Abstract—It is highly likely that classrooms of the future will feature robots to assist the human teachers. Tutor robots will be valued for their capacity to motivate learners and to provide affective support during learning activities, which will require from them to be able to understand the students’ affects and behaviours, and to respond to these through appropriate expressive motions. In this paper, we investigate the impact a robot teacher’s behaviour has on the students’ level of engagement. We outline research work we have carried out to tackle four of the challenges inherent to the effective deployment of social robots in classrooms: (1) sensing and understanding learners’ affective states and behaviours in class; (2) combining affect and behaviour understanding to capture classroom’s dynamics; (3) knowing what gestures a social robot should use as a learning facilitator; and (4) equipping the tutor robot with expressive and motivational capabilities.

Keywords—social robots; emotional support; motivational pedagogical gestures; affect and behaviour understanding

I. SOCIAL ROBOTS IN CLASSROOMS

Whether robots can replace the work done by humans on a large scale soon is a subject of heated debate. Based on how much and how rapidly education is changing (think of the Massive Open Online Courses movement which started less than 10 years ago and the swift move to online teaching as a result of the COVID-19 crisis), it is hard to predict how education will look like five to ten years from now. However, what we can observe today is that despite ethical concerns [23], social robots have started to play prominent roles in everyday education [19]. Social robots have primarily been used to provide language, science or technology education, where they have taken on the roles of tutor, tool and peer in the learning activities [18].

In 2009, Japan claimed the world’s first robot teacher, called Saya [10]. The humanoid robot was tested in a Tokyo primary classroom, and according to its creator: “The children were not fazed by Saya at all. They treated her like a real teacher.” Currently, institutions in Singapore, UK and France are among an increasing number of schools to have the SoftBank’s NAO and Pepper robots [25] offer learning support. At the Paul Bert school in Poitiers (France) for example, NAO teaches French grammar to children [14]. The teacher observes that in the presence of NAO, the pupils are more participative, enthusiastic and attentive than usual, and that they are unafraid of its opinion or judgments (see Fig. 1).

Fig. 1. The robot NAO [25] in a french classroom [14].

Robots can also function as stand-ins for students who are seriously ill, in long term hospital confinement and unable to come to class, increasing the options for distance learning. They can also act as stand-ins for teachers. In South Korea, small penguin-shaped robots called Engkey (see Fig. 2) teach English to kindergarten students, while being operated remotely by teachers in the Philippines [28]. Students with special requirements are also benefiting from the use of social robots. For example, NAO [25] is helping students with autism to learn social skills: it is easy to understand and has been shown to help reduce anxiety [12].

Fig. 2. Engkey (English jockey) a stand-in robot that teaches English [28].

So social robots have indeed entered classrooms and have started to play prominent roles in everyday education [4]. However, important key research questions remain. In this paper, we investigate the impact a robot teacher’s behaviour has on the students’ levels of engagement.
II. PROMISES, LIMITS AND CHALLENGES

Educational social robots’ true promise is in opening up teaching options and improving teachers’ and learners’ effectiveness. However, robots are still quite a way away from being implemented in classrooms autonomously, mostly due to technological limitations, but also because teachers have revealed a certain ambivalence regarding classroom robots [24]. While robots can handle a variety of specific tasks, and even provide students with emotional feedback by expressing happiness, satisfaction, or disappointment, they still lack the empathy and ability to inspire that teachers bring to the classroom [5].

According to [22], more than its capability to transfer knowledge, it is indeed the social supportive behaviour of a robotic tutor that can contribute most to increasing the learning efficiency of students. In other words, to be helpful, robot tutors should be able to correctly interpret the social cues that indicate learner’s task engagement, confusion, or attention. For example, one can imagine that a social robot equipped with speech and gesture recognition technology can collect and analyse information about the students and provide teachers with real-time information that could show them why the students might not be learning at full capacity. This would also allow the robotic tutor to adapt its own supportive behavioural strategies to the changing students’ needs and emotions.

This is a challenging proposition though. Although progress has been made in the constituent technologies, from perception (e.g. speech recognition and advanced sensing technologies for reading emotions, gestures, posture, and gaze) to action selection and production of behaviours, the integration of these technologies and their application in educational settings remain open challenges. Social robot tutors struggle to accurately interpret the learners’ social behaviours and choose appropriate support strategies, such as deciding when to take a break or encouraging help-seeking behaviour. Social robot tutors also must be able to implement these strategies through combinations of appropriate attention-guiding gestures, expressions, and gaze behaviours.

In the rest of the paper, we outline research work we have carried out to tackle four of the challenges inherent to the deployment of social robots in classrooms: (1) sensing and responding to learners’ affects and behaviour in class; (2) combining affect and behaviour understanding to capture classroom’s dynamics; (3) knowing what gestures a social robot should use as a learning facilitator; and (4) equipping the tutor robot with expressive capabilities.

The scenario we are considering is that of a telepresence application where a social robot is used to allow a university teacher to teach simultaneously in two different locations, and therefore in front of two separate audiences. The social robot is an embodiment of the human lecturer. It is in a partial master/slave relationship with her in that it is driven by her behaviour but also has a certain degree of autonomy. We seek to optimise the behaviour of the robot to make it more efficient: the robot should ultimately become better communicator than its human model. Moreover, the two teachers (human and robot) do not face the same audience, the robot must be able to adapt its behaviour to its own context and to its own audience’s level of engagement.

III. SENSING STUDENT ENGAGEMENT

The first challenge faced by robot teachers is to understand through sensing the learners’ affects and behaviour in class as these reflect and impact their levels of engagement.

A. Affect Recognition from Face Images

Students who are in a positive emotional state before working on a task view the task as more interesting than individuals who are in a negative or neutral mood [5]. Positive emotions lead to better performance and deeper satisfaction along with greater efforts; whereas negative emotions such as boredom and sadness negatively influence learning efficiency.

Application Programming Interfaces (API) for facial emotion recognition (e.g. Microsoft Face API [17]) have become readily available. So, recognising emotions of individual faces using a social robot as sensor has become possible [15], but interpreting these emotions in an educational context remains challenging, and the availability of audience data remains scarce. To add to the challenge, cultural differences are large. For example, students in China tend to be very quiet, displaying emotions which are largely neutral, but that don’t necessarily indicate a lack of engagement. Multicultural datasets will thus be necessary to advance the research in this area.

We have asked a small group of university students to act the following four affective states: thinking (normal), confused, distracted, and engaged (see Fig. 3) that were identified as relevant in a teaching environment. For each affect, 50 samples were collected and labelled accordingly. The Microsoft face API [17] was then used to recognise the facial emotions of each individual student. For each face detected, the API returns intensity values for eight of the Ekman’s basic emotions (Happiness, Surprise, Disgust, Sadness, Anger, Fear, Contempt and Neutral) [9]. These values are then used to train a K-Nearest Neighbours (K=4) classifier to infer an affective state (see Fig. 4). The final affective state is determined using a three video-frame sliding window in order to avoid rapid and irrelevant expression changes.

Fig. 3. The four students’ affects detected by the proposed system in a classroom: confused, interested (engaged), distracted and normal (thinking).
We found that when the students were acting “confused” or “distracted”, their facial emotions revealed either contempt or sadness, both negative emotions; when they were acting “engaged”, their facial emotions were uniformly recognised as happiness; whereas “thinking” (normal) revealed a surprising mixture of positive and negative emotions (see Fig. 5). We also found that “normal” and “distracted” were easily confused. A new metric based on eye movements had to be introduced to further differentiate these two states: a straight gaze was associated with “normal”.

B. Behaviour Recognition from Body Poses

To infer students’ behaviour from their body poses, we use the BERI (Behavioural Engagement Related to Instruction) protocol [13]. BERI describes various patterns of students’ behaviours that define engagement (e.g. listening) or lack of engagement. For example, the “listening” behaviour is described as follows: “eye contact is focused on the instructor or activity and the student makes appropriate facial expressions, gestures, and posture shifts”. Lack of engagement is described as follows: “eyes are closed or not focused on instructor or lecture material, student is slouched or sleeping, and student’s facial expressions are unresponsive to instructor’s cues”. Inspired by the BERI protocol, we designed and implemented a system that recognises four upper body poses: “sitting still”, “raising hand”, “lowering head” and “supporting chin” (see Fig. 6). A K-Nearest Neighbour (KNN) classifier was developed for classifying 16 skeleton coordinates into one of the four poses (see Fig. 7). Data collection (200 images) was achieved using simultaneously four cameras scattered in four different locations in the classroom. To avoid being too obtrusive, the size of the cameras is small, and each camera was placed on a desk with three to four students in its field of view. The cameras we used have a 16 Mega pixels resolution with a 170° wide angle and support 1080p videos, which provide enough data quality for image analysis. Skeleton detection is done using the Face++ API. The absolute coordinates retrieved from the API need to be normalised to account for students being pictured at different sizes depending on their location with respect to the camera.

To evaluate this system, 39 participants (21 students and 18 teachers) were invited to judge students’ behaviours from their body poses. Each participant evaluated 10 images from the dataset, these images showed body poses only with no facial expressions (to avoid facial expressions influencing the participants’ judgements). For each image, the participants were asked to choose one of the four poses (sitting still, raising hand, lowering head and supporting chin), which best described the student’s pose in the image. The levels of agreement between system’s predictions and participants’ judgements are high: 97.4% for lowering head; 94.9% for raising hand and supporting chin; 92.3% for sitting still. These ratios are calculated using confusion matrices. For a given pose, they correspond to the number of images for which the
subjective judgement and the system prediction match, divided by the total number of system predictions for that pose.

IV. MANAGING THE AUDIENCE AFFECTIVE STATE

The next challenge, having recognised individual or small groups of students’ affective states and behaviours, consists in combining these to capture classroom’s dynamics and devise appropriate responses from the robot teacher.

A. Combining Affect and Behaviour Recognition

Sensing small groups of students’ affect and behaviour is one thing but dealing with entire and potentially large classrooms (a pre- and hopefully post-COVID-19 situation) is the next challenge. In order to address this challenge, we integrated affect and behaviour recognition.

Because affect and behaviour analysis are done in two separate sub-systems, a “people mapping” module is needed to match face and body pose for each student detected in the multiple camera video frames. This mapping process, depending on the position of a student in the camera image and the accuracy of the respective detection algorithms (face and upper body), results in one of three possible cases: (1) only the face is detected, (2) only the upper body is detected, and (3) both face and body are detected (see Fig. 8).

For students whose face only or body only can be found in an image, their final predicted state is based on either affect or behaviour alone. However, for students whose both face and body are detected and successfully mapped, the results of affect and behaviour recognition are integrated to predict their final state. The behaviour is considered first. If the detected body pose is either “lowering head” or “raising hand”, confidence in these results is strong enough to conclude that the student is sleeping, distracted or wants to ask a question, without considering the affect state. If, however, the body pose is “sitting still” or “supporting chin”, both ambiguous postures, the affect takes priority and determines the student’s final state (see Fig. 9).

In an experiment, the system’s predictions were compared with participants’ subjective evaluations. The same 39 participants (21 students and 18 teachers) were invited to judge a person’s state from their face only, their body pose only, and from both. Each participant evaluated ten pictures of faces, body poses and combinations of both (30 images in total per participant). The 30 images were presented to the participants in a random order. A confusion matrix for the facial expressions (see Table I) shows good results for the interested and confused states (84.6% and 76.9% respectively); whereas the normal and distracted states show less agreement between the system’s predictions and the subjective judgements (71.8% and 66.7% respectively). When the subjective judgements are based on images showing both the face and the body pose, the scores are higher (see Table II): from 82.1% for normal, up to 92.3% for interested (confused is 84.6% and distracted is 87.2%). Here again, the ratios correspond to the number of images for which the subjective judgement and the system prediction match, divided by the total number of system predictions for each affective state.

Table I. Confusion Matrix for Affect (Faces Only)

<table>
<thead>
<tr>
<th>Subjective Judgement</th>
<th>Confused</th>
<th>Interested</th>
<th>Distracted</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confused</td>
<td>76.9%</td>
<td>0%</td>
<td>15.4%</td>
<td>7.7%</td>
</tr>
<tr>
<td>Interested</td>
<td>5.1%</td>
<td>84.6%</td>
<td>7.7%</td>
<td>2.6%</td>
</tr>
<tr>
<td>Distracted</td>
<td>15.4%</td>
<td>66.7%</td>
<td>17.9%</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>2.6%</td>
<td>15.2%</td>
<td>10.2%</td>
<td>71.8%</td>
</tr>
</tbody>
</table>

Table II. Confusion Matrix for Affect (Faces and Body Poses)

<table>
<thead>
<tr>
<th>Subjective Judgement</th>
<th>Confused</th>
<th>Interested</th>
<th>Distracted</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confused</td>
<td>84.6%</td>
<td>0%</td>
<td>0%</td>
<td>15.4%</td>
</tr>
<tr>
<td>Interested</td>
<td>5.1%</td>
<td>92.3%</td>
<td>2.6%</td>
<td>0%</td>
</tr>
<tr>
<td>Distracted</td>
<td>10.3%</td>
<td>0%</td>
<td>87.2%</td>
<td>2.5%</td>
</tr>
<tr>
<td>Normal</td>
<td>0%</td>
<td>7.7%</td>
<td>10.2%</td>
<td>82.1%</td>
</tr>
</tbody>
</table>

Fig. 8. Audience management system architecture: state analysis, people mapping and affective state recognition.

Fig. 9. Audience affective state recognition system, which combines affect and behaviour recognition.
The system can finally decide which robot’s response and behaviour are appropriate. The final output of the system, after considering and weighting every student’s states, is a suggestion to the robot teacher about how to address the audience (see Fig. 10). For example, if one or more students are raising their hand or if more than 20 percent of the students are in a confused state, the robot pauses and asks the audience for questions (see Fig. 11). If more than 20 percent of the students are in a distracted state, the robot gives a warning (e.g. “please pay attention”) and adopts a “motivational behaviour” as explained later in his section. In all other cases, the robot “keeps going”.

We evaluated the affective state recognition system’s outputs (the suggestions to the robot teacher) with the same 39 participants, using audience images such as the ones shown in Fig. 12. They were asked to formulate their own suggestions for the robot teacher, and these were compared with the system’s outputs. Each participant was shown four images simultaneously, based on which they had to judge the classroom’s dominant affective state and suggest a robot’s action for remediation, if needed. The levels of agreement are from high (75.7% for “warning”) to moderate (70.8% for “questions?” and 65.6% for “keep going”). These ratios correspond to the number of matching participants’ suggestions and system’s outputs, divided by the number of participants. More work is needed on mapping audience affective state and engagement level with the robot’s responses and behaviour.

When a negative state (“distracted” or “confused”) is identified the robot teacher should aim to change the audience affective state to “interested” or “normal”, i.e. to a positive state. Through observing video lectures, a series of desired, i.e. motivational lecturer behaviours were designed and synthesised as possible robot’s responses to students’ negative emotional states. To determine these motivational behaviours, we studied the strategies adopted by human lecturers. We invited small groups of students to watch online lectures and video recorded them while doing so. The students’ affects were measured using our affect recognition system. The lecturers’ behaviours that triggered “interested” affective states in most of the student audience were then extracted as short video clips. These clips served as models to program by demonstration a set of fifteen behaviours for Pepper to display in response to students showing a “distracted” affective state. These behaviours consist in series of postures copied from the human lecturers’ postures, including torso and arms’ orientation and directions, designed using the Choregraphe software, which allows control of the joint values to move the torso and limbs of the robot.

When the “distracted” state is detected, we programmed Pepper to randomly display one of the fifteen motivational behaviours. This strategy was evaluated using a within-participant experimental design, where a total of 16 participants were invited to listen to a short lecture delivered by Pepper. There were three conditions to which each group of 4 participants were assigned to in a random order. An 11-minute TED talk was divided into 3 sections to serve as lecture material. Under condition 1, Pepper is unaware of the audience affective state and uses few and random gestures while delivering its speech. Under condition 2, Pepper is also unaware of the audience affective state, but every 15 seconds it randomly displays one of the motivational behaviours. Under condition 3, Pepper is aware of the audience affective state and performs a randomly picked motivational behaviour each time the audience is in a “distracted” affective state. After each condition, participants were invited to fill out a
short questionnaire, which asked them about the content of the lecture and about the behaviour of Pepper. Results show that the robot’s behaviour has no significant effect on the understanding of the speech content. However, the robot’s motivational behaviours, when displayed at appropriate times, have a positive effect on students’ engagement. When the students are asked about their experience, most of them say that condition 3 offers the best experience.

V. THE ROLES OF GESTURES IN TEACHING AND LEARNING

We showed in the previous section, that motivational behaviours from the robot teacher positively impact students’ engagement. In particular, we showed that to improve the robot’s interaction with its student audience, the use of carefully designed postures and gestures is crucial. However, the roles of gestures in teaching and learning is complex, and identifying appropriate gestures still constitutes a challenge.

Fig. 13. Three “important” pedagogical gestures: pointing, circle and ball.

Fig. 14. Students distribution of visual attention while watching a video lecture.

Fig. 15. A robotic tutor displaying multimodal behaviour to direct students’ visual attention.

Fig. 16. Definition of presentation.

Gestures are central to human cognition, yet there exists little educational research that focuses on the role of gestures in teaching and learning [16][21]. In addition to supporting the teacher in the delivery of her speech and to serving as a form of scaffolding [3], gestures contribute to the shaping of students’ perceptions of the teacher and keep them motivated and engaged in the classroom [11]. Robots depicting several appropriate gestures may lead towards efficient learning and engagement and provide appropriate responses to detected student’s affective states [2]. Some gestures in particular have the ability to reinforce the pedagogical significance of concomitant spoken messages.

We have conducted an experimental study into the relationship between the gestures performed by lecturers and the pedagogical significance of the corresponding parts of the lecture [26]. We found that three types of gesture (“pointing”, “circle” and “ball”) (see Fig. 13) indicate that the corresponding lecture part is of pedagogical significance, with pointing being judged the most important gesture.

We then tested the gestures through a small experimental study where virtual pedagogical agents were added to voice-over-slide learning materials [29]. One avatar was generated that reproduces exactly all the lecturer’s upper-body movements; a second avatar displayed low amplitude movements, giving an impression of natural presence; a third avatar performed only the lecturer’s pointing gestures as observed in the recorded lecture; and finally, a fourth avatar displayed random “lecture-like” gestures. The experimental results indicate that the avatar that performs the lecturer’s pointing gestures only is the most effective at facilitating learning, whereas an agent that displays a more complex and divers set of gestures, even though the behaviour is an exact representation of the lecturer’s, is distractive. This corroborates the findings presented in [11] where robot deictic gestures are shown to consistently predict users’ information recall, and where all types of gestures are shown to affect user perceptions of the robot’s performance as a narrator.

We also used eye-tracking technology and data visualisation techniques to collect and analyse students’ distribution of visual attention in relation to the instructor’s speech and body language [30]. Fig. 14 shows a fixation count heat map calculated on a 5.32 second clip during which the instructor is performing a pointing gesture, looking at the slide and delivering speech that refers to the slide content. With the three behaviours combined, the student’s visual attention is clearly directed towards the slide. Our study suggests that the instructor’s gaze alone or a pointing gesture alone, is not enough to shift viewers’ attention, however, the combination of both is effective. A robot teacher should thus adopt a multimodal behaviour, combining gaze, postures and gestures, to effectively direct the learners’ visual attention toward the relevant learning material, as illustrated in Fig. 15.

VI. EQUIPPING THE ROBOT WITH EXPRESSIVE CAPABILITIES

In the previous studies, we demonstrated the pedagogical significance of some lecturing gestures, as well as their effectiveness at directing students’ attention and increasing engagement. However, motivational behaviour should also contribute to promote positive emotions in students, which requires the generation of complex robot’s expressive capabilities which go beyond gestures. Promoting positive emotions (i.e. “emotional induction”) while engaged in a learning task may lead the learner to develop a positive attitude towards this task [5]. The social robot should be able to be expressive, through head poses, body postures and tone of voice which are likely to promote positive emotions in students.

A large corpus of work exists that make use of facial expression including gaze to convey emotions [1]. In our
work, we rather focus on whole body movements. In order to generate whole body expressions, the most conventional techniques are based on selecting some features manually together with mimicking and retargeting techniques, or some machine learning technique is used to learn these features [27]. We proposed to use expert features coupled with control techniques. The features are chosen to be physically realisable on the robot: such as velocity, gaze, smoothness, etc. The emotion is expressed as a low priority task in a null space decomposition of the controller, i.e. the robot can convey an emotion if its degrees of freedom during the task are redundant [7]. The robot can successfully achieve its task and convey emotions without interference, both at the same time. We have used this technique for different tasks using our robot Pepper [25] and user studies have shown that this method successfully conveyed emotions such as sadness and happiness, and that such behaviours were affecting the perception the students had of the robot [8].

VII. CONCLUSION

Teaching is a complex endeavour and there is no clear consensus on what “good” teaching is, neither a clear metrics of long-term effects of teaching. Nevertheless, it can generally be said that adapting delivery to students’ changing affective states and using appropriate gestures and signals contribute to effective teaching by improving students’ learning experience and behaviours. In this paper, we have described work to equip a social robot with the ability to sense and respond to learners’ emotions and behaviour in a classroom environment. Work is ongoing to improve the robot’s sensing capabilities and to map audience affective state and engagement level with the robot’s responses and behaviour.

As robots are increasingly permeating society, children who have experienced learning with social robots may expect robot teachers to accompany them throughout their educational journey. Many challenges and questions remain though about the contribution and roles of social robots in education. Now is a good time for carrying research on how to build up their sensing and reasoning capabilities in order to improve their social supportive behaviour, which is most needed to enhance the students’ learning experience and to increase learning efficiency and promote sustained usage.

ACKNOWLEDGEMENTS

This research is partially funded by the Japanese Society for the Promotion of Science, grant in aid for scientific research (C) 19K12167.

REFERENCES
