Towards Affective TV with Facial Expression Recognition

PEDRO A. VALENTIM, FÁBIO BARRETO, DÉBORA C. MUCHALUAT-SAADE,

MídiaCom Lab - Universidade Federal Fluminense, Brazil

Facial recognition techniques, fantasized in fiction movie classics, have already become reality. Such technology opens up a wide range of possibilities for different kinds of systems. From the point of view of interactive applications, facial expression as input data may be more immediate and more trustworthy to the user's sentiment than the click of a button. For interactive television, facial expression recognition could be used for bringing broadcasters and viewers closer, enabling TV content to be personalized by the user sentiment. In fact, not only facial expression recognition, but any interaction that enables affective computing. In this work, we call this concept *Affective TV*. In order to support it, this work proposes facial expression recognition for digital TV applications. Our proposal is implemented and evaluated in the Ginga-NCL middleware, a digital TV standard used in several Latin American countries.

Additional Key Words and Phrases: Facial Recognition, Affective Computing, Digital Television, Multimedia Applications, Affective TV

ACM Reference Format:
Pedro A. Valentim, Fábio Barreto, Débora C. Muchaluat-Saade. 2021. Towards Affective TV with Facial Expression Recognition. In *LIQUE 2021: Life Improvement in Quality by Ubiquitous Experiences Workshop, together with IMX 2021: ACM International Conference on Interactive Media Experiences. June 21-23, 2021.* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

Affective computing [12] is a branch of computer science that meets psychology and cognitive science. It is based on the idea of embedding emotion-like behaviors to machines or, in other words, emulating empathy on machines. It is generally done by creating a way of receiving as input some aspect of the user emotional state during a certain interaction and interpreting it as data. Once this is done, the machine can respond accordingly.

All kinds of digital applications can benefit from affective computing. For instance, video games can create scenarios to ease or harden a player's experience based on their facial expression or based on the pressure they apply when pushing buttons. Another example is the usage in healthcare, in the absence of workers, for both physical and mental care, evaluating the emotion of a patient based on what they say or by other stimuli [3].

It is also a powerful concept in the sense that it creates a new dimension for human-computer interaction. For every application a user interacts with, that application can provide a more suitable environment for that user. Currently, however, there are only few work taking advantage of affective computing for television [1, 5]. To the best of our knowledge, none of them conceptualizes a framework or a generalization of how to provide affective computing for interactive TV applications.

Aiming at filling this gap, this work proposes a multipurpose solution to employ affective computing for TV using facial recognition and multimodal interaction. From the point of view of interactive applications, facial expression as input data may be more immediate and more trustworthy to the user's sentiment than a button click [7].

For interactive television, facial expression recognition could be used for bringing broadcasters and viewers closer, enabling TV content to be personalized by the user sentiment. In this work, we call this concept *Affective TV*. In order to support it, this work proposes facial expression recognition for digital TV applications to be used as a new interaction mode. Our proposal is implemented and evaluated in the Ginga-NCL middleware [8], a digital TV standard used in several Latin American countries. It is important to highlight that our proposal can be implemented using any kind of input mode that can be correlated to affective computing, such as gesture, eye movement, heartbeat, etc.

The remainder of this paper is structured as follows. Section 2 discusses related work about facial recognition, affective computing and sentiment analysis. Section 3 presents our proposal for affective TV. Section 4 details the implementation of multimodal interactions in the Ginga-NCL middleware and reports evaluation results. Finally, Section 5 brings closing remarks and ideas for future work.

2 RELATED WORK

Facial recognition does not demand unusual devices. Fundamentally, any video capture device (e. g. webcam, smartphone, etc.) is enough to fetch the input data. Therefore, the cost of implementing this technology integrated to a TV may be considerably low, comparing it to the wide applicability. In the context of facial recognition, there is the recognition of facial expressions. For this specialized form, there are the challenges of determining what defines a face, identifying a face and the specific challenges of recognizing expressions, such as identifying which elements of a face define emotions and which emotion it is indicating. It is important to note that for each step there stands fewer consensus and more room for new proposals.

The work of [4] focused on the usage of affective computing for security and based the emotion analysis on facial expressions. It also discusses the ethic implications of that concept. Cohn et al. [6], before reviewing the aspects of automated face analysis, wrote that "while open research questions remain, the field has become sufficiently mature to support initial applications in a variety of areas". McDuff et al. [11] presented AFFDEX SDK, demonstrating that the relation of affective computing and facial expression recognition was undoubtable. In fact, that work proposed it as market-ready.

3 AFFECTIVE TV PROPOSAL

In order to make Affective TV feasible, it is necessary to have a standard architecture for implementing affective computing concepts on interactive television (ITV) applications. Currently, to implement those functionalities, a specific script is needed to setup every aspect of the application interaction, from managing an external device (camera, mic, etc.) and processing the received data to trigger a response on the ITV middleware. This is not ideal, because, for every new application in which its author wants to use any interaction of that nature, they must create a script like that. That script, however, can easily be mostly generalized and the core of its behavior can be incorporated directly into the middleware. We propose those interactions to become events, which will be handled by the declarative language used by the ITV middleware. In many standards, this language is HTML5 – in a similar fashion of more tradition forms of interaction, such as mouse click events. In the ITU-T H.761 standard for IPTV systems, the authoring language is NCL (Nested Context Language) [8], which we used for validating our proposal, as it will be seen in Section 4.

As illustrated in Figure 1, we propose a recognition unit to be incorporated into the ITV middleware in order to provide a set of natively-handled events. Using this architecture, there is no need for the author (application developer) to implement any script in order to use an interaction mode already provided by the middleware. The way our proposal works is conceptually simple. As the application runs, the capture devices fetch audio and video from the user and sends them to the recognition units. As this data is processed by the units, they send the results to the middleware. Lastly, the ITV application receives those results and can behave accordingly. We provide different interaction modes, such as face recognition, facial expression recognition, voice recognition, gesture recognition, etc., and this set could be extended as desired [2].

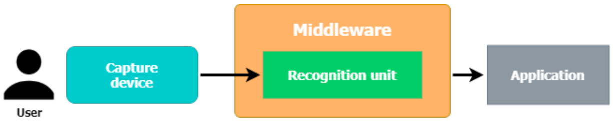

Fig. 1. Proposed architecture to integrate multimodal events for affective TV.

In this paper, we focus on facial expression recognition in order to understand viewer's sentiments. In a broader sense, the ITV application would be able to trigger an action when a certain sentiment is captured for a media content being presented. As an example, angry facial features for a show may imply that the viewer is not enjoying it, then the system could suggest another TV program.

4 IMPLEMENTATION AND EVALUATION

To implement our contribution, we used the Ginga-NCL middleware, to which applications are written in the NCL declarative language. NCL is based on XML, with the event-based paradigm for spatio-temporal synchronization and user interactions. NCL supports scripting using Lua. NCL and Lua could be compared to HTML5 and JavaScript in other middleware standards. Currently, the standard Ginga middleware does not support either facial expression recognition or voice recognition events. It only supports selection events (key press or mouse click).

In our implementation, the communication between the recognition units and the middleware is done via MQTT, a publish-subscribe network protocol that transports messages between devices. It was chosen for being a very simple lightweight protocol. The facial expression recognition unit used in our implementation is based on convolutional neural networks. This choice was made following what seems to be a trend in the field [9, 13]. Our facial expression recognition unit implementation is capable of identifying seven facial expressions: "anger", "disgust", "fear", "happiness", "neutral", "sadness" and "surprise", as illustrated in Figure 2.

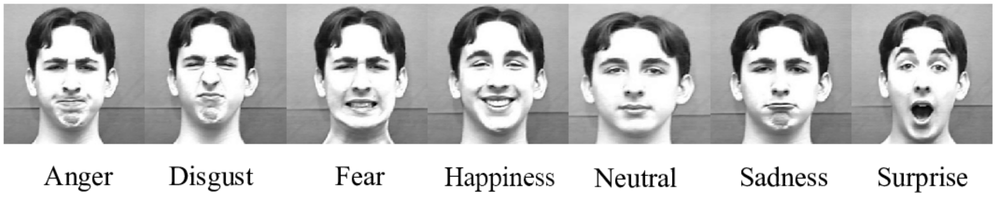

Fig. 2. Labeled sample images of the recognizable facial expressions. Examples collected from the CK+ dataset [10].

In fact, we have made two different implementations in order to compare their performance. The first one is based on the current standard of the middleware, which relies on Lua scripting for recognizing facial expressions. The second one, which actually implements the proposal of this paper, is based on a modified version of the middleware [2]. Figure 3 illustrates the difference between the implementations. The key point to notice is that the NCLua script is completely dropped in the proposed middleware implementation, and in its place a facial expression recognition unit, integrated into the middleware, does the job.

To evaluate our proposal and measure performance, we designed experiments to trace the delay from the recognition unit sending an interaction response to the middleware and the application reacting to the interaction. We ran the tests 100 times for both the proposed architecture and the current standard implementations and calculated the average

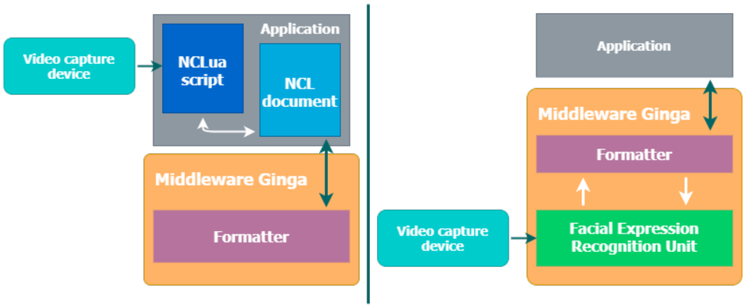

Fig. 3. On the left, the Ginga-NCL standard architecture; on the right, the proposed architecture.

delay. The average delay is considerably reduced using the proposed architecture (59 ms, with a standard deviation of 21 ms) over the current Ginga-NCL standard (222 ms, with a standard deviation of 12 ms). It is important to notice that, although this experiment was conducted using Ginga-NCL as the ITV middleware and Lua for scripting, the principles are similar to implement the proposal for ITV systems that are based on HTML5 and JavaScript.

5 CLOSING REMARKS

This work proposed a facial recognition module for affective TV, allowing applications to easily use facial expression recognition to better understand the viewer's sentiment. The proposal was implemented using the Ginga middleware, where applications are developed in NCL. This work also presented a quantitative comparison of our proposed architecture and the current Ginga-NCL middleware standard. Having those interactions as built-in functionalities in the middleware works better, in criteria of reducing response delay, than having to rely on application scripting.

As future work, it would be worth to have some user-experience-oriented evaluation for affective television. It seems needless to say that to think the ethics involved in having affect-based interaction with television applications and programs, such as thinking of privacy, is a must.

REFERENCES

[1] Sandra Baldassarri, Isabelle Hupont, David Abadía, and Eva Cerezo. 2015. Affective-aware tutoring platform for interactive digital television. *Multimedia Tools and Applications* 74, 9 (01 May 2015), 3183–3206. https://doi.org/10.1007/s11042-013-1779-z

[2] Fábio Barreto, Raphael S. de Abreu, Eyre Brasil B. Montevecchi, Marina I. P. Josué, Pedro A. Valentim, and Debora C. Muchaluat-Saade. 2020. Extending Ginga-NCL to Specify Multimodal Interactions With Multiple Users. In *Brazilian Symposium on Multimedia and the Web*. ACM.

[3] Simon H Budman. 2000. Behavioral health care dot-com and beyond: Computer-mediated communications in mental health and substance abuse treatment. *American Psychologist* 55, 11 (2000), 1290.

[4] Joseph Bullington. 2005. 'Affective'computing and emotion recognition systems: the future of biometric surveillance?. In *Proceedings of the 2nd annual conference on Information security curriculum development*. 95–99.

[5] Konstantinos Chorianopoulos and Diomidis Spinellis. 2004. Affective Usability Evaluation for an Interactive Music Television Channel. *Comput. Entertain.* 2, 3 (July 2004), 14. https://doi.org/10.1145/1027154.1027177

[6] Jeff F Cohn and Fernando De la Torre. 2015. Automated face analysis for affective computing. (2015).

[7] U Dimberg, M Thunberg, and K Elmehed. [n.d.]. Unconscious facial reactions to emotional facial expressions. *Psychological science* 11, 1 ([n. d.]).

[8] ITU. 2014. Nested Context Language (NCL) and Ginga-NCL. http://www.itu.int/rec/T-REC-H.761. ITU-T Recommendation H.761.

[9] K Liu, M Zhang, and Z Pan. [n.d.]. Facial expression recognition with CNN ensemble. In *2016 international conference on cyberworlds*.

[10] P Lucey, J F Cohn, T Kanade, J Saragih, Z Ambadar, and I Matthews. [n.d.]. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-workshops*.

[11] D McDuff, A Mahmoud, M Mavadati, M Amr, J Turcot, and R Kaliouby. 2016. AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. 3723–3726.

[12] Rosalind W. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, MA, USA.

[13] Minchul Shin, Munsang Kim, and Dong-Soo Kwon. 2016. Baseline CNN structure analysis for facial expression recognition. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 724–729.