Analysing User Visual Implicit Feedback in Enhanced TV Scenarios

Ioan Marius Bilasco[1], Adel Lablack[1(✉)], Afifa Dahmane[1,2], and Taner Danisman[1]

[1] Laboratoire d'Informatique Fondamentale de Lille, Université de Lille 1, Villeneuve-d'Ascq, France
[2] Université des Sciences et Technologies Houari Boumediene, Algiers, Algeria
{marius.bilasco,adel.lablack,afifa.dahmane,taner.danisman}@lifl.fr

Abstract. In this paper, we report on user behaviors by analyzing visual clues while users are watching various TV broadcast in pilot settings. We detail the first results of the empathic analysis of viewers watching four distinct videos in dedicated recording sessions. Viewers are sitting in front of a TV set in unconstrained position (free postures, free head poses and free body movements) on a chair and recorded by a regular webcam at both low and high resolutions. We have extracted metrics related to: head and global movement, changes in head orientation and facial expressions (happy, angry, surprise). We have conducted preliminary studies about how the extracted metrics can be employed in order to detect the interest, the amusement or the distraction of a viewer.

Keywords: Facial expressions · Emotions · Moods · Global body movement · Head pose

1 Introduction

The explosion of available multimedia contents makes interesting the study of the behavior of users while they are accessing to these contents on their computers, tablets or smartphones. Thus, the use of webcams when the user is facing his access device allows identifying visual cues and constitute an implicit visual feedback in response to the access of multimedia content.

A lot of research is performed into systems that are able to offer a personalized content stream to media consumers taking into account user preferences and context. In order to offer a content suggestions experience to a user in a natural way, the sensed mood and state of mind of the user should be captured in real time. Thus, the user's bodily and behavioral reactions should be translated to indicators using reliable motion/emotion assessment techniques in order to propose an empathic system that incorporate the end-user behavior, reactions and responses to provide content.

The analysis of the user visual implicit feedback in enhanced TV scenarios to understand and respond to user intentions and emotions could improve the experience. It is performed using a set of metrics extracted from non-verbal cues such as facial expressions, body posture, head and hand gestures. Unfortunately, the

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part I, LNCS 8925, pp. 315–324, 2015.
DOI: 10.1007/978-3-319-16178-5_22

current available datasets might not generalize well to the real world situations in which such systems would be used [7].

2 Related Work and Background

Advancements in TV technology bring new interfaces and functionalities such as in smart and connected TVs. It allows people to watch various programs such as sports games, reality shows, movies, etc. This experience could be improved by taking into account the user visual implicit feedback. Most of the proposed systems focus on the recognition of emotional state of the user to trigger actions to enhance the user experience. In order to develop a video summarization tool, Joho et al. [5] have proposed an approach to detect personal highlights in video contents based on the analysis of facial activities of the viewer. Their analysis suggests that the motion vectors in the upper part of human face are more likely to be indicative of personal highlights than the lower part of the face. Abadi et al. [1] describes a multimodal approach to detect viewers' engagement through psychophysiological affective signals. This study aims to understand which channels and combinations thereof are effective for detecting a viewer's level of engagement. They notice that EEG and GSR responses seem to contribute similarly to the engagement classification task under study; moreover, Facial Motion features seem to provide complementary information and the psychophysiological features employed to assess the viewers' state of engagement seem to indicate high inter-subject variability. Soleymani et al. [9] have shown the feasibility of an approach to recognize emotion in response to videos. They have proposed a user-independent emotion recognition method using participants' EEG signals, gaze distance and pupillary response as affective feedbacks. Hanjalic and Xu [3] introduced "personalized content delivery" as a valuable tool in affective indexing and retrieval systems by selecting video and audio content features based on their relation to the valence-arousal space that was defined as an affect model.

Nowadays, there is a move away from the automatic inference of the basic emotions proposed by Ekman towards the inference of complex mental states such as attitudes, cognitive states, and intentions. This shift to incorporate complex mental states alongside basic emotions is necessary to build affect sensitive systems as part of an empathic technology. There is also a move towards analyzing natural expressions rather than posed ones since the Action Unit amplitudes and timings differ in spontaneous and acted expressions. These differences imply that recognition results reported on systems trained and tested on acted expressions could not be generalized to spontaneous ones. We focus on approaches used to analyze the behavior of people through an analysis of their mood which is generally highly correlated with their facial expressions and body movement in spontaneous expressions datasets.

3 Pilot Settings

We start by presenting the pilot settings since our analysis has been conducted according to these specific settings. Our Pilot videos have been recorded using

regular cameras placed at the bottom of a TV screen which is distant from the user at approximately 2m. People sitting on a chair are following 4 fragments from "Everybody's Famous" series broadcasted by VRT. The fragments correspond to stories having different primary-intents (adrenaline: jumping in wingsuit, compassion and interest: prime minister's confidences, amusement: easter bunny, neutral: pupils on a school trip). The Figure 1 shows some screenshots from the video recorded as well as the overview of the scenario timing.

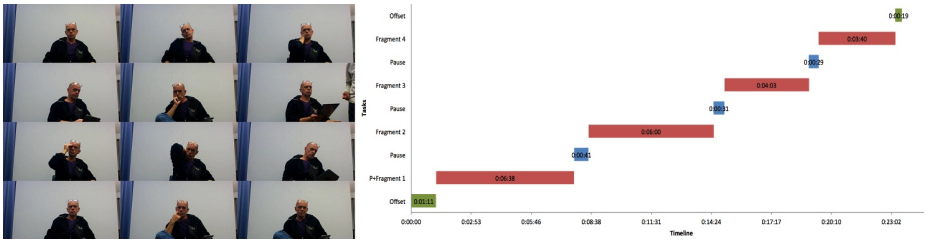

Fig. 1. Screenshots of a user watching TV and interacting during the recording session vs Video fragments and timeline in the recording sessions

During the pilot phase, we have captured 43 complete captures having an average of 22 minutes per session. We have recorded 37 high resolution videos (960×540) where the average face size is 107px and 6 in low resolution videos (320×240) where the average face size is 47px. The users were not instructed about how they should behave in order to be collaborative with the capturing system. They were informed that the capturing devices are on and that they will have to look at the TV and fill in a questionnaire at the end.

Hence, in the pilot analysis we are facing to big challenges : small faces and natural face and body poses as illustrated in Figure 1. The fact that the users were not especially collaborative yields to numerous occlusions (hand, hair) but also high pitch in faces orientation. In order to deal with these conditions we have selected a mix of local and global analysis solutions that track face expressions, but also movements. The analyzers are detailed in the following section.

4 The Architecture of the Analyzer

We present in Figure 2 an overview of the complete analyzing process. The first part of the process consists of locating the user within the scene and extracting basic information about the body and the face. At this point, we characterize the inner face movements and the global movements by exploring difference images and studying optical flow. Once the face is detected, we study the eyes region and extract specific eye movement patterns. The eye position is further used in order to normalize the face and estimating head orientation. We estimate metrics related to the level of interest and the positive (happy) / negative (anger) emotions in presence of frontal faces with limited pitch and pan, and eventually

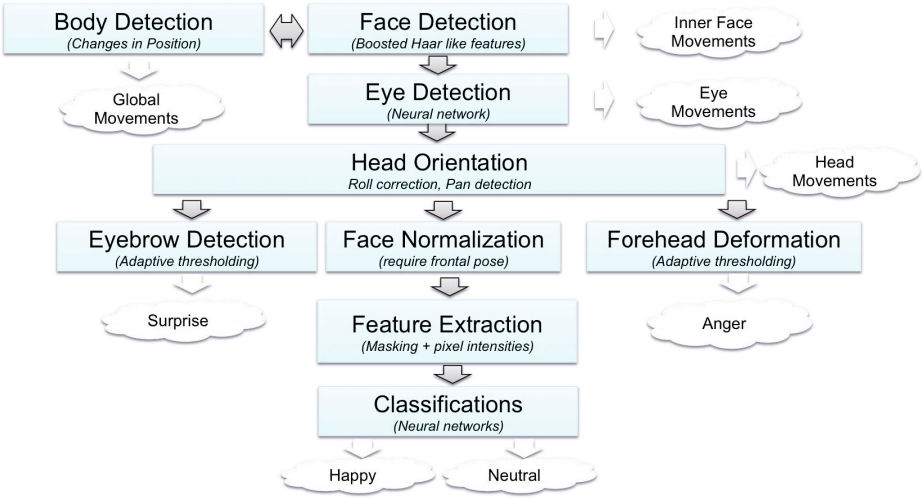

Fig. 2. Global architecture of the system

specific occlusions. We use continuous characterization for surprise and anger, and a discrete classifier for happy detection.

In the following, we detail the different metrics that we have extracted from the video streams. These metrics are then used to infer viewers mood that are actionable in a TV scenario.

4.1 Movement Based Analysis

This section evaluates the changes in the position and the global movements of the person. We take advantage of the scene configuration and we consider that globally all changes (except noise due to subtle illumination changes) reflect the user activity. However, we have reinforce the area of the study, by taking into account the head position (in case of a successful detection of face and/or eyes) in focusing the area of study in the neighboring regions (below and on sides). All metrics are normalized with regard to the region considered : whole image vs delimited region with regard to the head position.

We consider that there is a strong link between the generated metrics and the arousal of the person. However, we have to differentiate where the person keeps interest in viewing the TV (moving and still watching the TV set) or his engaged in actions not related to this activity (peeking up the tablet). Hence, in presence of body movements and the presence of continuous eye detection, we might infer that the person presents an arousal directly linked to the quantity of movement in the scene. Whereas in situations where the person lost eye contact with the screen (leaning forward for picking up the tablet), the body metric might not be directly correlated with the arousal metric.

In Figure 3, we present (on a scale of one to 100) the global body movement for a person while watching the second (on the left) and the third (on the right) segment of a viewing session.

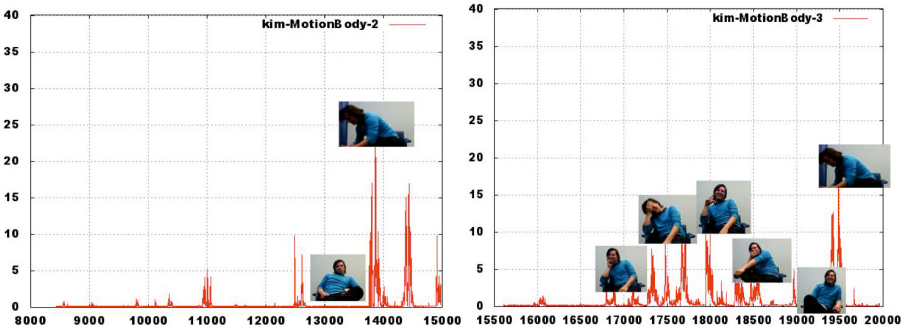

Fig. 3. Global movement analysis for the 2nd and 3rd segments

4.2 Preliminary Study on Face and Eye Detection

Considering the pilot conditions, most of the time the person is staying in the middle of the image. The spatial region containing the face can be inferred from the settings as well as personal characteristics. However, in order to have a more precise localization, we are using face detector from state of the art, such as, Boosted Haar classifiers. This kind of classifiers presents some drawbacks since they support only a limited set of head-poses and fail in case of strong occlusions. With the knowledge that the user is watching the video fragments, each frame of the video contains a single face and the absence of a detected face is a sign for either non supported head-poses (e.g. looking somewhere else) or partial occlusions. However, in order to distinguish between these two situations, we introduced a face tracker enhanced with eye location information.

As soon as a first face is detected, we use a dedicated neural network to detect the eye positions. In presence of small faces, we artificially increase their size to perform eye detection on high resolution images. Mean-shift trackers are then registered for each eye. In case of face lost, we take advantage of the tracked eye position in the following frames to infer the facial region. We have chosen the usage of eye tracker as eye regions are more specific with regard to the whole face and mean-shift trackers are more appropriate for this selection. In order to guarantee the coherency of the eye tracker, we observe the evolutions in terms of the Inter Pupillary Distance (IPD) and the associated roll angle. We also use this information in order to compute the position of a non-detected eye with regard to the position of the detected eye and the previous known IPD and roll angle values. When the regular face detector produces valid detections, the eye trackers are updated accordingly. This procedure allowed us to largely improve the face detection rate on the entire pilot data. We still keep the information about the faces missed by the regular face detector as an indicator of the user interest with the regard to the content.

Detected faces are then processed in order to analyze the head pose. We are using the eye positions in order to infer the roll. Then the roll is used in order to put the face in up-right position. On the latter image, we exploit the bilateral

symmetry of the face in order to infer the yaw orientation class. We use the size and the orientation of the symmetrical area of the face to estimate yaw poses by the mean of Decision Tree model. As illustrated in Figure 4, when the face is in front of the camera, the symmetry between its two parts appears clearly and the line which passes between the two eyes and nose tip defines the symmetry axis. However, when the head performs a motion, for example, a yaw motion, this symmetry decreases.

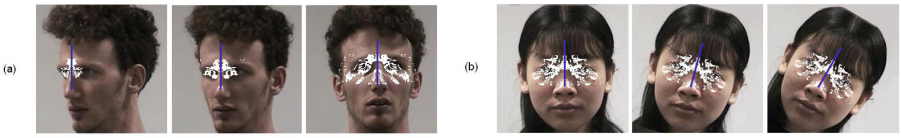

Fig. 4. (a) Variations in the size of the symmetrical region during yaw movement. (b) Variation in the angle of the symmetry axis during roll movement.

This approach doesn't require the location of interest points on the face and is robust to partial occlusions. The estimated pose is coarse but sufficient to infer general gaze direction. Symmetrical region is defined by analyzing pixels intensity. The intensity of one pixel on the right side of the face is more similar to its mirror pixel than another pixel in the image. An experiment has been conducted in [2] which indicate that the use of facial symmetry as a geometrical indicator for head pose is still reliable when local geometric features (such as eyes, nose or mouth) are missed due to occlusions or wrong detections.

At this stage, we are collecting the following information: is face detected or tracked? Are eyes detected or tracked? As well as, their positions and relative changes, is the face frontal or not? This latter information is used afterwards in order to guide and adapt the face expression analysis process, by filtering non frontal faces.

4.3 Face Expression Analysis

Our approach regardless of the emotion that we are seeking is divided into two stages. The first one consists in the different processing steps that allow extracting a normalized face from the input data. The second one consists in locating selected ROI from the face region and applying a specific filtering to each region to indicate the presence of an anger or surprise emotions, while happy emotion is extracted using a neural network.

Image Pre-Processing
This is an important step that aims to obtain images that have normalized intensity, are uniform in size and shape, and depict only the face region.

- Face and eye detection : As presented in the preliminary study of the face and eye detection, we have used the Boosted Haar like features method proposed by Viola and Jones [10] to detect the face. Then a neural network based approach is used to locate the positions of pupils. We derive only the eye detection code from the STASM library [8] which is variation of Active Shape Model of Coote's implementation that performs better on frontal upright faces with neutral expressions.
- Up-right face : We estimate the orientation of the face using the vertical positions of the two eyes and/or the facial symmetry. If the eyes are not in the same position, we compute the angle between these two pupil points and correct the orientation by setting the face center as origin point and we rotate the whole frame in opposite direction.
- Face normalization : We use histogram equalization to normalize image intensity by improving its contrast. It aims to eliminate light and illumination related defects from the facial area.

Emotion Extraction

Happiness detection is considered as a two-class classification problem where happiness is the positive class and the neutral is the negative class. We used MPLab GENKI-4K [4] dataset as a training set for happy/neutral classification. We used lower part of the normalized face which maximizes the accuracy for this particular classification problem. A backpropagation neural network having two hidden layers (20 and 15 neurons) used to train pixel intensity values obtained from the selected ROI. Input layer has 200 neurons and output layer has two neurons representing the happy and neutral classes.

In order to detect anger emotion, we focus on the ROI located in the upper part of the face and include the variations of AU4 of FACS where eyebrows are lowered and drawn together [6]. We apply Gabor filter to this region of face. In the literature, 2D Gabor filters have been used for texture analysis and classification. Gabor filters have both frequency and orientation selective properties. Therefore a 2D Gabor function is composed of a sinusoidal wave of specified radial frequency which is the spacing factor between the kernels in the frequency domain and orientation which is modulated by a 2D Gaussian. Gabor representation of a face image is computed by convolving the face image I (x, y) with the Gabor filter. Majority of AUs samples associated to anger emotion face images has vertical lines above the nasal root. So, we choose vertical orientation for the Gabor filter with a frequency of $\sqrt{1.3}$, Mu equal to 0, σ equal to π and Nu equal to 3 as Gabor parameters. Then real and imaginary responses are added together to find the magnitude response. After a binary thresholding, the sum of the total pixels in the magnitude response of the filter, just above the nasal root is examined by a threshold value to detect a negative emotion. Brighter pixels in the magnitude responses are used as an indicator of anger.

We use the same kind of approach used for anger detection to detect the emotion surprise. An adaptive thresholding is applied in the region of the face that includes the AU1 and AU2 to detect eyebrows movements.

5 Towards Inferring Viewers Moods

From a TV producer point of view, the scope of the analysis is not on measuring only the primary emotions expressed by the user, but also on moods that are actionable [11]. The primary emotion might be closely linked to the content (smile when watching a funny scene, disgust or fear when watching an horror scene) and not to the level of immersion of the user in the viewing experience. The moods targeted in this study are: bored, amused and interested. These moods were considered interesting since in their presence actions might be trigger in order to enhance the user experience. In presence of interest a complementary content could be proposed to the user. In presence of amusement an equivalent content might be proposed in order to keep the person active in watching. In presence of a bored user, the content presented might be changed by proposing to the user alternative contents.

We have tried to elaborate ad-hoc hypothesis for characterizing the mood of a user while watching a content that could convey various emotions (joy, sadness, etc.). We are lacking training corpuses for this context, as the annotation part is subjective and hard to collect. In addition, we are not able to leave the charge of recognizing moods to classifiers and other implicit solutions. In this work, we have proposed some explicit measurements that can be extracted and adapted to specific scenario requirements. These hypothesis were constructed taking into account the behavior exhibited by the user, as well, as the subjective evaluation of the content and the viewing experience done by the viewer itself at the end of the pilot.

In the following, we detail the hypothesis used for each of the moods and illustrate them using crossed-metrics situations where the moods addressed might be encountered. The common consideration of this entire hypothesis is that more importance is given to the global behavior of the user reflected by movement metrics. It allows to characterize the fact that the user is actively involved (amused, interest, frustrated) or passive (bored) in the experience. Then, we use the detected emotions to distinguish between moods having a similar level of involvement. For instance, bored persons (lacking of interest) might be subject to more frequent changes in the head pose (pan, pitch) and limited frontal poses as they might use secondary devices and often look aside. The global movement metric might be an indicator of whether we are in the presence of an active person to whom an alternative content might be proposed to get his attention.

High global movement metric, is not necessary a sign of lack of interest. For instance, in segment 3 (presented in Figure 5 - 17500 to 18500), the viewer moves a lot, but he maintains the eye contact with the TV. On the contrary, in the latter part of the segment (when he bends down for the tablet), we are in the presence of high movement metric, but the face and eye contact are lost frequently over an important time lapse denoting the lack of interest for the content displayed on the TV. In presence of viewer involvement, we explore the occurrences of positive expressions on the viewer face. We estimate that frequent occurrences with frequent movements are sign of amusement, whether the interest is signified by more calm viewing patterns although positive occurrences might arise.

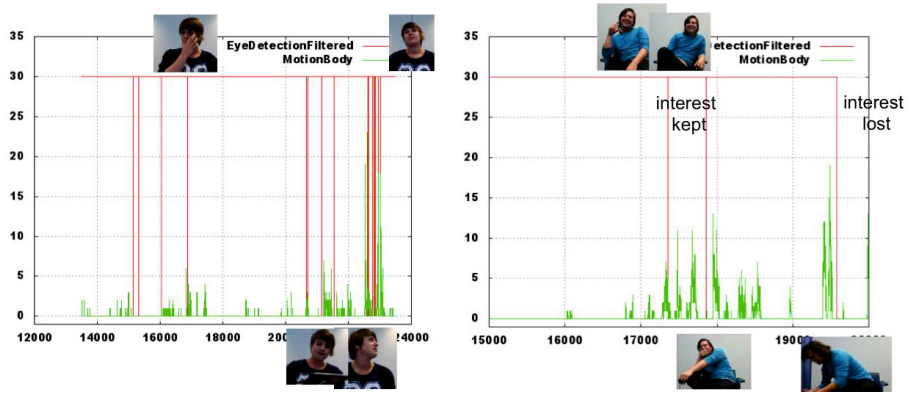

Fig. 5. Crossed Global movement and Eye Detection metrics to measure interests

In Figure 6, we present metrics measured over fragment 2 (left) and fragment 3 (right). In fragment 3 (frames 17500 to 18500), we can observe that the high arousal corresponds to frequent body movement (leaning forward and backward, covering face with hands, etc.), as well as, large intervals where the person is happy, can be considered as strong sign of amusement. The extracted metrics corresponding to segment 2 show moreover a calm viewing pattern, except the last part which corresponds to the person bending and taking the tablet. From time to time, the viewer exhibits some positive expressions sign of enjoyment of the content. For this latter case, the valence is not of major importance since we can observe the same pattern during negative experiences. What we underline here is the fact that in presence of calm viewing patterns, the observation of significant expressions distributed over time can reinforce the perceived interest.

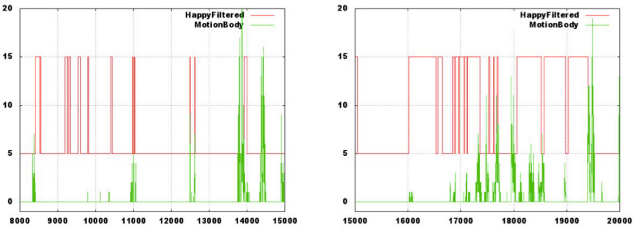

Fig. 6. Crossed Global movement metric and Happy detections to measure amusement

These hypothesis are applied to the pilot data by considering moving frame-windows. Mean values for various metrics are taken into account for the cross-metric analysis. For detecting amusement settings, we reinforce the analysis by considering the density of happy detection over larger time intervals.

6 Conclusion

In this paper, we have proposed ad-hoc metrics and rules for detecting actionable moods in a TV viewing scenario. We have employed both local and global approaches to deal with small resolutions, natural poses and occlusions. An annotation of the whole pilot corpus is performed in order to validate the inferring moods process, as well as, the precision of the proposed metrics. Further on, we will investigate deeply the fuzzy nature of the analyzers inputs in the inferring process. We have already made a first step in this direction by filtering faces that seem inappropriate for happy classification purposes.

Acknowledgments. This work was conducted in the context of the ITEA2 "Empathic Products" project, ITEA2 1105, and is supported by funding from the French Services, Industry and Competitivity General Direction.

References

1. Abadi, M.K., Staiano, J., Cappelletti, A., Zancanaro, M., Sebe, N.: Multimodal engagement classification for affective cinema. In: 5th Conference on Affective Computing and Intelligent Interaction (ACII) (2013)
2. Dahmane, A., Larabi, S., Djeraba, C., Bilasco, I.M.: Learning symmetrical model for head pose estimation. In: 21st International Conference on Pattern Recognition (ICPR), pp. 3614–3617 (2012)
3. Hanjalic, A., Li-Qun, X.: Affective video content representation and modeling. IEEE Transaction on Multimedia **7**(1), 143–154 (2005)
4. The MPLab GENKI Database, GENKI-4K Subset (2011)
5. Joho, H., Staiano, J., Sebe, N., Jose, J.M.: Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents. Multimedia Tools and Applications (MTAP) **51**(2), 505–523 (2011)
6. Lablack, A., Danisman, T., Bilasco, I.M., Djeraba, C.: A local approach for negative emotion detection. In: 22nd International Conference on Pattern Recognition (ICPR) (2014)
7. Mahmoud, M., Baltrušaitis, T., Robinson, P., Riek, L.D.: 3D Corpus of Spontaneous Complex Mental States. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part I. LNCS, vol. 6974, pp. 205–214. Springer, Heidelberg (2011)
8. Milborrow, S., Nicolls, F.: Locating Facial Features with an Extended Active Shape Model. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 504–513. Springer, Heidelberg (2008)
9. Soleymani, M., Pantic, M., Pun, T.: Multimodal emotion recognition in response to videos. IEEE Transactions on Affective Computing (TAC) **3**(2), 211–223 (2012)
10. Viola, P., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2001)
11. Willaert, K., Matton, M.: Empathic media personalization based on actionable moods. In: 1st Workshop on Empathic Television Experiences (EmpaTeX) (2014)