SensiTV - Smart EmotioNal System for Impaired people's TV

Diana Affi
HumanTech Institute
University of Applied Sciences
Fribourg, Switzerland
diana.affi@hes-so.ch

Elena Mugellini
HumanTech Institute
University of Applied Sciences
Fribourg, Switzerland
elena.mugellini@hes-so.ch

Joël Dumoulin
HumanTech Institute
University of Applied Sciences
Fribourg, Switzerland
joel.dumoulin@hes-so.ch

Omar Abou Khaled
HumanTech Institute
University of Applied Sciences
Fribourg, Switzerland
omar.aboukhaled@hes-so.ch

Marco Bertini
MICC
University of Florence
Florence, Italy
marco.bertini@unifi.it

Alberto Del Bimbo
MICC
University of Florence
Florence, Italy
alberto.delbimbo@unifi.it

TVX'15, June 03-05, 2015, Brussels, Belgium
ACM 978-1-4503-3526-3/15/06.
http://dx.doi.org/10.1145/2745197.2755512

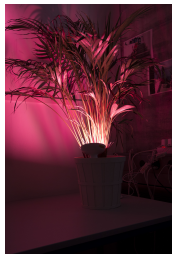
Abstract
In this paper, an innovative solution is presented: a smart emotional system for impaired people's TV. It aims to accompany the cognitive information contained in a movie, with the affective content. The affect is then communicated to the movie viewers in ways compatible for people with hearing and/or visual impairments, to let them experience all of the sensations offered by the movie. To do so, emotion recognition techniques are used to classify movie scenes into seven basic emotions. These emotions are then represented, in realtime, while the movie is playing, to the viewers, using environmental lights, emotional subtitles and a second screen application that integrates vibrations, emoticons and background music.

Author Keywords
Smart TV; Affective Computing; Emotion Recognition

ACM Classification Keywords
H.5.1 [Information interfaces and presentation (e.g., HCI)]: Multimedia Information Systems—video; K.4.2 [Computers and society]: Social Issues—Assistive technologies for persons with disabilities; H.5.3 [Image Processing and computer vision]: Feature Measurement—Feature representation; I.5.2 [Pattern recognition]: Design Methodology—Classifier design and evaluation,Feature evaluation and selection

(a) Sensi TV setup

(b) Light projection from over the plant for positive emotions



(c) Light projection from under the plant for negative emotions

Figure 1: Emotion communication via light. SensiTV setup

Introduction

People suffering of hearing or visual impairments are not being able to enjoy a movie in all of its aspects, and the percentage of these people cannot be neglected. Some solutions already exist to provide these people with the semantic content of the video, using for instance subtitles for the hearing impaired, and audio-description for the visually impaired. But solutions that bring them the emotions contained in the movies are not so common yet. Even though the current solutions help the hearing and visually impaired people to understand the content of the corresponding movie, they deliver raw and static cognitive information lacking all the affective level which is one of the essential factors in delivering the desired TV experience: *a)* by enhancing the ability to process and comprehend the language[1] *b)* and by immersing the viewer in the movie and helping him identify himself with the actors[11]. Finding a way to introduce empathic TV experiences in visually and hearing impaired people's living rooms will bring them the missing information in the current way of consuming visual media: the emotions.

Related work

This work focuses on two research aspects: emotion recognition in audio-visual content, as well as exploring different communication modalities that can be used to translate these emotions to the viewers. In both aspects some work has already been done.

Most of recent works on video analysis have focused on the extraction of video semantics, while the recognition of affective information is less explored. Emotion recognition in audio has dealt with music and with emotion recognition in prosody, leading to the development of toolkits such as OpenSMILE [8]. The linguistic content of speech has also been used for emotion recognition [3] as well as the accompanying text (subtitles, tags, comments etc.) of images [2]. Visual features have also been subject to emotion recognition researches, specially the facial expressions [7]. In this paper's case emotions are detected from movie scenes, which contain visual objects, audio aspects, as well as cinematographic techniques that do contribute in defining the global contained emotion. Hanjalic et al. [10] proposed to extract and model the affective content of the video using both audio and video features, and called this approach *Affective content analysis*. Studies followed Hanjalic, using multiple classification techniques (hidden Markov Models, Partial Least Squares, Support Vector Machines etc.). In [14] an accuracy of 50.37% have been obtained by using both audio and advanced visual features (deep learning) applied on faces and classified using Partial Least Squares.

Concerning the hearing and visually impaired, the approaches currently used to provide them the emotional information conveyed in a movie, rely on audio description and subtitles. [15, 9] have dealt with possible ways to deliver emotions to impaired people through different techniques relying on their complementary functional senses. The limitation of these techniques is that *i)* they rely on the presence of emotional meta-data related to the videos - that are not commonly available - and *ii)* they are not applied in home environments.

SensiTV concept

After going through the state of the art of affective data retrieval from audio-visual content and the study of emotion communication techniques to people with hearing or visual impairment, it was clearly revealed that there is some considerable work to add. SensiTV is developed in order to fill these shortcomings. The system will contain two main modules: a module for emotion recognition from audio-visual content as well as a module for communicating these de-

Emotion	Vibration pattern
Anger	
Disgust	
Fear	
Happiness	
Sadness	
Surprise	

Figure 2: Vibration patterns for each emotion

tected emotions to the viewers in realtime while they are watching the concerned video.

Emotion recognition in movies

Feature extraction

It has been proven that both audio and visual content are important when trying to depict the affective content of a specific scene. The cinematographic techniques, used to awaken certain emotions, are also the guide in finding adequate features but they represent high level features that are hard to detect programmatically (ex. filming from behind the character). In the following we will describe the extracted features in our system for emotion classification.

Audio features: The audio features are extracted using OpenSMILE[8] which enables specifying the framing and windowing over an audio file, as well as it gives the possibility to apply functions on the retrieved features. The audio features extracted are inspired of the ones used in the INTERSPEECH 2010 paralinguistic challenge [16]. Over these features the following functions are applied over a window of one second: Linear regression, Range, Skeweness, Kurtosis, Standard deviation, Minimum, Maximum, Mean and Delta regression.

Visual features: One frame per second is used to extract the following visual features: brightness, shot cut density, edges, color histogram, arousal and the biconcept[1].

After the feature extraction process, the audio and visual features are concatenated, normalised and a Linear Discriminant Analysis (LDA) is applied on them in order to reduce their dimensionality.

[1]The biconcept is a feature vector of size 1200, each dimension is associated to an adjective-noun couple expressing sentiment, developed in the visual sentiment ontology study [2].

Classification

Support Vector Machines (SVM) have proven their capabilities for emotion features classification from audio-visual content as seen in the state of the art. This led the choice of an SVM to be used for classification. In our case, we need to classify seven emotions so we are using the one-to-one approach for multi-class classification. This approach tends to train one classifier for each pair of classes. The SVM is based on an *RBF* kernel in order to treat nonlinear data.

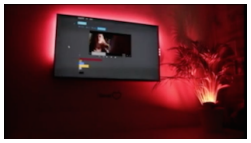
Communicating emotions

Once the movie is affectively annotated, its progress is tracked while the user is watching it, and all the emotion communication modalities are synchronised with it: at each second of the movie, a command for expressing the corresponding emotion is sent to the chosen modalities detailed in the following section.

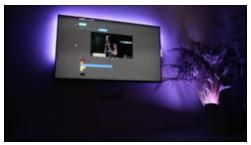
Implemented Modalities

The implemented modalities are inspired by communication techniques used by hearing and visually impaired people as well as by studies related to emotion and its expression through different cues. The main constraint is that the chosen modalities cannot be intrusive and should be applicable in the home environment.

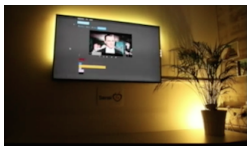
Lights: The lights have been always used to set the mood in houses, concerts, restaurants etc. using three modalities: light intensity, light projection and colours. We have profited from these findings by controlling two Philips hue LEDs and one Philips strip led. The setup of SensiTV is shown in Figure 1. It shows the strip LED attached to the borders of the TV as well as the Philips hue lights that are placed one aiming to lighten up the plant from above and the other from below in order to differentiate between negative and posi-



(a) Anger

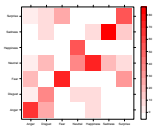


(b) Fear

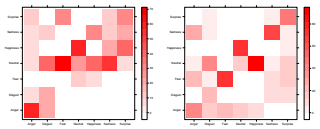


(c) Happiness

Figure 3: Emotion communication via lights and colours



(a) Full Movies



(b) Only audio **(c)** Only video

Figure 4: Confusion matrices of emotion recognition by viewers

tive emotions using colours and light projection type (see Figure 3).

Vibrations: Touch has been shown to be an alternative communication technique with both hearing and visually impaired people[12]. From this logic came the idea of using mobile phone vibrations for conveying emotions to these viewers. Vibrations are conceptualised to play the role of stimuli that will notify the viewers about an uprising emotion. They trigger the viewers attention and try to give hints about the current emotion by using a different vibration pattern for each emotion. In Figure 2 they are presented in a morse code way.

Emoticons: Emoticons are world widely known to convey emotions when using text messaging solutions. We display emoticons corresponding to the movie affective state on the mobile application as well. The viewer can check his second screen to know the current emotion.

Background mood music: We have added to the mobile application a module responsible for playing the corresponding mood music to the current emotion. The music's volume is very low so the viewer can still enjoy his movie and listen to the speech.

Subtitles: We chose the subtitles as an emotion communication modality due to their minimal intrusion level since they are widely used and already accepted by the viewers. The main idea is to personalise the font, the colour, and the size of the subtitles' text according to the emotion associated with the current movie scene (see Figure 5).

Preliminary results

Preliminary tests are conducted to verify the feasibility of the system and to provide a basis for the choice of emotion communication modalities. The user experience related tests (importance of audio-visual content in movies and communication modalities tests) are conducted using an online survey on people not suffering from any impairments. Since the visually and/or hearing impaired do convey and recognise emotions in similar ways as non impaired people [13], but with some delays[6], tests on people not suffering from any impairments are relevant to find emotion communication modalities and easier to organise. The survey was taken by 126 men and 87 women majorly from Switzerland and Lebanon, which ages are spread between 17 and 60 (avg=24,δ=3).

Importance of audio-visual content in movies
A test was conducted to explore the importance of the two dimensions of a movie (audio and video), in the viewer's perception. Movie clips are shown to viewers who are asked to select the corresponding emotion between anger, disgust, fear, happiness, sadness, surprise and neutral. The viewer is showed several video clip types: *i)* normal movies, *ii)* movies with a muted sound; and *iii)* audio clips from movies. The survey comes in three formats in order to shuffle the movie presentation type for each viewer.

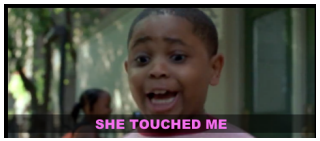
The results of emotion recognition accuracy by viewers are shown as confusion matrices (see Figure 4) where the diagonal line of the confusion matrix is best seen when the movie is presented fully to the viewers, which means that both audio and video streams are important in emotion transmission. Also, these results show that the video has more relevance than audio in distinguishing emotions.

Emotion recognition tests
For training and testing we have used the Acted Facial Expression in the Wild (AFEW) dataset [4, 5]. It consists of two separated movie clip sets: training data and test data. It contains short movie scenes (three seconds) showing a

(a) Anger

(b) Disgust

(c) Happiness



(d) Surprise

Figure 5: Showing emotions through subtitles

specific emotion among the following emotions: anger, disgust, fear, happiness, sadness, surprise and neutral.

The tests are performed using different features' combinations, and for each test, GridSearch is used to test all possible combinations of SVM's parameters in order to find the best solution. The best found estimator is an SVM with an *RBF* kernel, a *penalty parameter of the error term* (C) equal to 1000 and a *kernel coefficient* (Gamma) equal to 0.0001. The combination of audio and visual features have proven to be the best specially when dimensionality reduction techniques are applied such as LDA (number of components = 6) or SVD (number of components = 45). The first results are indicating that the biconcept visual feature is not adapted for this dataset. The confusion matrix associated with the best classification results is shown in Figure 6.

Communication Modalities Tests
The tests were conducted by displaying some of these modalities during the online survey and asking people to guess which emotion corresponds to which encoding.

One of the tested modalities was the colour. People were asked to select a colour from a list of predefined colours (red, green, blue, grey, yellow and purple) to represent the basic emotions. Results are shown in Figure 7a. Anger,

(AV_LDA)

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	53%	12%	6%	8%	11%	5%	5%
Disgust	24%	10%	3%	30%	20%	5%	8%
Fear	23%	4%	23%	13%	20%	9%	7%
Happiness	18%	13%	4%	30%	30%	1%	4%
Neutral	9%	13%	3%	26%	37%	9%	3%
Sadness	4%	15%	6%	29%	29%	13%	4%
Surprise	9%	11%	4%	31%	27%	9%	9%

Figure 6: Confusion matrix for emotion recognition (AV-LDA: audio visual features with LDA applied)

sadness and disgust show a unique response from the users as can be seen. Concerning the other emotions, *fear* will be associated with *blue* since the *grey* is more dominant for the *sadness* emotion. *Happiness* and *surprise* are tricky since both *purple* and *yellow* are associated to them.

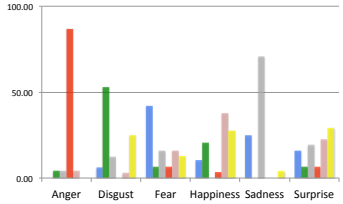
Another test was performed to make sure the used emoticons were a universal standard. Results from the survey part concerning the emoticons can be seen in Figure 7b. The results have assured that the emoticons are a very straightforward modality for emotion communication.

Future work

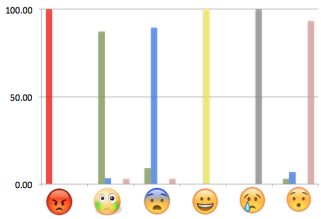
First of all, user experience tests will be conducted on the target audience (visually and/or hearing impaired) to gather their feedback and adjust the system for their needs. According to the results, improvements and adjustments will be done regarding the emotion communication part. Some new channels will as well be considered in a later step, through smart objects for instance, like animated paintings or vibrating seats, in order to provide an even more intense experience. As for the classification part, it will be continuously improved, by adding new features (mainly visual), by enhancing the features selection, and by using more advanced classification techniques (e.g. deep learning).

Conclusion

This paper presented and analysed the problem related to the empathic TV experience of people with hearing and/or visual impairments. To this end, a solution to fill this gap is proposed, by implementing a complete system covering the emotion recognition from movies, as well as the communication of these emotions to TV viewers. The empirical results of emotion classification show high accuracy for some emotion classes and low for others, while the communication process rely on some well proven modalities in



(a) SensiTV survey: association of colours with emotions



(b) SensiTV survey: association of emoticons with emotions (red: anger, green: disgust, blue: fear, yellow: happiness, grey: sadness, pink: surprise)

Figure 7: Communication modalities survey test results

emotion expression. Both of these modules will be objects to enhancements for better performance and better user experience.

References

[1] C. Becker, S. Kopp, and I. Wachsmuth. 2001. Why emotions should be integrated into conversational agents. (2001).

[2] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proc. of ACM MM*. ACM Press, 223–232. http://dl.acm.org/citation.cfm?doid=2502081.2502282

[3] J. De Silva and P. S. Haddela. 2013. A term weighting method for identifying emotions from text content. In *Proc. of IEEE ICIIS*. 381–386.

[4] A. Dhall, R. Goecke, J. Joshi, and T. Gedeon. 2014. Emotion Recognition In The Wild Challenge 2014 : Baseline , Data and Protocol Categories and Subject Descriptors. (2014).

[5] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. 2012. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia* 19, 1 (2012), 34–41. DOI:http://dx.doi.org/10.1109/MMUL.2012.26

[6] M. J. Dyck, C. Farrugia, I. M. Shochet, and M. Holmes-Brown. 2004. Emotion recognition/understanding ability in hearing or vision-impaired children: do sounds, sights, or words make the difference? *Journal of Child Psychology and Psychiatry* 45, 4 (2004), 789–800. DOI:http://dx.doi.org/10.1111/j.1469-7610.2004.00272.x

[7] P. Ekman. 2003. *Emotions Revealed*.

[8] F. Eyben, F. Weninger, F. Gross, and B. Schuller. 2013. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In *Proc. of ACM MM*. 835–838. DOI:http://dx.doi.org/10.1145/2502081.2502224

[9] S. Firdus, W. Fatimah, W. Ahmad, and J. B. Janier. 2012. Development of Audio Video Describer Using Narration to Visualize Movie Film for Blind and Visually Impaired Children. In *Proc. of IEEE ICCIS*. 1068–1072.

[10] A. Hanjalic, L. Xu, C. D. Delft, A. Park, M. Heath, and I. Ip. 2001. User-oriented Affective Video Content Analysis. In *Proc. of IEEE CBAIVL*. 50–57.

[11] A. Hanjalic and L.-q. Xu. 2005. Affective Video Content Representation and Modeling. 7, 1 (2005), 143–154.

[12] M. J. Hertenstein, D. Keltner, B. App, B. a. Bulleit, and A. R. Jaskolka. 2006. Touch communicates distinct emotions. *Emotion (Washington, D.C.)* 6, 3 (Aug. 2006), 528–33. DOI:http://dx.doi.org/10.1037/1528-3542.6.3.528

[13] R. Hiraga, N. Kato, and T. Yamasaki. 2006. Understanding emotion through drawings comparison between hearing-impaired people and people with normal hearing abilities. In *Systems, Man and Cybernetics, 2006. SMC '06. IEEE International Conference on*, Vol. 1. 103–108. DOI:http://dx.doi.org/10.1109/ICSMC.2006.384366

[14] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen. 2014. Combining Multiple Kernel Methods on Riemannian Manifold for Emotion Recognition in the Wild. In *Proc. of ICMI*. 494–501.

[15] J. Ohene-djan and R. Shipsey. 2006. E-Subtitles : Emotional Subtitles as a Technology to assist the Deaf and Hearing-Impaired when Learning from Television and Film. In *Proc. of ICALT*. 2–4.

[16] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, M. Christian, and S. Narayanan. 2010. The INTERSPEECH 2010 Paralinguistic Challenge. In *Proc. Interspeech*. 2794–2797.