Estimating Emotional Intensity from Body Poses for Human-Robot Interaction

Mingfei Sun
Computer Science & Engineering
Hong Kong University of Science and Technology
mingfei.sun@ust.hk

Yiqing Mou
Psychology Department
Hong Kong University
myq.maxine@gmail.com

Hongwen Xie
Tencent Inc.
hongwenxie@tencent.com

Meng Xia
Computer Science & Engineering
Hong Kong University of Science and Technology
iris.xia@connect.ust.hk

Michelle Wong
Computer Science
Smith College
mwong46@smith.edu

Xiaojuan Ma
Computer Science & Engineering
Hong Kong University of Science and Technology
mxj@cse.ust.hk

Abstract—Equipping social and service robots with the ability to perceive human emotional intensities during an interaction is in increasing demand. Most of existing work focuses on determining which emotion(s) participants are expressing from facial expressions but largely overlooks the emotional intensities spontaneously revealed by other social cues, especially body languages. In this paper, we present a real-time method for robots to capture fluctuations of participants' emotional intensities from their body poses. Unlike conventional joint-position-based approaches, our method adopts local joint transformations as pose descriptors which are invariant to subject body differences as well as the pose sensor positions. In addition, we use a Long Short-Term Memory-Recurrent Neural Network (LSTM-RNN) architecture to take the specific emotion context into account when estimating emotional intensities from body poses. The dataset evaluation suggests that the proposed method is effective and performs better than baseline method on the test dataset. Also, a series of succeeding field tests on a physical robot demonstrates that the proposed method effectively estimates subjects emotional intensities in real-time. Furthermore, the robot equipped with our method is perceived to be more emotion-sensitive and more emotionally intelligent.

I. Introduction

There is an increasing demand in Human-Robot Interaction (HRI) for real-time perception of participants' emotional intensities [1], i.e., how emotions fluctuate over time. Human participants may express their feelings and intents through subtle emotional intensities in HRI [2]. For example, when users become unsatisfied with robots' performance, they may frown. If agitated by robots' inappropriate responses, they may shrug to show their strong disappointments. If the robot is capable of detecting such fluctuations of emotional intensities, it can then fine-tune its reactions in a timely manner to lower participants' discomfort [3], to satisfy their preferences [4], and consequently to gain more acceptance [5].

In spite of such demand, most of existing work focuses on determining which emotion(s) participants are expressing from facial expressions [6] but largely overlooks the emotional intensities spontaneously revealed via other social cues, especially via body languages. Particularly, when emotions get
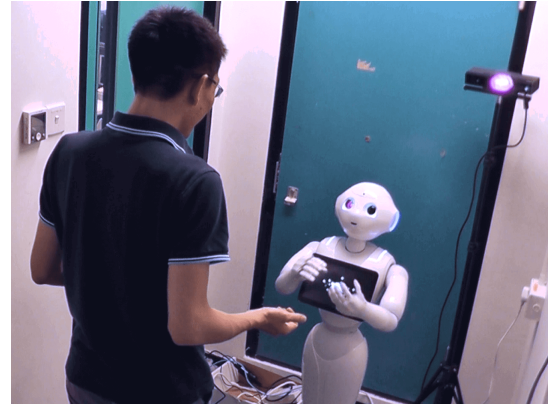

Fig. 1. A subject shows his high intensity of surprise by spreading out his hands.

more intense, the discriminative power of facial expressions for emotional intensities will degrade whereas the body cues will become dominant [7]. Nevertheless, a dominance increase of body poses is not necessarily an indicator of more emotional intensity, since the latter is usually emotion-specific [8]. Take Fig. 1 as an example. Open arms may just be an ordinary gesture if the person is speaking calmly. But when he is with a surprised facial expression, the pose may imply more surprise than that when his arms are in resting positions. Hence, in order to get a more complete picture of the emotional intensity from body poses, it is better to integrate emotion types into the intensity estimation process. Furthermore, most of previous studies on estimating affective status from body poses directly take absolute joint positions [9] or other physical quantities derived from these positions, e.g., joint speeds [10], body expansions [11] etc., as pose descriptors. Regardless of the reported effectiveness, these handcrafts, in general, are subject to individuals' body differences (such as heights, body shapes etc.) as well as their positions to pose sensors, which may possibly limit or even deprive their practical usages.

In this paper, we present a novel method to estimate

emotional intensities from body poses under various affective status. Specifically, we adopt the local joint transformations to describe body poses, which are invariant to subjects' body differences as well as their positions against sensors. Furthermore, instead of deriving handcrafted features, we use a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) to directly model the pose sequences with variable lengths. This has two advantages: First, the LSTM-RNN captures complex temporal correlations among pose sequences and thus provides a more appropriate modeling of fluctuations of emotional intensities; Second, through proper design of neural network structures, we integrate emotional contexts into the intensity estimation, and obtain more accurate predications of emotional intensities. The quantitative evaluations on a dataset demonstrate that the proposed method accurately predicts emotional intensities with an averaged Pearson correlation score of around 0.8, outperforming the previous method which achieves only 0.46. We also deploy the proposed method to a physical robot (Pepper robot) and conduct a series of field tests to evaluate its practical usage. The results imply that the robot equipped with our method senses subjects' emotional intensities effectively and in real-time. Also, the subjects' ratings suggest that the robot with the proposed emotional intensity estimation is perceived to be more emotional-sensitive and emotionally intelligent.

II. RELATED WORK

Emotional intensity is defined as the psychological states of being activated by some evoking stimulus [12], which can be measured by bio-electric signals, such as heart rate, blood pressure, skin conductance etc [13], [14], [15]. For example, Kulic and Croft leveraged heart rate, perspiration rate, and facial muscle contraction to estimate participants' emotional intensities in terms of anxiety, calm and surprise towards robot arm motions [16]; Swangnetr and Kaber used heart rate and galvanic skin response signals to estimate the intensity of joy and excitement for the elderly patients during medical Patient-Robot Interactions [17]. Saulnier *et al.* exploited the natural muscle tensions to estimate participants' stress levels for domestic robots [18]. However, all these studies require putting sensor on participants, e.g., via special wearable devices [3] or electro-physiological monitoring devices [16], [4], it may be difficult and sometimes impossible to deploy such biological methods for social or service robots in real-time HRI.

Non-contact affective intensity sensing is thus proposed, which usually adopts different modalities to estimate the emotional intensity. Among all modalities, the body pose has gained increasing popularity in HRI due to the following two reasons. First, body poses are becoming readily accessible in HRI due to the real-time motion capture sensors and robust vision estimation techniques; Second, more and more psychological results demonstrate the strong possibility of using body poses to perceive emotions [19], [20], [8], [21]. Although whether body poses are indicative of specific emotions is still debatable [8], psychologists have reached a consensus that body poses can effectively convey people's

emotional intensity [22]. Thus many researchers employ body poses as an important indicator of participants' emotional intensity. For example, McColl and Nejat designed a model to recognize the elderly people's pleasure intensity based on their upper body movements during a dining process with a robot assistant [11]. Their results indicated that, compared with the specific emotion types, the emotional intensities can be more accurately perceived from body movements. In addition, Xu *et al.* employed an expressive Nao robot to emotionally influence participants' body poses [10]. Their results also indicated that the emotional arousal can be effectively conveyed via body poses. Other studies have also identified the strong link between a subject's emotional intensities and the body movements [23] as well as the head positions [24]. Furthermore, the body poses are also adopted to perceive other affective and cognitive process, e.g., accessibility level [25], engagement dynamics [26], [27]. However most of the aforementioned works adopt the absolute joint positions or handcrafted features, which could result in two problems. First, the features can be affected by irrelevant physical body differences. For example, joint speeds[10], joint accelerations [28], joint distances [11], [23], body expansion[11] and silhouette motion images [29] are closely related to subjects' heights and body shape (e.g., bone lengths). Thus, for subjects with different heights and body shapes, the corresponding features will have distinguishable differences even if subjects perform the same set of expressive body motions. In addition, the joint-position-based features is also dependent on camera positions. For example, the joint angles, and the joint structures will change dramatically if the pose sensors are relocated to another position. Furthermore, as pointed out in [8], the emotional intensity expressed via body poses are specific to emotion types, which is also largely overlooked by previous studies.

III. METHOD

In this section, we describe the proposed method in detail.

A. Body pose representation

The body pose is usually denoted as a skeletal tree with parent-child structures, as shown in Fig. 2(a). Each node in the tree is associated with a body joint, and its coordinates represent the joint position with respect to the sensor coordinate frame. This representation is widely adopted for affective estimation in most of the existing literatures. However, as aforementioned, the body pose represented in this way is dependent on the sensor positions and contains individual body differences, e.g., bone lengths(the distance between two joints), which could limit the generalization of the proposed methods in previous work.

Nevertheless, human body poses can also be described via a series of local joint transformations, as widely used in motion capture[30] and motion re-targeting[31]. The basic idea is to separate relative joint displacements and motions in skeletal structures by using local transformations in joint coordinate frames. Specifically, each joint has its own coordination frame (a predefined right-handed frame) and all frames are ordered

based on the skeletal tree structure, forming multiple parent-child pairs, as shown in Fig. 2(b). The position of each joint is defined by a homogeneous matrix, which contains the joint rotation and translation with respect to its parental coordinate frame. This transformation representation describes how joints move against their parental joints, i.e., the joint relative motions rather than absolute positions, thus invariant to sensor positions. To get rid of subject-dependent body differences, we discard translations in the homogeneous matrix and only adopt rotations since joint translations are relevant to bone length. For simplicity, we use Euler angles $(\theta_r, \theta_p, \theta_y)$ in the order of roll, pitch and yaw to represent rotations in a homogeneous matrix. Consequently, each body pose is described by a vector of Euler angles, ordered according to the skeletal tree structure. In addition, we do not consider rotations of the root joint as it is in the sensor coordinate frame. For each body pose X_i in a pose sequence, we thus have following descriptions:

$$X_i = \begin{bmatrix} \theta_{r1} & \theta_{r2} & \theta_{r3} & \ldots & \theta_{r13} \\ \theta_{p1} & \theta_{p2} & \theta_{p3} & \ldots & \theta_{p13} \\ \theta_{y1} & \theta_{y2} & \theta_{y3} & \ldots & \theta_{y13} \end{bmatrix}$$

where $\theta \in [-\pi, \pi]$.

In most of existing pose extracting sensors, e.g., Kinect, the local joint transformation representation is readily available. Nevertheless, we can also easily convert the skeletal tree representation into the local transformations. The conversion between these two representations is essentially a set of transformations between different frames, which are usually described by forward/inverse kinematics in robotics. As illustrated in the example of Fig. 2(c), the conversion from local joint transformation T_i of joint#i to its positions G_i in skeletal tree can be found by:

$$G_i = T_1 T_{i_2} T_{i_3} ... T_i$$

where $G_i \in \mathbf{R}^4$ is the homogeneous coordinate of joint#i in the sensor coordinate frame, T_i is the transformation matrix (homogeneous matrix) of joint#i in its parental frame, and $i, i_1, i_2, ..., 1$ is a path from joint#i to join#1. To find the local transformation representation for joint#i, we can use:

$$T_i = T_i^{-1} T_{i_1}^{-1} T_{i_2}^{-1} ... T_1^{-1} G_i$$

In addition,

B. LSTM-RNN modeling

Owing to its capability of modeling complex temporal dependencies, Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) has been widely adopted for pose-related estimation, e.g., human pose estimation[32], action recognition[33] etc. Fig. 3(a) presents the internal structure of an LSTM memory block, which has a single cell and three multiplicative gates (input gate, forget gate and output gate). The detailed multiplicative relationships between these gates can be referred to [34]. LSTM-RNN can capture temporal flows along thousands or even millions of time steps in sequences with variable lengths, which is usually hard
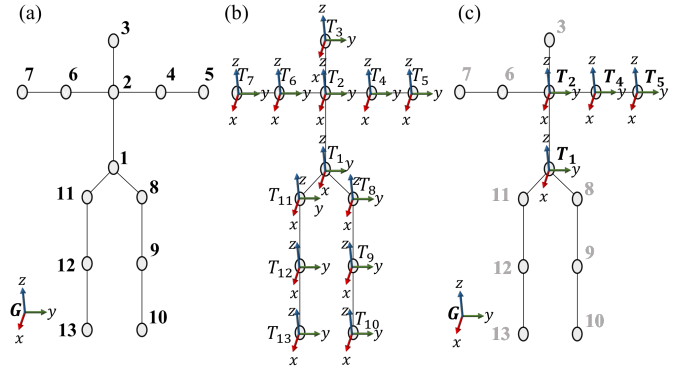

Fig. 2. (a) Skeletal tree structure with the joint order number; (b) Each joint has its own coordination frame and all frames are ordered; (c) To find the transformation T_5 for joint#5 with position G_5, we first compute forward kinematics along transformation tree: $G_5 = T_1 T_2 T_4 T_5$, and then $T_5 = T_4^{-1} T_2^{-1} T_1^{-1} G_5$
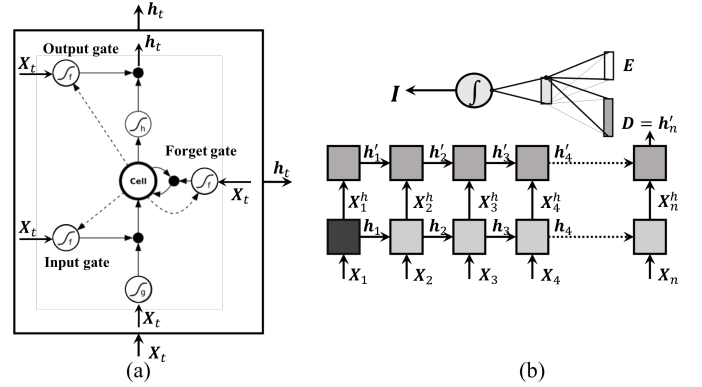

Fig. 3. (a) LSTM memory block; (b) the proposed LSTM architecture

to achieve for conventional sequence models, e.g., Hidden Markov Model. These advantages make LSTM-RNN a good choice for processing the complicated, variable-length and intra-correlated body pose sequences.

Inspired by the work [35] that uses a two-layered LSTM-RNN to model a sequence of structured features, we adopt an LSTM-RNN architecture as shown in Fig. 3(b). It is composed of two LSTM layers and one fully-connected layers. The LSTM layers take pose sequences $[X_1, X_2, X_3, ..., X_n]$ of variable lengths n as input and output a pose description vector $D = h'_n$ with fixed size. The fully-connected layer fuses the emotion vector E (emotional context of body poses) with pose descriptor D and outputs a estimated scalar I of emotional intensity by a sigmoid function.

IV. EXPERIMENTS & RESULTS

In order to evaluate our system, we conduct two studies: empirical evaluation on a public dataset and a field test on a physical robot.

	Pearson correlation coefficient					Avg.
	#1	*#2*	*#3*	*#4*	*#5*	
Proposed	0.7998	0.7958	0.8028	0.8201	0.8192	0.8075
SVM+Feature	0.4660	0.4394	0.4555	0.4656	0.4812	0.4615

TABLE I
FIVE-FOLD CROSS VALIDATIONS ON THE MPI DATASET

A. Dataset evaluation

We use the body expression dataset[1] for the evaluation of the proposed method. This dataset contains close-to-natural emotional body expressions recorded by motion capture devices when subjects were expressively narrating stories in front of the camera. Each pose is described by 23 body joints in 3D with recording rate at $120Hz$, and each body joint$\#i$ has a local transformation T_i. The dataset has two labels: the intended emotion and the perceived emotion. The intended emotion is what the narrator is required to express, i.e., E in the input of LSTM-RNN, while the perceived emotion is a set of emotion labels annotated by experts. We use the percentage of the intended emotion in the set of the perceived emotion as the ground truth for emotional intensity I based on the intuition that: if the emotion expressed via a pose sequence is intenser, then annotators are more likely to assign the same and correct label to it. The dataset contains 1447 sequences of natural emotional body poses, and each sequence is stored in local transformation format.

Since the performance of a neural network depends heavily on the size of training dataset, we augment the training data to obtain more data in two steps. First, we double the dataset by swapping left and right side of body poses since this swap will not change the expressed emotional intensity. Second, we sample from raw pose sequences (120Hz) at a fixed rate(30Hz) and get four samples from every pose sequence, then further quadrupling the dataset. Finally, we have around 12,000 body sequences.

The LSTM RNN is trained with following setups: $h \in \mathbf{R}^{64}$, $h' \in \mathbf{R}^{128}$ for hidden units; 100 epochs with batch size 256 for one complete training. We use RMSprop as the optimizer with mean absolute error as the loss.

To show the effectiveness of the proposed method, we use the SVM together with handcrafted features proposed in [36], [37] as the baseline, which adopts joint speeds, joint accelerations, joint angles and body expansions to estimate emotion from postures. The Pearson correlation coefficient[2] is used as the evaluation metric. We use 5-fold cross-validation to compare the results.

The results in Tab.I demonstrate that the proposed method effectively estimates the emotional intensities. Compared with the baseline method (SVM with handcrafted features), our method consistently achieves higher Pearson correlation coefficients on all tests. In addition, the average coefficient of

[1]Provided by the Max-Planck Institute for Biological Cybernetics; downloadable: http://ebmdb.tuebingen.mpg.de/

[2]The definition can be found in: https://en.wikipedia.org/wiki/ Pearson_correlation_coefficient

the proposed method is almost 75% higher than the baseline, which further confirms that the adopted pose descriptors are superior to the joint-position-based handcrafted features. And LSTM-RNN is also effective for intensities estimation from body poses.

B. Field testing

To further test the practical usage, we deploy the proposed method on a physical robot with the task to detect participants' emotional intensities. However, during a real Human-Robot Interaction (HRI), the ground-truths of subjects' emotional intensities are hard to obtain, thus making the evaluation difficult. On the other hand, subjects themselves are well aware of what emotions and intensities they are expressing when interacting the robot. Thus, if the robot can show the estimation results in real-time to the subjects, then the subjects can online evaluate whether the results are accurate or not. Instead of directly showing/telling the detected emotional intensities, we use different robot behaviors to imply the estimates in order to make the interaction more natural and interactive. Specifically, for the easy of robot behavior design, we threshold the emotional intensities into two levels (weak and strong) and only consider three common emotion types (joy, surprise and sadness). Based on this, we design two types of robot behaviors: *none* and *expressive*. In *none*, the robot has no emotional intensity estimation, only reacting with random fillers, e.g., oh and hmm, regardless of subjects' emotions and the intensities. By contrast, in mode *expressive*, the robot will show behaviors based on the detected emotional intensities in real-time. If low intensity is detected, the robot will provide speech feedback, e.g., "you look so happy!" for joy, "why are you surprised?" for surprise, and "what happened to you?" for sadness. If high intensity is detected, the robot will provide gestural feedback, as shown in Figure 5. The set of speeches and postures varies for different emotions.

The field testing was conducted on a Pepper robot with a Kinect sensor, as shown in Figure 4. The subject's body poses were captured by the Kinect Sensor in 25 joints, each with a 3D position in the camera coordinate frame and a 4D rotation quaternion in a local coordinate frame, which are converted to Euler angles based on these formulas[3]. The subjects' faces are also detected and captured by the Kinect sensor, which are then analyzed by an facial analysis tool [4] to obtain facial expressions.

We recruit 14 subjects with their informed consents to participate in the field test. In the process, subjects were required to express three emotions (joy, sadness and surprise) via their faces and body poses in two levels of intensities. Subjects were also told about the purpose of the experiment to encourage them to change their emotions and intensities. Each experiment consists of three sessions, and, in each session, the subject was required to perform at least one emotion for around 60s. After each session, subjects need to rate the robot's

[3]https://msdn.microsoft.com/en-us/library/hh973073.aspx
[4]https://azure.microsoft.com/en-us/services/cognitive-services/emotion/
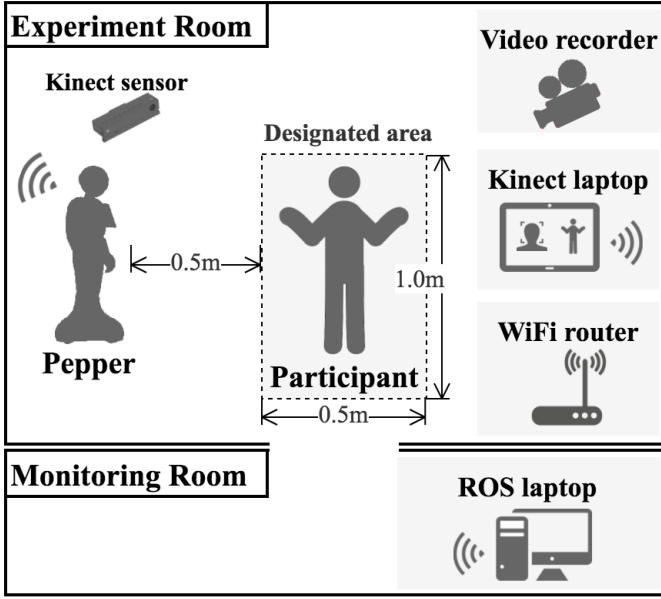
Fig. 4. Illustration of experiment setup

Fig. 6. Ratings of: (a) robot emotion perceptions(1 means very poor and 7 means very good); (b) robot other performances(1 means very poor and 7 means very good). ($*: p < .05$ and $**: p < .01$)
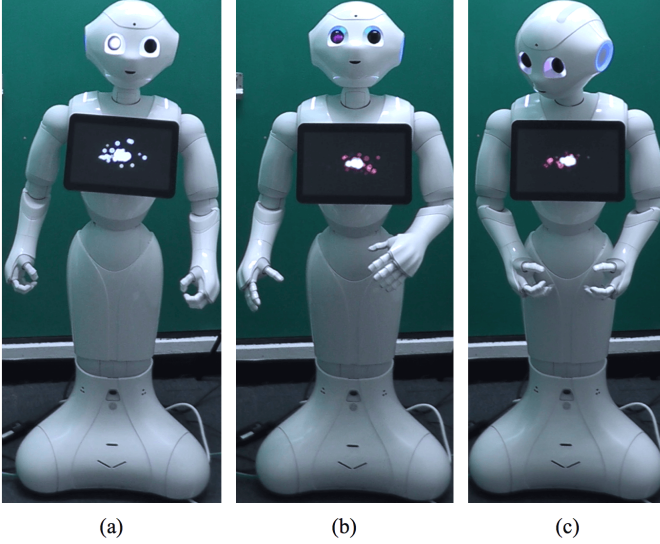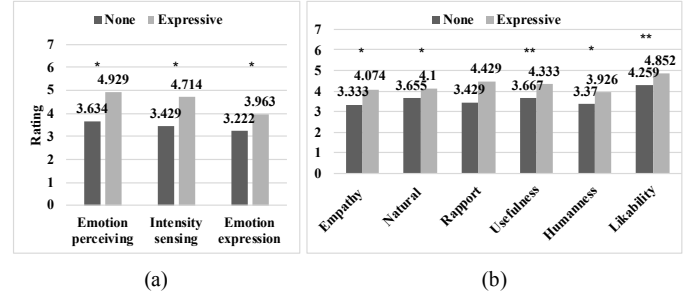

Fig. 5. Body expressions for Pepper robot when it detects high level of emotional intensities for: (a) joy; (b) surprise; (c) sadness

emotion-perceiving ability in terms of "emotion perceiving" and "arousal sensing" on a 7-point Likert scale (1 means very poor and 7 means very good). In addition, they need to rate the robot's other performances derived from [38], [39]. The whole system was implemented in Robot Operating System (ROS) and runs in real-time with Ubuntu system.

Ratings are shown in Fig.6. To spot the statistic significance, we use the repeated measures ANOVA to analyze these ratings. First, the manipulation check on Peppers expressiveness shows that the manipulation is effective; repeated measures ANOVA, $F(2, 52) = 9.427, p < 0.01, \eta^2 = .266$; Bonferroni post-hoc test $p < 0.01$. The robots using the our method ("expressive": M = 4.296, SD = 0.183) are indeed perceived to be more

expressive than the baseline version ("none": M = 3.222, SD = 0.252). As for the robot emotion perception, all subjects rated the emotion perceiving, intensity sensing and emotion expression in mode "expressive" significantly higher than the correspondences in mode "none". Specifically, for emotion perceiving, mode "expressive" (M = 4.929,SD = 0.165) significantly exceeds mode none (M = 3.634, SD = 0.372) with repeated measures ANOVA, $F(2, 26) = 4.225, p < 0.05, \eta^2 = .247$ and Bonferroni post-hoc test $p < 0.05$, and, for intensity sensing, mode "expressive" (M = 4.714,SD = 0.221) is significantly higher than mode "none" (M = 3.429, SD = 0.388) with repeated measures ANOVA, $F(2, 26) = 4.381, p < 0.05, \eta^2 = .252$ and Bonferroni post-hoc test $p < 0.05$. In summary, the proposed method is effective for real-time application in HRI. Furthermore, subjects also gave significantly different ratings on robots' other performances, such as robots' ability to show empathy, to behave naturally, to demonstrate usefulness, to perform human-like and to gain popularity. Overall, ratings in mode "expressive" on various aspects are significantly higher than that in mode "none". These results suggest that the proposed method is useful for a real HRI, and the robot equipped with our model can dramatically change subjects' views of its performances.

V. CONCLUSION

We propose a method to estimate emotional intensities from body poses under various affective status. This method adopts the local joint transformations to describe body poses, which are invariant to subjects' body shape differences and their positions against sensors. Moreover, we propose an Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) architecture to model pose descriptors without any handcrafted features. Our quantitative evaluations on a dataset imply that the method accurately predicts emotional intensities with a very high correlation score. The field tests on a physical robot demonstrate that the proposed method can be well applied for practical usage as it enables a humanoid robot (Pepper) to sense subjects' emotional intensities effectively in real-time. Also, subjects reported that the robots with our method are more emotional-sensitive and outperform the robots without in all aspects.

REFERENCES

[1] B. Gonsior, S. Sosnowski, M. Buß, D. Wollherr, and K. Kühnlenz, "An emotional adaption approach to increase helpfulness towards a robot," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. IEEE, 2012, pp. 2429–2436.

[2] D. McColl, A. Hong, N. Hatakeyama, G. Nejat, and B. Benhabib, "A survey of autonomous human affect detection methods for social robots engaged in natural hri," *Journal of Intelligent & Robotic Systems*, vol. 82, no. 1, pp. 101–133, 2016.

[3] P. Rani, C. Liu, and N. Sarkar, "Affective feedback in closed loop human-robot interaction," in *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. ACM, 2006, pp. 335–336.

[4] C. Liu, K. Conn, N. Sarkar, and W. Stone, "Online affect detection and robot behavior adaptation for intervention of children with autism," *IEEE transactions on robotics*, vol. 24, no. 4, pp. 883–896, 2008.

[5] R. Sorbello, A. Chella, C. Calí, M. Giardina, S. Nishio, and H. Ishiguro, "Telenoid android robot as an embodied perceptual social regulation medium engaging natural human–humanoid interaction," *Robotics and Autonomous Systems*, vol. 62, no. 9, pp. 1329–1341, 2014.

[6] F. Cid, J. A. Prado, P. Bustos, and P. Nunez, "A real time and robust facial expression recognition and imitation approach for affective human-robot interaction using gabor filtering," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 2188–2193.

[7] H. Aviezer, Y. Trope, and A. Todorov, "Body cues, not facial expressions, discriminate between intense positive and negative emotions," *Science*, vol. 338, no. 6111, pp. 1225–1229, 2012.

[8] H. G. Wallbott, "Bodily expression of emotion," *European journal of social psychology*, vol. 28, no. 6, pp. 879–896, 1998.

[9] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, 2007.

[10] J. Xu, J. Broekens, K. Hindriks, and M. A. Neerincx, "Robot mood is contagious: effects of robot body language in the imitation game," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 973–980.

[11] D. McColl and G. Nejat, "Determining the affective body language of older adults during socially assistive hri," in *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*. IEEE, 2014, pp. 2633–2638.

[12] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant." *Journal of personality and social psychology*, vol. 76, no. 5, p. 805, 1999.

[13] S. Khalfa, P. Isabelle, B. Jean-Pierre, and R. Manon, "Event-related skin conductance responses to musical emotions in humans," *Neuroscience letters*, vol. 328, no. 2, pp. 145–149, 2002.

[14] B. M. Appelhans and L. J. Luecken, "Heart rate variability as an index of regulated emotional responding." *Review of general psychology*, vol. 10, no. 3, p. 229, 2006.

[15] G. D. James, L. Yee, G. A. Harshfield, S. G. Blank, and T. G. Pickering, "The influence of happiness, anger, and anxiety on the blood pressure of borderline hypertensives." *Psychosomatic Medicine*, vol. 48, no. 7, pp. 502–508, 1986.

[16] D. Kulic and E. A. Croft, "Affective state estimation for human–robot interaction," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 991–1000, 2007.

[17] M. Swangnetr and D. B. Kaber, "Emotional state classification in patient–robot interaction using wavelet analysis and statistics-based feature selection," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 1, pp. 63–75, 2013.

[18] P. Saulnier, E. Sharlin, and S. Greenberg, "Using bio-electrical signals to influence the social behaviours of domesticated robots," in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. ACM, 2009, pp. 263–264.

[19] H. G. Wallbott and K. R. Scherer, "Cues and channels in emotion recognition." *Journal of personality and social psychology*, vol. 51, no. 4, p. 690, 1986.

[20] W. H. Dittrich, T. Troscianko, S. E. Lea, and D. Morgan, "Perception of emotion from dynamic point-light displays represented in dance," *Perception*, vol. 25, no. 6, pp. 727–738, 1996.

[21] F. E. Pollick, H. M. Paterson, A. Bruderlin, and A. J. Sanford, "Perceiving affect from arm movement," *Cognition*, vol. 82, no. 2, pp. B51–B61, 2001.

[22] P. Ekman and W. V. Friesen, "Detecting deception from the body or face." *Journal of personality and Social Psychology*, vol. 29, no. 3, p. 288, 1974.

[23] A. Kleinsmith and N. Bianchi-Berthouze, "Recognizing affective dimensions from body posture," *Affective computing and intelligent interaction*, pp. 48–58, 2007.

[24] M. M. Gross, E. A. Crane, and B. L. Fredrickson, "Effort-shape and kinematic assessment of bodily expression of emotion during gait," *Human movement science*, vol. 31, no. 1, pp. 202–221, 2012.

[25] D. McColl and G. Nejat, "Affect detection from body language during social hri," in *RO-MAN, 2012 IEEE*. IEEE, 2012, pp. 1013–1018.

[26] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva, "Automatic analysis of affective postures and body motion to detect engagement with a game companion," in *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*. IEEE, 2011, pp. 305–311.

[27] M. Sun, Z. Zhao, and X. Ma, "Sensing and handling engagement dynamics in human-robot interaction involving peripheral computing devices," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 556–567.

[28] S. Saha, S. Datta, A. Konar, and R. Janarthanan, "A study on emotion recognition from body gestures using kinect sensor," in *Communications and Signal Processing (ICCSP), 2014 International Conference on*. IEEE, 2014, pp. 056–060.

[29] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 71–82.

[30] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer vision and image understanding*, vol. 104, no. 2-3, pp. 90–126, 2006.

[31] M. Gleicher, "Retargetting motion to new characters," in *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*. ACM, 1998, pp. 33–42.

[32] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 816–833.

[33] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *International conference on machine learning*, 2015, pp. 843–852.

[34] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, ser. Studies in Computational Intelligence. Springer Berlin Heidelberg, 2012. [Online]. Available: https://books.google.com.hk/books?id=wpb-CAAAQBAJ

[35] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[36] A. Kapur, A. Kapur, N. Virji-Babul, G. Tzanetakis, and P. F. Driessen, "Gesture-based affective computing on motion capture data," in *International conference on affective computing and intelligent interaction*. Springer, 2005, pp. 1–7.

[37] S. Piana, A. Staglianò, F. Odone, and A. Camurri, "Adaptive body gesture representation for automatic emotion recognition," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 6, no. 1, p. 6, 2016.

[38] M. Strait, L. Vujovic, V. Floerke, M. Scheutz, and H. Urry, "Too much humanness for human-robot interaction: exposure to highly humanlike robots elicits aversive responding in observers," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 3593–3602.

[39] M. K. Lee, S. Kielser, J. Forlizzi, S. Srinivasa, and P. Rybski, "Gracefully mitigating breakdowns in robotic services," in *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 2010, pp. 203–210.