*Multimodal Technologies and Interaction*

MDPI

*Article*

# Detecting Emotions from Illustrator Gestures—The Italian Case

**Daniele Fundarò [1], Vito Gentile [1], Fabrizio Milazzo [2] and Salvatore Sorce [3,*]**

[1] synbrAIn srl, 90128 Palermo, Italy; daniele.fundaro@synbrain.ai (D.F.); vito.gentile@synbrain.ai (V.G.)
[2] Dipartimento di Ingegneria, Università degli Studi di Palermo, 90128 Palermo, Italy; fabrizio.milazzo@unipa.it
[3] Facoltà di Ingegneria e Architettura, Università degli Studi di Enna "Kore", 94100 Enna, Italy
* Correspondence: salvatore.sorce@unikore.it

**Abstract:** The evolution of computers in recent years has given a strong boost to research techniques aimed at improving human–machine interaction. These techniques tend to simulate the dynamics of the human–human interaction process, which is based on our innate ability to understand the emotions of other humans. In this work, we present the design of a classifier to recognize the emotions expressed by human beings, and we discuss the results of its testing in a culture-specific case study. The classifier relies exclusively on the gestures people perform, without the need to access additional information, such as facial expressions, the tone of a voice, or the words spoken. The specific purpose is to test whether a computer can correctly recognize emotions starting only from gestures. More generally, it is intended to allow interactive systems to be able to automatically change their behaviour based on the recognized mood, such as adapting the information contents proposed or the flow of interaction, in analogy to what normally happens in the interaction between humans. The document first introduces the operating context, giving an overview of the recognition of emotions and the approach used. Subsequently, the relevant bibliography is described and analysed, highlighting the strengths of the proposed solution. The document continues with a description of the design and implementation of the classifier and of the study we carried out to validate it. The paper ends with a discussion of the results and a short overview of possible implications.

**Keywords:** emotion recognition; gesture recognition; illustrator gestures; RNN



**Citation:** Fundarò, D.; Gentile, V.; Milazzo, F.; Sorce, S. Detecting Emotions from Illustrator Gestures—The Italian Case. *Multimodal Technol. Interact.* **2022**, *6*, 56. https://doi.org/10.3390/mti6070056

Academic Editors: Cristina Portalés Ricart, Ester Alba Pagán, Jorge Sebastián Lozano, Valeria Seidita, Marcos Fernández Marín, Maurizio Vitella, Georgia Lo Cicero and María del Mar Gaitán Salvatella

Received: 31 May 2022
Accepted: 14 July 2022
Published: 17 July 2022



## 1. Introduction

The evolution of computers seen in the last few years has given a boost to research techniques aimed at easing human–computer interaction. Such techniques tend to mimic the dynamics of the human–human interaction process, which is, in turn, based on our innate ability to understand emotions from other humans [1].

Automatic Emotion Recognition is widely adopted today, and many actual applications have demonstrated its effectiveness in improving human–computer interaction. For example, novel multimedia user-centric applications are becoming able to reveal affective behaviours from users. As a result of this ability, the information provided by the extraction, analysis, and encoding of user preferences can be enriched by their affective states [2]. Affective avatars have been successfully used to motivate elderly citizens to remain active in their life at home [3]. Lie detection systems may be used to help judges in legal trials [4]. Online games, which allow for interaction with avatars, pets, or similar things, may adapt the behaviour of the virtual elements to the feelings of the human user. E-tutoring systems may exploit emotion detection to tune timing and content [5].

The past research in Emotion Recognition (ER) has focused only on the cues conveyed by facial expressions. This was due to the earliest psychological studies of Ekman [6], who noted that people look primarily at facial expressions to understand other people's emotions. Further studies widened this vision and proved that the human ER process is

multi-modal and involves complex evaluations of body/hand gestures, speech, intonations, and facial expressions, which are mutually correlated [7,8].

The performance of Multi-modal ER algorithms depend on the capabilities of the underlying unimodal algorithms. In particular, the impact of body movements is comparable or even more decisive than the other channels [9–11].

Focusing on this path, many studies attempted to discover the body motion cues which drive humans in the perception of emotions [12]. The state-of-the-art approaches use low-level kinematic cues provided by 2D/3D motion data, such as motion energy, hands distance, etc. [13,14]. Psychology claimed that temporal dynamics cues are likewise important in human perceptions of emotions [15]. However, there is not a firm agreement in the scientific community about the most relevant cues for discriminating emotions by body gestures, which, in turn, Ekman et al. identified into five categories [16]:

1. Manipulators: One part of the body manipulates another part of the body or an object.
2. Regulators: Regulate the flow of the conversation (e.g., "continue" or "repeat") among humans.
3. Emotional: Involuntary movement, which signals that an emotion is occurring in the performer.
4. Emblems: Replace a word in speech (e.g., the "victory sign").
5. Illustrators: Tightly related to the performer's culture, this movement illustrates speech.

Gestures belonging to categories 1–4 have a well-specified and universally accepted meaning (usually shared across different cultures). It seems that the human ability to recognize emotions from such categories is not influenced by cultural differences. In contrast, Illustrators are used to represent a visual image of a spoken sentence, and, even worse, the same gesture may illustrate different sentences across different cultures.

The question then arises: "Can emotions be recognized from illustrator gestures by using only the body motion cues?". The answer is probably "no", and this can be clarified by an example.

Figure 1 represents an illustrator gesture, which means: "be patient" in Egypt, "What do you mean?" in Italy, and "that's perfect" in Greece [17]. The Egyptian and Greek meanings lead, almost surely, to a positive performer's affective state, whereas, in Italy, the same gesture is usually associated with a negative state. More generally, the performed movement is a "consequence" of the sentence it is trying to illustrate, and the container of the emotion is the sentence. It means that body cues associated with illustrators are culturally dependent, and the same body movement may express contrasting emotions.


**Figure 1.** A culturally dependent illustrator gesture.

Recognizing emotions from this category necessarily requires coupling the gesture meaning to the motion data. For this reason, approaches based solely on motion cues cannot work well with this category of gestures. This claim is proved in the experimental results in Section 6.

The purpose of this work is to demonstrate that good performance of ER from illustrator gestures can be achieved by using a unimodal approach, which uses the motion data that are coupled, in some way, to the gesture meaning.

We tried to translate the ER problem into an equivalent Gesture Recognition problem. For any given culture, the sentence associated with an illustrator gesture is fixed and unambiguous [17]. Therefore, it holds that the association between the body movement and the intended emotion is one-to-one. It follows that recognizing the performed movement leads to recognizing the performer's emotion.

The research findings of this work are related to the Italian illustrator gestures. The proposed methodology does not focus on a specific culture and may be applied to other illustrator gestures of different cultures. We do not intend to replace existing methods in emotion recognition from body gestures; rather, we aim to provide a complementary approach. A hypothetical algorithm which intends to switch between a state-of-the-art method and the one proposed here should only be able to detect that an illustrator gesture is going to be performed. Such an issue should be easy to solve, as illustrators are the only gesture category performed together with a spoken sentence.

The contributions of this research can be summarised in the following points:

1. An emotion elicitation study aimed at discovering emotions expressed by the Italian illustrator gestures (the reference dataset is described in the work [18]).
2. An emotion classifier trained and tested on the elicited dataset.
3. A series of optimisation techniques aimed at minimising time complexity and maximising the recognition performance of the classifier.

The rest of the paper is organized as follows: Section 2 revises modern motion-cue-based approaches for emotion recognition as well as some gesture recognition techniques; Section 3 defines the mathematics needed to describe the methodology of Section 4; Section 5 deeply analyses the results of the elicitation study as well as the performance of the proposed classifier; and Section 6 contains the conclusions of this work.

## 2. Related Works

Several studies investigated how body gestures can be categorised. The work of Kendon et al. [19] ranked gestures based on formality: gesticulations, language-like gestures, pantomiming, emblems, and sign language. The work of Karam et al. [20] reviewed the above taxonomy and provided a five-element categorisation: deictic gestures, gesticulations, manipulation, semaphores, and sign language. A similar work by Ekman [16] stated that human gestures are as follows:

- Manipulators: One part of the body manipulates another part of the body or an object.
- Regulators: Regulate the flow of the conversation (e.g., "continue" or "repeat") among humans.
- Emotional: Involuntary movement, which signals that an emotion is occurring in the performer.
- Emblems: Replace a word in speech (e.g., the "victory sign").
- Illustrators: Tightly related to the performer's culture, this movement illustrates speech.

An important claim by Ekman is that:

**Claim 1:** *Manipulators, regulators, emotional, and (for the most part) emblem gestures are universally understood. On the other hand, illustrators are intimately related to the speaker's speech [...] usually augmenting what is said* [16].

This confirms our intuition that recognizing emotions from illustrator gestures requires (in some way) considering the gesture's meaning. Emotions have been widely studied in psychology. A very basic bi-dimensional model (circumplex) was proposed by Russell [21]. Such a model classifies emotions according to the intensity and pleasure dimension. The work of Ekman on "Basic Emotions" [22] argued about the two dimensions of Russell and said that:

**Claim 2:** *Humans express six basic innate emotions, namely: anger, disgust, fear, joy, sadness, and surprise.*

Based on his cross-cultural studies of facial expressions, Ekman understood that the circumplex model is not able to describe other facets of the six basic emotions. Such observations led Parrott [23] to a categorisation that included more than 100 emotions, arranged in a three-layered poly-tree. The roots of the poly-tree are exactly the six basic emotions proposed by Ekman.

The six basic emotions are widely recognized and accepted, especially from the body gestures perspective [24]. The substantial agreement of the community led us to choose the Ekman categorisation as the basis of our emotion classifier for illustrator gestures.

### *2.1. Motion-Cue-Based Approaches for ER*

Some recent studies discussed the "universality" of body language in the expression of basic emotions and how emotions arise despite cultural factors [25]. On the other hand, there is strong evidence that emotions are controlled and expressed (by the body, face, voice, etc.) by psychological processes shaped by the performer's culture [26].

The work of Elfenbein et al. [27] tried to find a link between such two contrasting points of view. They said that:

**Claim 3:** *Emotions are expressed by a universal language, but different "dialects" arose across different cultures.*

A confirmation of such a claim was given by Hess et al. [28], which demonstrated that:

**Claim 4:** *ER is more accurate among people of the same cultural group.*

The universality of body language has led many researchers to study what motion cues influence accuracy in emotion recognition, both in humans and automatic systems. Almost all the recent works in the literature tend to present the same canvas: a (large) set of features are extracted from the 3D motion data; features are ranked according to their influence on the ER accuracy; and finally, a classifier is trained on feature vectors extracted from emotion-labelled data.

Glowinski et al., in [13], proposed a set of 25 expressive and dynamic features for emotion recognition from body gestures. The expressive features were the energy, spatial extent, symmetry, smoothness, and forward–backward leaning of the head; the dynamic features regarded some statistics such as the maximum, minimum, mean, peaks number, and duration (number of frames) of the expressive features. A quite similar approach was described in [29], which proposed a total of 87 features to annotate human gestures. They were divided according to four dimensions: the body (the relative positions of the skeletal joints), effort (the velocity and acceleration of the joints), shape (the way that the body changes during movement), and space (describes the movement in relation with the environment). Similar approaches have been proved to be effective in [30] and [31].

The main problem that we found in these works is the lack of clear categorisation of the analysed movements. Based on claims 3 and 4 of Elfenbein and Hess, we can deduce that a universal set of motion cues for emotion recognition may work only for culturally independent gestures.

The universality of body language arose from humans' innate factors [27]. However, illustrator gestures are the product of the specific socio-cultural context where people live. Indeed, as claimed by Ekman:

**Claim 5:** *Illustrators are learned when the language itself is learned [...]. Sicilian immigrants use very different types of illustrators, but these differences were not preserved in their offspring who assimilated into the mainstream New York City culture* [16].

This proves that ER for illustrator gestures must consider the gestures meaning. Here, we chose to face the problem by means of a gesture recognition methodology which merges both meaning and motion data.

### *2.2. Gesture Recognition for ER*

Gesture recognition aims to recognize human movements, involving the face, head, and body. Earlier works performed gesture recognition by means of RGB [32] and depth [33] cameras. The diffusion of Kinect-like devices [34], i.e., an integrated solution for RGB, depth, and skeletal information, has contributed to the development of more accurate algorithms [35].

Today, gestural motion data are modelled as temporal sequences of skeletal joints. Many mathematical frameworks have demonstrated their effectiveness over the years. Among them, the most popular are: dynamic time warping (DTW), Hidden Markov Models (HMM), and time delay neural networks (TDNNs). The authors of [36] proposed the use of a Gaussian Mixture Hidden Markov Model. Interesting research findings have been found in [37], which demonstrated that Dynamic Time Warping requires less training examples and is more accurate than Hidden Markov Models. Time delay neural networks (TDNNs) have been proved to be effective for the recognition of American Sign Language in [38].

How to choose the best mathematical model that fits best for an emotion classifier is quite a hard task. It involves long development times and endless validation sessions. The aim of this work was to develop a classifier that is able to maximise accuracy, to minimise the time effort of ER, and to check if the classifier is more capable than humans in ER.

First, based on [38], we deduced that Dynamic Time Warping is more accurate than HMMs and requires less training examples, i.e., less time effort. Moreover, [37] demonstrated that TDNNs are very effective in gesture recognition, but they do not allow incremental training.

We considered the properties of the problem and opted for a TDNN classifier.

In Section 6, we explain how to design and build the neural network.

## 3. Background and Definitions

In this section, we provide useful definitions that are needed to describe the proposed methodology for emotion recognition from illustrator gestures.

The basic assumption of our methodology is the existence of an original dataset of the gesture motion data paired with the illustrated sentence. We assume that motion data are acquired by a Kinect-like device that is able to provide the skeletal joint positions of the gesture performer.

**Definition 1.** *The original dataset is named O and contains N elements. Each element $O_i$ is a triple containing the RGB video V, the skeletal data of the gesture G, and the spoken sentence M:*

$$O = (V_1, G_1, M_1); \ldots; (V_N, G_N, M_N)$$

**Definition 2.** *Each gesture G is a temporal sequence lasting N time instances. For any fixed time instance t, G(t) is a static body posture of K skeletal joints represented by points in the 3D space:*

$$G(t) = ((x_1, y_1, z_1); \ldots; (x_K, y_K, z_K))$$

**Definition 3.** *Gesture G is a matrix:*

$$G = \begin{pmatrix} (x_{1,1}, y_{1,1}, z_{1,1}) & \cdots & (x_{1,K}, y_{1,K}, z_{1,K}) \\ \vdots & \ddots & \vdots \\ (x_{N,1}, y_{N,1}, z_{N,1}) & \cdots & (x_{N,K}, y_{N,K}, z_{N,K}) \end{pmatrix}$$

Gestures may be performed at different origins. In the following, we make them refer to the position of their very first joint.

**Definition 4.** *The 3D position of the very first joint of a gesture G is:*

$$J_{1,1} = (x_{1,1}, y_{1,1}, z_{1,1})$$

A normalized gesture is such that its origin is $J_{1,1}$:

$$\widetilde{G} = \begin{pmatrix} (0,0,0) & \cdots & (x_{1,K}, y_{1,K}, z_{1,K}) - J_{1,1} \\ \vdots & \ddots & \vdots \\ (x_{N,1}, y_{N,1}, z_{N,1}) - J_{1,1} & \cdots & (x_{N,K}, y_{N,K}, z_{N,K}) - J_{1,1} \end{pmatrix}$$

Knowledge of the motion and the spoken sentence is enough to determine the emotion to be associated with an illustrator. Emotions from the original dataset can be elicited (such steps require human intervention); therefore, a new gestural dataset S, containing motion data and emotion labels, can be produced. Sentences are no longer retained, as their informative content is subsumed by the knowledge of the emotion label.

**Definition 5.** *The gesture–emotion dataset is named S and contains N elements. Each element $S_i$ is a pair containing the normalized gesture skeletal data and the labelled emotion E:*

$$S = \left(\widetilde{G}_1, E_1); \ldots; \widetilde{G}_N, E_N\right)$$

**Definition 6.** *Emotions assume values in a set R, defined as follows* [22]:

$$R = \{anger,\ disgust,\ fear,\ joy,\ sadness,\ surprise\}$$

## 4. Methodology

The proposed methodology to set up a classifier to recognize emotions from illustrator gestures is depicted in Figure 2.

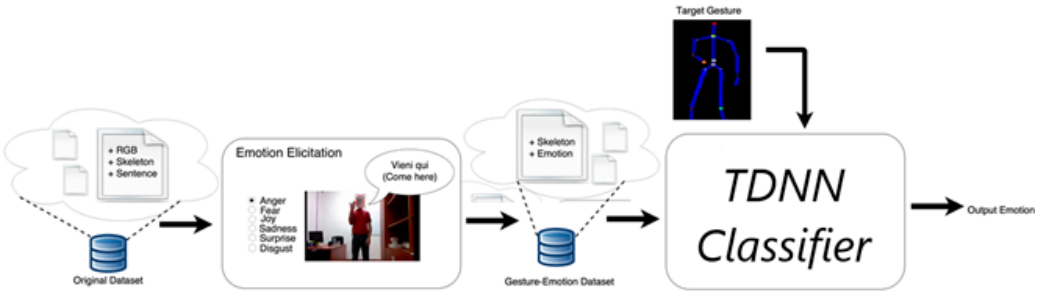



**Figure 2.** High-level representation of the proposed methodology.

The original dataset, denoted as O, contains the associations between the RGB video, the skeletal data, and the spoken sentence of each illustrator.

In order to obtain a dataset of gestures labelled with emotions, we set up an emotion elicitation study in which users were asked to see the videos of the original dataset O and to label the corresponding gesture with an emotion. The output of this study was the gesture–emotion dataset, denoted as S.

The new dataset S is then used to train, validate, and test the classifier, which, at the end, provides estimations of emotions for illustrator gestures that have never been processed before (called "target").

The following sections describe in more detail the emotion elicitation study and the design and testing of the classifier.

## 5. Emotion Elicitation Study for the Italian Illustrator Gestures

In order to build an NN for emotion recognition from gestures, we needed to have a gesture dataset labelled with the corresponding emotion. As a consequence, first, we had to build a suitable dataset for the training, validation, and testing of the NN.

To this end, as a side-step of this work, we designed an emotion elicitation study starting from a well-known gestural dataset of Italian illustrator gestures. Such a dataset was built in the "Chalearn 2013 competition", described in [18]. The download is available at http://sunai.uoc.edu/chalearn/ accessed on 13 July 2022.

The dataset contains 393 sessions corresponding to 7754 Italian illustrator gestures, each belonging to one among 20 categories (Figure 3), which are the meanings of the gestures. The gestures are performed by 27 different users in front of a Kinect (version 1) while the related sentence is spoken in a controlled environment. Sessions are annotated with the starting and ending frame of each gesture and the corresponding meaning.


**Figure 3.** Italian illustrator gestures and their meaning (Italian/English) used for emotion elicitation: (**a**) Basta/Stop it. (**b**) Buonissimo/Really good. (**c**) Che due palle!/What a bore! (**d**) Che hai combinato?/What have you done? (**e**) Che vuoi?/What do you want? (**f**) Cosa ti farei/What I would do with you. (**g**) È un furbo/He's an old fox. (**h**) Ho fame/I am hungry. (**i**) Le vuoi prendere?/Do you want to be beaten? (**j**) Sono messi d'accordo/They are conspired. (**k**) Non ce n'è più/There is no more. (**l**) Non me ne frega niente/I do not care. (**m**) Ok. (**n**) Perfetto/Perfect. (**o**) Sei pazzo?/Are you crazy? (**p**) Sono stufo/I am tired. (**q**) Tanto tempo fa/A long time ago. (**r**) Vanno d'accordo/They get along. (**s**) Vattene!/Go away! (**t**) Vieni qui/Come here.

Each session (sampled at 20 FPS) is composed as follows:

- RGB videos recorded at 8 bits, VGA resolution (640 × 480).
- Depth videos recorded at 11 bits, VGA resolution (640 × 480).
- Audio track in Italian, recorded with a Kinect microphone.
- Skeletal data containing 3D coordinates of 20 joints.
- Starting to ending frames of each gesture subsequence.

As a preliminary step, we processed the Chalearn gestural dataset (training version) and extracted 7754 gestural subsequences by using indications about the starting and ending frame. From such subsequences, we kept only the RGB and skeletal data and created the entries of the original dataset, which we named O.

Emotion elicitation was performed according to standard procedures, as depicted in [39]. In more detail, we first prepared an online survey which we spread through our personal and social contacts. The survey contained the following questions:

1. Indicate your gender.
2. Indicate your nationality.
3. Indicate your age.
4. (Only for Italian people) Indicate your region.
5. See the following videos of a user performing a gesture while a sentence is spoken. After that, please indicate the emotion you think the performer intends to express.

The first three questions were mainly intended to collect demographic information.

The fourth question aimed at checking the statistical representativeness of the sample compared to the Italian population distribution among regions.

In the fifth question, each user was asked to view different videos among the available 7754, randomly extracted from the original dataset O, and to choose an emotional label from the set R of emotions in Definition 6 to be associated with each video. The online survey was set to behave differently from one session to another. In particular, in one session, the user was asked to label 10 videos randomly picked up among the 7754, whereas in the next one, the user was asked to label 5 videos randomly picked up among the 7754, where the face of the actor was blurred, and so on with the following sessions.

In the first case, we aimed at having a more faithful and truthful estimate of emotions based also on the facial expressions. In the second case, we obtained a dataset that was labelled the same way that the NN would have labelled it.

At the end of the study, we had two datasets labelled with emotions: the first one was obtained starting from videos where both gestures and facial expressions were visible; the second one was obtained from videos with blurred faces, which was the same as that which the NN was required to perform. We used the first dataset to train, validate, and test the NN and the second one to compare the NN performance against humans for the emotion recognition task under the same conditions.

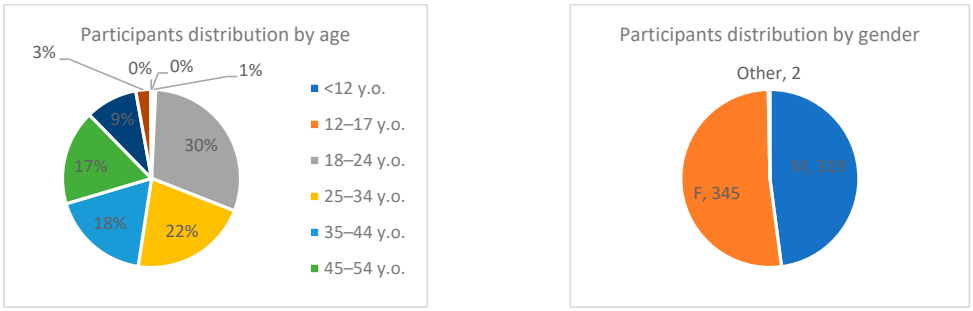
It is worth noting that, for each video, both with plain and blurred faces, we had different emotion labels corresponding to the perceived mood from the video by each user. We labelled each video with the most frequent emotion label. We knew that this could not lead to a 100% correct emotion label, but here, we aimed at testing the effectiveness of the classification based on the gesture; therefore, a mis-labelling would not be a problem, provided that a mis-labelled gesture is recognized accordingly. If the classifier is proven to be effective, it is enough to have a 100% correctly labelled gesture dataset to re-train the NN.

*Obtained Data Analysis*

The survey was sent throughout Italy, aiming at obtaining a judgement that was more culturally independent on a particular region as possible.

We collected 666 questionnaires, of which 330 were for videos with blurred faces, and 336 were for videos with plain faces. The age and gender distributions of all participants are represented in Figure 4a,b, and Figure 5 shows a comparison of participant distribution

along Italian regions with the Italian populaton density by region. Except for a few cases, it can be seen that the two trends are almost equivalent, thus confirming the overall validity of the collected data with respect to the region-specific culture neutrality.




(**a**) (**b**)

**Figure 4.** Demographic distribution of participants (**a**) by age and (**b**) by gender.

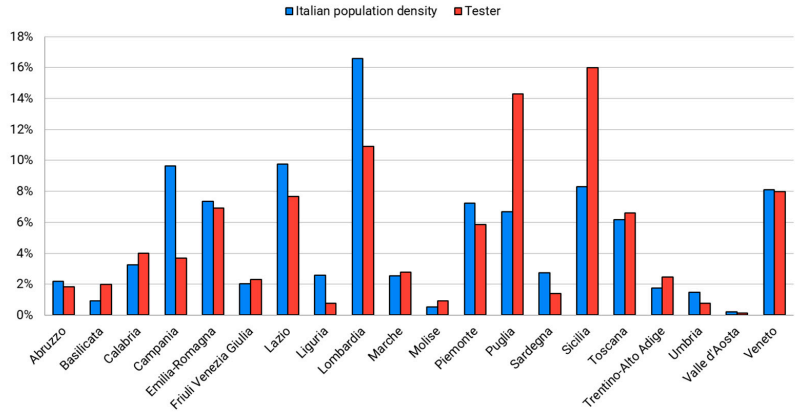



**Figure 5.** Comparison by region between population density and survey participants.

Figure 6 shows the label count for each category from Figure 3 obtained by the 336 participants who labelled the videos with plain faces. According to this, we labelled each gesture with the corresponding emotion, thus obtaining the output of the elicitation study, represented by the gesture–emotion dataset S, which is set out in Definition 5. This could then be used as the input for our classifier.

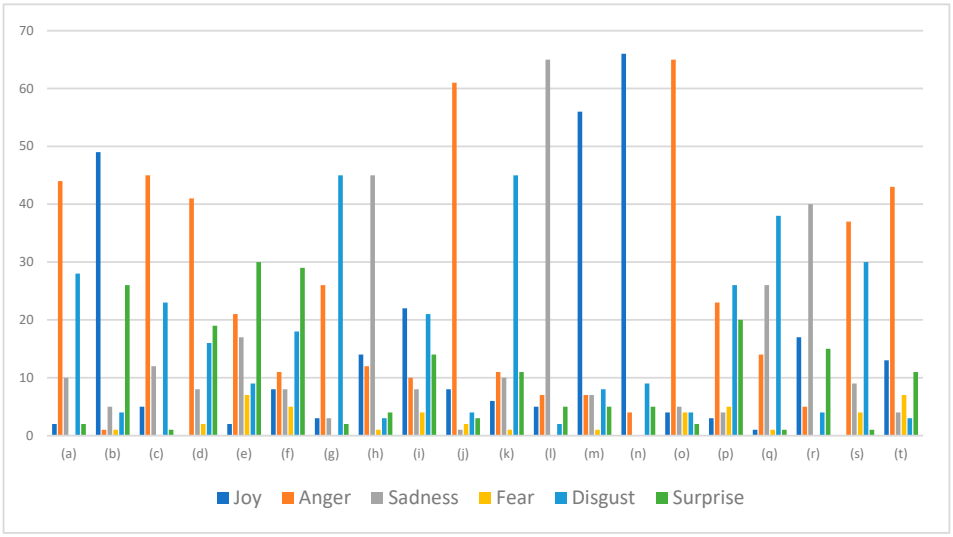



**Figure 6.** Label count for each gesture category.

**6. Recurrent Neural Network Design**

The neural network that was built is of the RNN type with LSTM units. The design question was about the hyperparameters that were to be used to generate a good neural network.

In our case, the hyperparameters that were used to let the NN classify gestures and assign them emotion labels were:

- The number of neurons in the input layer, which is equal to the number of inputs.
- The number of neurons in the output layer, which is equal to the number of classifications.
- The number of hidden layers.
- The number of neurons for each hidden layer. The number of LSTM units.
- The activation function of each layer.
- The loss type that the model must minimize during training.
- The optimizer that was used.
- The number of epochs: how many iterations are performed on all the data provided, to conclude the training.
- Batch size: the number of samples that were used before updating the gradient value, during training.

Analysing the nature of the data in the dataset, it is possible to calculate how many neurons there are in the input layer. The skeleton data of each video were used to train the neural network. Each skeleton was made up of 20 joints, and for each joint, we had information of the spatial coordinates *(x, y, z)* and quaternions *(x, y, z, w)*. The choice was made to consider only the spatial coordinates, thus ignoring the information on quaternions because the reliability of these values was not particularly high. The duration of the gestures was variable and ranged from 1 to 4 s approximately. To have a homogeneous duration, a simple sampling filter was created, which allowed us to obtain gestures of the same length of 1 s, therefore being 20 frames (or 20 skeletons) for each gesture. At this point, it is easy to calculate the number of neurons in the input layer, which is equal to:

$$Input\ layer = 20\ frames \cdot 20\ joints \cdot 3\ coordinates = 1200\ neurons$$

The number of neurons in the output layer is equal to the number of classifications that the network is enabled to make. For the purpose of the study, six basic emotions were identified; therefore, the possible classifications and therefore the number of neurons in the output layer is equal to six.

Except for the number of neurons in the input and output layers, all the other hyperparameters can be decided "at will". This implies that there are an infinite number of choices to correctly combine the hyperparameters from which a good neural network can be obtained, if the problem admits one. The decision is often guided by the designer's intuition and the experience that he or she has accumulated.

It can immediately be said that, with regard to the hyperparameters listed below, the most common or default values were chosen:

- Activation function. For the LSTM network, the default values were left unchanged; therefore, the choice fell on tanh. For the hidden layers, the choice was based on two possible functions: relu and sigmoid. For the output layer, we agreed on the softmax function.
- Loss: mean_squared_error. The mean squared error (MSE) measures the average of the squares of the errors, i.e., it indicates the mean squared difference between the values of the observed data and the values of the estimated data.
- Optimisation: Adam.
- Period: 10,000. It is an exaggeratedly large number that can waste a lot of time in training. To avoid this inconvenience, a check was made on the accuracy value of the network, which interrupts training early if it notes that, for a certain period, there are no more substantial improvements.
- Batch size: 100.

For the choice of how many units the LSTM network would be composed of, how many hidden layers the neural network would form, and how many neurons each layer would have, in the literature, there is no standard and commonly accepted method, but there are useful guidelines [40].

There are many empirical methods for determining the correct number of neurons used in hidden layers, such as the following:

- The number of hidden neurons should be between the size of the input layer and output layer neurons.
- The number of hidden neurons should be two-thirds of the size of the input layer, plus the size of the output layer.
- The number of hidden neurons should be less than double the size of the input layer.

These three rules provide a starting point to consider. In the end, the selection of an architecture for the neural network is a result of trial and error. Using these empirical rules, the choice of the number of neurons per hidden layer was limited to a value between 8 and 1024. We also agreed that each hidden layer had at most half the number of neurons compared to the number of neurons in the previous level.

When the classification problem and the data to be treated are simple, simple intuition is enough to define a good neural network. If the data are more complicated, it is not possible to rely on intuition, and unfortunately, since there are no "rules" to create and know if a neural network is correct or if it can provide good results, an alternative is to create and evaluate multiple networks with different hyperparameters, eventually choosing the best solution.

### *6.1. Neural Network Development*

For this study, the choice that was made was precisely the generation of multiple neural networks with different hyperparameters evaluated with different measures, and the best model was selected. The research on which hyperparameters to choose has focused on:

- The number of hidden layers in the neural network.
- The number of neurons for each hidden layer.
- The hidden layer activation function.
- The number of LSTM units.

For the number of neurons in the hidden layers, we defined only a minimum and a maximum value, exactly between 8 and 1024. To limit all possible combinations, we opted for neurons equal to a power of two to avoid the risk of creating an exaggeratedly large number of neural networks to evaluate. Therefore, we obtained the following set:

$\#neurons \in \{8, 16, 32, 64, 128, 256, 512, 1024\}$ Once the potential neurons were obtained, we operated the same mechanism to restrict the number of hidden layers in the network. Having already agreed that each hidden layer had at most half the number of neurons compared to the number of neurons in the previous level, it was easy to deduce a possible range:

$\#hidden\ layers \in [1, 8]$ For the activation function of the hidden layers, as already mentioned, the possible choice was limited to:

$f(x)_{activation} \in \{relu(x), sigmoid(x)\}$ The last hyperparameter that was to be defined was the number of LSTM units. For the latter, a reduced set of the number of neurons was chosen, namely:

$\#units \in \{32, 64, 128, 256, 512\}$ The possible structure of the neural network is shown below:

- Input layer with 1200 neurons and tanh activation function.
- Output layer with 6 neurons and softmax activation function.
- LSTM units belonging to the set {32, 64, 128, 256, 512}.
- The number of hidden layers included in the range [1, 8].
- The number of neurons of each hidden layer belonging to the set {8, 16, 32, 64, 128, 256, 512, 1024} and relu or the sigmoid activation function.

### *6.2. Network Creation Procedure and Evaluations*

We implemented an algorithm capable of creating and training different neural networks independently. The simplest algorithm to design for this purpose was the brute force algorithm. The procedure is shown in Algorithm 1. Once started, it took about 60 days to obtain all the neural networks.

Once the neural networks were obtained and trained, we proceeded to evaluate their performances using four metrics, which were based on the count of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) in the predictions of the classifier.

**Algorithm 1** Brute force net creation procedure

```
procedure BFC(S)
  #neurons = [8, 16, 32, 34, 128, 256, 512, 1024]
  #units = [32, 64, 128, 256, 512]
  activeFunction = [relu, sigmoid]
  Ln = length(#neurons)
  Lu = length(#units)
  outLayer = output_layer(6, softmax)
  for i = 1 to Lu do
      lstmUnits = create_lstm(#units[i], tanh)
      for i = 1 to Ln do
      hiddens = combination(#neurons, activeFunction, i) ▷ combination without repetition
      net = create(lstmUnits, hiddens, outLayer, loss = mqe, optimizer = adam,
metrics = accuracy, epoch = 10000, batch = 100)
      train(net, S)
      save(net)
    end for
  end for
end procedure
```

**Precision score.** This is a performance metric closely related to the relationship of true positives with the sum of true negatives and positives. It reaches its best value at 1 and the worst score at 0. The accuracy is intuitively the ability of the classifier to not label a negative sample as positive:

$$PRE = \frac{TP}{TP + FP}$$

**Recall score.** This is a performance metric closely related to the relationship of true positives with the sum of false negatives and true positives. It reaches its best value at 1 and the worst score at 0. The recall is intuitively the classifier's ability to find all positive samples:

$$REC = \frac{TP}{FN + TP}$$

**F1 score.** The $F1$ metric can be interpreted as a weighted average of precision and recall, where $F1$ reaches its best value at 1 and the worst score at 0. The relative contribution of precision and recall at $F1$ are equal. The formula for $F1$ is:

$$F1 = 2\frac{PRE * REC}{PRE + REC}$$

**Cohen's kappa score.** This is a statistical coefficient that represents the degree of accuracy and reliability in a statistical classification. It is an index of concordance that considers the probability of random agreement. It is expressed by the formula:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

where $Pr(a)$ is given by the sum of the first diagonal of the matrix divided by the total of the judgements and represents the percentage of judgement that is, in fact, agreed between the judges. $Pr(e)$ is the product of the positive totals added to the negative ones, all divided by the square of the total of the judgements and represents the probability of agreement randomly. There are different "degrees of agreement", based on which we can define whether Cohen's kappa is poor or excellent:

- k assumes values lower than 0, and there is no concordance.
- k assumes values between 0 and 0.4, and the agreement is poor.
- k assumes values between 0.4 and 0.6, and the concordance is discrete.
- k assumes values between 0.6 and 0.8, and the agreement is good.
- k assumes values between 0.8 and 1, and the agreement is excellent.

## 7. Experimental Results

Once all the neural networks were evaluated, it was possible to decree the best one. Figure 7 shows the structure of the best one, which obtained the best scores for all metrics. The network was made up of the following hyperparameters:

- An input layer with 1200 neurons and the tanh activation function.
- 512 LSTM units.
- Three hidden layers with 1024, 512, and 128 neurons, with the relu activation function for the first two and the sigmoid for the third.
- An output layer with 6 neurons and the softmax activation function.
- Type of loss: Mean squared error.
- Optimizer: Adam.
- Epochs: 10,000.
- Batch size: 100.

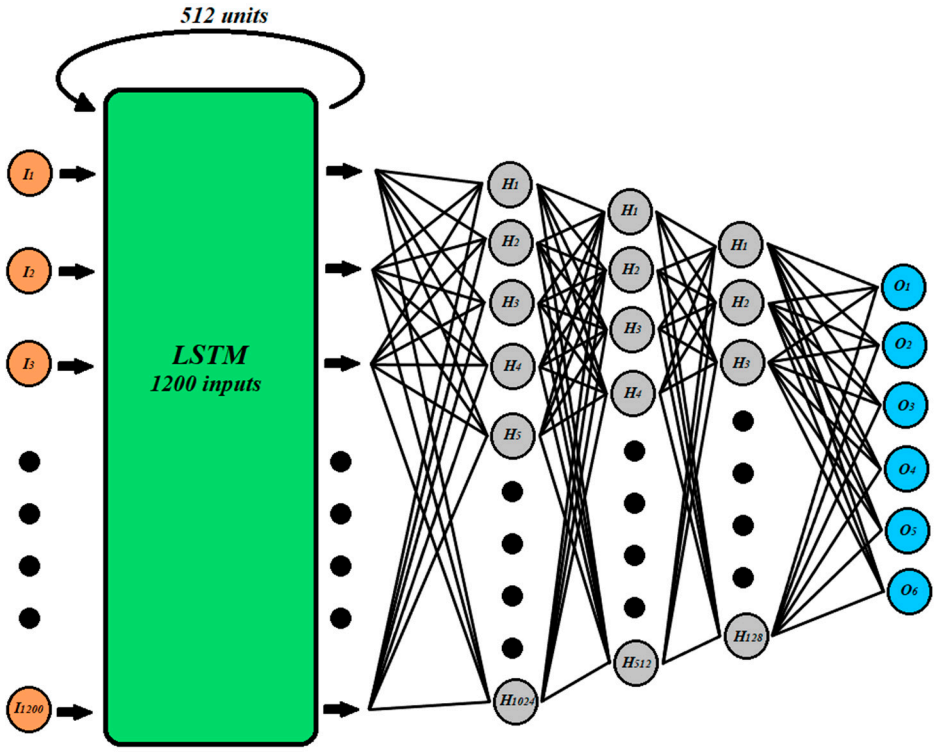



**Figure 7.** Visual representation of the neural network for gesture–emotion classification.

Once the classifier was obtained, we moved on to the last phase of the study, which was to verify if the model managed to classify correctly (i.e., according to our labelling) the gestures with the corresponding emotions, starting from videos with no audio and blurred faces. Apart from this, since we had the emotion labels for the same type of videos (blurred faces) made by humans, we also wanted to compare the emotion estimation between humans and the NN model.

For the evaluation of the model, we used the confusion matrix and the metrics listed and described above. Such metrics were also used for the comparison of the classifications between humans and machines. The results are shown in Tables 1 and 2.

**Table 1.** Comparison of evaluations obtained by the neural network vs. human beings.

| Type | Precision | Recall | F1 | Cohen's Kappa |
|---|---|---|---|---|
| Neural Net | 0.98209 | 0.98198 | 0.98196 | 0.97650 |
| Humans | 0.62847 | 0.64721 | 0.62336 | 0.51742 |

**Table 2.** Model's confusion matrix.

| | Anger | Disgust | Fear | Joy | Sadness | Surprise |
|---|---|---|---|---|---|---|
| **Anger** | 590 | 0 | 0 | 0 | 2 | 0 |
| **Disgust** | 4 | 329 | 0 | 3 | 1 | 3 |
| **Fear** | 0 | 0 | 0 | 0 | 0 | 0 |
| **Joy** | 5 | 0 | 0 | 343 | 0 | 1 |
| **Sadness** | 3 | 2 | 0 | 2 | 251 | 0 |
| **Surprise** | 3 | 1 | 0 | 1 | 0 | 177 |

By comparing the measurements of human beings with those of the classifier, we can see that the model performs better than humans in recognizing the emotions connected only to gestures, with no further clues from facial or vocal expressions. Such results were possible and predictable because the network was trained using the skeleton data provided by the Kinect device, which did not include information on the actor's face. On the contrary, humans base their assessments on the emotions transmitted by relying heavily on facial expressions and the tones of voices.

## 8. Conclusions

In this work, we created a model to perform the emotional classification of gestures belonging to a well-known dataset of Italian illustrator gestures. We mainly focused on the design of the neural network and on the choice of hyperparameters. Once we obtained the model, it was evaluated with the task of recognizing the emotions of an illustrator gesture only from the gesture itself, with no further information about facial or vocal expressions. For this task, we found that the model had an overall accuracy of 91.4%.

We also compared the model's performance against humans for the same task. In this case, the model performed better than humans in recognizing the right emotion. This was highly likely because people heavily base their assessments on facial expressions and on voice tones, whereas the model based its recognitions only on gestures.

Although we believe that these results are per se interesting and useful for the design and implementation of classifiers for the recognition of emotions, the study we carried out offers several different cues for future research, including: checking if there is a better combination of hyperparameters that allows one to have an even higher assessment measure, using, for example, genetic algorithms for the creation of a population of neural networks; creating a model that uses only video frames to free oneself from the need of specific hardware such as a Kinect; creating a model capable of analysing videos of variable duration, thus eliminating the need for having them all with the same number of frames; considering a similar study, taking advantage of a bigger dataset with more gesture categories and classifying them using a greater number of emotional labels; conducting a

similar study in different cultural and/or less controlled environments; and carrying out a qualitative study for the explainability of the results.

As a final remark, which may be useful for any further work on natural environments and live video streams, the classifier only takes around 4 ms on a medium to low range device to perform a classification.

**Author Contributions:** Conceptualisation, D.F., V.G., F.M. and S.S.; formal analysis, D.F., V.G. and S.S.; investigation, D.F. and F.M.; methodology, D.F. and V.G.; supervision, S.S.; validation, D.F.; visualisation, V.G.; writing—original draft, D.F. and S.S.; writing—review and editing, S.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the gesture elicitation study.

**Data Availability Statement:** Raw and aggregated data are available (in Italian) at https://docs.google.com/spreadsheets/d/1OEu_tHPGw_l510OJQDSKJQ6gJiFeKK2GZwVprJwWML0/edit?usp=sharing accessed on 13 July 2022.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Fragopanagos, N.; Taylor, J.G. Emotion recognition in human-computer interaction. *Neural Netw.* **2005**, *18*, 389–405. [CrossRef] [PubMed]
2. Daras, P.; Achilleopoulos, N.; Panebarco, M.; Mayora, O.; Stollenmayer, P.; Williams, D.; Pennick, T.; Magnenat-Thalmann, N.; Guerrero, C.; Pelt, M.; et al. User Centric Media of the Future Internet. In Proceedings of the 2008 The Second International Conference on Next Generation Mobile Applications, Services, and Technologies, Cardiff, UK, 16–19 September 2008; pp. 433–438. [CrossRef]
3. Andreou, P.; Georgiadis, D.; Pamboris, A.; Christophorou, C.; Samaras, G. Towards a backend framework for supporting affective avatar-based interaction systems. In Proceedings of the European Conference on Ambient Intelligence 2015, Athens, Greece, 11–13 November 2015.
4. Davatzikos, C.; Ruparel, K.; Fan, Y.; Shen, D.G.; Acharyya, M.; Loughead, J.W.; Gur, R.C.; Langleben, D.D. Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *Neuroimage* **2005**, *28*, 663–668. [CrossRef] [PubMed]
5. Picard, R.W. *Affective Computing*; The MIT Press: Cambridge, MA, USA, 1997; Volume 167, p. 170.
6. Ekman, P.; Friesen, W. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*; Consulting Psychologists Press: Palo Alto, CA, USA, 1978.
7. Kleinsmith, A.; Bianchi-Berthouze, N. Affective Body Expression Perception and Recognition: A Survey. *IEEE Trans. Affect. Comput.* **2013**, *4*, 15–33. [CrossRef]
8. Dael, N.; Mortillaro, M.; Scherer, K.R. Emotion expression in body action and posture. *Emotion* **2012**, *12*, 1085–1101. [CrossRef] [PubMed]
9. Patwardhan, A.; Knapp, G. Multimodal affect recognition using kinect. *arXiv* **2016**, arXiv:1607.02652.
10. Castellano, G.; Kessous, L.; Caridakis, G. Emotion Recognition through Multiple Modalities: Face, Body Gesture, Speech. In *Affect and Emotion in Human-Computer Interaction*; Peter, C., Beale, R., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2008; Volume 4868. [CrossRef]
11. Gunes, H.; Piccardi, M. Bimodal emotion recognition from expressive face and body gestures. *J. Netw. Comput. Appl.* **2007**, *30*, 1334–1345. [CrossRef]
12. Kipp, M.; Martin, J. Gesture and emotion: Can basic gestural form features discriminate emotions? In Proceedings of the 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam, The Netherlands, 10–12 September 2009; pp. 1–8. [CrossRef]
13. Glowinski, D.; Dael, N.; Camurri, A.; Volpe, G.; Mortillaro, M.; Scherer, K. Toward a Minimal Representation of Affective Gestures. *IEEE Trans. Affect. Comput.* **2011**, *2*, 106–118. [CrossRef]
14. Senecal, S.; Cuel, L.; Aristidou, A.; Magnenat-Thalmann, N. Continuous body emotion recognition system during theater performances. *Comp. Anim. Virtual Worlds* **2016**, *27*, 311–320. [CrossRef]
15. De Gelder, B. Towards the neurobiology of emotional body language. *Nat. Rev. Neurosci.* **2006**, *7*, 242–249. [CrossRef] [PubMed]
16. Ekman, P. Emotional and Conversational Nonverbal Signals. In *Language, Knowledge, and Representation*; Larrazabal, J.M., Miranda, L.A.P., Eds.; Philosophical Studies Series; Springer: Dordrecht, The Netherlands, 2004; Volume 99. [CrossRef]
17. Krishnan, H.; Raj, U. Cross Cultural Communication. 2014. Available online: https://www.slideshare.net/harikrishnann2/crosscultural-communication-42061337 (accessed on 15 July 2022).

18. Escalera, S.; Gonzàlez, J.; Baró, X.; Reyes, M.; Lopes, O.; Guyon, I.; Athitsos, V.; Escalante, H. Multi-modal gesture recognition challenge 2013: Dataset and results. In Proceedings of the 15th ACM on International Conference on Multimodal Interaction, (ICMI '13), Sydney, Australia, 9–13 December 2013; Association for Computing Machinery: New York, NY, USA, 2013; pp. 445–452. [CrossRef]
19. Kendon, A. How gestures can become like words. In *Cross-Cultural Perspectives in Nonverbal Communication*; Poyatos, F., Ed.; Hogrefe & Huber Publishers, 1988; pp. 131–141.
20. Karam, M.; Schraefel, M.C. A Taxonomy of Gestures in Human Computer Interactions. 2005. Project Report. Available online: http://eprints.soton.ac.uk/id/eprint/261149 (accessed on 15 July 2022).
21. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161–1178. [CrossRef]
22. Ekman, P. Basic Emotions. In *Handbook of Cognition and Emotion*; Dalgleish, T., Power, M.J., Eds.; John Wiley & Sons: Hoboken, NJ, USA, 1999; pp. 45–60. [CrossRef]
23. Parrott, W.G. (Ed.) *Emotions in Social Psychology: Essential Readings*; Psychology Press: Philadelphia, PA, USA, 2001; ISSN 1531-2569.
24. Cornelius, R.R. Theoretical approaches to emotion. In Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, Newcastle, UK, 5–7 September 2000.
25. Elfenbein, H.A.; Mandal, M.K.; Ambady, N.; Harizuka, S.; Kumar, S. Cross-cultural patterns in emotion recognition: Highlighting design and analytical techniques. *Emotion* **2002**, *2*, 75–84. [CrossRef] [PubMed]
26. Mehrabian, A.; Friar, J.T. Encoding of attitude by a seated communicator via posture and position cues. *J. Consult. Clin. Psychol.* **1969**, *33*, 330–336. [CrossRef]
27. Elfenbein, H.A.; Ambady, N. On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychol Bull.* **2002**, *128*, 203–235. [CrossRef]
28. Hess, U.; Senecal, S.; Kirouac, G. Recognizing emotional facial expressions: Does perceived sociolinguistic group make a difference? *Int. J. Psychol.* **1996**, *31*, 18486.
29. Aristidou, A.; Charalambous, P.; Chrysanthou, Y. Emotion Analysis and Classification: Understanding the Performers' Emotions Using the LMA Entities. *Comput. Graph. Forum* **2015**, *34*, 262–276. [CrossRef]
30. Fourati, N.; Pelachaud, C. Relevant body cues for the classification of emotional body expression in daily actions. In Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 21–24 September 2015; pp. 267–273. [CrossRef]
31. Carreno-Medrano, P.; Gibet, S.; Marteau, P. End-effectors trajectories: An efficient low-dimensional characterization of affective-expressive body motions. In Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 21–24 September 2015; pp. 435–441. [CrossRef]
32. Cutler, R.; Turk, M. View-based interpretation of real-time optical flow for gesture recognition. In Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 416–421. [CrossRef]
33. Shotton, J.; Sharp, T.; Kipman, A.; Fitzgibbon, A.; Finocchio, M.; Blake, A.; Cook, M.; Moore, R. Real-time human pose recognition in parts from single depth images. *Commun. ACM* **2013**, *56*, 116–124. [CrossRef]
34. Gentile, V.; Sorce, S.; Gentile, A. Continuous Hand Openness Detection Using a Kinect-Like Device. In Proceedings of the 2014 Eighth International Conference on Complex, Intelligent and Software Intensive Systems, Birmingham, UK, 2–4 July 2014; pp. 553–557. [CrossRef]
35. Ren, Z.; Yuan, J.; Meng, J.; Zhang, Z. Robust Part-Based Hand Gesture Recognition Using Kinect Sensor. *IEEE Trans. Multimed.* **2013**, *15*, 1110–1120. [CrossRef]
36. Song, Y.; Gu, Y.; Wang, P.; Liu, Y.; Li, A. A Kinect based gesture recognition algorithm using GMM and HMM. In Proceedings of the 2013 6th International Conference on Biomedical Engineering and Informatics, Hangzhou, China, 16–18 December 2013; pp. 750–754. [CrossRef]
37. Carmona, J.M.; Climent, J. A Performance Evaluation of HMM and DTW for Gesture Recognition. In Proceedings of the CIARP 2012: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Buenos Aires, Argentina, 3–6 September 2012; Alvarez, L., Mejail, M., Gomez, L., Jacobo, J., Eds.; Lecture Notes in Computer Science. Springer: Berlin/Heidelberg, Germany, 2012; Volume 7441. [CrossRef]
38. Yang, M.H.; Ahuja, N. Recognizing Hand Gestures Using Motion Trajectories. In *Face Detection and Gesture Recognition for Human-Computer Interaction*; The International Series in Video Computing; Springer: Boston, MA, USA, 2001; Volume 1. [CrossRef]
39. Harmon-Jones, E.; Amodio, D.M.; Zinner, L.R. Social psychological methods of emotion elicitation. In *Handbook of Emotion Elicitation and Assessment*; Coan, J.A., Allen, J.J.B., Eds.; Oxford University Press: Oxford, UK, 2007; pp. 91–105.
40. Heaton, J. *Introduction to Neural Networks for Java*, 2nd ed; Heaton Research, Inc.: Chesterfield, MO, USA, 2008.