Body Gestures and Spoken Sentences: a Novel Approach for Revealing User's Emotions

Vito Gentile, Fabrizio Milazzo,
Salvatore Sorce, Antonio Gentile
Università degli Studi di Palermo
Dipartimento dell'Innovazione Industriale e Digitale (DIID)
Viale delle Scienze, Building 6, Palermo, Italy
Email: {firstname.lastname}@unipa.it

Agnese Augello, Giovanni Pilato
Consiglio Nazionale delle Ricerche
Istituto di Calcolo ad Alte Prestazioni (CNR-ICAR)
Via Ugo La Malfa 153, Palermo, Italy
Email: augello@pa.icar.cnr.it, giovanni.pilato@cnr.it

***Abstract*—In the last decade, there has been a growing interest in emotion analysis research, which has been applied in several areas of computer science. Many authors have contributed to the development of emotion recognition algorithms, considering textual or non verbal data as input, such as facial expressions, gestures or, in the case of multi-modal emotion recognition, a combination of them. In this paper, we describe a method to detect emotions from gestures using the skeletal data obtained from Kinect-like devices as input, as well as a textual description of their meaning. The experimental results show that the correlation existing between body movements and spoken user sentence(s) can be used to reveal user's emotions from gestures.**

***Keywords*-Emotion Recognition, Gesture Recognition, Sentiment Analysis**

I. Introduction and Background

Nowadays, there is a growing interest in the field of emotion recognition, which can be defined as the process of identifying *human emotions* by different modes of expressions. According to the popular categorization proposed by Ekman [1], human emotions are basically six, namely *anger, disgust, fear, joy, sadness* and *surprise*.

Indeed, many computer scientists and researchers have significantly contributed to the development of innovative algorithms of sentiment and emotion analysis [2]. Emotion recognition has a practical use in several areas, such as social media marketing and analysis [3], human-robot interaction [4] and automobile safety systems [5]. Until now, the algorithms devoted for such specific task have mainly focused on the analysis of speech and facial expressions. Nevertheless, there is a notable evidence that affective states are largely communicated also through body movements and gestures, in some cases better than the other communication channels [6]. The interest towards this modality of expression is now growing [7], and several works have recently investigated the motion cues that mainly convey affective information [8].

Multi-modal emotion recognition tries to combine different communication channels simultaneously to improve the recognition performance. There are several works which implement such an approach, and have shown that the combination of even just two modalities of expression increases significantly the performance of the related emotion classifier [9], [10], [11], [12]. In particular, speech and gesture channels are internally coordinated towards conveying communicative intentions [13] and make a unified meaning with the verbal part of an utterance [14]. As a consequence, there should exist a sort of "correlation" between gestures performed by a user and what s/he feels, thinks and says while moving his/her body: if a gesture and a sentence are simultaneous, then they can be considered somehow mutually correlated [15]. Assuming that this correlation exists, then it is questionable whether there is a correlation between an emotion recognized from a sentence spoken while performing a gesture (via some textual emotion recognition algorithm), and the emotion recognized using only that gesture as input.

Starting from this assumption, we propose a system for recognizing emotions from body gestures. At this purpose, Kinect-like data (i.e. RGB-D videos, audio and joint sequences [16]) can be used for extracting significant features for gesture recognition [17]. For the purposes of this paper, we evaluated some available datasets for emotion recognition [18], [19] and [20] and found that several of them provide data taken from a Kinect-like device. In particular, the one proposed in *Chalearn multimodal gesture recognition dataset* [21] also offered gesture information along with their audio/text description in Italian and for this reason has been used to assess the performance of the proposed system.

We assume that the gestures are associated to a textual description (i.e. a *sentence* representing the gesture meaning), spoken by the user while s/he is performing the gesture. Then, our system associates emotion labels to body gestures by applying a sentiment analysis algorithm to their associated sentence. After the labeling phase, new (unseen) body gestures can be labeled with an emotion by analyzing only their skeletal data.

978-1-5090-4284-5/17 $31.00 © 2017 IEEE
DOI 10.1109/ICSC.2017.14

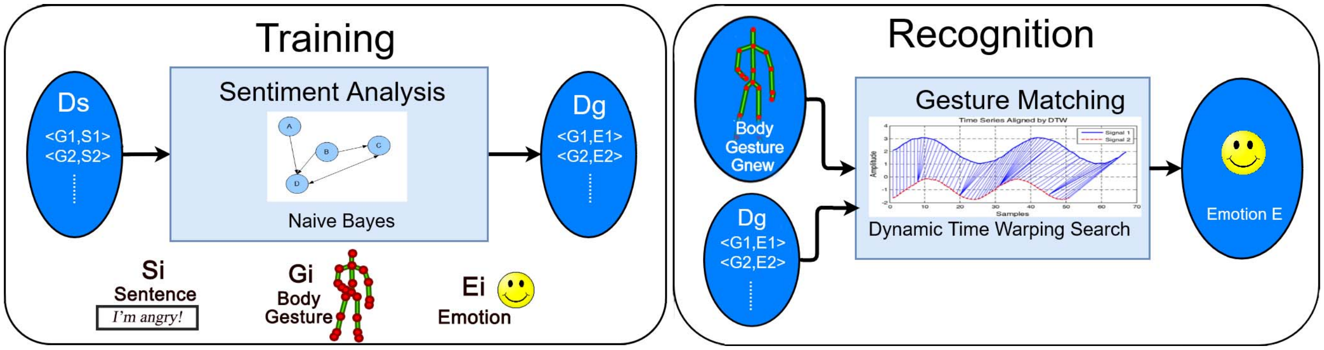

Figure 1. The block scheme of the proposed system.

II. System Description

Our emotion recognition system is based on body gestures as data input, and uses a mapping which exploits the existing knowledge about emotion recognition from textual data. Figure 1 reports the main building blocks of the proposed system and depicts the internal logical information flow. The system is composed of two modules, namely *Training* and *Recognition*.

In the design of the Training module, we assume the availability of a *Gesture-Sentence* input dataset, namely D_s. Each of its records has the form $< G_i, S_i >$, i.e. a pair which links a gesture G_i (i.e. a temporal skeletal joints sequence) to a sentence S_i (i.e. the text spoken by the performing user). Its output is a *Gesture-Emotion* dataset D_g, which contains gesture-emotion pairs in the form $< G_i, E_i >$, where a gesture G_i is associated to the emotion E_i.

The Recognition module accepts a new (unseen) gesture G_{new} as well as the background knowledge D_g, and outputs the emotion E derived from the gesture G_{new}.

The following subsections describe the implementation details of the two system modules.

A. Training

The Training module is based on the assumption that the meaning of a gesture G is expressed by a sentence S. In order to compute the mapping between text and emotion, we have used a lexical-semantic approach based on sentiment analysis, according to the belief that it is possible to infer emotion properties from the *emotion words* [22].

We used a Naive Bayes classifier trained on the emotions lexicon obtained from the Word-Net Affect Lexicon. The lexicon associates a *synset* of WordNet to a set of affective labels in order to mark the specific synsets as representatives of an affective concept. The lexicon is composed by a set of emotional words defined in the Semeval 2007 context [23].

Let us define the variable $x \in \{$*anger, disgust, fear, joy, sadness, surprise*$\}$ indicating a possible emotion label.

Each sentence S is thus mapped into a vector $\vec{V}(x)$, which belongs to an emotional space whose dimensions are the six basic emotions given by Ekman. In other words, each sentence is coded as an *emonoxel*, analogously as the *knoxel* in the conceptual space paradigm [24].

The components of the generic emonoxel $\vec{V}$ are computed from the Naive Bayes as the probability to have the sentence S given a certain emotion:

$$\vec{V}(x) = p(S|Emotion = x) \tag{1}$$

From the emonoxel $\vec{V}$, the largest component E is selected, which represents the predominant emotion expressed by sentence S:

$$E = \arg\max_{x} \{\vec{V}(x)\} \tag{2}$$

As a consequence, being a gesture paired with a sentence, it is possible to consider an indirect mapping between a gesture and an emotion, through the sub-symbolic coding of the sentence. Starting from this assumption, the system generates a gesture-emotion dataset D_g as described in Algorithm 1.

Algorithm 1 Gesture-Emotion dataset generation algorithm.

```
Define a new empty database $D_g$
**foreach** $< G_i, S_i >$ **in** $D_s$
  compute the Emonoxel $\vec{V_i}$ using Equation (1)
  compute the Emotion $E_i$ using Equation (2)
  add a new entry $< G_i, E_i >$ to $D_g$.
```

B. Gesture Recognition

The most adopted mathematical frameworks which implement *Gesture Matching* are Hidden Markov Models (HMMs) and Dynamic Time Warping (DTW). We chose DTW because it has been demonstrated to be more appropriate for applications involving gesture recognition [25].

DTW is a method for computing the distance between two differently sized sequences. From a geometric point of view, it aligns the sequences by making use of some simple operations accounted in the computation of the distance. Let us define the following quantities:

- G_{new}: a $N \times T$ input gesture, where N is the number of components of each frame, and T is the number of frames;

- G: a $N \times W$ gesture belonging to D_g, where W is the number of frames;
- DTW: a $T \times W$ distance matrix, where $DTW_{T,W}$ will represent the distance between G_{new} and G;
- $d(\cdot, \cdot)$: a distance metric whose input are two $N \times 1$ gesture frames.

Then, each element $DTW_{i,j}$ is computed as in the Algorithm 2.

If we think to the Recognition module as a black-box F, then it computes $E = F(G_{new}, D_g)$, i.e. it outputs the emotion E to be associated to the incoming gesture G_{new} by using the background knowledge D_g.

Figure 1 instead describes the module as a white box: the inputs D_g and G_{new} are processed by the Gesture Matching block, which computes Dynamic Time Warping distance between G_{new} and each gesture G contained in the records of D_g. Then it applies a simple nearest neighbor classifier to compute the closest gesture G^* as:

$$G^* = \arg\min_G \{DTW(G_{new}, G)\}, \forall G \in D_g. \qquad (3)$$

The related output emotion E^* is retrieved from the record $< G^*, E^* >\in D_g$.

Algorithm 2 Dynamic Time Warping matrix computation.

```
initialize first row and column of $DTW$ to $\infty$
$DTW_{1,1} \leftarrow 0$
for $i \leftarrow 1$ to $T$
    $G_{new,i} \leftarrow$ the $i$-th column of $G_{new}$
    for $j \leftarrow 1$ to $W$
        $G_j \leftarrow$ the $j$-th column of $G$
        $cost \leftarrow d(G_{new,i}, G_j)$
        $m \leftarrow min(DTW_{i-1,j}, DTW_{i,j-1}, DTW_{i-1,j-1})$
        $DTW_{i,j} \leftarrow cost + m$
$Distance \leftarrow DTW_{T,W}$
```

III. Experimental assessment

We used the database described in [21], made of 7663 gestures, represented by RGB videos, depth information, skeletal frames (each one represented by a set of (x, y, z) coordinates of 20 joints, i.e. a 20×3 matrix), and a textual description of the gesture meaning.

The Training module was run to extract emotions from the textual description of each gesture. These emotions were assumed as ground-truth for the rest of the experiment. We thus generated a gesture-emotion dataset D_g which counted 1153 samples for *anger*, 1155 for *disgust*, 1142 for *fear*, 1539 for *joy*, 767 for *sadness*, and 1907 for *surprise*.

After that, we tested the performance of the Recognition module: we created different training sets by using different percentages of the whole dataset (between 50% and 95% with 5% steps, with the addition of 99%), and used the remaining samples as test set. We resampled the training set 350 times for each split percentage, in order to get more reliable results. Moreover, we have run Dynamic Time Warping using three different metrics $d(\cdot, \cdot)$ which were: i) *Manhattan*, ii) *Cosine* and iii) *Euclidean* distances. For implementation needs, we linearized each 20×3 frame matrix into a $N = 60$-dimensional vector.

For each run, we computed the 6-by-6 confusion matrix, whose rows represent true emotions, and columns represent the output of the recognition process. Thus, we were able to compute the accuracy, precision and recall metrics of the recognition module [26]. Figure 2 shows the results obtained by averaging the performance metrics over the 350 runs for each training percentage. The Recognition module performs quite well in all the cases, although it seems that Euclidean distance achieves better results. This is probably due to the fact that it represents the shortest path between two points in space while Manhattan is only a rough approximation. As regards the Cosine distance, its performance is very near to the Euclidean distance and may be more convenient to be used in a real-time context as it requires less computational effort (provided that the vector representing each gesture frame is normalized). The computed metrics ranged approximately between 73% and 78% for Manhattan distance, and between 75% and 81% for Cosine and Euclidean. It is also quite interesting noting that as the number of samples used in the training set increases, then the related performance increases as well.

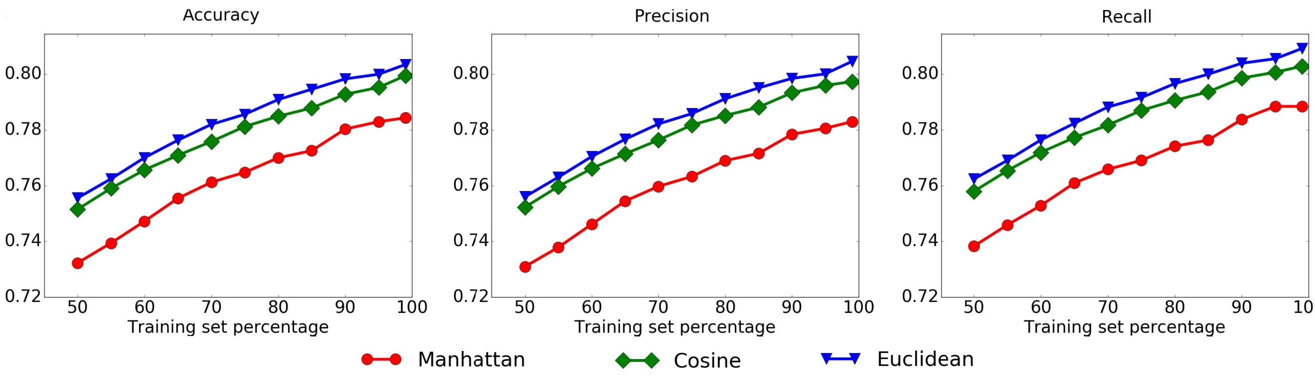

Figure 2. Recognition accuracy (a), precision (b) and recall (c) using Manhattan, Cosine and Euclidean distances.

This is not surprising, as a greater number of samples in the training set increases the chances to match the correct gesture for DTW. From a qualitative point of view, we think that performance metrics values are limited by the nearest neighbor classifier which is very sensitive to noise. A better solution would use a training set made up of very few examples (prototypes) selected by using a suitable clustering algorithm: this would certainly reduce the noise and lower the computational burden (due to the comparisons needed by DTW), thus increasing the possibilities for our system to be used in a real-time context.

IV. Conclusion

In this paper we described a system for mapping body gestures to emotions. If gestures are associated to spoken sentences (which can be mapped into emotions), then it is possible to recognize emotions directly from gestures. The emotion recognition task have shown good performance despite it is carried out by a simple one nearest-neighbor classifier. We are thus planning to improve our system by using more sophisticated algorithms. This should lower misclassifications due to the high sensitivity to noise of the proposed classifier. Finally, we would be interested in testing our system by including facial expressions and vocal prosody data, testing its real-time capabilities and check how it generalizes to different datasets.

References

[1] P. Ekman, *Basic Emotions*. Wiley & Sons, 2005, pp. 45–60.

[2] L. Canales and P. Martínez-Barco, "Emotion Detection from Text: A Survey," *Processing in the 5th Information Systems Research Working Days (JISIC 2014)*, p. 37, 2014.

[3] E. D'Avanzo and G. Pilato, "Mining Social Network Users Opinions to Aid Buyers Shopping Decisions," *Computers in Human Behavior*, vol. 51, pp. 1284–1294, 2015.

[4] K.-H. Park, H.-E. Lee, Y. Kim, and Z. Z. Bien, "A Steward Robot for Human-Friendly Human-Machine Interaction in a Smart House Environment," *IEEE Transactions on Automation Science and Engineering*, vol. 5, no. 1, pp. 21–25, 2008.

[5] C. Nass, I.-M. Jonsson, H. Harris, B. Reaves, J. Endo, S. Brave, and L. Takayama, "Improving Automotive Safety by Pairing Driver Emotion and Car Voice Emotion," in *CHI'05 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2005, pp. 1973–1976.

[6] A. Kleinsmith and N. Bianchi-Berthouze, "Affective Body Expression Perception and Recognition: A Survey," *IEEE Trans. on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2013.

[7] V. Gentile, A. Malizia, S. Sorce, and A. Gentile, "Designing Touchless Gestural Interactions for Public Displays In-the-Wild," in *International Conference on Human-Computer Interaction*. Springer, 2015, pp. 24–34.

[8] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer, "Technique for Automatic Emotion Recognition by Body Gesture Analysis," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'08)*. IEEE, 2008, pp. 1–6.

[9] G. Pilato and U. Maniscalco, "Soft Sensors for Social Sensing in Cultural Heritage," in *2015 Digital Heritage*, vol. 2. IEEE, 2015, pp. 749–750.

[10] L. Kessous, G. Castellano, and G. Caridakis, "Multimodal Emotion Recognition in Speech-based Interaction Using Facial Expression, Body Gesture and Acoustic Analysis," *J. on Multimodal User Interfaces*, vol. 3, no. 1-2, pp. 33–48, 2010.

[11] M. Ditta, F. Milazzo, V. Ravì, G. Pilato, and A. Augello, "Data-driven relation discovery from unstructured texts," in *Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on*, vol. 1. SCITEPRESS, 2015, pp. 597–602.

[12] S. Sorce, A. Augello, A. Santangelo, G. Pilato, A. Gentile, A. Genco, and S. Gaglio, "A multimodal guide for the augmented campus," in *Proceedings of the 35th annual ACM SIGUCCS fall conference*. ACM, 2007, pp. 325–331.

[13] Z. Yang and S. S. Narayanan, "Analysis of Emotional Effect on Speech-Body Gesture Interplay." in *INTERSPEECH*, 2014, pp. 1934–1938.

[14] A. Özyürek, "Hearing and Seeing Meaning in Speech and Gesture: Insights from Brain and Behaviour," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 369, no. 1651, p. 20130296, 2014.

[15] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago press, 1992.

[16] V. Gentile, S. Sorce, and A. Gentile, "Continuous Hand Openness Detection Using a Kinect-like Device," in *Eighth International Conference on Complex, Intelligent and Software Intensive Systems (CISIS)*. IEEE, 2014, pp. 553–557.

[17] S. Sorce, V. Gentile, and A. Gentile, "Real-Time Hand Pose Recognition Based on a Neural Network Using Microsoft Kinect," in *Eighth International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA)*. IEEE, 2013, pp. 344–350.

[18] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: A Comprehensive Multimodal Human Action Database," in *IEEE Workshop on Applications of Computer Vision (WACV)*. IEEE, 2013, pp. 53–60.

[19] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing People for Training Gestural Interactive Systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 1737–1746.

[20] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A Multimodal Dataset for Human Action Recognition Utilizing a Depth Camera and a Wearable Inertial Sensor," in *Int. Conf. on Image Processing (ICIP)*. IEEE, 2015, pp. 168–172.

[21] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, and H. Escalante, "Multi-Modal Gesture Recognition Challenge 2013: Dataset and Results," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 445–452.

[22] C. Strapparava and A. Valitutti, "WordNet Affect: an Affective Extension of WordNet," in *LREC*, vol. 4, 2004, pp. 1083–1086.

[23] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: Affective text," in *Proceedings of the 4th International Workshop on Semantic Evaluations*. Association for Computational Linguistics, 2007, pp. 70–74.

[24] P. Gärdenfors, *Conceptual Spaces: The geometry of Thought*. MIT press, 2004.

[25] J. M. Carmona and J. Climent, "A Performance Evaluation of HMM and DTW for Gesture Recognition," in *Iberoamerican Congr. on Pattern Recognition*. Springer, 2012, pp. 236–243.

[26] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.