Implicit Image Tagging via Facial Information

Jun Jiao
Department of Computing,
Imperial College London,
SW7 2AZ, United Kingdom
jun.jiao09@imperial.ac.uk

Maja Pantic
Department of Computing,
Imperial College London, UK/
EEMCS, University of Twente, The Netherlands
m.pantic@imperial.ac.uk

ABSTRACT

Implicit Tagging is the technique to annotate multimedia data based on user's spontaneous nonverbal reactions. In this paper, a study is conducted to test whether user's facial expression can be used to predict the correctness of tags of images. The basic assumption behind this study is that users are likely to display certain kind of emotion due to the correctness of tags. The dataset used in this paper is users' frontal face video collected during an implicit tagging experiment, in which participants were presented with tagged images and their facial reactions when viewing these images were recorded. Based on this dataset, facial points in video sequences are tracked by a facial point tracker. Geometric features are calculated from the positions of facial points to represent each video as a sequence of feature vectors, and Hidden Markov Models (HMM) are used to classify this information in terms of behavior typical for viewing a correctly or an incorrectly tagged image. Experimental results show that user's facial expression can be used to help judge the correctness of tags. The proposed is effective in case of 16 out of 27 participants, the highest prediction accuracy for a single participant being 72.1%, and the highest overall accuracy being 77.98%.

Categories and Subject Descriptors

I.5.4 [**Computing Methodologies**]: Pattern Recognition—*Applications*; J.m [**Computer Applications**]: Metrics—*complexity measures, performance measures*

General Terms

Algorithms

Keywords

implicit tagging, image tagging, dynamic classification, facial feature extraction, ensemble learning

SSPW'10, October 29, 2010, Firenze, Italy.

1. INTRODUCTION

The past decade has seen a fast growth in the number of multimedia resources such as images and videos. For example, in the case of photos alone, there are many popular online photo sharing websites such as Flickr, SnapfishHow and WebShot. The number of images hosted by these sites are as follow: over 2 billion on Flickr, 1 billion on SnapfishHow, 500 million on WebShots, and 3.8 billion on PhotoBucket [6]. These new media sites show that the Web is transforming to a participatory medium in which users can be actively involved in creating, distributing and evaluating information. How to efficiently manage and locate these resources has become an increasingly important problem in computer science.

In multimedia management and sharing, tags assigned to the content play an important role in search and retrieval. With the help of tags, multimedia contents, such as, images, videos and web pages, can be efficiently classified, indexed and searched. The tagging technique currently adopted by most social media websites is explicit tagging, in which images are annotated based on the image content by human, either experts or amateurs. Although explicit image tagging has been successfully applied in a wide range of WEB 2.0 applications these days, it still has several drawbacks [12]. First, the tags obtained by explicit tagging are likely to be inaccurate in practice. When tagging, people usually behave according to their individual interpretation of the content, personal and social needs without considering whether the tags can be used to improve the performance of retrieval. Second, it requires extra effort from users. Explicit tagging is sometimes really inconvenient for users since adding tags to images may take a large amount of time and energy.

In contrast to explicit tagging, implicit tagging is the technique to annotate multimedia data based on user's nonverbal reactions, such as facial expression and head gesture [12]. The data is tagged in a implicit way because it is based on user's reaction without explicitly requesting a user to associate tag with the data. One possible application of implicit tagging is to assess the correctness of explicit tags. For example, if a user smiles or nodes the head while watching a tagged datum, this may imply that the tag is correct; however, if the user displays an expression of surprise or disappointment, the tag is likely to be incorrect. Another possible application of implicit tagging is to assign new explicit tags. For instance, if the user laughs while watching an untagged datum, it is likely that it contains some funny elements, thus can be labelled as "funny" or "interesting".

A number of previous works are considered to be implicit

tagging related. There are a couple of previous works adopting user's emotional reactions to help retrieve and organise multimedia data. Arapakis et al. [1] conduct a study to investigate the role of emotions in the information seeking process. The conclusion shows that user's emotion feedback can be used as good predictors of document relevancy. Arapakis et al. [2] propose a method to enrich user profiling using affective information collected from users. A series of works have been proposed to use physiological or EEG signals to tag multimedia data [16][15][7][8][21]. Yamamoto et al. [20] design a system to acquire user's preference of TV shows from users behaviour. However, none of them tries to implicitly annotate image data.

This paper focuses on the task of using human behavior to assess the correctness of explicit tags of images. The two most essential problems here are how to represent the user's reactions and how to model the relationship between these reactions and the correctness of tags. Since facial expression is the most common way for people to express their emotion, including agreement and disagreement [3], it can be used to judge user's agreement with the associated tag. The usually extracted features for representing facial expressions can be classified into two classes: geometric features and appearance features [22]. Geometric features characterise the shapes of facial components based on the locations of facial salient points such as corners of the eyes and mouth. Typical works that use geometrical features are those of Pantic et al. [9][10][11][18][17], in which movements of a set of salient facial points are used as features. Appearance features use facial textures to represent facial expression. Typical examples are Valstar et al. [19], and Chang et al. [4]. Modelling human nonverbal behaviour is generally viewed as a machine learning problem, in which either frame-based static classifier or sequence-based dynamic classifier can be used [14]. Support Vector Machine (SVM) and Neural Networks (NN) are the most popular static classifiers, and the most commonly used dynamic classifiers are Hidden Markov Models (HMM) and its variations, such as Coupled Hidden Markov Models (CHMM).

In this paper, we use user's facial reactions to predict the correctness of tags of images. Facial points are tracked by using Patras-Pantic particle filtering [13]. Geometrical features are calculated based on the positions of facial points to represent each video clip as a sequence of feature vectors, and Hidden Markov Models (HMM) are used to classify these sequences in terms of correctly or incorrectly tagged data. To this end, weak classifiers for each individual are combined. Experimental results show that a relationship exists between user's facial expression and the tag's correctness, and that the proposed method is able to capture such relationships.

The rest of paper is organised as follows. Section 2 introduces the dataset that we used in this paper. Section 3 presents how the facial features are extracted. Section 4 introduces our method. Section 5 reports the experimental results. Finally, section 6 concludes the paper.

2. DATASET

The dataset used in this study consists of video sequences collected during an Implicit Tagging Experiment, in which the behavioral reactions of 27 volunteers, 14 male and 13 female, were recorded. Each participant was presented with 28 stimulus images. Among these, the tags presented for 14 images were correct and those for other 14 images were incorrect. For each stimulus image, the following was presented:

Figure 1: An example frame of the captured video of frontal face

(1) Untagged Stimulus Image: The untagged stimulus image was displayed for 5 seconds. This allows the subject get to know the content of the image, and thus react on the tag associated with the image rather than on the content of the image.

(2) Tagged Stimulus Image: The same image with tag was displayed for 5 seconds. The subject's behavior shown in this period should contain his or her reaction to the correctness of the tag.

(3) Question: A question was displayed on the screen to ask whether the subject agree with the suggested tag. The answer was indicated by pressing the left mouse button for yes or the right mouse button for no. The question remained on the screen for a non-limited time until a selection was made by the subject. Afterwards, the program moved to the next image, in which the same three steps are repeated.

The dataset used in this paper is colour video data of subjects' frontal faces captured during the experiment. Fig. 1 shows an example of the used video data. The length of each session is about 11 seconds. The resolution of all videos is 780×580, and the frame rate of video sequence is 30f/s. The video for each participant was segmented to 28 small clips according to the images presented. Each clip corresponds to the period from the time point when annotated image appears to the time point when the user gives a feedback. The total number of clips is 756 and the average length of each small clip is about 5 seconds.

3. FACIAL FEATURE EXTRACTION

To extract facial features, Patras-Pantic particle filter [13] is used to track 19 facial points, as illustrated in Fig. 2. The tracked points are outer and inner eyebrow (4 points), eye corners and eyelids (8 points), left and right nostrils (2 points), mouth corners and lips (4 points), and chin (1 point). The initial positions of these points for each participant's video were manually labeled, then automatically tracked for the rest of the sequence. After tracking, each frame of the video data was represented as a vector of the facial points' $2D$ coordinates.

For each frame, geometric features $f1 - f20$ were then extracted based on the positions of the facial points. Those features are the following.

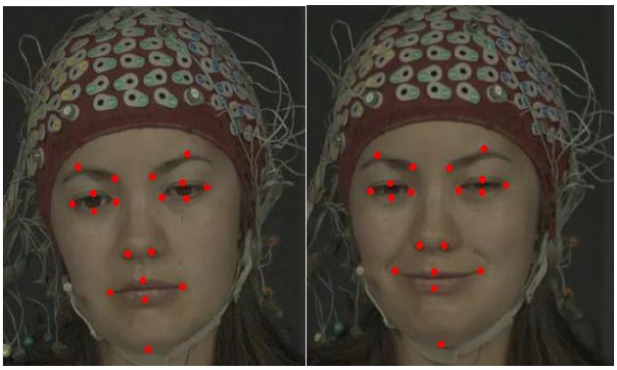

Figure 2: An example frame of the captured video data of frontal face

Figure 3: The stats topology used in

Eyebrows: (1) f1, f2: angles between the horizontal line connecting the inner corners of the eyes and the line that connects inner and outer eyebrow. (2) f3, f4: the vertical distance from the outer eyebrow to the line that connects the inner corners of the eyes. (Fig. 3)

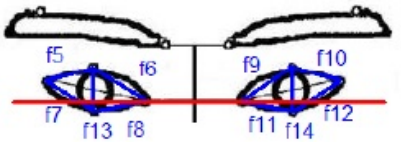

Figure 4: The stats topology used in

Eyes: (1) f5, f9: distance between the outer eye corner and the upper eyelid. (2) f6, f10: distance between the inner eye corner and the upper eyelid. (3) f7, f11: distance between the outer eye corner and the lower eyelid. (4) f8, f12: distance between the inner eye corner and the lower eyelid. (5) f13,f14: vertical distance between the upper eyelid and the lower eyelid. (Fig. 4)

Mouth:(1) f15(f16): distance between the upper lip and the left (right) mouth corner. (2) f17(f18): distance between the lower lip and the left (right) mouth corner. (3) f19: distance between the left and the right mouth corner. (4) f20: vertical distance between the upper and the lower lip. (Fig. 5)

The line that connects the inner eye corners was used as a reference line since the inner eye corners are stable facial points(i.e. changes in facial expression do not induce any changes in the position of these points) and they are also the two most accurately tracked points. For each sequence, the above-listed 20 features have been calculated for the first frame and the frame at time t. The difference in $f1 - f20$ was used in further processing. Therefore, each frame of the captured video was represented as a 20-dimensional vector.

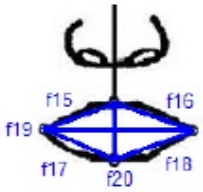

Figure 5: The stats topology used in

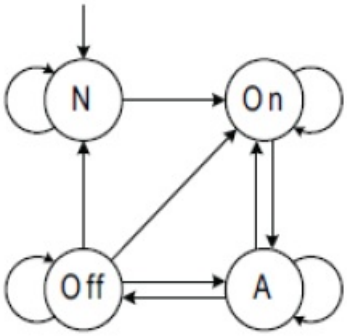

Figure 6: The stats topology used in

4. FACIAL BEHAVIOUR ANALYSIS

4.1 Hidden Markov Model

Since Hidden Markov Model(HMM) is commonly used for modelling dynamic sequences, we chose to use HMM as the classifier for captured facial expressions. As shown in [14][5], a temporal facial movement consists of four steps: (1) neutral - there are no signs of muscular activation; (2) onset - the muscular contraction begins and increases in intensity; (3) apex - a plateau where the intensity reaches a stable level; (4) offset - the relaxation of muscular action. Based on this, the number of modeled states in the HMM is four. Three different HMM topologies were tried out.

(1)Ergodic: All the states are connected with each other. This means that the four states, neutral, onset, apex and offset, can change from any state to any other state.

(2)Sequential: The four states are connected sequentially such that the model is only allowed to stay in its current state or change to the next state.

(3)The topology used in [14], as shown in Fig. 6.

The parameters used for the HMM are the same as those proposed in [14], and can be summarized as follows:

•Number of States: 4 (neutral, onset, apex, offset),
•Initial State Probabilities: Randomly generated,
•Initial Transition Probabilities: randomly generated,
•Density: Gaussian,
•Number of Gaussians: 5,
•Covariance Type: Diagonal.

For each participant, two HMMs were trained: one trained on participant's facial data displayed when he or she was seeing a correctly tagged image, and one trained on his or her facial data displayed when an incorrectly tagged image was shown. To predict whether a newly displayed image is correctly tagged based on this participant's facial data displayed when seeing the tagged image, these two HMMs are used and the image is labeled as correctly depending on which of the two HMMs gives the higher likelihood.

4.2 Combining Individual Classifiers

If we view the HMMs trained for each individual as a weak classifier, these weak classifiers of different users can be combined to form a stronger classifier, which can be used to predict the correctness of tags more reliably. Three different techniques to combine HMMs trained subject-wise are proposed and tested in this paper.

(1) Highest Probability: For the HMMs of each participant, the two output likelihoods are normalized to sum to 1. We can regard these normalized likelihoods as the probability of the tag being correct or incorrect. Let (P_i^1, P_i^0) denotes the probabilities output by the HMMs of each of n participants, then the label L(incorrectly/correctly tagged image) of a new sequence can be calculated as follows:

$$L = \begin{cases} 1 & max\{P_i^1 | i = 1..n\} \geq max\{P_i^0 | i = 1..n\}, \\ 0 & max\{P_i^1 | i = 1..n\} < max\{P_i^0 | i = 1..n\}. \end{cases} \quad (1)$$

(2) Average: For each class(incorrectly/correctly tagged image), linear averages of the probability from all HMMs are calculated. The label L can be calculated as follows:

$$L = \begin{cases} 1 & (\sum_{i=1}^{n} P_i^1)/n \geq (\sum_{i=1}^{n} P_i^0)/n, \\ 0 & (\sum_{i=1}^{n} P_i^1)/n < (\sum_{i=1}^{n} P_i^0)/n. \end{cases} \quad (2)$$

(3) Top Average: Similar to the method (2), but only the top m (m<n) probabilities are averaged. The assumption behind this method is that for each stimulus image only some participants display a facial reaction. Averaging over all HMMs might bias the results towards people who display no facial reaction at all. Considering only the outputs with top probabilities might help. Let (P_i^1, P_i^0) denotes the probabilities output by the HMM of each participant, and both P_i^1,P_i^2 are sorted so that $i \leq j \rightarrow P_i^1 \geq P_j^1, P_i^0 \geq P_j^0$ then the label L can be calculated as follows:

$$L = \begin{cases} 1 & (\sum_{i=1}^{m} P_i^1)/m \geq (\sum_{i=1}^{m} P_i^0)/m, \\ 0 & (\sum_{i=1}^{m} P_i^1)/m < (\sum_{i=1}^{m} P_i^0)/m. \end{cases} \quad (3)$$

5. EXPERIMENTAL RESULTS

5.1 Individual Classifiers

Two HMMs explained above are first trained for each participant. Then, for each participant, 10-fold cross validation was conducted. The result of a single run of the 10-cross validation is a 27×28 matrix, in which the element at row i, column j is the predicted correctness of image j based on the behavior shown by participant i. The accuracy of predictions for each participant and for each image can then be calculated from this matrix based on the ground truth labels. Since the initial parameter for the two HMM models are randomly chosen, the result obtained for each cross validation will be slightly different. To reduce the randomness of the obtained result, the above 10-cross validation was conducted 30 times. The averaged results are presented in Table 1, in which each entry is the accuracy of predictions based on the facial reactions of each participant over 28 images.

Among the 27 tested participants, we can see that 16 of them display facial reactions such that distinction can be made between facial reactions shown when a correctly tagged image is displayed and those shown when an incorrectly tagged image is displayed, resulting in a prediction accuracy of over 50%. The highest individual accuracy is achieved for P15 with the Ergodic topology. The overall average achieved accuracy is higher than 50%, which means that the captured facial reactions do reveal users opinion on the correctness of tags. However, the performance for 11 participants is poorer than a random guess. The reason for this is that many participants did not display any facial reaction during the whole process.

Table 1: Predicting accuracy on each participant

Participant	Topology		
	Ergodic	Sequential	Petridis
P1	0.630	0.595	0.600
P2	0.549	0.598	0.581
P3	0.531	0.622	0.628
P4	0.521	0.514	0.549
P5	0.585	0.537	0.508
P6	0.460	0.490	0.459
P7	0.448	0.533	0.523
P8	0.466	0.498	0.437
P9	0.709	0.713	0.703
P10	0.387	0.379	0.380
P11	0.563	0.555	0.534
P12	0.585	0.560	0.582
P13	0.406	0.381	0.409
P14	0.478	0.505	0.456
P15	0.721	0.689	0.711
P16	0.711	0.714	0.711
P17	0.448	0.441	0.445
P18	0.424	0.517	0.505
P19	0.539	0.535	0.556
P20	0.521	0.578	0.540
P21	0.418	0.525	0.524
P22	0.448	0.413	0.411
P23	0.642	0.614	0.668
P24	0.600	0.577	0.610
P25	0.563	0.497	0.493
P26	0.412	0.427	0.418
P27	0.672	0.699	0.650
Average	0.535	0.540	0.538

From the result we can see that there is no significant difference between the three topologies used for HMM. This phenomenon may be attributed to the following reason. In the dataset used, the number of times that facial movements occur is very limited. Except for blink, other facial movement seldom happened. Therefore, most of each sequence contains (nearly) expressionless face data, on which the three topologies will have very similar performance.

5.2 Combining Classifiers

The previously introduced three different classifier fusion methods were tested based on the results obtained from cross-validation process presented in section 5.1. The parameter m for Top Average was set to 3. Since there is no significant difference between the three HMM topologies, Ergodic is used for all HMMs in this experiment.

Table 2 shows the results for each method. As we can see, in case of 17 images the methods are able to predict the correctness of the suggested tag better than a random guess, which proves that our method can be used for implicit tagging. In addition, the number of images that are consis-

Table 2: Predicting

Image	Combining Methods		
	Highest	Average	Top Average
Image1	0.833	0.333	0.6667
Image2	0	0	0
Image3	0.833	0.833	0.833
Image4	0.333	0.333	0.333
Image5	0.666	0.833	0.833
Image6	0.833	1.000	1.000
Image7	0.333	0.833	0.833
Image8	1.000	1.000	1.000
Image9	0	0	0
Image10	0.833	0.833	0.833
Image11	0.833	0.166	0.166
Image12	0.833	1.000	1.000
Image13	0	0	0
Image14	0.333	0.833	1.000
Image15	1.000	1.000	1.000
Image16	1.000	1.000	1.000
Image17	0.333	0.333	0.333
Image18	0.333	0.333	0.333
Image19	0.166	0	0
Image20	0.333	0.166	0
Image21	1.000	1.000	1.000
Image22	0.500	0	0.500
Image23	1.000	1.000	1.000
Image24	1.000	1.000	1.000
Image25	1.000	1.000	1.000
Image26	1.000	1.000	1.000
Image27	0	0	0
Image28	1.000	1.000	1.000
Average	0.619	0.601	0.631

tently correctly classified, independently of the classifier fusion method applied, is 12, which is twice the number of images that are consistently misclassified. This indicates that our method can perform very well for some images. However, the overall performance of our method is only around 60%, which is not a very good result. This means that it is unadvisable to build implicit-tagging systems only based on facial reactions. Information from other modalities such as gaze and sound should be used in addition to facial expression data to achieve a better performance.

The best average performance was obtained by the Top Average approach. This should be attributed to its ability to rely only on relevant feedbacks. As introduced in the previous section, many participants show little or no facial reactions, thus data coming from these participants is more-or-less meaningless. This affects negatively methods that take into account data of all participants. On the other hand, high probability outputs from individual HMMs are more likely to be meaningful because the probability for an unexpressive participant to trigger a high-probability HMM output is low. The Top Average Approach considers only those meaningful HMM outputs. The parameter m used here is quite small, assuming that only a very small number of participants will have a facial reaction when seeing a given image. For the method that only considers the highest HMM output probability, although it is likely to pick out meaningful data, it is vulnerable to random errors of the classifiers.

6. CONCLUSION

From the experiment explained in this paper we can see that user's facial reactions convey some information about the correctness of tags associated with multimedia data. Training classifiers based on features extracted from user's facial expression can lead to predicting correctness of image tags that is better than a random guess. Although the experiment was conducted on the image data only, we believe that this finding can also be applied to other multimedia data as well, such as video and web pages. However, the experimental results show that the relationship between user's facial reactions and correctness of image tags is not very strong. The performance of classifiers trained on facial expression data of individual subjects is rather poor, being 50% in many cases. This is because many people seldom display facial reactions when viewing multimedia data. In addition, the subject's reaction is likely to be affected by the content of the data rather than by the associated tag. Combining these weak individual classifiers into a stronger classifier leads to a much better result. Among the 27 participants, we found that some are more likely to display facial reactions than others, thus HMMs trained on their data are more meaningful.

Some possible future work is as follows.

(1) The experiment can be conducted using a larger dataset. The currently used dataset contains only 28 images viewed by 27 subjects.

(2) The effectiveness of using other modalities such as head gesture, shoulder movement and gaze information can be tested.

7. ACKNOWLEDGMENTS

This work has been funded in part by the European Community's 7th Framework Programme [FP7/2007Ú 2013] under the grant agreement no 231287 (SSPNet). The work of Maja Pantic is also funded in part by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB).

8. REFERENCES

[1] I. Arapakis, J. M. Jose, and P. D. Gray. Affective feedback: an investigation into the role of emotions in the information seeking process. In *Proc. the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 395–402, 2008.

[2] I. Arapakis, Y. Moshfeghi, H. Joho, R. Ren, D. Hannah, and J. M. Jose. Enriching user profiling with affective features for the improvement of a multimodal recommender system. In *Proc. the ACM International Conference on Image and Video Retrieval*, 2009.

[3] K. Bousmalis, M. Mehu, and M. Pantic. Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools. *Proc. IEEE International Conference on Affective Computing and Intelligent Interfaces (ACII'09)*, 2, 2009.

[4] Y. Chang, C. Hu, and M. Turk. Probabilistic expression analysis on manifolds. *Proc. IEEE Int'l*

Conf. Computer Vision and Pattern Recognition (CVPR'04), 2:520–527, 2004.

[5] P. Ekman. About brows: Emotional and conversational signals. In *Human Ethology: Claims and Limits of a New Discipline: Contributions to the Colloquium*, pages 169–248, 1979.

[6] B. Elliott and Z. M. Ozsoyoglu. Annotation suggestion and search for personal multimedia objects on the web. In *Proceedings of the 2008 International Conference on Content-based image and video retrieval*, pages 75–84, 2008.

[7] H. Joho, J. M. Jose, R. Valenti, and N. Sebe. Exploiting facial expressions for affective video summarisation. In *Proc. the ACM International Conference on Image and Video Retrieval*, 2009.

[8] J. J. M. Kierkels, M. Soleymani, and T. Pun. Queries and tags in affect-based multimedia retrieval. In *Proc. IEEE International Conference on Multimedia and Expo*, 2009.

[9] M. Pantic and M. S. Bartlett. Machine analysis of facial expressions. *Facial Recognition*, pages 377–416, 2007.

[10] M. Pantic and I. Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments form face profile image sequences. *IEEE Trans. Systems, Man, and Cybernetics Part B*, 36(2):433–449, 2006.

[11] M. Pantic and L. J. M. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Trans. Systems, Man, and Cybernetics Part B*, 34(3):1449–1461, 2004.

[12] M. Pantic and A. Vinciarelli. Implicit human-centered tagging. *IEEE Signal Processing Magazine*, 26(6):173–180, November/December 2009.

[13] I. Patras and M. Pantic. Particle filtering with factorized likelihoods for tracking facial features. In *Proc. IEEE Interntional Conference on Automatic Face and Gesture Recognition*, pages 97–102, 2004.

[14] S. Petridis, H. Gunes, S. Kaltwang, and M. Pantic. Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities. In *Proc. 2009 international conference on Multimodal interfaces(ICMI'09)*, pages 23–30, 2009.

[15] M. Soleymani, G. Chanel, and J. J. M. Kierkels. Affective ranking of movie scenes using physiological signals and content analysis. In *Proc. the 2nd ACM workshop on Multimedia semantics*, pages 32–39, 2008.

[16] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun. Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses. In *Proc. Tenth IEEE International Symposium on Multimedia*, pages 228–235, 2008.

[17] M. F. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proc. Ninth ACM Int'l Conf. Multimodal Interfaces (ICMI'07)*, pages 38–45, 2007.

[18] M. F. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn. Spontaneous versus posed facial behavior: Automatic analysis of brow actions. In *Proc. Eighth Int'l Conf. Multimodal Interfaces (ICMI'06)*, pages 162–170, 2006.

[19] M. F. Valstar, M. Pantic, and I. Patras. Motion history for facial action detection from face video. *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics (SMC'04)*, 1:635–640, 2004.

[20] M. Yamamoto, N. Nitta, and N. Babaguchi. Automatic personal preference acquisition from tv viewer? behaviours. In *Proc. IEEE International Conference on Multimedia and Expo*, pages 1165–1168, 2008.

[21] A. Yazdani, J. S. Lee, and T. Ebrahimi. Implicit emotional tagging of multimedia using eeg signals and brain computer interface. In *Proc. the first SIGMM workshop on Social media*, pages 81–88, 2009.

[22] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, January 2009.