifically in 8 out of the 18 clips). Yet the HP localization clearly depended, in a reliable manner across viewers, on the viewed clip. For example, when viewing a car safety clip, the average HP started 14 seconds before its end, probably because a highly unpleasant violent sequence of car crashes had began a second earlier. We may conclude that the HP tends to focus around the clip's end most of the time and with very high agreement between viewers, but clip-specific analysis is preferable in order to localize it more precisely. The distribution of HPs is presented in Fig. 11.

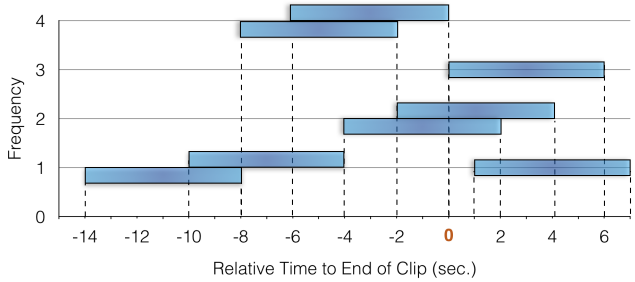


Fig. 11: Histogram of HPs relative to the clips' end time, which marks the origin of the $X-$axis ($\mu = -7.22$, $\sigma = 4.14$).

D. Discussion of results

Our prediction results (Section IV-A), obtained by the most powerful media tagging (IMT-2) model, are much better than those obtained by published state-of-the-art methods (e.g. [11]), and specifically the ability to tag media in VALR terms with accuracy rates ranging around 90% (Table IV) is unprecedented. That being said, it is important to note that our results were obtained using a newly composed database; since previous methods are for the most part not publicly available, it is not possible to evaluate their performance on this new database. While the difficulty of the tasks used in the evaluation of earlier methods does not seem to be higher than our benchmark task, this observation should be further verified. Finally, we note that these methods did not have access to our training data.

Our results notably show that media tagging (IMT) models achieve higher success rates than the affect prediction (AP-1) model. In particular, although IMT-1 and AP-1 both predict an affective score when given a single (familiar) viewer's facial response to a new clip, the first yielded noticeably higher success rates (average difference between methods $\mu = .131$). These results imply that it is easier (and more reliable) to predict the *expected* affective score from the viewer's facial expressions than it is to predict the viewer's own *reported* score. This observation is rather surprising, as the opposite may seem more plausible – that one's facial behavior will be better matched with her *subjective* emotional score than the score averaged over many different viewers.

TABLE V: Mean error of all viewers subjective scores and the predictions of AP-1, IMT-1 and IMT-2.

	Valence	**Arousal**	**Likability**	**Rewatch**
AP-1	.265	.370	.388	.371
IMT-1	.214	.262	.325	.341
IMT-2	.179	.239	.295	.306

Moreover, although seemingly AP-1 yields rather accurate predictions of affect, further inspection reveals that the predicted media tag is an even better predictor of individual affect. This is shown in Table V, which gives the mean error (ME) between the actual subjective affect scores and the predictions given by the 3 methods: AP-1, IMT-1 and IMT-2; clearly IMT-2 is most accurate, and AP-1 least accurate. Thus, if the media tags are known, it's preferable to train a model to predict these tags (IMT-1) and rely on this model alone, rather than on a model trained to predict the viewer's subjective scores. Moreover, if other viewers' data is also available, the predictions of IMT-2 give even more accurate predictors of the viewer's subjective scores. Thus, the AP-1 model is only valuable when it is not possible to perform non-specific media tagging of the viewed clip, e.g., in situations when it is not known what the viewer is looking at (as in most real life situations, like an interview).

Comparing the results reported in Fig. 9 and Fig. 10, we hypothesize that the temporal aspects of viewers' facial activity, while providing a good predictor for emotions in general, can also be used to distinguish between different viewers. This idea is in line with [17], where it was shown that these aspects are the most discriminative when comparing schizophrenia patients and healthy individuals.

V. SUMMARY AND DISCUSSION

The contribution of this work is two fold. First, we generated a database of video clips which give rise to strong predictable emotional responses, as verified by an empirical study, and made it available to the community. Second and more importantly, we described several algorithms that can predict emotional response based on spontaneous facial expressions, recorded by a single depth camera. Our method provided fairly accurate predictions for 4 scores of affective state: *Valence, Arousal, Likability*, and *Rewatch* (the desire to watch again). We achieved high correlation between the predicted scores and the affective tags assigned to the video. Moreover, our models are based directly on the automatically inferred activity of facial muscles over time, considering dozens of muscles, in a method which is blind to the actual video being watched by the subject (*i.e.* the model isn't aware of any of the stimuli details or attributes).

As expected, our results show that better prediction of media tag can be achieved using a group of viewers rather than a single viewer (compare the performance of the two media tagging methods, IMT-1 which used a single viewer, and IMT-2 which used a group of viewers). Hence, in real-life systems, it's preferable to rely on the facial behavior of a group of known viewers, if possible.

When using facial expressions for automatic affect prediction, we saw that it's easier to predict the *expected* affective state (*i.e.* implicit media tagging) than the viewer's *reported* affective state (*i.e.* affect prediction). Under the assumption that the participants in this study attempted to report their veridical emotional experience (as requested), and that human facial expressions are more faithful to one's true emotional state than their verbal report, this result may imply that media tags – as obtained by averaging over the reports of many viewers – provide a better predictor of one's emotional response to the video than the viewer's own report. These findings may appear rather surprising, but they stand in agreement with theories of emotional self-report bias. This bias could arise from many factors, including cultural stereotypes, or the wish to conform with the acceptable expected response.

Interestingly, when computing the period of strongest response in the viewing recordings, we saw high agreement between the different viewers. This tells us that facial expressions by themselves may be used to find the emotional peak of video clips, without the need of explicit report.

Further analysis revealed that different types of facial features are useful for the prediction of different scores of emotional state. For example, we saw that simply counting the viewer's smiles and blinks provided an inferior, yet significantly correlated, prediction of *Likability* and *Rewatch*. For commercial applications, these facial features can be obtained from the laptop's embedded camera. Furthermore, we found that the dynamical aspects of facial expressions contributed more to the prediction of viewers affective state than to the prediction of media tags.

In conclusion, the staggering amount of videos currently available poses many challenges to computer scientists and engineers. Since every user can upload videos as he or she desires, it becomes essential to be able to classify and annotate these videos for accurate and quick retrieval, ease of use, search engines and recommendation systems. It is therefore essential to develop assistive technology for the automatic tagging of video, by assigning descriptive labels that aid indexing and cataloging. With emotional tagging, we may want to predict the viewers' expected emotional response, to refine personal customization and to comprehend the effect of the video on the users. Automatic analysis of facial expressions, as demonstrated in this study, may be used to advance the state of the art in these domains.

REFERENCES

[1] http://www.faceshift.com/.

[2] X. Bao, S. Fan, A. Varshavsky, K. Li, and R. Roy Choudhury. Your reactions suggest you liked the movie: Automatic content rating via reaction sensing. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 197–206. ACM, 2013.

[3] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang. Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 433–436. ACM, 2016.

[4] M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[5] S. Carvalho, J. Leite, S. Galdo-Álvarez, and Ó. F. Gonçalves. The emotional movie database (emdb): A self-report and psychophysiological study. *Applied psychophysiology and biofeedback*, 37(4):279–294, 2012.

[6] P. Ekman, W. V. Friesen, and J. C. Hager. Facial action coding system (facs). *A technique for the measurement of facial action. Consulting, Palo Alto*, 22, 1978.

[7] H. Gunes, B. Schuller, M. Pantic, and R. Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 827–834. IEEE, 2011.

[8] D. Hadar. Implicit media tagging and affect prediction from video of spontaneous facial expressions, recorded with depth camera. *arXiv preprint arXiv:1701.05248*, 2017.

[9] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[10] P. J. Lang. The emotion probe: Studies of motivation and attention. *American psychologist*, 50(5):372, 1995.

[11] D. McDuff, R. El Kaliouby, T. Senechal, D. Demirdjian, and R. Picard. Automatic measurement of ad preferences from facial responses gathered over the internet. *Image and Vision Computing*, 32(10):630–640, 2014.

[12] R. Niese, A. Al-Hamadi, A. Panning, and B. Michaelis. Emotion recognition based on 2d-3d facial feature extraction from color image sequences. *Journal of Multimedia*, 5(5):488–500, 2010.

[13] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1113–1133, 2015.

[14] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic. Analysis of eeg signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing*, 7(1):17–28, 2016.

[15] M. Soleymani and M. Pantic. Human-centered implicit tagging: Overview and perspectives. In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3304–3309. IEEE, 2012.

[16] T. Tron, A. Peled, A. Grinsphoon, and D. Weinshall. Automated facial expressions analysis in schizophrenia: A continuous dynamic approach. In *International Symposium on Pervasive Computing Paradigms for Mental Health*, pages 72–81. Springer, 2015.

[17] T. Tron, A. Peled, A. Grinsphoon, and D. Weinshall. Facial expressions and flat affect in schizophrenia, automatic analysis from depth camera data. In *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 2016*, pages 220–223. IEEE, 2016.

[18] T. Vandal, D. McDuff, and R. El Kaliouby. Event detection: Ultra large-scale clustering of facial expressions. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015.

[19] A. Vinciarelli, N. Suditu, and M. Pantic. Implicit human-centered tagging. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 1428–1431. IEEE, 2009.

[20] S. Wang, Z. Liu, Y. Zhu, M. He, X. Chen, and Q. Ji. Implicit video emotion tagging from audiences facial expression. *Multimedia Tools and Applications*, 74(13):4679–4706, 2015.

[21] S. Yang, M. Kafai, L. An, and B. Bhanu. Zapping index: using smile to measure advertisement zapping likelihood. *IEEE Transactions on Affective Computing*, 5(4):432–444, 2014.

[22] S. Zhao, H. Yao, X. Sun, P. Xu, X. Liu, and R. Ji. Video indexing and recommendation based on affective analysis of viewers. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1473–1476. ACM, 2011.