2017 IEEE International Conference on Robotics and Automation (ICRA)
Singapore, May 29 - June 3, 2017

Emotional Intelligence in Robots: Recognizing Human Emotions from Daily-life Gestures

Mohammad Reza Loghmani
ACIN - Vienna Univ. of Technology,
Gußhausstraße 27-29,
1040 Vienna, Austria
loghmani@acin.tuwien.ac.at

Stefano Rovetta
DIBRIS - University of Genoa,
Via Dodecaneso, 35
16146 Genova, Italy
stefano.rovetta@unige.it

Gentiane Venture
Tokyo Univ. of Agri. and Technology,
2-24-16 Nakacho Koganei,
Tokyo 184-8588, Japan
venture@cc.tuat.ac.jp

***Abstract*— The rapid advancement of robotics poses the problem of a deep integration of robotic systems in human environments. In order to achieve this symbiosis between humans and robots, the artificial systems have to take into account one of the most important aspects in human life: emotions. The recognition and understanding of human emotions is crucial for robotic systems to behave in appropriate ways according to the situation and smoothly integrate with all the different aspects of human life. This paper proposes a novel algorithm which uses state-of-the-art techniques in Machine Learning, in particular Recurrent Neural Networks, to automatically infer emotional clues from non-stylized motions (i.e. motions which are not supposed to convey emotional information as primary goal). This algorithm recognized human emotions with an accuracy between 0.68 and 0.80, depending on the considered motion, and clearly overcomes human capacity in the same task for the considered cases studied. Since the implemented algorithm is able to perform online, its results can be used to allow a behavioural programming which gives the robot the flexibility to act in a more human-oriented way.**

I. Introduction

A. Motion and Emotional Intelligence

Motion is not only a displacement of some parts of the body aimed at performing a specific task, but is also one of the most important channels in human communication. In fact, non-verbal communication, i.e. every type of communication with no use of words, represents most of all communications in human interactions [1].

Differently from the early industrial robots, designed to accomplish repetitive tasks in isolated environments, robotics is now focused on designing robots capable of operating in contact with humans and to interact with them in the most natural way. A fundamental requirement for understanding people is the ability to perceive and interpret their actions. For this reason, a desirable characteristic for a modern robot operating in human environments is the ability to recognize and predict human gestures together with the one of extracting conscious or unconscious "messages" from those motions. The information carried by human motions can belong to a wide range of scopes, from healthcare [2] [3] [4] to affects [5] [6] [7], without forgetting identity [8] [9]. As a consequence, a robot able to identify different types of motion and extract information from them can be used in countless applications. This paper focuses on the emotional contents of human motions and how a subject feeling different emotions can perform the same movement differently. In the literature, several authors have dealt with this problem, considering both motions that convey emotion as a primary goal (*stylized motions*) [10] [11], and motions that convey emotional information as a side effect (*non-stylized motions*) [12] [5]. Unlike algorithms that recognize emotions from facial expressions which already present excellent performances [13], the research in emotion recognition from body gestures presents non-conclusive results for non-stylized motions.

In order to be able to deal with different motions and different emotions at the same time, a rich description of the motion is necessary. This requirement motivated the choice of a multi-sensor system, with a special attention to its cost. In fact, in order to create a system that is both compact enough and affordable for private users, low cost commercial sensors have been preferred. Furthermore, because of the complexity in analysing the emotional content involved in daily-life, non-stylized motions, which are the target of this research, techniques allowing a high level of abstraction are fundamental. For this reason, deep learning methods such as Recurrent Neural Networks (RNNs) are considered.

B. Related research

Since face-to-face interaction is dominated by facial expressions and speech, these are the modalities that have been predominantly studied for automatic extraction of affective information. Nevertheless, when we are dealing with situations in which the affective state is estimated from a distance or it is easier to communicate through motion, body movement is an effective modality for Emotion Recognition (ER).

This paper focuses on movements that are performed to accomplish a specific task, unrelated to the expression of affect [12] [14]. In this case expressiveness is secondary to function and affective states can only be expressed through a modulation of motion. Pollick et al. [5] have compared automatic affect recognition model performance with human recognition performance in distinguishing between angry and neutral in knocking, lifting, and waving actions. A Multi-Layer Perceptron (MLP) was used to distinguish between anger and neutral in the different motions. The results show that MLP outperforms humans in affect recognition. Bernhardt and Robinson have studied automatic recognition of happiness, anger, sadness and neutral in non-stylized motions

978-1-5090-4633-1/17/$31.00 ©2017 IEEE

by using Support Vector Machines (SVMs) with polynomial kernel. They show the improvement of performances by removing the personal biases from movements. Karg et al. [15] used affective whole body gait patterns to build automatic recognition models able to examine the differences between inter-individual and person-dependent recognition accuracies for emotion categories. This problem was also addressed by Bernhardt and Robinson [16] and by Gong et al. [17] who focused on the effect of personal biases of the considered models.

C. Contribution

The modern trend in psychology is to consider emotional thought as a contributor to logical thought and, as a consequence, a part of general intelligence [18]. Psychologists Mayer and Salovey introduced the concept of Emotional Intelligence [19] which is defined as the subset of Intelligence *"[...] that involves the ability to monitor ones own and others feelings and emotions, to discriminate among them and to use this information to guide ones thinking and action"*. This interpretation of emotions has led Artificial Intelligence (AI) researchers to consider the development of emotion-aware machines, giving rise to the AI sub-field of affective computing. The present research puts itself into this context and aims at discriminating between some basic emotions from the data of daily-life, non-stylized motions. In particular, this research focuses on distinguishing between *happiness*, *sadness*, *anger* and *neutral* by analysing the motions *clapping*, *drinking*, *throwing* and *waving*. The knowledge of the emotional content of motions is intended to significantly improve the interaction between humans and robots, allowing a behavioural programming that takes into account this fundamental aspect of human beings. Microsoft Kinect v2, Nintendo Wii Balance Board and IMU Shimmer r2 are used together to set up a unique affordable multisensory system.

II. Recurrent Neural Network

A. Deep Learning

The mainstream approach of overcoming the *"curse of dimentionality"* [1] [20] was to reduce the dimensionality of the data via human-engineered feature extraction, which is time consuming and highly application-dependent.

Recent neuroscience findings about neocortex led to a new trend in which raw data are directly input into a Neural Network (NN) with a complex hierarchy of modules which learns data regularities and allows a high generalization capacity [21] [22].

In many cases, including activity recognition and analysis, the temporal component also plays a key role. The meaning conveyed by a sequence of patterns may vanish when analysing isolated fragments of this sequence. For this reason, modelling the temporal component is among the fundamental goals of deep learning systems.

[1] In the context of pattern classification applications, the learning complexity grows exponentially with linear increase in the dimensionality of the data.

B. Gated Recurrent Unit

When dealing with data with a temporal component, a desired characteristic for our network is the ability to process the data in their temporal order, by "remembering" the crucial information of the past time steps. The main category of deep learning systems which satisfies these requirements is Recurrent Neural Networks. This technique addresses the problem of dealing with data temporal component by introducing a loop in the network that has the goal of passing the information about the previous time steps to those that are coming, effectively implementing a discrete-time, strongly nonlinear dynamical system. Due to this characteristic, RNN-based networks are often categorized as deep architectures even when containing few layers. In fact, unrolling the loop shows this kind of networks to be "deep" in the temporal dimension.

Considering vanilla RNN, the network contains a single unit applying a nonlinear function, such as $\tanh$, on the data. We can thus write the input-output relation as

$$h_t = \tanh(W \cdot [h_{t-1}, x_t] + b) \tag{1}$$

where x_t and h_{t-1} are the input at time step t and the hidden state at time step $(t-1)$, respectively, and W and b are the weight matrix and bias vector, respectively.

The simple structure of vanilla RNNs is not optimal in terms of memory persistence. The network is only able to track the information carried by the few last time steps, while taking into account long-term dependencies is impossible. In order to overcome this problem, many types of recurrent networks have been proposed. One of the most effective architecture is the Gated Recurrent Unit [23]. This architecture contains structures, called gates, that decide which information is important for the next time steps and which is not. First, the *reset gate* at time step t is computed as

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \tag{2}$$

where $\sigma(\cdot)$ is the logistic sigmoid function.

Then, the *update gate* at time step t is computed as

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \tag{3}$$

Finally, the hidden state at time step t is computed as

$$h_t = (1 - z_t)h_{t-1} + z_t\widetilde{h}_t \tag{4}$$

where the candidate hidden state $\widetilde{h}$ is

$$\widetilde{h}_t = \tanh(W \cdot [r_t h_{t-1}, x_t]) \tag{5}$$

Eqs. 2–5 can be roughly interpreted as follows: the reset gate r decides whether the previous hidden state h_{t-1} is ignored or not, while the update gate z determines whether the hidden state is to be updated with a new candidate hidden state $\widetilde{h}$.

It is worth noting that, for the purposes of this paper, capturing long-term dependencies is essential. Since the duration of the analysed motions is unconstrained, the use of vanilla RNNs do not guarantee a sufficient temporal persistency for effectively capturing emotional clues.

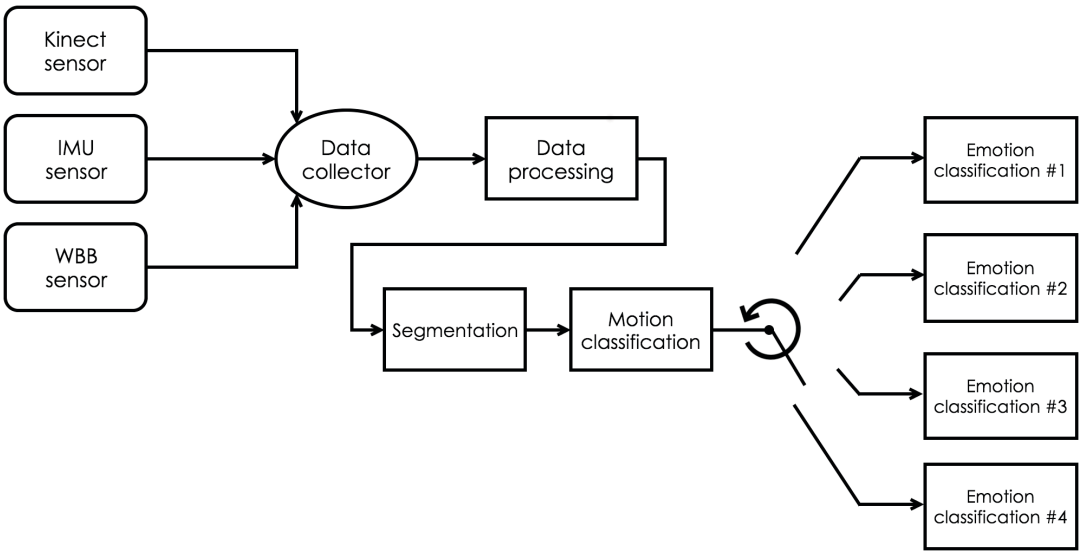


Fig. 1: High-level view of the algorithm's structure

III. Implementation

A. Algorithm

Since psychological studies do not offer a unified way for distinguishing between different emotions through body gestures yet, this work proposes a two-stage approach in which the recognition of the emotion is biased by the analysed motion. The scheme in Fig. 1 presents a high-level view of the proposed algorithm. The data provided by the sensors are collected and then normalized and co-registered to obtain inter-sensory synchronization. The processed data stream is then segmented and the portion corresponding to motion is selected and given as input to the motion classifier, which works as a multiplexer and selects the proper emotion classifier according to the recognized motion. Finally, the selected emotion classifier provides the final prediction for the emotion.

B. Data preprocessing

One of the most important aspects of the present work is the multi-sensoriality of the data. The use of three different sensors (Kinect, Wii Balance Board and IMU Shimmer) makes a large set of features available to the algorithm. Specifically, the three adopted sensors provide the following data: *Kinect v2* provides the 3-dimensional (3D) position of human skeleton joints; *IMU Shimmer r2* provides linear acceleration and angular velocity along the x-, y- and z-axis of one subject wrist, according to his/her handedness; *Wii Balance Board* provides the displacement of the Center of Pressure (CoP) along the x- and y-axis and the vertical ground reaction force.

This wealth of information allows the analysis of a wide range of motion involving different parts of the body. Nevertheless, the use of multiple heterogeneous sensors give rise to different problem for the parallel management of diverse data. Considering the difference in the format and in the sample frequency, it is important to deal with the problems of adapting and synchronizing the data. For this purpose, the data are subject to the following preprocessing steps:

- *Kinect normalization:* The data acquired by Kinect is made invariant with respect to translation in space by moving the reference frame from the camera to the hip position. Let K_j^q be the 3D position of joint q in the j^{th} sample. The normalized joint positions are obtained as:

$$\hat{K}_j^q = K_j^q - K_j^{HIP} \tag{6}$$

  with $j \in \{1, \ldots, n_{samples}\}$ and $q \in \{1, \ldots, n_{joints}\}$.
- *Feature scaling:* In order to make data comparable and avoid outliers to affect too much the training phase of the classifiers, a typical feature scaling that linearly maps all values in the range $[0, 1]$ is performed. The normalized input $\hat{\boldsymbol{x}}$ is then computed as

$$\hat{x} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{7}$$

  with x a generic input vector, x_{min} and x_{max} constant vectors computed offline on the training data representing the minimum and the maximum of the input, respectively.
- *Interpolation and resampling:* In order to deal with the difference in the sampling frequency and, at the same time, tackling the problem of variable frequency of the Wii Balance Board (the sampling frequency oscillates between 23 Hz and 27 Hz) interpolation and resampling are performed on the data.
- *PCA:* The high dimensionality of the data raises the necessity of a dimensionality reduction. For this purpose, one of the most common techniques is PCA and the number of principal components has been selected in such a way that the new coordinates keep 99% of the variance of the original data.

C. Segmentation

In order to isolate the data portions related to motions, a window-based segmentation module is implemented. It consists in a binary classifier that, for each windows, distinguishes between "motion" and "idle time." Each window groups all the data within a time interval Δt, similarly to the sequence-based window introduced in [24], and adjacent windows overlap by $\frac{\Delta t}{2}$. For this task, three different classifiers are considered: Support Vector Machine (SVM), Random Forest (RF) and K-Nearest Neighbour (KNN). Uniform grid search and Cross-Validation (CV) are used to explore their hyperparameter space and choose the architecture that performs better.

D. Offline classification

To test the performances of well-known classifiers in the considered tasks, the same classifiers used for segmentation, SVM, RF and KNN, are considered. A combination of gridsearch and CV allowed to find the best architecture possible in the range specified for the hyperparameters and to validate the obtained results for each of the classifiers present in the developed algorithm.

E. Online classification

To improve the prediction accuracy and to deal with the temporal dimension of the data, recurrent networks are adopted for motion and emotion classification. The general architecture of the networks used for classification is composed by an input layer, n_g consecutive GRU layers, a fully-connected layer with SoftMax activation function which produces a distribution over the 4 class labels. The choice of this activation function is fundamental for the online implementation of the algorithm since it is possible to define a threshold θ and stop the classification when the value of one of the outputs exceeds this value. This gives the user more flexibility, allowing to stop both motion and emotion classification before the end of the motion itself, according to the application. For example, in an application in which speed is crucial even at the expense of accuracy, a low threshold θ can be set. Once again, uniform grid search and Leave-One-Out CV are used to define the network architecture, as well as the other hyperparameters, for each classifier and validate the results. The implemented RNN-based classifiers present 2 to 3 GRU layers with 50 to 250 hidden neurons each and ReLU non-linearity, depending on the specific task. A learning rate of 0.01 and a dropout ratio of 0.5 are set. RMSProp optimizer and categorical cross-entropy loss function are used for training the weights of the network. The training phase is stopped after 70 epochs and the a batch size of 10 is considered at each iteration.

F. Rejection option

In some critical applications, it could be better not to act when the prediction is not reliable enough. In these situations, the system should have the additional option not to select between the analysed emotions and decide "not to decide." It is not possible to simply base this decision on the output values, since, in most of the cases, the network is forced to converge to one of the classes after some time steps. This behaviour is due to the high value of the weights in the fully-connected layer. By analysing the output of misclassified emotions, it appeared that "chaotic outputs" are more frequent than in correctly classified cases. A "chaotic output" is an output in which two or more classes continuously flip their position and temporarily prevail over the others. In order to take advantage of this characteristic, the area underlying the curves defined by the four outputs in their temporal evolution is computed and, together with final values of the output layer, is given as input to a RF classifier. This classifier assigns to the "rejection" class those samples in which the emotional information is unclear, in the sense defined above.

In order to evaluate the performances of the RF classifier used in this stage, two different quantities has been taken into account: the fraction of the test data correctly classified by the emotion classifier, but assigned to the "rejection" class (CR) and the fraction of the test data misclassified by the recurrent network and then assigned to the "rejection" class (WR). The improvement in the performance is thus computed as

$$performance\_improvement = WR - CR \tag{8}$$

G. Human baseline

In order to be able to interpret in the best way the results obtained for emotion recognition task, a baseline is determined by asking a group of people to recognize the emotion from the videos of the motions. Eighty samples from the dataset have been selected randomly but with equal distribution between the different actors and the different (motion, emotion) cases. The videos of these samples have been processed by blurring the face of the actors both to preserve their privacy and for preventing people from recognizing emotions from facial expressions. The set of videos has been divided in 5 groups of 16 videos, each group containing one sample for each different (motion,emotion) case. Eighteen people (5 females and 13 males) have been recruited for the experiment and each person has watched the videos and filled the related questionnaire. For each of the videos, the participant was asked to indicate the emotion characterizing the motion.

H. Online performances

In some applications, for example anything concerning security and safeness, a quick prediction of the emotions can be crucial. Consequently, it is important to provide some data related to the temporal performances of the system. As already discussed in the previous sections, the introduction of RNNs allows to deal with time series data in a natural way and the system is able to provide continuous response while the algorithm is running. So, in order to make the system work in real time, the different parts of the algorithm have been parallelized and inserted in a multi-process framework.

TABLE I: Gender, age, handedness and nationality of the actors enrolled for the experiments

SUBJECTS' INFORMATION				
Subject	Gender	Age	Handedness	Nationality
actor1	M	22	+1.0	Japanese
actor2	M	20	+0.9	Japanese
actor3	M	21	+0.7	Japanese
actor4	F	21	-0.9	Japanese
actor5	F	21	+1.0	Japanese

IV. Experiments

A. Experimental setup

The experiments were conducted in a 36 m^2 room. A desktop computer was connected to Microsoft Kinect v2 (wired USB 3.0 connection), Nintendo Wii Balance Board (wireless Bluetooth connection) and IMU Shimmer r2 (wireless Bluetooth connection). The Wii Balance Board was positioned approximately in the center of the room, while the Kinect was pointing to it at a distance of 1.8 m. The relative orientation between Kinect and Wii Balance Board was selected such that the coronal plane of the subject's body is recorded from the Kinect. In other word, the two sensors were arranged in such a way that the z-axis of the Kinect was orthogonal to the coronal plane of the subject. This choice is due to Kinect's performance limitations: the information related to the joints composing the skeleton structure provided by this sensor are corrupted or missing in other settings due to occlusions or lack of references. The IMU sensor was attached to the wrist of the subject according to his/her handedness using the "Edinburgh inventory" [25]. Finally, a camera was positioned close to the Kinect to record the performance of the subject.

B. Motions and Dataset

In order to acquire data for all the four motions (clapping, drinking, throwing and waving) performed in all the four emotional status (happy, sad, angry and neutral), acting school students has been enrolled. Data related to 5 actors of both genders between 20 and 22 years old has been collected. Tab. I shows the information about each single actor. One of the most important information for the experiment is the "handedness," because it determines the positioning of the IMU sensor. The value of this field in Tab. I indicates which hand (right or left) the subject prefers to use in a continuous scale $\in \{-1.0, +1.0\}$ with $+1.0$ indicating "totally right handed" and -1.0 indicating totally left handedaccording to the "Edinburgh inventory".

Each actor was asked to perform all the analysed motion in all the selected emotional status. Each combination was recorded three times for a total of 48 $(4 \times 4 \times 3)$ motions per subject. Due to acquisition problem, some of the motions were discarded: the final dataset is composed of 235 motions, with at least one sample per couple $(motion, emotion)$ for each actor.

During the experiments, the specific execution of these motion were not constrained in order to keep them as natural and genuine as possible. The only motion for which some guideline was given was *throwing*: the many different ways of throwing can be considered as totally different motions. For this reason, the subject were asked to throw a small plastic ball (approximately of the same dimensions of a billiard ball) diagonally, from the same height of his/her shoulder to the floor. For this motion the throwing motion considered in [26] is taken as reference. Except from that, the only constraint imposed to the motions is a system-related one due to the Wii Balance Board: the subject had to stand on the board thus couldn't freely move in the room.

C. Procedure

During all the experimental session, for each of the actors involved, the following procedure was followed:

- a brief general explanation of the project is given to the actor in order to make him/her understand what and why he/she is asked to do;
- the actor is asked to fill a form with some personal information and to take the "Edimburgh inventory" test for the handedness score;
- the actor is asked to perform the actions by acting in the different required emotional status;
- in order to perform naturally, the actors rely on two different methods: the Stanislavskij System, through the implication of actual, personal past experiences, and the use of different plausible real life scenarios (story telling) involving the desired couple $(motion, emotion)$. Both procedures foster the authenticity and believability of the portrayals as they discourage the use of stereotypical patterns [27];
- in order to prevent from collecting irrelevant data, the actor is asked to auto-evaluate himself/herself, thanks also to the recorded videos, to discard and eventually repeat poorly performed motions;

V. Results

A. Segmentation

As discussed in Section III, three different classifiers were tried to classify a data window as "motion" or "idle time": SVM, RF and KNN. Tab. II summarizes the results of the segmentation and the motion classification, validated with Leave-One-Out CV, for the different analysed classifiers. It can be noticed that the performances of all the considered classifiers are comparable and go above 0.9 accuracy. The best result is obtained with RF, which scores 0.93 and shows a slightly better result with respect to the other classifiers. From these results we can conclude that, thanks to the rich informational content guaranteed by the multi-sensory data, the classifiers can easily distinguish between the two classes. The main source of error in this task are the beginning and the end of the motion, which do not present a sharp change during the transition between "idle time" and "motion".

TABLE II: Summary of the results obtained for the segmentation and motion classification task with the different classifiers

SEGMENTATION/MOTION CLASSIFICATION ACCURACY				
TASK	SVM	RF	KNN	RNN
Segmentation	0.92	**0.93**	0.91	-
Motion classification	0.94	0.91	0.91	**0.99**

TABLE III: Summary of the results obtained for the emotion classification task with the different classifiers

EMOTION CLASSIFIERS' PERFORMANCES				
MOTION	SVM	RF	KNN	RNN
Clapping	0.58	0.53	0.54	**0.75**
Drinking	**0.81**	0.51	0.71	0.80
Throwing	0.44	0.405	0.36	**0.68**
Waving	0.48	0.46	0.35	**0.68**

As a consequence, the windows at the beginning and at the end of the motion are sometime misclassified as "idle time". Nevertheless, an absolute accuracy in defining the boundaries of the motion is out of the scope of this paper and segmentation is only a task functional to the final emotion classification task.

B. Motion classification

The results of the motion classification in Tab. II show that it is possible to classify these motions with data collected by the three commercial sensors selected. All the classifiers' performances goes above 0.9 of accuracy, with SVM which obtains the best score (0.94) among the non-recurrent classifiers. The recurrent motion classifier presents an accuracy of 0.99, outperforming the best non-recurrent classifier (accuracy = 0.94) of 0.05. By analysing the misclassification cases, no common pattern occurs.

C. Emotion classification

From the results presented in Tab. III it can be noticed that the accuracy of emotion recognition task for all the motions goes above chance (0.25). Nevertheless, among the results obtained with SVM, RF and KNN, the only motion that presents relevant results is "drinking", where the emotion classification accuracy reaches 0.815 with SVM classifier. These results highlight the difficulty of the problem of emotion recognition, especially in this case where no manual feature extraction is performed. In addition, as already discussed in the previous sections, the fixed size of the inputs required by the non-recurrent classifiers does not allow to obtain online prediction.

It can be noticed that, analogously to the results of the non-recurrent classifiers considered in Section 7.5, the easiest motion to analyse remains "drinking", while the most difficult ones seem to be "throwing" and "waving". For "clapping", "throwing" and "waving" the recurrent classifiers outperform the accuracy of the non-recurrent ones of 0.17, 0.23 and 0.21 respectively. The only motion in which the recurrent network shows slightly poorer performances is "drinking" in which the SVM classifier presents an accuracy 0.01 higher. These results also overcome the recognition rate established by the human baseline. In fact, the performed experiments showed that human participants were able to recognize different emotions with an accuracy of 0.45.

Fig. 2a–2d show the confusion matrices related to the recurrent emotion classifiers. The analysis of this matrices highlights some interesting points about the classification of the different emotions. Fig. 2a shows that neutral and angry clapping are easier to distinguish with respect to happy and sad clapping. In fact, by analysing the videos of the experiment, it can be noticed that happiness and sadness are the two emotions that present the higher interpersonal variance: each actor has interpreted these two emotions in a very different way. Fig. 2b shows that the emotion that is difficult to recognize from drinking motion is anger. This fact is linked to the nature of the motion itself: since the person in handling a full glass/bottle, this emotion does not present the jerkiness that characterizes it in other motions and it is confused with the other options. Fig. 2c shows that, as expected, the emotion recognized with highest accuracy from throwing motion is anger. Instead, happy and sad throwing seems to be difficult to recognize. In fact, sad throwing is often confused with neutral throwing since they look very similar, while happy clapping presents a high interpersonal variance. Finally, Fig. 2d shows that angry waving is often confused with the other options. This can be explained by the infrequent occurrence of this motion-emotion combination in real life.

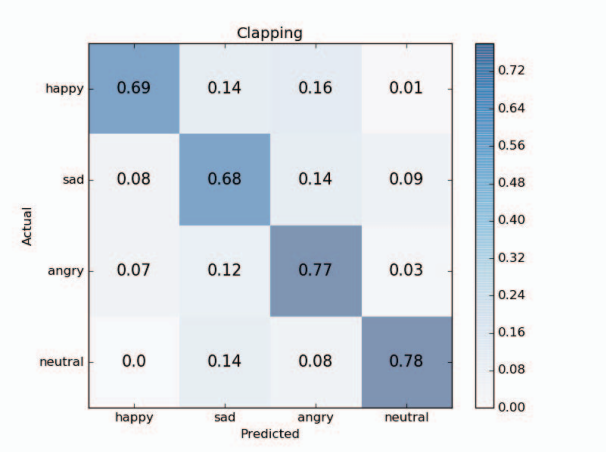
D. Rejection option and temporal performance

When considering applications in which a false positive in the final predicted emotion implicates a high cost, the addition of a "rejection" class improves the performances for all the analysed motions. In this context, the rejection of misclassified emotion is precious and thus to be intended as a correct result. The details of this improvement in the performances in shown in Tab. IV.

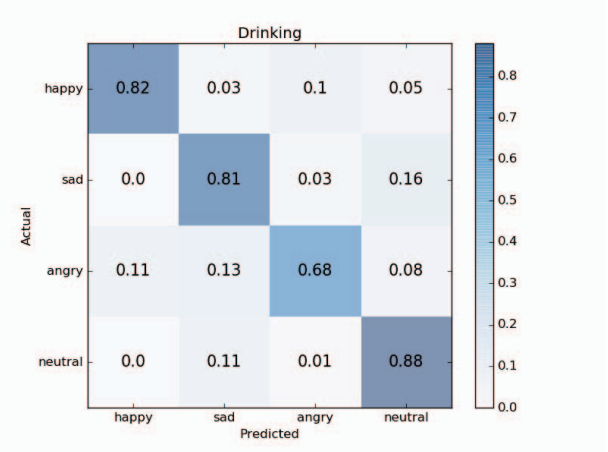
Concerning the temporal performances of the developed algorithm, several tests showed that the final prediction in achieved is average in 4.9s from the beginning of the motion. Since the average duration of the motions in the dataset is 6s, the final prediction is usually given before the end of the motion itself.

VI. Conclusion

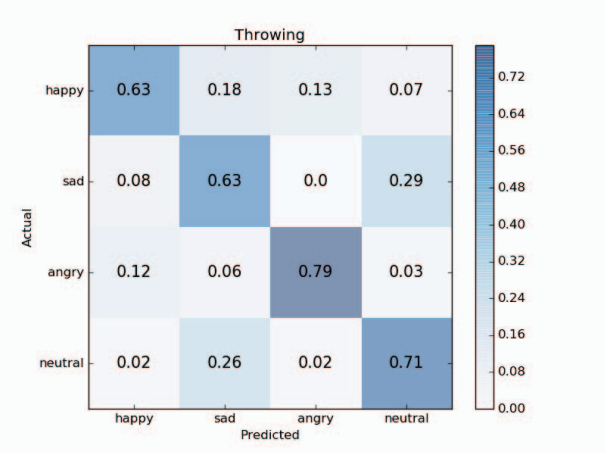
In this paper, we addressed the issue of distinguishing between *happiness*, *sadness*, *anger* and *neutral* by analysing the non-stylized motions *clapping*, *drinking*, *throwing* and



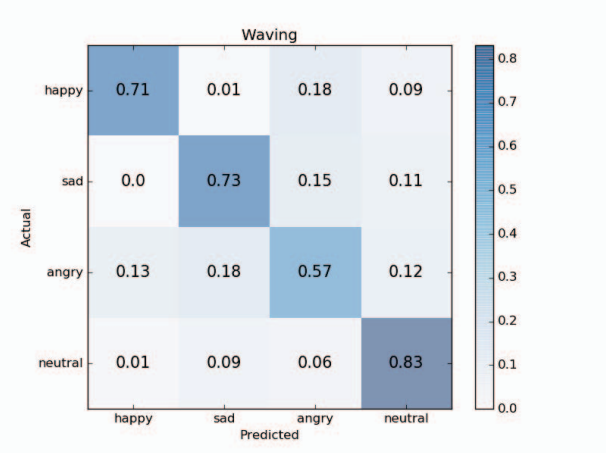
(a) Clapping



(b) Drinking



(c) Throwing



(d) Waving

Fig. 2: Confusion matrices representing the result of the emotion recognition task for the four different motions.

TABLE IV: For each motion, a different "rejection" classifier is trained in order to catch possible motion specific misclassifications. The values reported in the table have to be intended as an improvement (addition) with respect to the RNN-based results in Table III

EVALUATION OF THE "REJECTION" CLASSIFIER				
Motion	Clapping	Drinking	Throwing	Waving
Performance improvement	0.05	0.10	0.03	0.06

waving. Because of the task complexity, a recurrent network based on GRU layers was used as classifiers to first recognize the performed motion and successively predict the involved emotion. The performed tests demonstrated the validity of the proposed algorithm that was able to predict the emotional status of the subject with an accuracy between 0.68 and 0.80, depending on the performed motion. We show that these results both outperform other Machine Learning classification techniques as SVM, RF and KNN and overcome the human capacity of performing the same task for the selected case studies.

For applications in which a wrong response from the algorithm can lead to a high cost, an additional option not to select any of the considered options in the uncertain cases has been implemented. The insertion of this "rejection" class has brought an improvement in the performance between 0.034 and 0.101 depending on the motion.

Finally, since for certain applications a fast response is required, the performances of the whole algorithm has been tested also in terms of response time. The tests pointed out that the algorithm is generally able to make a prediction about the emotion involved in the observed motion before the end of the motion itself.

ACKNOWLEDGMENT

Work partially funded by the Japanese Society for the Promotion of Science Grant in Aid for Challenging Exploratory research K115K12124, the Japanese student service organization scholarship SSSV and the European Community, Horizon 2020 Programme (H2020-ICT-2014-1), under grant agreement No. 676157, ACROSSING.

REFERENCES

[1] R. Baron, *Social psychology*. Pearsonnonverbal, 2009, no. 12th ed., ch. "Social perception", pp. 79 – 109.

[2] M. Allison, J. Maya, and I. Henrik, "Medical and health-care robotics," *IEEE Robot. Automat. Mag.*, vol. 17, no. 3, pp. 26–37, 2010.

[3] P. Kazanzides, G. Fichtinger, G. D. Hager, A. M. Okamura, L. L. Whitcomb, and R. H. Taylor, "Surgical and interventional robotics: Core concepts, technology, and design," *IEEE Robot. Automat. Mag.*, vol. 15, no. 2, pp. 122–130, 2008.

[4] S. Micera, M. C. Carrozza, L. Beccai, F. Vecchi, and P. Dario, "Hybrid bionic systems for the replacement of hand function," *IEEE*, vol. 94, no. 9, p. 17521762, 2006.

[5] F. Pollick, V. Lestou, J. Ryu, and S. Cho, "Estimating the efficiency of recognizing gender and affect from biological motion," *Vision research*, vol. 42, no. 20, pp. 2345–2355, 2002.

[6] A. Camurri, B. Mazzarino, and G. Volpe, "Expressive interfaces," *Cognition, Technology & Work*, vol. 6, no. 1, pp. 15–22, 2004.

[7] A. Kapur, A. Kapur, N. Virji-Babul, G. Tzanetakis, and P. Driessen, "Gesture-based affective computing on motion capture data," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2005, pp. 1–7.

[8] C. Tomasi and T. Kanade, *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ. Pittsburgh, 1991.

[9] T. Zhang, M. Karg, J. F. S. Lin, D. Kulic, and G. Venture, "Imu based single stride identification of humans," in *IEEE Int. Work. Robot Hum. Interact. Commun.* IEEE, 2013, p. 220225.

[10] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 64–84, 2009.

[11] A. Camurri, I. Lagerlöf, , and G. Volpe, "Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques," *International journal of human-computer studies*, vol. 59, no. 1, pp. 213–225, 2003.

[12] G. Venture, H. Kadone, J. G. T. Zhang, A. Berthoz, and H. Hicheur, "Recognizing emotions conveyed by human gait," *International Journal of Social Robotics*, vol. 6, no. 4, pp. 621–632, 2014.

[13] B. Mishra, S. Fernandes, K. Abhishek, A. Alva, C. Shetty, C. Ajila, D. Shetty, H. Rao, and P. Shetty, "Facial expression recognition using feature based techniques and model based techniques: A survey," in *Electronics and Communication Systems (ICECS), 2015 2nd International Conference on*. IEEE, 2015, pp. 589–594.

[14] Y. Ma, H. Paterson, and F. Pollick, "A motion capture library for the study of identity, gender, and emotion perception from biological motion," *Behavior research methods*, vol. 38, no. 1, pp. 134–141, 2006.

[15] M. Karg, K. Kuhnlenz, and M. Buss, "Recognition of affect based on gait patterns," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 40, no. 4, pp. 1050–1061, 2010.

[16] D. Bernhardt and P. Robinson, "Detecting affect from non-stylised body motions," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 59–70.

[17] L. Gong, C. Wang, F. Liu, F. Zhang, and X. Yu, "Recognizing affect from non-stylized body motion using shape of gaussian descriptors," in *Proceedings of the 2010 ACM Symposium on Applied Computing*. ACM, 2010, pp. 1203–1206.

[18] R. Leeper, "A motivational theory of emotion to replace 'emotion as disorganized response'." *Psychological Review*, vol. 55, no. 1, p. 5, 1948.

[19] P. Salovey and J. Mayer, "Emotional intelligence," *Imagination, cognition and personality*, vol. 9, no. 3, pp. 185–211, 1990.

[20] R. Bellman, *Dynamic Programming*. NJ: Princeton Univ. Press, 1975.

[21] T. Lee and D. Mumford, "Hierarchical bayesian inference in the visual cortex," *JOSA A*, vol. 20, no. 7, pp. 1434–1448, 2003.

[22] T. Lee, D. Mumford, R. Romero, and V. Lamme, "The role of the primary visual cortex in higher level vision," *Vision research*, vol. 38, no. 15, pp. 2429–2454, 1998.

[23] K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1724–1734.

[24] B. Babcock, M. Datar, and R. Motwani, "Sampling from a moving window over streaming data," in *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2002, pp. 633–634.

[25] R. C. Oldfield, "The assessment and analysis of handedness: the edinburgh inventory," *Neuropsychologia*, vol. 9, no. 1, pp. 97–113, 1971.

[26] D. Bernhardt, "Emotion inference from human body motion," Ph.D. dissertation, University of Cambridge, 2010.

[27] K. Scherer and T. Bänziger, "On the use of actor portrayals in research on emotional expression," *Blueprint for affective computing: A sourcebook*, pp. 166–176, 2010.