Journal of Future Robot Life 1 (2020) 91–97
DOI 10.3233/FRL-200007
IOS Press

Emotional manifestations of humanoids of the future

Raphaël C.-W. Phan [a,*] and Elizabeth Sheppard [b]
[a] *Monash University, Malaysia*
[b] *University of Nottingham, UK*

Abstract. This article puts forth a vision of what *emotionally manifesting* capabilities the robots of the future will have, and then details the latest state-of-the-art techniques in the current computer science, engineering and psychological literature that will form the enabling blocks to realizing this future vision.

Towards this goal of emotional manifestation, robots of the future would be better known as humanoids as they will not just mimic human behaviour but in fact be ideally *indistinguishable* from humans. These types of humanoids will have the capability to manifest emotions like any normal human in such a manner that another human will not be able to tell if they are engaging with a human or a humanoid. In fact, if this indistinguishability is really exhibited well, not even other AI-enabled machines will be able to tell the difference between human vs humanoid.

To manifest this emotional capability requires the humanoid to be capable of both *recognizing* human emotions as well as *expressing* human-like emotions in reaction to the human with whom it is interacting.

The article will detail the latest breakthroughs in allowing machines powering such humanoids to recognize human emotions including subtle emotions, and then discuss the recent scientific advances enabling the *synthesis* and *transfer* of human emotions onto other potentially synthetic non-expressive faces.

Underpinning these discussions will be the need to *map* human expressions to underlying emotions being felt by humans, and the issues associated with doing so will be examined from a psychological perspective.

This article will conclude by highlighting prospective challenges that will need to be addressed in order to achieve the future vision of emotionally manifesting humanoids.

1. INTRODUCTION

The world as we experience now is one of mixed reality; where the physical world is realistically blended with the cyber realm that interacts with and reacts to real-world stimuli.

The trend started some decades ago with virtual reality, where a digital representation of the real world is so constructed as to give the human a real world-like experience within that digitally created realm. The virtual experience was decent, and yet virtual reality did not really take off as a transformational technology; it was hard to realistically mimic the real world entirely with digital counterparts. The next and more recent innovation was super-imposing digitally constructed content over the human's real-world view, so-called augmented reality. This is nice, and yet it is still evident from this combined view of the surroundings as to which is real and which is digitally constructed. Furthermore, this union between real and cyber is confined to one's view of the world rather than encompassing the human's multiple senses beyond his/her sense of sight.

In contrast, the future we envision, and as championed by this article, is one where humans and human-like machines i.e. humanoids will co-exist, and in such a manner that they will be indistinguishable from each other. That is the key paradigm shift in how humans should really immerse themselves in a mixed reality experience of the future.

[*] Corresponding author. E-mail: raphael.phan@monash.edu.

Central to humanoids being able to behave like humans is the ability to manifest their emotions as they interact with real humans, such that they can sense human emotions and react emotionally to them in a naturalistic human-like manner.

2. ENABLING TECHNOLOGIES

Key to machines being emotionally aware of their surroundings is firstly the ability to sense as much information as possible about the human whose emotions are being manifested or the situations that trigger such emotions; and three main categories of enabling technologies have led to this capability:

- *Sense: Internet of everything* (*IoE*)
  - ∗ The proliferation of diverse types of built-in sensors in basically any human-owned device, due to the so-called internet of everything (IoE) phenomenon, now means that data can be sensed anywhere at anytime, often with the human owner oblivious to the fact. From smart meters within smart grids sensing our home's energy usage, to smart refrigerators and spying smart TVs, data of various modalities (textual, visual, audio) could be sensed from remote locations. As we walk the streets, our actions may be within the view of CCTV cameras, or even be captured in the background of passers by's selfie shots or livestreams; and as we use mobile phones for our daily activities, we are oblivious to the many built-in sensors in such devices including accelerometers, gravity sensors, gyroscopes, magnetometers, GPS, proximity sensors, ambient light/temperature sensors, air humidity sensor, barometers, photometers, pedometers, thermometers, and touchscreen sensors. Whatever we do with the mobile device, our physical interaction with it may result in certain behavioural analysis of our present moods, perceptions and preferences.
- *Communicate: High-speed and high-capacity networks*
  - ∗ With 5G looming upon us, we can now send vast amounts of data across any distances on Earth, in real time, no matter how much we have sensed about our surroundings and irrespective of what modalities they are sensed in. This means that it is now possible to channel all the sensed data immediately to a remote server for detailed analysis of what is going on. The situation is exacerbated by the fact that diverse IoE devices with built-in sensors are often automatically connected to the internet, thereby the speed with which any sensed data can be shared.
- *Compute: Advances in computational power*
  - ∗ Current machines have the capability to process millions of operations in parallel, thereby realizing the dream of deep neural networks where huge amounts of training data can be processed to build machine learning models that are able to recognize patterns and perform accurate predictions. This has led to significant breakthroughs in machine vision i.e. how machines can replace humans in understanding what is going on.

2.1. Emotion recognition

Humans learn the ability to recognize emotional patterns after having made sufficient number of informed observations of what different types of emotions look like, e.g. happy vs sad faces. It is with this principle that the essence of machine-enabled emotion recognition, which is a special case of pattern recognition, has been so designed (Picard, 2000).

In particular, during the so-called *training* phase, the machine is shown samples of different categories of emotions. Each emotional sample is processed by a *feature extraction* process in order to extract a concise and most precise representation of the sample. For instance, we recognize emotions shown on faces based on certain cues observed on the face, e.g. smiley eyes, upturned mouth corners indicate a positive emotion. In analogous fashion, each output feature from this process is a concise representation of the different categories of emotion samples. These features are stored in a database, ready for use for subsequent recognition.

During the *recognition* phase, when a sample is observed for which the manifested emotion is unknown, the sample is put through a similar feature extraction process, and then a *classification* (i.e. matching) function is performed that compares the currently extracted feature with the features stored in the database. The feature in the database that is the closest match to the current feature will lead to the conclusion that the current unknown sample belongs to the same emotion category as that matching feature in the database. This is the gist of the emotion recognition process as handled by the most advanced machine learning and AI techniques in the literature, to date able to achieve accuracies of over 97%.

Thus far, researchers have demonstrated that machines are able to recognize emotions when given data of various modalities, including facial videos, human speech, gestures such as keystrokes and key presses, and even gait (how we walk). Indeed, the science has advanced to even recognizing hidden emotions, exhibited as so-called micro-expressions. These typically are only observable within a fraction of a second on a human's face because s/he would tend to immediately suppress such facial expressions in a bid to conceal the felt emotion.

2.2. Emotional manifestation: Synthesis vs emotions transfer

To complete the emotional manifestation capability of machines, beyond just recognizing emotions, there is a need to be able to react to the sensed emotions, so-called affective computing (Picard, 2000). What is more realistically convincing is if the machine has the capability to reciprocate the emotional expressions.

Research into emotional synthesis includes enhancing speech with emotional effects (Schröder, 2001). While efforts had been made to directly synthesize such emotional speeches from scratch, the initial attempts sounded mechanical and therefore unrealistic. The situation is quite different now with the latest advances of deepfakes, which refers to using AI techniques to fabricate realistic fake images, video and audio. In particular (Floridi, 2018), researchers synthesized what was supposed to have been JF Kennedy's last speech if he had not been assassinated.

The main technology underlying deepfakes is known as *generative adversarial networks* (GANs) (Goodfellow et al., 2014), which can be seen as a specific type of *deep learning*, the in-trend term used to refer to neural networks with multiple inner layers of neurons. GANs are so powerful as they comprise two types of networks essentially. A generative network whose aim is to generate realistic outputs (images, video, audio) from random data, guided by the responses from a discriminative network that aims to distinguish between whether inputs given to it are real or fake. The *generator* learns from the responses of the *discriminator*, and they take turns until they reach *equilibrium*, i.e. no further changes in each side's actions will affect the optimality of its outcome. Since 2017, GANs and GAN-enabled deepfakes have taken the world by storm, with numerous victims in celebrities, public figures and even people on the streets who have had deepfakes generated of their images or videos shown in compromising situations.

Beyond swapping faces, these types of technologies also enable facial style transfer (Thies et al., 2019), including transplanting facial expressions (which infer emotions) from one to another, leading to a way to manifest emotions despite the original facial video not exhibiting such emotions. Transferring the facial expressions results in more realistic videos than synthesis from scratch as in this case one starts with an already realistic facial video except that no desired facial expressions are exhibited yet.

Facial expression transfer techniques mean that it is now possible to also imbue neutral-faced animated avatars with human-like emotional expressions. In the physical realm, if humanoid faces, e.g. Sophia's (Raymundo, 2016) had motorized facial landmarks, then these can be triggered to move in synch with the facial expressions of the originating human subject, further contributing to the emotional realism of humanoid faces.

2.3. Mapping facial expressions to underlying emotions

In order to enable machines to recognize human emotions, there is a need to map the observed facial expressions to underlying emotions felt by the human. Psychology tells us that this mapping is no simple issue. Before one can even attempt to measure the accuracy with which an entity (human or machine) can read the emotion of another human, it is necessary to know the true mental state or emotion of that individual (henceforth referred to as 'the target'), which (West and Kenny, 2011) refer to as the 'truth criterion'. Establishing the truth criterion when it comes to emotions is challenging, and perhaps for this reason much of the research on emotion recognition in psychology has focused on questions of process rather than accuracy (Zaki and Ochsner, 2011).

Very often, then, emotion recognition has been investigated by presenting people with photographs (or, more recently, videos) of actors who have been asked to pose different emotional states. Accuracy in this context can be established either by examining the extent to which observers' judgments match with the emotion that target was asked to pose, or by observer consensus whereby the 'correct' emotion is the one agreed by observers; e.g. (Baron-Cohen et al., 1997). While such methods can usefully shed light on the processes involved in making judgments about emotional states, it is not clear that they tell us anything about the actual mental states of the target in question or how they relate to the expressions produced. Indeed, the target who is posing the emotion may not feel any emotion at all or may be feeling an entirely different emotion from the one he/she is posing. If we want to map emotional expressions onto internal states it seems apparent that this needs to be done using natural displays of emotion rather than posed stimuli.

Suppose that, rather than asking the target to pose specific emotions, a video has been recorded of a target who is naturally engaged in a conversation, throughout which the target's face is animated. Probably in this context the target is experiencing some genuine emotional states, offering opportunity to carry out such mapping. But how might we know what emotional state(s) the target is experiencing? Perhaps the simplest solution is to ask the target him/herself. This technique was devised by (Ickes, 2001; Ickes, 2009), who asked targets to view videos of themselves engaged in conversation and asked them to state what emotions they were experiencing throughout the interaction. Later Ickes presented the same videos to other observers who were asked to infer the emotions of the target and accuracy was operationalised as the agreement between the observers' judgments and the target's self-report. Using this method, known as the empathic accuracy task, Ickes reports high levels of agreement between targets and observers, suggesting a close mapping between inner mental states and expressions that observers are able to read.

However, the technique is open to criticism in that when viewing the video playback, the target may not recall or may not even know what emotion he/she was experiencing at the time of the conversation, so arguably may make an inference based on observable behaviour in much the same way as any other observer (the accuracy of which would be not known (Teoh et al., 2017)). Further difficulty arises when we acknowledge that in real life situations, people are likely to experience emotions that are fleeting and subtle, and blends of different emotions and mental states, which may be difficult for a target to report. Finally, self-reports of emotions or mental states are constrained by the labels that have developed within language to describe these states. These may not perfectly match the wide range of states that people experience over time.

On the other hand, if the aim is to create machines that are indistinguishable from humans in their emotion recognition capabilities, it is not necessary to know whether the robot is accurate in its recognition: it merely needs to make judgments in the same way as humans do. This means that mapping emotions onto real internal states may not be necessary – it is just necessary to know what mental states people infer in these situations.

3. CHALLENGES FOR A FUTURE MIXED REALITY OF HUMAN-LIKE HUMANOIDS

Evidence suggests that people already can and do respond to the emotions of robots and artificial agents. For instance, participants were found to subjectively evaluate training of a robot more positively when the robot provided facial feedback during the training (Pais et al., 2013). Several studies have reported that human observers feel empathy towards robots expressing negative emotions such as sadness or pain (Scheeff et al., 2002; Weiss et al., 2009), although this may be stronger in children than adults. On the other hand, it has also been found that people can behave aggressively or even cruelly towards robots, even when they display distress (Bartneck and Hu, 2008; Scheeff et al., 2002). More research will be needed to identify what factors determine whether humans respond positively or negatively to displays of emotion in robots.

As our real human world becomes more intertwined with humanoids with human-like capabilities, e.g. as the focus of this article, capabilities to perceive and manifest emotions thereby existing as an emotional and sentient being, we will find it increasingly challenging to differentiate between what is real and what is machine, and what is human and what is not. This indistinguishability problem could give rise to social issues as humanoids, though apparently sentient, have no heart i.e. human conscience; a virtue that differentiates humans from beasts.

Therefore humanoids if they are indistinguishable from humans, pose a huge risk for the society. If they are exploited by malicious individuals, they could be weaponized to perform actions that adversely affect humans, and since there is no built-in self consciousness, malicious instructions could be followed blindly without any thought of refusal.

On the other extreme, the possibility that human-like humanoids could exist may also trigger misbehaving individuals to leverage on this fact to deny involvements in any of their past actions, citing that they were likely to have been performed by humanoids bearing resemblance to them. Therefore there is a need to irrefutably link any actions or incidents to humans, i.e. before linking is done, to check that the involved party is human and not just human-like, else it might have been a humanoid.

Ultimately, one of the main challenges is how we can continue to separate the creator (human) from the creature (humanoid); such that they remain distinguishable, as otherwise the issues described in the preceding paragraphs will lead to highly adverse societal problems.

4. AUTHOR BIOGRAPHIES

Raphaël spends most of his time contemplating malicious thoughts including different ways of subverting systems vs how one could defend against adversarial affronts. He views socio-technological advances such as the robotic+AI revolution with skepticism, and is researching on ways to minimize the adverse effects that such technologies would have on humankind.

Lizzy has a wide range of research interests relating to mind reading and visual perception. At the moment she is particularly interested in how people make inferences about other people's mental states and personality traits. Conversely, how do typical adults form impressions of people such as those with atypical development like autism? What type of perceptual information do people rely on to make these judgments?

REFERENCES

Baron-Cohen, S., Jolliffe, T., Mortimore, C. & Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with autism or Asperger syndrome. *Journal of Child psychology and Psychiatry*, *38*(7), 813–822. doi:10.1111/j.1469-7610.1997.tb01599.x.

Bartneck, C. & Hu, J. (2008). Exploring the abuse of robots. *Interaction Studies*, *9*(3), 415–433. doi:10.1075/is.9.3.04bar.

Floridi, L. (2018). Artificial intelligence, deepfakes and a future of ectypes. *Philosophy & Technology*, *31*(3), 317–321. doi:10.1007/s13347-018-0325-3.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C. & Bengio, Y. (2014). Generative adversarial nets. In *NIPS 2014* (pp. 2672–2680).

Ickes, W. (2001). Measuring empathic accuracy. *Interpersonal sensitivity: Theory and measurement*, *1*, 219–241.

Ickes, W. (2009). Empathic accuracy: Its links to clinical, cognitive, developmental, social, and physiological psychology. *The social neuroscience of empathy*, 57–70.

Pais, A.L., Argall, B.D. & Billard, A.G. (2013). Assessing interaction dynamics in the context of robot programming by demonstration. *International Journal of Social Robotics*, *5*(4), 477–490. doi:10.1007/s12369-013-0204-0.

Picard, R.W. (2000). Affective perception. *Commun. ACM*, *43*(3), 50–51. doi:10.1145/330534.330539.

Raymundo, O. (2016). Meet Sophia, the female humanoid robot and newest SXSW celebrity, *PCWorld*. Available online at: https://www.pcworld.com/article/3045299/meet-sophia-the-female-humanoid-robot-and-newest-sxsw-celebrity.html.

Scheeff, M., Pinto, J., Rahardja, K., Snibbe, S. & Tow, R. (2002). Experiences with sparky, a social robot. *InSocially Intelligent Agents* (pp. 173–180). Boston, MA: Springer.

Schröder, M. (2001). Emotional speech synthesis: A review. *INTERSPEECH*, *2001*, 561–564.

Teoh, Y., Wallis, E., Stephen, I. & Mitchell, P. (2017). Seeing the world through others' minds: Inferring social context from behaviour. *Cognition*, *159*, 48–60.

Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C. & Nießner, M. (2019). Face2Face: Real-time face capture and reenactment of RGB videos. *Commun. ACM*, *62*(1), 96–104. doi:10.1145/3292039.

Weiss, A., Wurhofer, D. & Tscheligi, M. (2009). “I love this dog” – children’s emotional attachment to the robotic dog AIBO. *International Journal of Social Robotics*, *1*(3), 243–248.

West, T.V. & Kenny, D.A. (2011). The truth and bias model of judgment. *Psychological review*, *118*(2), 357.

Zaki, J. & Ochsner, K. (2011). Reintegrating the study of accuracy into social cognition research. *Psychological Inquiry*, *26*(3), 159–182. doi:10.1080/1047840X.2011.551743.