Emotions and Interaction with Virtual Agents and Robots

Emotional Body Gesture Recognition

Nour Eldin Badr

Bielefeld University, Bielefeld, Germany

September 15, 2020

Abstract

Emotions can be recognized through the face, can be recognized through speech and can be recognized through body gestures. We chose to discuss emotions recognized through body gestures as it is not as easy as recognizing it through facial and speech as well as it is more challenging as the same emotion can be expressed by different poses.We will also show that body gestures can increase the accuracy of emotion recognition if combined with facial expression.
In this essay we will discuss how can humans express emotions through body languages as well as the steps or the pipeline to to recognize emotions through body gesture, starting from how to detect humans in an image or a video passing through the state of the arts of pose estimations ending by emotion recognition.Finally, we will discuss the importance of using emotional body gesture recognition in some applications such as violence detection in hockey games.

1 Introduction

In the seminar of "Emotions and Interaction with Virtual Agents and Robots", students have presented many interesting topics,starting from defining emotions, ending to the computational models of emotion. Before stating the definition of emotion, let us have an example to understand the definition better.
Let us assume that you are having a walk and you accidentally hit someone on the shoulders. He can have two momentary reactions, he can quietly smile and accept your apology or he can hit and punch you, all depends on his on going state of mind which we can call it "the mood". If he got a salary raise, then he is happy and when you hit him, he can just ignore the hit. But if he got fired from his job and his ongoing mood is really bad and he is angry, then his momentary reaction will be aggressive. After understanding this example, we can easily understand the definition from [15] which we can summarize it as a momentary reaction in a current situation which depends on the ongoing state of mind which we can call it mood.
These emotions we have discussed can be expressed in some action modalities such as face expression, voice expression and body movement or gestures. These modalities can be expressed by humans and nowadays scientists are trying to express it on robots such as in virtual agents and humanoids(e.g. Geminoid).Through this expressions we can express emotions like anger, happiness, sadness and surprise.
But how can we recognize these emotions in humans? In this essay we will mainly focus on Emotional Body Gesture Recognition.As we will discuss later in section 2, Emotional Body Gesture Recognition consists of three main components; Human detection, body pose estimation and emotion recognition.later on, through discussing two experiments, we will show the importance of using body gesture in emotion recognition.

2 Emotional Body Gesture Recognition

This section discusses the steps to recognize emotions through body gestures.In each step we will discuss briefly one or two examples of the techniques used in this method, please refer to the references for more information.
Before discussing the steps, we need to understand how humans express emotions through body language.

2.1 Emotions expression through body language

We humans in order to communicate and express our emotions, feeling and thoughts, we usually do body movements and gestures,these gestures usually are referred to as body language[14].
Although expressing emotions through body language can be different from culture to culture and from gender to gender, but there are six basic emotions that seem to be common and universal and according to[9], they follow general movements protocols.

Figure 1: Emotional Expression through body language[16]

As we can see in figure 1, anger for example can be expressed by spreading the body and putting hands on hips or waist. Disgust cab be expressed by baking hands and covering the neck.Fear can be expressed by crossing legs and arms and dragging elbows inward. Happiness can be expressed by opening the arms and stretching legs apart. By dropping body and bending head we can express sadness. Finally, by moving towards head we can express surprise[9].

2.2 Steps of Emotional Body Gesture Recognition

In order to recognize emotions through body gesture,there are main three steps should be taken in consideration.In this section each step will be discussed briefly. First step is to detect humans from an image or a video.Secondly we estimate the body pose and finally we recognize emotions .

2.2.1 Human Detection

To detect humans in an image or video, there are so many approaches have been used in this field, but in this survey we are going to focus on two of them.

2.2.1.1 Dalal and Triggs Technique

This technique is also called (HOG + SVM) which is (Histograms of Oriented Gradients + Support Vector Machine) and it consists of the following [4]:

1. crop 64x128 slide window from original image as shown in figure 2.

Figure 2: Detection window(sliding window) of size 64x128[1]

2. Divide that cropped image into 8x8 cells in which in each cell we will calculate HOG.

3. for each grid we calculate gradient direction and gradient magnitude,check [4] .

4. As shown in figure 3 we have a vector of 9 bins divided on angles from 0 to 180 degrees. Vectors will be filled with the value of grad magnitude that is corresponding to each angle in grad direction.

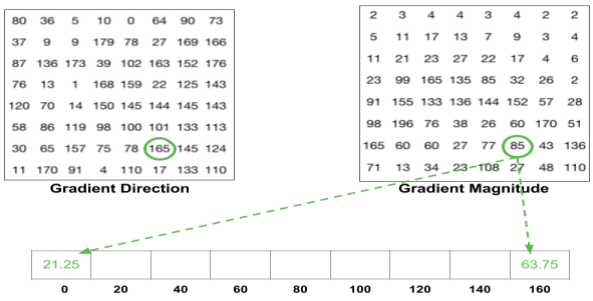

Figure 3: Creating the 9 feature vectors[12]

5. Now each grid in the cropped image has 9 feature vectors.Create a sliding window within the cropped part of the image of 4x4 to concatenate each 4 grids together which will form 36 feature vectors.

6. Now normalize feature vectors so that it wouldn't be affected by the change of light intensity.

7. Apply SVM classifier [3] to detect if it is human or not.

2.2.1.2 Deformable Part Models

Another model to detect humans in an image or a video is DPM(Deformable Part models).This technique depends on detecting the whole human body then dividing the human body into parts and detecting them individually as shown in figure 4.

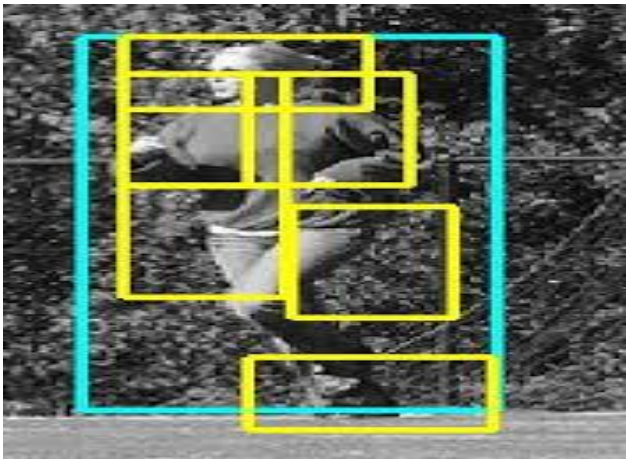

Figure 4: DPM(whole body + body parts)[7]

To elaborate on that more, DPM consists of two paths as it is explained in [6]:

1. First path is to extract features map using HOG method and then multiplying it with SVM coefficient of the whole body.

2. In the second path we will twice the resolution and extract the feature maps using HOG method too. Then we multiply it with the SVM coefficient of face and the other five parts which are the right and the left upper chest, the torso and the feet.

Finally we add all of them together as shown in figure 5.

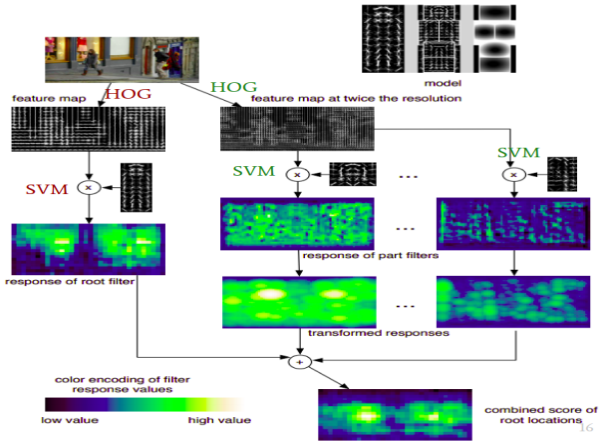

Figure 5: DPM steps[6]

Please note that DPM is quite similar to Dalal and Triggs Technique, but its main advantage is that it increases the accuracy and it gains efficiency as it is not only working on the whole body, but the body parts too. To elaborate more we can have an example. If one person is standing behind another person with a small lag between them then Dalal and Triggs model gets confused, which leg related to the human it should detect, But DPM solves this problem, as it detects the whole body then its parts, so it will annotate that this leg is related to that body not the other one.

2.2.2 Pose Estimation

After detecting humans, pose needed to be estimated to have the ability to recognise emotions. In pose estimation, we try to estimate a certain pose through key points in the human body like elbows, wrists, torso, neck, shoulders and knees.In this section two state of the art methods used for pose estimation are going to be explained briefly, which are Deep pose and Duel source Deep Neural Network.

2.2.2.1 Deep pose

Deep pose paper is one of the first papers to apply deep learning on human pose estimation and they have proved that CNN can be used for localization rather than used for classification only [17].

Deep pose consists of two main stages[17]:

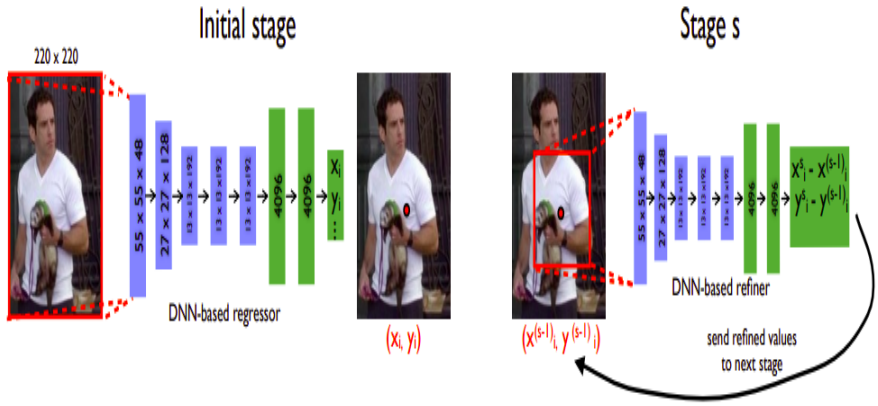

Figure 6: Deep pose stages[17]

1. first stage as shown in figure 6 is the initial stage.In this a joint will be detected on the image using DNN but it will not be very accurate, that's why we have stage two after the initial stage.

2. second main stage which is called Stage S as shown in figure 6, is to refine the the detected joint to get better result and repeat stage S over and over till getting a better result.

In the following example figure 7, we can see the differences between stage 1 and stage 3, and how the model is improved a bit in stage 2 then more in stage 3 and that is done by refining the joints detected in each stage.

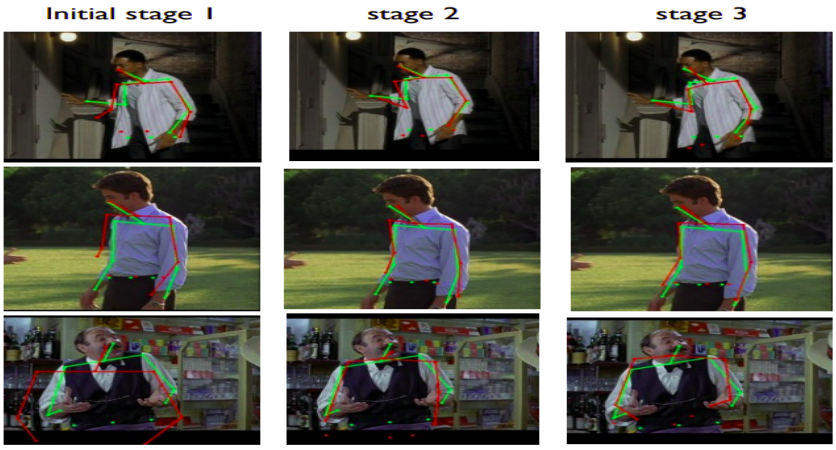

Figure 7: Example after applying deep pose to estimate poses and detect joints. [17]

2.2.2.2 Dual source deep neural network

Dual source Deep Neural Network is kindly the same idea as DPM. As shown in figure 8, firstly it Detects the whole body and and a body part. Then feed it to Duel source convolutional neural networks (DS_CNN). This DS_CNN is consisting of 5 convolutional layers in which layers 1 and 2 are followed by max pooling then 3 fully connected layers and softmax layer. DS_CNN has two outputs, the detected joint and it is represented with a heat map and the joint location represented in x and y coordination [5] .

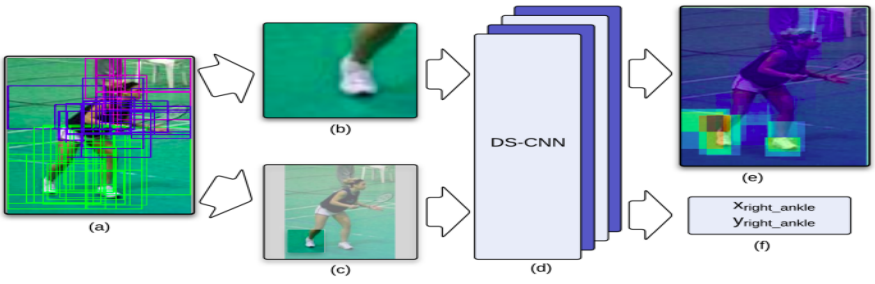

Figure 8: Dual source Deep Neural Network illustration.The output is the detected joint and the joint location [5]

In [5] they have made an experiment on Leeds Sports Pose (LSP) dataset to compare the probability of correct pose(PCP) for different methods such as the Deep pose which have discussed and the DS_CNN. As shown in the following table 1 DS_CNN got really good accuracy compared to Deep Pose.
.

Method	Arm		Leg		Torso	Head	Avg.
	Upper	Lower	Upper	Lower			
DS-CNN	**0.80**	**0.63**	**0.90**	**0.88**	**0.98**	0.85	**0.84**
DeepPose	0.56	0.38	0.77	0.71	-	-	-

Table 1: Experiment results on LSP dataset [5]

2.2.3 Emotion Recognition

After detecting human in an image or a video and estimate his pose,the next step is to recognize the emotion. To recognize emotion, classification algorithms in machine learning are being used.In this section we will give a brief introduction to two of them as they will be used in the experiment in the discussion section.

2.2.3.1 Decision trees

It is a method more like a flow chart in a tree shape where the leaf which is the end node is a decision or the class that we predict and the roots are the inputs[10]. The decision in decision trees is probabilistic and it is depending on the distributions of the frequency of the likely events. One of the examples of decision trees is C4.5 in which its main advantage is that it works with discrete data and it can handle incomplete data well[10].

2.2.3.2 Bayes net

Bayes net is also graphical model but it is not flow chart, it is acyclic graph.

It is depending on conditional probability (i.e. it is representing set of variables and their conditional dependencies)[11]. One of its advantages is that it can be used for small data set. By this we have discussed briefly each step to recognize emotion.

3 Discussion

In this section we would like to discuss briefly two experiments that show the importance of using body gesture in emotion recognition. we will also discuss an application of using body gesture to recognise emotions.

3.1 Experiments

As shown in the figure 9 in [8], there are two inputs which are the face input and the body input.In the mean while there are two outputs which are the feature level fusion and decision level fusion.

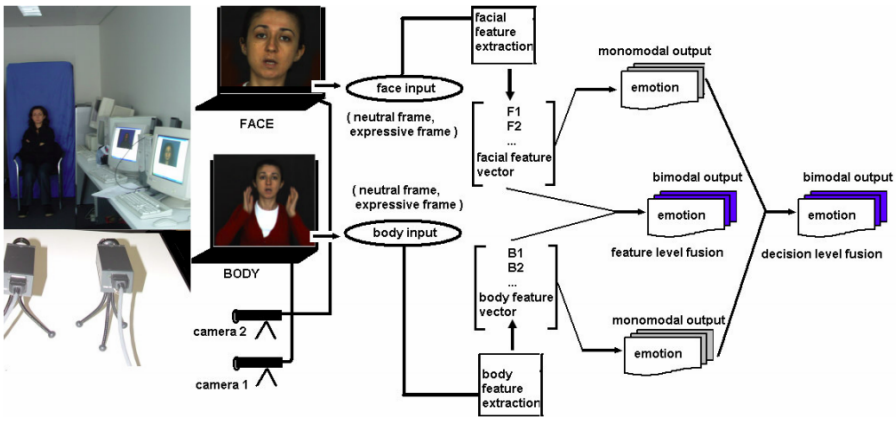

Figure 9: monomodal and bimodal experiment [8]

In decision level fusion (monomodal output), each modality is classified independently (i.e. emotion is recognised separately from facial and body feature vectors).Then calculate the maximum posterior probability for each class(anger, sadness, happiness ,etc. ..) with respect to face feature vector and body feature vector [8].

In feature level fusion(bimodal output), features are being extracted from each modality (i.e. facial feature vector and body feature vector) then they are being concatenated into a large vector [8].
In this experiment, when each modality was trained independently (i.e. monomodal), the accuracy of recognising emotions from facial expressions was 78% while in body gestures was 90%. But when Body gesture was add to facial expressions in the feature level fusion, the accuracy has increased to 96% [8]. As a summary, from this experiment we can notice that adding body gestures to facial expression has increased the accuracy by 8% more.
In another experiment in another paper [2], i have noticed a very interesting point.In this paper they have used three modalities, speech, facial expressions and body gestures.what caught my attention here is that the accuracy of recognising emotion from body gestures(monomadal) was the highest accuracy with 83.2% while the facial and the speech were 59.6% and 70.8% respectively.The accuracy of the overall multimodal system(feature level fusion) is 89.4%.
After reading these two experiments, I got the curiosity to search why they got better results in body gestures than the other two modalities. Unfortunately they didn't state why but i would like to do more research in this part and to know the reason behind it.
As a summary, from my own point of view and from what i have noticed from the above experiments, recognising emotion from body gestures might have a better accuracy than other modalities like facial and speech and facial.It can also be big advantage and a boost for a multimodal or bimodal system .

3.2 Application

One of the applications that can be used for recognizing emotions through body gestures is "violence detection", one of the papers that has applied that is [13], in which they have tried to analyze fights videos and detect from it violence. To quickly summarize the steps done here,an image will be represented as set of features by bag of words to extract histograms to be classified using support vector machines(SVM) [13]. Results were quite good in which they have achieved an accuracy of 90% after training model on 1000 clips from the hockey dataset.

4 Further Discussion

This survey has sparkled two ideas that i think one of them could be a master thesis. These ideas will help children with autism and their parents too. Recognizing emotion through face sometimes is difficult so recognizing it through body gestures might help a bit. To clarify it more let us discuss the two ideas.

1. I thought of creating an animated live virtual agent that can communicate with a kid with autism and be his friend. There would be multiple cameras and a microphone to analyze the current emotion of the kid. This analysis will be an input to the virtual agent. The virtual agent react according to this emotion which we can get help how to react from the psychology department. The agent should also teach the kid how to express and recognise the right emotion. The agent can also see the kid's emotion expression through body and try to correct it.

2. The second idea to detect violence and depression in kids with autism in a kinder garden for example and analyse the kids emotion to get better understanding and to help teachers how to react and deal with them.

5 Conclusion

Emotions can be recognised through monomodal recognition system using input such as speech,facial expression or body gestures. It can be also recognised through multimodal recognition system using the combination of the three inputs or two of them. As we have discussed in the experiments in the last section, accuracy from body gesture in monomodal could be higher than using other inputs, thus it can be very beneficial if we added it to a multimodel recognition system and it could boost the overall accuracy.

References

[1] Richa Agrawal. *person-detection-in-various-posture-using-hog-feature-and-svm-classifier*. https://medium.com/@richa.agrawal228/person-detection-in-various-posture-using-hog-feature-and-svm-classifier-2c3a3991022c. Oct. 2018.

[2] George Caridakis et al. "Multimodal emotion recognition from expressive faces, body gestures and speech". In: *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer. 2007, pp. 375–388.

[3] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297.

[4] Navneet Dalal and Bill Triggs. "Histograms of oriented gradients for human detection". In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. IEEE. 2005, pp. 886–893.

[5] Xiaochuan Fan et al. "Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1347–1355.

[6] Pedro F Felzenszwalb et al. "Object detection with discriminatively trained part-based models". In: *IEEE transactions on pattern analysis and machine intelligence* 32.9 (2009), pp. 1627–1645.

[7] Ross Girshick. *Deformable part models*. http://vision.stanford.edu/teaching/cs231b_spring1213/slides/dpm-slides-ross-girshick.pdf. 2013.

[8] Hatice Gunes and Massimo Piccardi. "Affect recognition from face and body: early fusion vs. late fusion". In: *2005 IEEE international conference on systems, man and cybernetics*. Vol. 4. IEEE. 2005, pp. 3437–3443.

[9] Hatice Gunes and Massimo Piccardi. "Fusing face and body gesture for machine recognition of emotions". In: *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.* IEEE. 2005, pp. 306–311.

[10] Badr Hssina et al. “A comparative study of decision tree ID3 and C4. 5”. In: *International Journal of Advanced Computer Science and Applications* 4.2 (2014), pp. 13–19.

[11] Chen Jing and Fu Jingqi. “Fire alarm system based on multi-sensor bayes network”. In: *Procedia Engineering* 29 (2012), pp. 2551–2555.

[12] Satya Mallick. *Histogram of Oriented Gradients*. https://www.learnopencv.com/histogram-of-oriented-gradients. 2016.

[13] Enrique Bermejo Nievas et al. “Violence detection in video using computer vision techniques”. In: *International conference on Computer analysis of images and patterns*. Springer. 2011, pp. 332–339.

[14] Fatemeh Noroozi et al. “Survey on emotional body gesture recognition”. In: *IEEE transactions on affective computing* (2018).

[15] Franklin E Payne. “A definition of emotions”. In: *Journal of Biblical Ethics in Medicine* 3.4 (1989), pp. 1–9.

[16] Konrad Schindler, Luc Van Gool, and Beatrice De Gelder. “Recognizing emotions expressed by body pose: A biologically inspired neural model”. In: *Neural networks* 21.9 (2008), pp. 1238–1246.

[17] Alexander Toshev and Christian Szegedy. “Deeppose: Human pose estimation via deep neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1653–1660.