Is automatic facial expression recognition of emotions coming to a dead end? The rise of the new kids on the block

Gunes, H; Hung, Hayley

DOI
10.1016/j.imavis.2016.03.013

Publication date
2016

Document Version
Final published version

Published in
Image and Vision Computing

Image and Vision Computing 55 (2016) 6–8

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Is automatic facial expression recognition of emotions coming to a dead end? The rise of the new kids on the block☆

Hatice Gunes[a,*], Hayley Hung[b]

[a]Computer Laboratory, University of Cambridge, UK
[b]Technical University of Delft, The Netherlands

ARTICLE INFO

Article history:
Received 31 January 2016
Accepted 29 March 2016
Available online 8 April 2016

Keywords:
Opinion paper
Facial expression recognition
New emotional and social signals

1. The success of facial expression recognition

A facial expression is displayed by moving the muscles beneath the skin of the face. Facial expressions convey social and emotional information between humans, and according to some researchers, they are the primary means of non-verbal communication. Over the last 2 centuries, many researchers from Darwin to Duchenne, investigated how humans feel, express and recognize emotions. Over 50 years ago, Ekman and his colleagues conducted various experiments of human judgement on still photographs of deliberately displayed facial behaviour and concluded that six basic facial expressions of emotion can be recognized universally: happiness, sadness, surprise, fear, anger and disgust. To provide a more complete description of the facial behaviour, Ekman and Friesen later on developed the Facial Action Coding System (FACS) for coding fine-grained changes in the face, which are related to facial muscle activations.

Automatic facial expression recognition (often abbreviated to A/FER) is a multidisciplinary research field that spans across computer vision, machine learning, neuroscience, psychology and cognitive science. In automatic facial expression recognition research, the most common approach is to classify continuous expressive facial displays according to specific labels, categories or dimensions. Ekman's theory of basic emotions is the most commonly used scheme when creating vision-based systems that attempt to recognize facial expressions of emotions and analyse human affective behaviour. The main assumption is that emotions that are felt inside the body are displayed externally via the face, and these in turn can be universally mapped into the six categories of happiness, sadness, surprise, fear, anger and disgust. In reality though felt emotions are not always so visibly manifest because the experience is subjective, nor do they map cleanly to Ekman's six categories. Another limitation of this approach is that expressive facial signals are highly context dependent and will communicate different things in different context — emotions, cognitive load, back-channelling, turn-taking, etc.

In the early 1990s, a number of facial expression recognition researchers had a motivation of revolutionising the way we interact with technology [1] by enabling it to become more human like. By being able to analyse human emotions through the displayed facial expressions and responding to these in an appropriate and meaningful way, machines would become more intuitive and emotionally and socially intelligent. This paved the way for novel computer vision techniques for analysing people's facial expressions. It has been over 15 years since [1] was published in the IEEE Transactions on Pattern Analysis and Machine Intelligence. Since then, AFER, and in particular recognising the six categories of emotions, have received a lot of attention in both the computer vision research community and the press.

In addition to the computer vision researchers, by now AFER has received considerable attention from machine learning researchers, which is understandable. For many problems, where the (input) sensing conditions and output labels are more or less standardised,

☆ This paper has been recommended for acceptance by Sinisa Todorovic, PhD.

* Corresponding author.
E-mail addresses: Hatice.Gunes@cl.cam.ac.uk (H. Gunes), H.Hung@tudelft.nl (H. Hung).

http://dx.doi.org/10.1016/j.imavis.2016.03.013

nearly frontal faces, constant/acceptable illumination conditions and the six emotion categories in this case, researchers without expertise in the relationship between felt emotion and displayed expression can come in and apply their different techniques to solve this input–output problem on publicly available datasets.

This trend has caused many people outside the AFER research field, and in particular the media, to believe that facial expression recognition is a solved problem. However, AFER researchers have frequently reported that while expression classification works reasonably well for posed expressions, such as posed smiles, their performance drops quite dramatically on spontaneous expressions elicited during natural conversations and day-to-day interactions [2,3]. One of the biggest issues is working out how to obtain ground truth labels for spontaneous expressions and modelling the fact that individuals have subjective and idiosynchratic ways and scales of expressing emotions.

As announced recently by the Wall Street Journal, Apple has just bought Emotient [4], 'a startup company that utilises artificial intelligence to analyze facial expressions and read emotions'. With the acquisition of Emotient by Apple, we can confidently state that the biggest success of AFER research field has been the spin out companies such as Affectiva and Emotient in the USA, and CrowdEmotion in the UK. These companies mainly deliver market research related output, i.e. analysing how much viewers smile while watching an advert or a movie clip. Another 'lighter' application has been the smile detector embedded in digital cameras, and mobile apps that enable someone's facial expression to be modified and morphed, possibly for sharing with their social network for fun and entertainment.

2. Coming to a dead end?

On the one hand it has been great to see the growth in research in this domain — recognising Ekman's six categories in clean conditions is now a solved problem. On the other hand, we can ask whether this has led to a sufficient growth in the AFER area as most of the new researchers that are coming in from outside assume that the inputs and outputs are already a well understood phenomenon.

As mentioned earlier, we know that the six categories of emotions have no use for the majority of everyday applications. This simplification of the task, while serving us well in the early days, needs to change significantly. This forces us to move into uncomfortable territory where we have to ask ourselves the more fundamental questions like: what is the contemporary definition of emotion in this technologically-driven fast-changing world that is very different from that of Darwin's? How are these emotions represented in facial expressions? How do we do the labelling (in time and also type — frames, intervals, FACS, dimensional, etc.)? Recently a number of researchers have been arguing that the continuous and dimensional approaches match better with reality these days, but how many people are working on that compared to using simplistic data sets acquired under simplified and controlled conditions (e.g., the Cohn–Kanade or MMI Database)?

Emotient's acquisition by Apple coupled with the statement made by Andrew Moore, the dean of computer science at Carnegie Mellon, that 2016 is the year when machines learn to grasp human emotions, should in theory excite all of us researchers that have been working in this challenging field for some time. However, as insiders we are rather apprehensive about this news. Moore's statement regarding the spreading trend across the industry in emotion recognition technology is indeed correct. However his statements about computers doing a better job than humans in accessing emotional states and humanity getting to a stage where we will be having more meaningful dialogue with computers is debatable. Moore is right however, in pointing that emotion recognition technology can be used for many everyday applications including mental heath, security, determining patient pain, and tracking how shoppers react to products in stores.

Despite the dream described by Moore, the current state of the AFER domain seems to indicate that AFER researchers no longer know what their work is really about. The most prominent researchers in the field appear to be constantly proposing more elaborate and complex machine learning or computer vision approaches, aiming to publish at conferences such as International Conference on Computer Vision (ICCV) or International Conference on Computer Vision and Pattern Recognition (CVPR), losing track of what they are really trying to do. What are AFER researchers really trying to achieve? What is the real research problem in AFER? What is the dream that was/is being sold?

3. A new age of expression recognition: the new kids on the block?

While attempting to answer the abovementioned questions, we need to keep in mind that since the publication of the PAMI surveys in 2000 [1] and 2009 [2], our understanding of how humans and technology interact has changed considerably as social media and mobile phones have become the predominant ways in which we interact with technology.

With the huge increase in mobile phone usage, we interact with technology mostly in dynamic and noisy environments, often while being on the move. This shift from the *personal computer* to the *portable computer* has led to a change in the human–computer interaction paradigm. This shift forces us to face the challenging question of whether the visual understanding of human emotions and social behaviour is still the primary modality of interest for researchers in this domain. We already know that not all aspects of emotions can be measured using the same sensors; for instance, the arousal dimension is known to be better communicated with nonvisual signals such as voice or with physiological signals [2]. So, are we as a research community, moving with the shift in people's relationships with technology? Or have we become stuck in solving problems for technologies of a bygone age?

Let us look at a prominent application domain that keeps on receiving an ever-growing amount of research funding — health care. With a growing and aging population, there is an increasing demand, as well as political and social pressure to revolutionise health care around the world, particularly in the wealthier countries such as the USA, the UK and Japan. What has the automatic facial expression recognition technology delivered in health care and autism domains to date? Is it convincing to say that the promise has already been delivered by other modalities that (i.e., the new social signals) we refer to as *the new kids on the block*. Simple bio signals such as Electrodermal activity (EDA) have been covering much more ground and delivering practical, realistic and life changing solutions (such as early seizure prediction and warning). These coupled with the myriad ways the mobile sensing technology provides (location sensing, acceleration sensing heart rate monitoring) readily in our pockets, has revolutionised the way intuitive and ecologically valid sensing can be done and integrated into daily life without the need for the analysis of face and facial expressions.

4. Issues for the future

4.1. Moving from vision-only to multimodal emotion sensing

As we already know, different emotions can be better expressed by one modality rather than the other. The most incremental transition from vision-only AFER systems is to include the audio modality. This is particularly needed to correctly analyse and differentiate the

facial deformations caused by expressions from facial deformations caused by speech. Some researchers have started to work more in this area but the community effort is still small.

In the day to day use of technology, the usage of other alternative sensing modalities such as touch, rgbd, bio signals, and other wearables, has been taking over. However, in the AFER research world we still see the dominance of the vision modality, which is clearly the default choice for people who have dedicated their years and careers working in this domain. We need to then ask ourselves, are other data sets for these new modalities not available for low entry level research? Is the whole community shooting itself in the foot by not enabling more low-entry level research in the areas where progress is really needed? Are we ready to accept that other sensing modalities (e.g. audio, keyboard usage, phone call use, heart rate, and GSR acceleration) in fact are acting as a game changer?

4.2. Educating the next generation

As mentioned already the old style of AFER using Ekman's six categories is totally out-dated. However, to investigate what the underlying problem is, would require the collection of new data and a new way of thinking about how to label the data. To do that, we need to be training more people who understand the relationship between facial expression and emotions, and affective computing.

If we see the number of affective computing or social signal processing courses in the world, we can probably already see the issue. The number of machine learning and computer vision courses are likely to significantly outnumber those. But how can we possibly train people to solve problems when they do not understand what the problem is? What is more, this goes far away from the safety of making simplified assumptions that can be nicely formulated into an easy optimisation problem. Once the notion of ground truth starts changing, who is qualified to help question that and develop and refine that notion so that we can go beyond those killer six categories or the two dimensions of arousal and valence?

4.3. Moving deeper into the wild

The current picture shows that majority of the AFER researchers are actually doing computer vision, with the aim of solving the expectations of yesterday. We no longer can define the goal to be facial expression recognition for personalised computing, because computing itself has been transformed. Instead of having a machine that is portable and understands us intimately, i.e., what we are feeling right now, the current problem is understanding the true emotions in the wild in real life contexts.

The prevalance of mobile and wearable technology shows that predicting or perceiving our needs is the way to go, i.e. the personal butler/assistant applications such as Google Now — a digital companion that knows all about you, does not share that information with others, and can help facilitate all the needs the user has in life from socio-emotional needs to career ambitions to health. To get to that stage, the idea that a video camera will be pointed towards our face anytime and anywhere is unlikely. Therefore, the biggest question we need to ask ourselves is whether the visual understanding of human emotions and social behaviour is still the primary modality of interest for researchers in this domain.

The really fascinating new problems arise when we try to estimate the sentiment of experience in the multi-sensorial real world of today. 15 years ago, smart phones did not exist. Now they have revolutionised not just how we live but also how we think. The challenge is addressing how we can link the spontaneous behaviour that we exhibit as we navigate through our every day lives and how this relates to real emotions and feelings. How do we label these? Can we rely on clean labels? Probably not. We will end up with a multitude of noisy labels that could be associated with all sorts of activities, embedded in a whole load of short term and long term contexts. This is an extremely challenging problem but one that is interestingly fundamental to computer science, and yet, not sufficiently tackled. Perhaps because of that, we are all looking forward to see what Apple will do with the emotion recognition technology of Emotient. Will Apple be able to use its renown creativity to find the killer app that the AFER field has been waiting for? Or is this yet another hype that will soon pass and leave us AFER researchers to face the questions of tomorrow? We shall see.

Acknowledgments

Hatice Gunes' work is partially supported by the EPSRC under its IDEAS Factory Sandpits call on Digital Personhood (Grant ref: EP/L00416X/1). Hayley Hung was partially supported by the Dutch national program COMMIT, by the European Commission under contract number FP7-ICT-600877 (SPENCER), and is affiliated with the Delft Data Science consortium.

References

[1] M. Pantic, L.J.M. Rothkrantz, Automatic analysis of facial expressions: the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) (2000) 1424–1445.
[2] Z. Zeng, M. Pantic, G. Roisman, T. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2009) 39–58.
[3] E. Sariyanidi, H. Gunes, A. Cavallaro, Automatic analysis of facial affect: a survey of registration, representation and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (6) (2015) 1113–1133.
[4] Apple buys emotion-reading ai company emotient, http://www.wired.co.uk/news/archive/2016-01/08/appleemotient-ai-emotions.