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
The University of Hong Kong
myq.maxine@gmail.com

Hongwen Xie
Tencent Inc.
hongwenxie@tencent.com

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

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

***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 and the pose sensor positions. In addition, we use an Long Short-Term Memory-Recurrent Neural Network (LSTM-RNN) architecture to take the specific emotion context into account when estimating the intensities from body poses. Through dataset evaluations, we show that the proposed method delivers good performances on 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. And 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 more intense, the discriminative power of facial expressions for emotional intensities will degrade whereas the body cues will become dominant [7]. On the other hand, a dominance increase for 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 if he is with a surprised facial expression, the pose may imply more intensity of surprise than when 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 take the absolute joint positions [9] or other derived mechanic features, e.g., joint speeds [10], body expansions [11] etc., as pose descriptors. They might be effective for limited cases, however, in general, these handcrafts are subject to individuals' irrelevant physical differences (such as heights, body shapes etc.) as well as their positions to pose sensors, which may possibly limit or even deprive their potentials of practical applications.

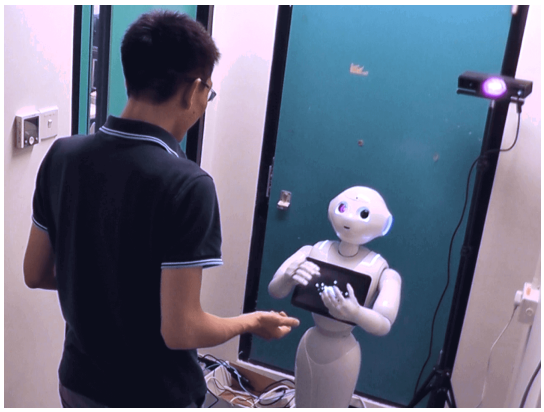


Fig. 1. A subject shows his high intensity of surprise by spreading out his hands.

In this paper, we propose a method to estimate emotional

intensities from body poses under various affective status. Different from aforementioned conventional approaches, we adopt the local joint transformations to describe body poses. Such descriptions are invariant to subjects' body shape differences, as well as their positions against sensors. Moreover, instead of deriving handcrafted features, we propose an Long Short-Term Memory-Recurrent Neural Network (LSTM-RNN) architecture to directly integrate the pose descriptors with emotion contexts for intensity estimation. The quantitative evaluations on a dataset demonstrate that the proposed method accurately predicts emotional intensities, achieving a correlation score of around 0.8. In the meantime, a series of succeeding field tests on a physical robot also implies that the proposed method enables a humanoid robot (Pepper) to sense subjects' emotional intensities effectively in real-time. Also, the robots equipped with our method are reported by subjects to be more emotional-sensitive and emotionally intelligent.

II. Related Work

Estimating affective status from body motions or gestures has been studied for years as psychological results report that body poses can encode emotions [12], [13], [8]. Some researchers thus investigated the possibility of recognizing emotion types from body motions under specific scenarios or together with other modalities. For example, Glowinski *et al.* used the trajectories of head and hands to predict valence and arousal. Gunnes and Piccardi classified body gestures into 6 emotions with naive representation from upper body gestures. To achieve a higher accuracy of emotion recognition, body poses are often integrated with other modalities, such as speeches [14], facial expressions [15], [9], [16] or both [17].

The body poses have also been used to estimate emotional intensities. 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 [18] as well as the head positions [19]. Furthermore, the body poses are also adopted to perceive other affective and cognitive process, e.g., accessibility level [20], engagement dynamics [21], [22]. 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 [23], joint distances [11], [18], body expansion[11] and silhouette motion images [24] 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, as shown in Fig. 2(a). Each node in the tree is associated with a body joint, and its position represents the joint position in the sensor coordinate system.

The local joint transformation representations are widely used in motion capture[25] and re-targeting[26]. The basic idea is to separate relative joint displacements in skeletal structures by local transformations in joint coordinate frames, rather than global positions in the sensor coordinate frame. Specifically, each joint has its own coordination frame (a predefined right-handed frame) and all frames are ordered based on the skeletal tree to form a parent-child structure, as shown in Fig. 2(b). The position of each joint is defined by a homogeneous transformation in its parental coordinate frame, i.e. a rotation and translation. The conversion from a local frame to the global frame, i.e., the sensor coordinate frame, is described by forward kinematics, and the inverse conversion by the corresponding inverse kinematics, as illustrated in the example of Fig. 2(c). Basically, the forward kinematic is as follows:

$$G_i = T_i T_{i_1} T_{i_2} ... T_1$$

where $G_i \in \mathbf{R}^4$ is the 3D 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 transformation for joint#i, we can use:

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

This series of transformations along the tree structure describe how joints move against their parental joints, and capture joint motions rather than absolute positions, thus invariant to sensor positions. To get rid of subject-dependent body differences, we further adopt only rotations to describe body poses since the joint translations are related to bone length (the distance between two joints). For simplicity, we use Euler angles $(\theta_r, \theta_p, \theta_y)$ in the order of roll, pitch and yaw to represent rotations rather than complex rotation matrices.

Consequently, each body pose is described by a vector of Euler angles, ordered strictly based on the skeletal tree structure. In addition, the rotation angles for the root joint is in the sensor coordinate frame, and will not be counted into

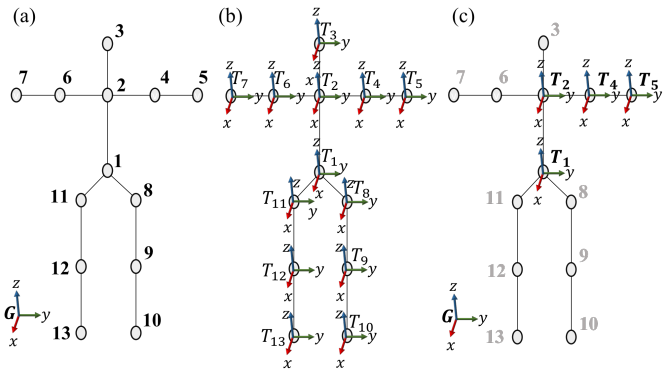


Fig. 2. (a) Skeletal tree structure with joint order numbers; (b) Each joint has its own coordination frame and all frames are ordered; (c) To find the transformation T_5 for joint#5 with global position G_5, first find forward kinematics along transformation tree: $G_5 = T_5T_4T_2T_1$, then $T_5 = G_5T_1^{-1}T_2^{-1}T_4^{-1}$

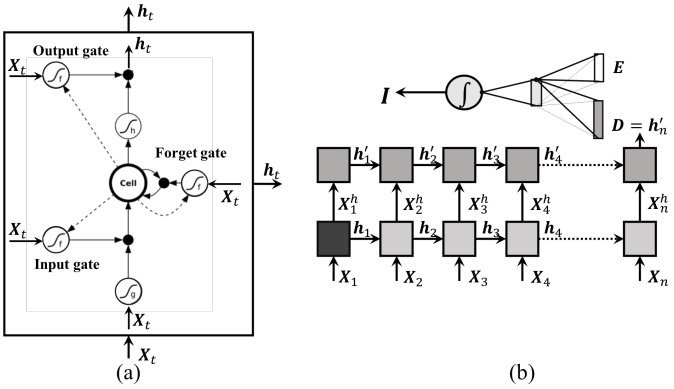


Fig. 3. (a) LSTM memory block; (b) the proposed LSTM architecture

the body pose feature vector. For each body pose X_i in a pose sequence, we have following descriptions:

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

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

B. LSTM-RNN modeling

Long Short-Term Memory-Recurrent Neural Network (LSTM-RNN) has been widely used for pose-related estimation, e.g., human pose estimation [27], action recognition [28] etc., owing to its powerful sequence modeling. Fig. 3(a) illustrates one LSTM memory block with a single cell, and three multiplicative gates, i.e., the input gate, the forget gate and the output gate. Their transformations and activation formulas can be found in [29]. Due to the trainable gates, LSTM-RNN can capture temporal flows along thousands or even millions of time steps in sequences with variable lengths, which is usually hard to achieve for conventional sequence model, e.g., Hidden Markov Model. These advantages make the LSTM-RNN a naturally good model to process body pose sequences since they are often intra-correlated over time and of variable lengths.

Inspired by work [30] which uses two-layered LSTM-RNN to describe a sequence of visual features, we also adopt a similar LSTM-RNN architecture as shown in Fig. 3(b). The architecture is composed of two LSTM layers and one fully-connected layers. The LSTM layers are to take pose sequences $[X_1, X_2, X_3, ..., X_n]$ of variable lengths as input and output a pose description vector $D = h'_n$ with fixed size. Then the full connected layer fused the emotion type vector E together with D before being processed by a sigmoid function to obtain the emotional intensity I.

IV. Experiments & Results

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

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, while the perceived emotion is a set of emotion label annotated by annotators. We use the percentage of the intended emotion in the set of the perceived emotion as the ground truth for emotional intensity based on the intuition that if the emotion expressed via a pose sequence is stronger, then the annotators are more likely to assign the same 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 need to further augment the training data. First, we double the dataset by exchanging left and right side of body skeletons since such exchange will not affect the emotion perception. Second, we sample from the raw pose sequence (120Hz) at fixed interval (30Hz) and then quadrupling the dataset. Finally, we have around 12,000 body sequences.

To show the effectiveness of the proposed method, we use the SVM together with handcrafted features proposed in [] 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

[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

	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

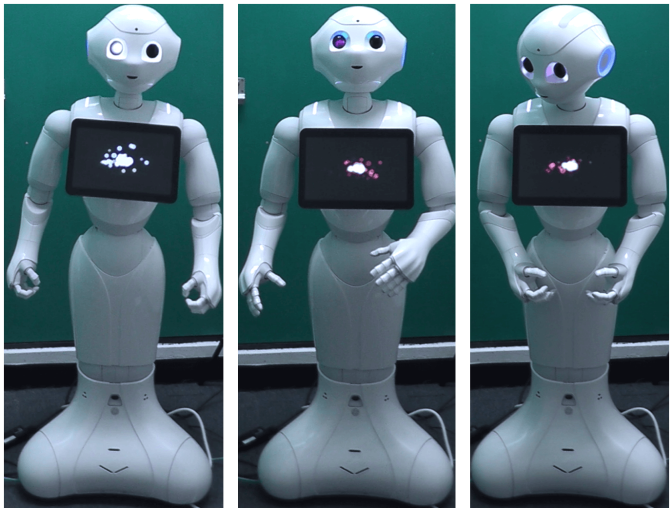
TABLE I
FIVE-FOLD CROSS VALIDATIONS ON THE MPI DATASET

the baseline method (SVM with handcrafted features), our method consistently achieves higher Pearson correlation coefficients in all tests. In addition, the average coefficient of 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 its participants' emotional intensities. However, during a real Human-Robot Interaction (HRI), the ground truths of subjects' emotions are hard to obtain, thus making the comparison evaluation difficult to achieve. On the other hand, the subjects themselves are well aware of what emotions and the associated intensities they are expressing when they are interacting the robot. Thus, if the robot can show the estimation results in real-time to the subjects, then the subjects can evaluate whether the results are accurate or not. Instead of directly showing/telling the detected emotions and their associated intensities, we use different robot behaviors to imply the estimation results in order to make the interaction process 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 the type *none*, the robot has no emotion detections or intensity estimation, and, during the interaction, it only reacts with random filler speeches, e.g., oh and hmm, regardless of subjects' emotions and the intensities. By contrast, the type *expressive* adopts the proposed method to manipulate robot behaviors. Specifically, in mode *expressive*, the robot behaviors are triggered by users' emotions and the intensities in real-time. If low intensity of emotion is detected, 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. Furthermore, the robot will provide more gestural feedback once the subject's emotion intensity becomes high, as shown in Figure 4. The set of speeches and postures for emotions are different to each other.

The field testing was conducted on a Pepper robot with a Kinect sensor. The subject 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



(a) (b) (c)

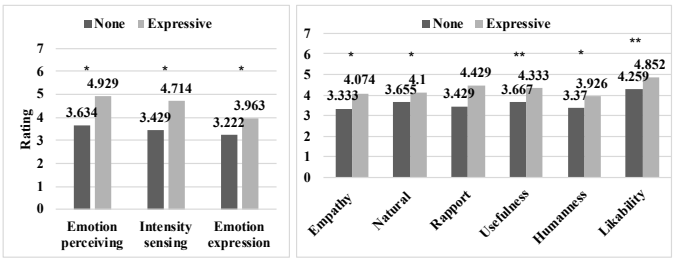
Fig. 4. Body expressions for Pepper robot when it detects high level of emotional arousals: (a) joy; (b) surprise; (c) sadness

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. Accordingly, we only choose three common facial expressions (happiness, surprise and sadness), which are then used as the specific emotion contexts for estimating the intensities from body poses.

We recruit 14 subjects with their informed consents to participate in the field test. In the test, subjects were required to express three emotions (joy, sadness and surprise) via their faces and body poses to a robot in two intensity levels. Subject were also told about the purpose of the experiment and were encouraged to change their emotions and the 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, the subjects need to rate the robot's affective-related ability in terms of "emotion perceiving" and "arousal sensing" on a 7-point Likert scale. In addition, they also need to rate the robot's other performances through a set of 7-point Likert scale questions derived from [31], [32]. The whole system was implemented in Robot Operating System (ROS) and run in real-time with Ubuntu system (Intel Core i7, GeForce 920M and 8GB RAM).

The rating results are shown in Fig.5. 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

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



(a) (b)

Fig. 5. The 7-point ratings of: (a) robot emotion perceptions ; (b) robot other performances. ($* : p < .05$ and $** : p < .01$)

"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. In addition, subjects also gave significantly different ratings on robots' other performances in terms of robots' ability to show empathy, to be natural, to be useful, to be human-like and their likability. Overall, all the ratings on different aspects in mode "expressive" are higher than that in mode "none". All these results show that the proposed method is useful for real HRI, and the robot equipped with the ability to perceive emotional intensities can dramatically change subjects' views on it.

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, as well as their positions against sensors. Moreover, we also 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 with ground truth labels. 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 in all aspects.

Acknowledgements

This project is sponsored by WeChat-HKUST Joint Laboratory on Artificial Intelligence Technology (WHAT LAB). We thank WeChat team for their contributions to this paper.

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] 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.

[13] 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.

[14] H. A. Vu, Y. Yamazaki, F. Dong, and K. Hirota, "Emotion recognition based on human gesture and speech information using rt middleware," in *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*. IEEE, 2011, pp. 787–791.

[15] H. Gunes and M. Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," in *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, vol. 4. IEEE, 2005, pp. 3437–3443.

[16] A. Psaltis, K. Kaza, K. Stefanidis, S. Thermos, K. C. Apostolakis, K. Dimitropoulos, and P. Daras, "Multimodal affective state recognition in serious games applications," in *Imaging Systems and Techniques (IST), 2016 IEEE International Conference on*. IEEE, 2016, pp. 435–439.

[17] L. Kessous, G. Castellano, and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, pp. 33–48, 2010.

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

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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

[30] 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.

[31] 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.

[32] 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.