AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

ROADMAP ON AI TECHNOLOGIES & APPLICATIONS FOR THE MEDIA INDUSTRY

SECTION: "AFFECTIVE ANALYSIS"

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu

Authors	**Ioannis Patras** (Queen Mary University of London) **Niki Foteinopoulou** (Queen Mary University of London) **Ioannis Maniadis Metaxas** (Queen Mary University of London) **Ioanna Ntinou** (Queen Mary University of London) **James Oldfield** (Queen Mary University of London) **Christos Tzelepis** (Queen Mary University of London) **Georgios Zoumpourlis** (Queen Mary University of London)

This report is part of the deliverable D2.3 - "*AI technologies and applications in media: State of Play, Foresight, and Research Directions*" of the AI4Media project.

You can site this report as follows:

F. Tsalakanidou et al., Deliverable 2.3 - AI technologies and applications in media: State of play, foresight, and research directions, AI4Media Project (Grant Agreement No 951911), 4 March 2022

This report was supported by European Union´s Horizon 2020 research and innovation programme under grant number 951911 - AI4Media (A European Excellence Centre for Media, Society and Democracy).

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf.

Affective analysis

Current status

Over the last two decades, there has been an increasing interest towards modelling human affect through the field of ***Affective Computing***. Affective Computing is *computing that relates to, arises from or deliberately influences emotion or other affective phenomena*[1]. One of the incentives of such research is the creation of ***empathetic machines***, i.e. machines that understand and interpret a human's emotional state and adopt their behaviour to give responses corresponding to our emotions and moods. Human affective and cognitive mental states are pivotal to human experience, and hence, the creation of empathetic machines can possibly influence mental health care through automatic depression detection[2], pain estimation[3] or post-traumatic stress disorder identification[4]. Other uses include automotive industry[5], education[6,7] and media as it will be discussed below.

Figure 1: Face of different identities and expressions.[8]

Key for all of the aforementioned applications is reliable estimation of human affect. Most of the existing works on ***automatic emotion prediction*** rely on two psychological theories that can generally be distinguished according to the way emotion is modelled: to be categorical or continuous. In the *continuous approach*, which is coined ***Valence-Arousal-Dominance (VAD)***

[1] R. Picard, "Affective Computing" MIT Technical Report #321 (Abstract), 1995
[2] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, and D. P. Rosenwald. Social risk and depression: Evidence from manual and automatic facial expression analysis. In FG, 2013
[3] Xin X, Lin X, Yang S, Zheng X. Pain intensity estimation based on a spatial transformation and attention CNN. In PLoS One, 2021
[4] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency. Automatic nonverbal behavior indicators of depression and ptsd: Exploring gender differences. In ACII, 2013.
[5] C. Busso and J. J. Jain. Advances in Multimodal Tracking of Driver Distraction. In Digital Signal Processing for in Vehicle Systems and Safety, pages 253–270. 2012
[6] B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp, and a. Graesser. Facial Features for Affective State Detection in Learning Environments. 29th Annual meeting of the cognitive science society, pages 467–472, 2007
[7] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. IVC, 2010
[8] Figure taken from: O. Ignatyeva, D. Sokolov, O. Lukashenko, A. Shalakitskaia, S. Denef, T. Samsonowa, "Business Models for Emerging Technologies: The Case of Affective Computing", In 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) 2019 Sep 3 (pp. 350-355). IEEE.

Model[9], emotions are described based on three dimensions in the range [-1,1], with *Valence* referring to the level of positiveness (-1 being negative and +1 being positive), *Arousa*l referring to the level of activation (-1 Being deactivated And +1 being activated), and *Dominance* referring to the level of control a person feels over a given situation (-1 being submissive and +1 being in-control (Figure 2). On the other hand, the *categorical approach* refers to ***seven basic emotions***, defined by Paul Ekman[10], namely Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Neutral. Additionally, Ekman and Friesen[11] proposed another system for emotion analysis, which is based on the exclusive analysis of facial expressions, namely ***Facial Action Coding System***. This system defines emotional states as combinations of atomic facial muscle movements, coined *Action Units (AUs)*.

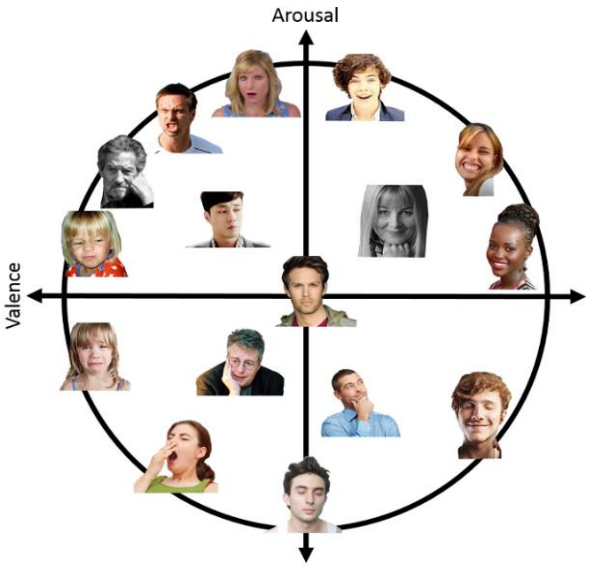

Figure 2: Illustration of Russell's dimensional emotion modelling scheme (Arousal-Valence).[12]

A significant body of work on emotion prediction is on unimodal affect prediction and in particular on facial expression analysis. This happens because facial signals are among the most important channels of nonverbal communication, thus offering discriminative cues for affective prediction. Based on this, there has been a significant development of datasets on the task of facial expression analysis with extensive Action Unit and VAD annotations. Other single modalities explored in the literature include voice and speech expressions[13], body gestures[14],

[9] A. Mehrabian, "Framework for a comprehensive description and measurement of emotional states." Genetic, social, and general psychology monographs, 1995.
[10] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." Journal of personality and social psychology, vol. 17, no. 2, p. 124, 1971
[11] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," Palo Alto, 1978
[12] Figure taken from: A. Mollahosseini, B. Hasani, M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild", IEEE Transactions on Affective Computing, 2017 Aug 21;10(1):18-31
[13] K. Scherer, T. Johnstone, and G. Klasmeyer. Vocal expression of emotion. Handbook of affective sciences, pages 433–456, 2003
[14] C. Navarretta. Individuality in communicative bodily behaviours. In Cognitive Behavioural Systems, pages 417– 423. Springer, 2012

gait[15], and physiological signals such as respiratory and heart cues[16], and more recently general context and scene background[17].

Although most of the works in affect prediction are focused on facial expression analysis, making predictions from a single modality can be challenging, especially because single sensory observations are ambiguous or incomplete. Indeed, many researchers advocate towards ***Multimodal Emotion Recognition***, i.e. the fusion of multiple modalities, based on two main arguments: a) *Accuracy increase*: different modalities can complement each other resulting in more accurate inference and, b) *Reliability of predictions*: single sensor observations can be corrupted leading to ambiguous, noisy or incomplete data. Hence, combining observations from multiple sources can potentially mitigate such challenges. To combine different modalities three different approaches have been investigated: feature-level (early) fusion, decision-level (late) fusion and model-level fusion. The main objective in such approaches is figuring out *when,* i.e. which stage of the training process, and *how,* i.e. with what weights, different modalities can be fused to leverage information from the most reliable ones. Overall, multimodal recognition has significantly advanced the task of emotion prediction in the wild in datasets like CMU-MOSEI[18] and IEMOCAP[19].

Research challenges

Despite the significant advances of affective computing in the last two decades, there are still numerous challenges that need to be addressed in the future. A number of them is discussed below.

Affect modelling across cultures. One recurrent debate in affective computing is whether emotion expression is a universal, biologically based construct or a social construct, i.e. affect expression changes across cultures. Paul Ekman argued that emotion is biologically defined, and hence, emotions are experienced and interpreted similarly across cultures. Indeed, this theory proposes six basic emotions (e.g. happiness, sadness, disgust) as the basic feelings someone can experience. Contrary to the basic emotions theories of Ekman, Russell and Barrett suggest that emotions change across cultures and proposed that emotions can be defined by three-independent dimensions: pleasant-unpleasant, tension-relaxation, and excitation-calm, i.e. the Valence-Arousal-Dominance model. In the field of affective computing, some datasets are annotated with Ekman's models (6 basic emotions or AUs), others in the Valence-Arousal-Dominance model and others in a less structured manner with emotional words. There is a challenge in developing computational models that can learn with all of those annotations and

[15] T. Randhavane, A.t Bera, K. Kapsaskis, U. Bhattacharya, K. Gray, and D. Manocha. Identifying emotions from walking using affective and deep features. arXiv preprint arXiv:1906.11884, 2019

[16] B. Knapp, J. Kim, and E. Andre.´ Physiological signals and their use in augmenting emotion recognition for human–machine interaction. In Emotion oriented systems, pages 133–159. Springer, 2011

[17] R. Kosti, J.M. Álvarez, A. Recasens and A. Lapedriza, "Context based emotion recognition using EMOTIC dataset", IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2019

[18] A. Zadeh, P.P. Liang, S. Poria, E. Cambria and L.P. Morency. "Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph." ACL (2018): 10.18653/v1/P18-1208

[19] The Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database: https://sail.usc.edu/iemocap

learn mappings between them that are informed by theories in the field of psychology – those will be potentially conditioned on cultural and other contextual factors.

Dataset bias and inconsistent annotations. A shared problem across several domains is the fact that data collection conditions are inevitably limited, and hence the available training and testing datasets are biased. This is evident during cross-dataset experiments where we notice a significant discrepancy in performance when an algorithm trained on a dataset generalises well on the test set of the same data but performs poorly on a different dataset. This problem is more prominent in the field of affective computing, as emotion annotations can be very subjective thus can vary a lot between different annotators. This means that even if annotations are consistent within a specific dataset, it will be very hard to be consistent across different data. Given this, merging multiple datasets for affect prediction is a challenging task.

Class imbalance. One common challenge in datasets that use categorical emotions is class imbalance. This is traced back in the data acquisition process where collecting positive feelings is much easier than collecting data that correspond to negative feelings like anger or disgust. Overall, capturing or collecting data associated with negative emotions is a challenging task that needs to be addressed.

Design choices in multimodal affect recognition. While combining data from different sensors has advanced the field of affective computing, it comes with new challenges. To design a multimodal system, we first need to decide which modalities can be combined and how. Some modalities are co-occurring while others are not. For example, IEMOCAP and CMU-MOSEI datasets hold a similar set of modalities (facial expressions with the corresponding text and audio) and are typically employed together in the literature. Additionally, one critical design challenge is the way different modalities can be combined together, typically coined fusion. As stated above, there are three main fusion techniques, feature-level (early) fusion, decision-level (late) fusion and model-level fusion; however, it is not clear how modalities can be effectively fused given that there is inherent asynchronicity in the different data streams, interactions at potentially large time frames, and lack of large-scale data that would allow examining many different architectural choices. A challenge in this area is to design architectures that model the interplay between the different modalities at the appropriate level.

Context modelling. When trying to estimate the emotions that visual stimuli (e.g. images) elicit to human subjects, parsing the entire image and directly performing inference, is an approach with inherent limitations. Not all image regions contribute equally to emotion elicitation processing, and also individual objects/scene parts might not contribute alone to emotions, but through their interactions/relationships with other objects/scene parts. Stronger insights can be obtained through machine learning models that are capable of reasoning about object-to-object and object-to-scene interactions[20]. A first step towards this direction, is the development of techniques that achieve spatial localisation of regions that influence sentiment evoking, producing soft sentiment maps[21]. Also, the exploration of architectures that can learn emotion-

[20] J. Yang, X. Gao, L. Li L, X. Wang, J. Ding, "SOLVER: Scene-Object Interrelated Visual Emotion Reasoning Network", IEEE Transactions on Image Processing. 2021 Oct 19;30:8686-701

[21] D. She, J. Yang, M. M. Cheng, Y. K. Lai, P. L. Rosin, L. Wang, "WSCNet: Weakly supervised coupled networks for visual sentiment classification and detection", IEEE Transactions on Multimedia, 2019 Sep 5;22(5):1358-71

related representations, taking into consideration the context around image regions[22]. Finally, the currently existing affective datasets do not suffice to develop robust affective models with capabilities as those described above. Some of the problems that the existing datasets have, are the following: i) small sample cardinality, ii) lack of fine-grained emotion annotations, iii) lack of multi-label emotion annotations, and iv) lack of spatially localised annotations. Thus, the curation of affective datasets with such content and annotations, will help to build stronger machine learning models.

Causality. In scenarios where multimodal media content is used as stimuli for emotion elicitation to users, estimating the temporal causality from the various stimulus modalities is a difficult task. For example, considering a movie clip as a stimulus, it may contain varying sources of information, such as actor facial expressions, actor speech, background music, the scenery that is depicted, the text that is articulated by the actors, etc. Until now, emotion recognition approaches may receive multimodal inputs and infer affective states, but the problem of determining which specific cues of the multimedia content caused that specific affective state, is far from solved[23].

Societal and media industry drivers

Vignette 1: Using affective analysis to avoid sensationalism in news coverage

Anna is an editor in a big national newspaper that covers the civil war in another country. While her aim is to ensure the public is aware of the issues, she wants to ensure that the material is objectively portraying the situation without use of sensationalist material that would be against the ethical code of conduct. As a large number of articles come through every day, a tool that helps understand where each piece stands in terms of emotional neutrality would help her more accurately curate the material. Furthermore, if such a tool could also localise the emotionally heavy text extracts and suggest less charged language it would greatly help editors in the paper to maintain the standard set by the ethical code. Similarly, with video reports of events automatic analysis of content could help either broadcasters or independent regulators ensure sensationalism and emotionally charged news do not air at the expense of accuracy.

Vignette 2: Producing media content that captures audience engagement and attention

Kate is a director working for a media production company that focuses on educational content for children. Given their intended audience, it is essential for the content they produce to be highly engaging in order to maintain children's attention. This is a very difficult challenge, given that the content must also be informative and leverage attention to achieve educational objectives. Furthermore, it is important that the emotions the content stimulates in the

[22] Z. Xu, S. Wang "Emotional Attention Detection and Correlation Exploration for Image Emotion Distribution Learning", IEEE Transactions on Affective Computing, 2021.
[23] T. Mittal, P. Mathur, A. Bera, D. Manocha "Affect2MM: Affective analysis of multimedia content using emotion causality", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021 (pp. 5661-5671)

audience must not be harmful or traumatising, which is a concern when subject matter is sensitive in nature.

This problem can be addressed with focus groups consisting of the audience in question (in this case children), or by judgment calls by adults. The latter method is, naturally, less dependable, while the former can be expensive, time consuming, and imprecise, depending on the size of the focus group and how accurately it represents the target audience. AI tools measuring the engagement and affective response could facilitate this process by providing estimates of the impact of each piece of content on the audience, and, ideally, identify ways to retain attention and invoke engagement that are also productive and educationally beneficial.

Future trends for the media sector

Below, we highlight some of the emerging opportunities for research in the field of affective analysis to be translated into applications in the media sector:

- Creating tools that can identify and flag content that might trigger intense or undesirable emotional responses from the audience. Such tools could be used for **content moderation** across a variety of platforms and content modes (audio, visual and text), and could help shield the audience from problematic content that might otherwise evade other filters.
- Integrating affective analysis tools in **interactive media** (e.g. video games) would open the possibility for content producers to adjust their creations to the responses of the audience in real time and in a personalised manner. For example, video games might adjust their gameplay and/or visual and audio based on a player's emotional feedback to make them feel more challenged, be engaged, or to increase their immersion in the game's narrative.
- Developing more fine-grained and complex **multimedia tagging systems**, taking into consideration recent research findings about the multi-dimensional distribution of emotional states evoked by multimedia stimuli.[24,25]
- Measuring the audience engagement and tracking behavioural changes during the presentation of multimedia content (e.g. movies), can be used to train ***models with predictive capabilities***, so as to estimate future ratings of a movie[26].
- Building **recommendation systems** for multimedia content, based on user preferences that are estimated from neurophysiological signals[27].

[24] A. S. Cowen, D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients", Proceedings of the National Academy of Sciences, 2017 Sep 19;114(38):E7900-9

[25] J. J. Sun, T. Liu, A. S. Cowen, F. Schroff, H. Adam, G. Prasad, "EEV: A Large-Scale Dataset for Studying Evoked Expressions from Video", arXiv preprint arXiv:2001.05488

[26] R. Navarathna, P. Carr, P. Lucey and I. Matthews, "Estimating audience engagement to predict movie ratings", IEEE Transactions on Affective Computing, 2017, 10(1), pp.48-59.

[27] S. Koelstra, et al., "DEAP: A Database for Emotion Analysis; Using Physiological Signals". IEEE Trans. Affect. Comput. 3(1): 18-31 (2012)

- Generation of novel media content (alternatively to the selection of pre-existing content) to be streamed based on user preferences, combining neurofeedback techniques and generative machine learning approaches[28].

Goals for next 10 or 20 years

The long term goal of Affective Computing in analysis and understanding of affect requires progress in several fronts, including in the theory of emotions and in understanding of the role they play in the way that we perceive the world and ourselves, in the way that we interact with others and in the way we act in the world. Affect manifests in widely varied ways, depending on the context - this varies from prototypical facial expressions, to subtle changes in the behaviour and neurophysiological responses. In this sense a holistic, multimodal approach that models the interplay between different human signals is required. To this end progress in ***sensing and data collection equipment at large scale*** *is* essential. In this direction, technological developments that will provide equipment capable of capturing modalities such as MEG and fMRI in out-of-lab environments[29, 30] would be needed in the next 5 years. Beyond that, a major goal in this direction for the next 5-10 years would be data collection at large scale, and the development of models that learn from this data collectively, so as to build personalised models in a way that respects privacy.

Recognition of emotions in others requires a theory of mind; that is an understanding of people's goals and their perception of the world. In this sense, progress in the field of Affective Computing requires progress in ***Machine Perception***, including among others in the field of computer vision, natural language processing and audio and speech analysis so as to model the world and the people in it. A goal with a 10-20 years horizon would be the development of artificial intelligence systems that would be able to ***model people's goals and intentions*** and be able to assist them in achieving them.

[28] M Spapé et al., "Brain-computer interface for generating personally attractive images", IEEE Transactions on Affective Computing, 2021, PP. 1-1.
[29] T. Horikawa, A. S. Cowen,D. Keltner, Y. Kamitani "The neural representation of visually evoked emotion is high-dimensional, categorical, and distributed across transmodal brain regions", iScience, 2020 May 22;23(5):101060
[30] M. Liu, N. van Rijsbergen, O. Garrod, R. Ince, R. Jack, P. Schyns, "Semantic Decoding of Affective Face Signals in the Brain is Temporally Distinct", Journal of Vision, 2021, Sep 27;21(9):2589

AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu