Emotion Recognition from Body Motion

by

Laleh Ebdali Takalloo

A Report Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Engineering

in the Department of Electrical and Computer Engineering

Supervisory Committee
Emotion Recognition from Body Motion

by

Laleh Ebdali Takalloo

Supervisory Committee

Dr. Kin Fun Li, Supervisor
(Department of Electrical and Computer Engineering)

Dr. Lin Cai, Member
(Department of Electrical and Computer Engineering)

ABSTRACT

Understanding human emotions has become a research trend and practical topic in recent decades. There are much research having good result on detecting human emotions with facial expressions, speech and text. Recognizing emotions from body posture or movement is an emerging area of research, and it has shown progressive results in the past few years. This study presents an overview of recent research in this field. The relationship between emotion and body movements is discussed. The factors that affect this relation are presented. Based on recent advanced research, an integrated and comprehensive structure for the automatic detection of emotions based on body movements is introduced. Each component of this structural process is considered. In particular, body movement models, their evaluation, and available datasets are examined.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank:

My supervisor, Dr. Kin Fun Li, for his constant guidance, support and patience.

My parents for their unconditional support.

My husband whose constant love and support keep me motivated and confident.

DEDICATION

Dedicated my dissertation work to my family and my husband for their support, motivation and guidance.

Chapter 1

Introduction

In addition to hearing, humans interact with visual expressions to better convey concepts and meanings, especially in these days of virtual communications. Body language, including movement, posture, facial expression, and eye movement, is an important vehicle to express one's emotional state. The common emotions are angry, disgusted, fearful, happy, sad, surprised, and neutral [45]. The subject of body language based on facial expressions and sounds has also been studied in 19century studies. One study found that 65% of the emotions are transmitted through nonverbal communication [25]. Emotion may be read from different parts of the body simultaneously: face, hands, head position, and torso. In fact, for the correct interpretation of body language as an emotional state, different parts of the body must be considered simultaneously.

A study by Silva showed that body posture features such as head and elbow flexion, arm and shoulder distance, and chest position are statistically important indicators of emotions [22]. To better understand the subject, figure 1.1 shows an example of body posture. There are many other similar studies in various fields that highlight the importance of recognizing body-language based emotion in human life.

The relationship between body posture and emotions has been studied in various fields. Darwin first gave evidence for this connection between emotions and body condition in 1872 [12]. The human body has evolved to react to the events around with its body movements (Such as conversation between people, dancing, daily activities and baby movements), and this movement of the body will be a feeling towards that event. Atkinson [9], for instance, indicated in a study that static form (body posture) and dynamic movement characteristics play an important role with emotions. It is also important in making animations and they have used various researches in this

Figure 1.1: Body language and emotions [45]

field. In fact, all these researches and issues are a confirmation of the main subject of this survey, which is the possibility of recognizing emotions from the state and movements of the human body.

Body posture movement can include motion of the hands, head, legs and other parts of the body. These movements and gestures can be innate, acquired, or the body's natural physical reaction. Emotions can have different definitions from different perspectives. In general, emotions are the short-term response of part or all of the body in response to the observation and evaluation of an external or internal event and it often causes changes in behavior [12]. Today, computers and hardware are embedded in various parts of human life, and this issue is expanding and progressing every day, making lives more enjoyable and controllable. With the advancement of technology, its complexities have increased and with the increase in the number of technological tools, their shrinkage and expansion for public use, the traditional interaction of the past can no longer be generalized to the world of future technology and needs to be changed. Therefore, one of the approaches is to have more machines intelligent to give machines more human-like abilities to communicate with users, and this shows the importance of automatic detection of emotions. Smart machines in order to be really smart and interact with human naturally, should have the ability to recognize, understand and express emotions.

This work studies the research on recognizing emotion based on body movement the current literature. In the following sections, the details of an automatic emotion

recognition system, challenges, relevant databases and different applications of this topic will be discussed.

1.1 Project Objective

Recognition of emotions based on face, voice and body signals has been studied in many studies and has yielded good results. Also, research has been done in the field of recognizing emotions based on the combination of face, voice, etc. with the title of multiple modalities. The best accuracy for recognizing emotion in a multi-modal mode (Face, voice, body, etc.) is obtained which human subconsciously uses this approach and understands the feelings of other people using combination of verbal and non-verbal cues. However, in the machine world, there are some conditions and restrictions that make it impossible to use multi-modal mode (for example, not having proper access to voice and face information in video surveillance systems in public places).

Although much research has been done on emotion recognition, most of it has focused on face-based diagnosis, and less attention has been paid to body movements. In recent years this issue has received more attention. A 2009 study found that 95% of facial recognition studies were based on facial expressions and only 5% were based on other body elements such as sound, with the lowest number of studies based on body language [20]. Today, emotion recognition based on body movement is more welcomed because of its benefits in application as well as multi-modal diagnosis. Some of the advantages and reasons for using this approach are listed below.

- According to psychological research, some emotional states can only be recognized through body movements [26].
- In some applications, such as video surveillance in public places, it is not possible to detect emotion through face and voice remotely, and body condition makes this recognition possible [26].
- Creating a dummy emotion on face and voice can be controlled, but it is much more difficult to create a dummy feeling through body posture. Therefore, recognizing emotion based on body posture and movements is more reliable than face and sound [41].

1.2 Outline

Chapter 1 contains a problem statement, followed by introducing emotion detection, and project objective.

Chapter 2 describes the structure of an emotion recognition based on body movement system.

Chapter 3 focuses the dataset perspective, its acquisition tools, modality, and proposes available datasets.

Chapter 4 is where the experiments and the methodology for them are described.

Chapter 5 discusses the project objective and results. It also enumerates avenues of future work.

Chapter 2

Process of emotion recognition from body movement

Examining the components and issues related to recognition of emotions based on posture and body movements is the goal of this study. Figure 2.1 shows an emotion recognition system with its components based on current research in this field. The input is typically a camera input as an image or video frame. The output is the estimated or classified emotion. The essentials of a system for recognizing emotions based on the human body and its components are discussed in this section.

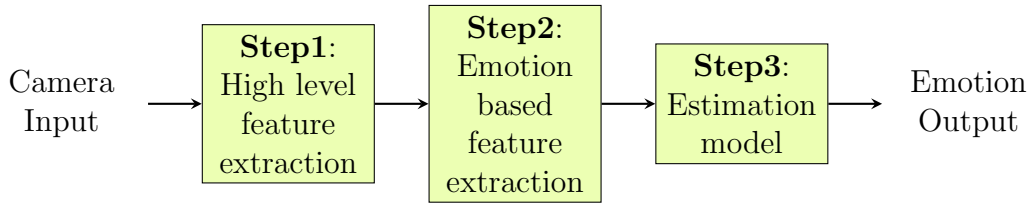

Figure 2.1: Structure of an emotion recognition based on body movement

After receiving the input data, the first step is to extract high-level features such as human diagnosis, key points and skeleton of the body. The next step is to obtain or derive meaningful and distinctive features, while the third step is to use a trained model (classification or regression) to estimate emotion based on the features extracted in the previous step. The final output is the predicted or estimated emotion. However, in methods based on deep learning, the steps of feature extraction and estimation are performed in a single step.

The input data can be in the form of a two-dimensional or a three-dimensional image from various types of sensors. The first processing step is to extract the high-level features and model the input. Since the goal is to recognize the feeling of a human, it is necessary to detect the human in the frame as an object in the input image, localized in a bounding box. Most of the approaches to object and human recognition include the extraction of candidate areas, the classification of objects in the candidate areas, and decision-making [49]. There are various methods of human diagnosis from traditional approaches as well as deep neural networks.

The next step is to extract the human from the output of the previous module and then obtain the features to feed the learning model. The third and last part of the process is the learning model or estimator that has to estimate or classify the correct emotions. This model can have a different structure depending on the type of input and output data. It can be considered as a classification or regression problem. The modeling methods used in existing works are discussed in more detail in the following sections. It should be noted that the emotion output can be discrete or continuous.

2.1 Data acquisition

The first part of an emotion recognition system based on body movement is data input. The data acquisition tool and the type of input data are two factors that may affect performance. To obtain the structure and shape of the body, the input data could be an image or a signal or a combination of the two. To get such images, ordinary RGB cameras, 3D cameras such as Kinect or a combination of multiple cameras with markers on the body, can be used. Each of these approaches has certain advantages and disadvantages, depending on the purpose of the system. Further details on data types and data acquisition tools are provided in detail in the Dataset section. Training data, in either a public or private dataset, should be used to teach the learning model.

2.2 Human detection

Human detection is an object detection methodology that localizes predefined objects within an image or video sequence, which often means identifying a rectangle or bounding box that surrounds a human. Due to the flexible nature of the body and the

diverse appearance of humans, coupled with the uncontrolled environments, human detection has many challenges.

Methods to detect humans can be divided into two categories: traditional methods and deep neural networks. The AdaBoost learning algorithm is used to construct extremely efficient classifiers. One of the other traditional methods of body appearance based human detection is Histogram of Gradient (HOG) [30] that has no prior knowledge of the structure of the human body and uses gradient-based features for description. Most traditional human detection approaches act as a classifier based on features extracted from images [19].

Recent methods of human detection are often based on deep neural networks. With the growth of hardware accelerators such as GPUs and the widespread availability of image data, training for deep learning methods have become much more enriched. Contemporary deep learning approaches can identify humans in an image more accurately than traditional methods. These in-depth learning methods include the steps of extracting potential candidate areas, showing areas, classifying areas, and integrating results.

One of the most commonly used recognition methods is Covnet, which is a CNN network. The CNN based object detector can be classified into single-stage and double-stage. Single-stage based object detector consists of one-step regression framework, which uses a CNN to predict object locations directly. The double-stage based object detector, on the other hand, extracts a set of region proposals in the first step, and in the second step applies CNN to generate the object location and desired class label. The modern object detection techniques are characterized on the basis of the region proposal stage: R-CNN, Fast R-CNN, and Mask R-CNN follow double-stage pipelining, while YOLO and SSD fall under single-stage pipelining [38].

There are two other categories of human recognition methods: Body Appearance Based and Motion Features Based, which are not commonly used in emotion recognition articles. For further study of human diagnostic methods and a classification of different approaches, interested readers are referred to [7]. 2.2 shows a complete classification of different approaches in this field.

2.3 Extraction of the human model

As described in the previous section, the human has been detected in the image and now it has to be modeled properly in order to extract the feature. Due to the flexibility

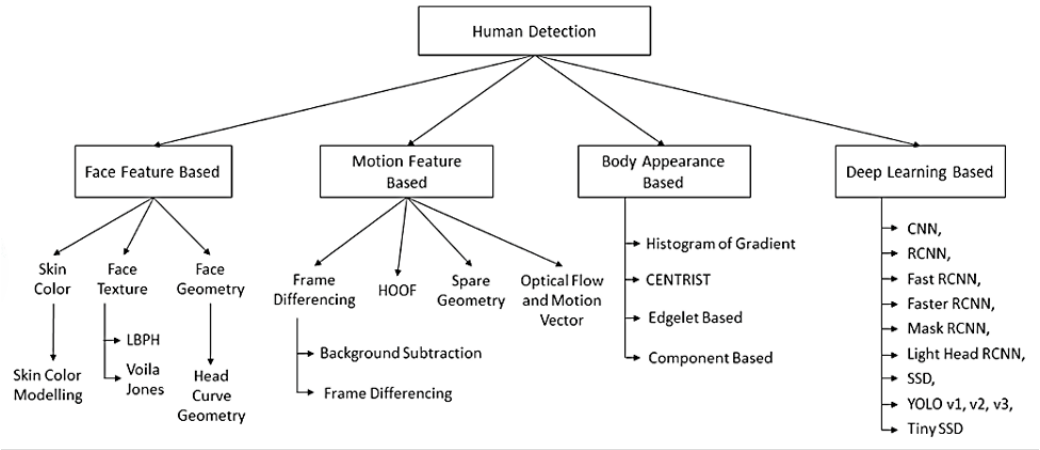

Figure 2.2: Classification of human detection techniques [7]

and freedom of the human body, conditions and changes in the environment, it is a challenging task to select and extract an appropriate and robust display model. The most significant models of displaying the human body are discussed in the following.

2.3.1 Human display model

One of the important issues in the recognition process is how to model the input (human body) and output (feeling). This modeling and representation has a big impact on system performance. The human body is a flexible and complex non-rigid object. It has many specific characteristics such as kinematic structure, body shape, surface texture, body joints position, etc. A workable model does not necessarily have to include all these attributes, though, it must meet the requirements of emotion recognition. According to various application scenarios, there are three common types of human body models: Skeleton based model, Contour-based model, and Volume based model [16] as shown in figure 2.3.

The skeleton-based model represents a set of joint (between 10 to 30) locations that show the skeletal structure of the human body. This model is very simple and flexible and is widely utilized to show two-dimensional and three-dimensional human pose in existing datasets. The main drawback of this method is the lack of texture information of a human body, thus width and contour information is not available for further processing.

In contrast, contour-based models contain rough width and contour information of body limbs and torso, and have been used extensively in earlier methods. In these

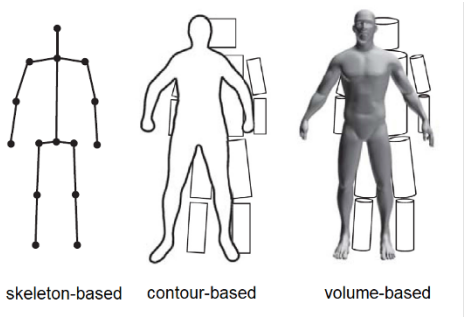

Figure 2.3: Human body models

models, human body parts are represented with rectangles or boundaries of a person's silhouette. Volume-based models represent the entire the body in 3D with geometric shapes such as cylinders or conics. Modern volume-based models are represented in mesh form and captured with 3D scans.

The main purpose of this step is to extract a predefined template for expressing a person's (or persons') body position in the input image. Based on the various models of human pose estimation, the proposed methods are classified into two categories. Regression-based methods attempt to learn a mapping from input image to kinematic body joint coordinates. In detection-based methods, body parts locations or joints are predicted using a sequence of rectangular windows, providing more stable results than regression-based ones [16]. Most of the works in the field of emotion recognition are based on the skeleton-based model.

2.3.2 Body skeleton extraction

There are several methods in skeleton detection due to its wide range of applications. Skeleton detection means identifying and classifying key points or joints in the human body. The major issues being that the human body is flexible and different parts of the body can be placed in different positions relative to each other.

Body posture detection approaches can be divided into classic method and deep learning. From another perspective, these methods can be divided into categories of 2D and 3D methods. The 2D approaches estimate the location of key points in 2D space with values of the x and y axes. 3D posture detection converts a human in an image to 3D by adding a z dimension to the prediction. There is also a distinction between detecting one or more people in an image. These two approaches are called single-state (single-person) and multi-state (multi-person) estimates.

Modeling the human body is computationally expensive and requires a large set of training data. However, modern technological advances have solved some of the issues and enable deep learning. The first method in this area is the DeepPose approach, which has made extensive changes and thus improvements in traditional methods [50]. The most popular methods to estimate 2D single-state skeleton of the human body in one image [50] or a sequence of images [15]are multi-state/single-skeleton estimation [39] and 3D skeleton estimation [6]. Both are based on deep learning and have provided good results.

Since the location and the number of people in an image is unknown, it is more difficult to estimate the pose of several people than the pose of single person. Multi-Person pose estimation methods have two approaches: (1) top-down and (2) bottom-up. Top-down approaches first detect a person and then estimates parts and calculate the pose for each person, while bottom-up approaches first detect all body parts of every person (key points) in the image, followed by associating parts belonging to distinct persons. Each approach has its advantages and disadvantages [38]. Figure 2.4 shows an overview of these two approaches [1]. In the emotion recognition, various approaches based on deep learning have been used and it is difficult to state which approach performs better.

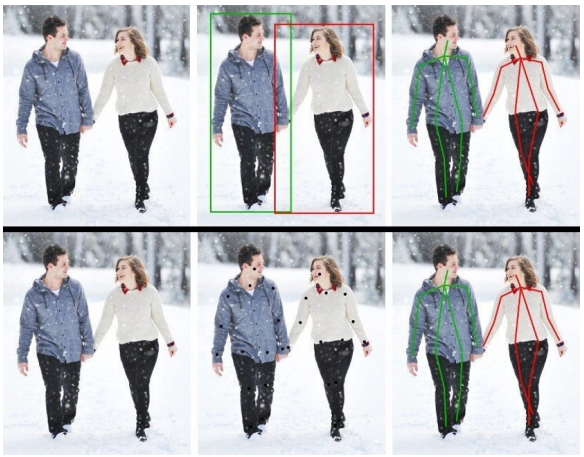

Figure 2.4: Top: Top-Down approach; Bottom: Bottom-Up approach

2.4 Feature extraction

The second stage of a emotion recognition system is the extraction of features from the human representation model as the input to the learning model. Depending on the

input and the selected model, the feature extraction method can vary. The concept of time is very important at this stage. Emotion recognition can be based on an image or a sequence of continuous images of a movement. For instance, motion-based emotions can be described by displacement, distance, speed, acceleration and time. Psychologically, the relationship between emotions and both body posture and body movements has been studied and confirmed [18]. Some of the prominent methods of feature extraction in recognizing emotions based on body motion are discussed here.

Gunes et al. [29] used skin color information of the upper torso to detect emotions. They considered spatial changes of the hands and face relative to the neutral state as their extractive features. Movement protocols are used as a distinction between emotions. For example, in a neutral position there is no movement in the upper body, but in a happy or sad state the arms move towards the head. Maret et al. [37] used the skeletal representation model of the body, and spatial, velocity, and acceleration characteristics as characteristics to distinguish emotions based on arm movements. A genetic algorithm is also used to find the best system parameters (feature selection) to achieve a higher detection rate. In another approach, [41] used a data-driven algorithm and model to estimate emotions based on the type of gait. They also used LSTM extraction features in addition to movement and posture features.

Most other research has focused on different spatial and temporal information from key points. Ahmed et al. [26] used a two-layer framework for feature selection, the input of which is a comprehensive list of all motor features including information derived from human emotion-related movement. In this approach, the first layer is used to remove unrelated traits using statistical information. The second layer is used to select the best traits using genetic algorithm. In their work, most of the features used by other researchers are given, as well as a comprehensive classification of different possible features based on the body skeleton in emotion recognition. At this stage, the model of representing the human body in an image or sequence of images becomes a comprehensible representation for classification, the next stage of the system.

2.5 Emotion recognition learning model

The last major part of the emotion recognition system is the learning model, a pattern recognition model that can determine the most likely class for each input. Since the data is labeled, it is considered as supervised learning, or regression. In most research,

emotions have been considered as the output of the system in a discrete way and the learning model has been proposed as a classification problem.

Ahmed et al. [26] used a combined approach for the learning model. The basic classification methods include SVM, LDA, DT, GNB and KNN and performance of score-level and rank-level fusion are compared. They achieved 90% accuracy for emotion recognition during walking, 96% during sitting action sequences, and 86.66% in action-independent cases. In many other works, traditional approaches to classification have been used in different ways [32] [31].

Some researchers have also used deep learning methods for the learning model. [43] used different CNN, RNN, and RNN-LSTM neural networks for the learning model, and the best accuracy is obtained by the RNN-LSTM method at 69%. [33] also used deep learning algorithms such as RBM and stacked RBM and compared with SVM and Naïve Bayse methods. The stacked RBM method is found to have the best accuracy.

2.6 Emotion display model

In emotion recognition systems, a representation of human emotion is needed so that the label of training samples, the type of estimator model and the model output type are determined accordingly. Often labels such as happy, sad, and angry are used to express emotions. However, emotion is a complex phenomenon that is constantly changing. This leads to different expressions for categorization and distinction of emotions. In general, the methods of representation or display of emotions can be divided into two forms, discrete and dimensional.

In discrete representation, emotions are represented separately in several categories. The number of emotions can be determined in two or more categories according to the application including happiness, sadness, fear, anger, surprise, hatred. Figure 2.5 shows six basic emotions expressed by the body [17]. This kind of representation is widely used in this field due to its simplicity and comprehensiveness.

In the continuous representation, human emotions are displayed in a multidimensional space. Certain points of space are considered as the main emotions (in the discrete state) and the rest of the space is considered based on a combination of these basic emotions. Figure 2.6 represents Russell's model with two dimensions: valence – arousal [40]. This approach can describe more subtle and complex emotions. Of course, this approach is more difficult to use for automated detection systems.

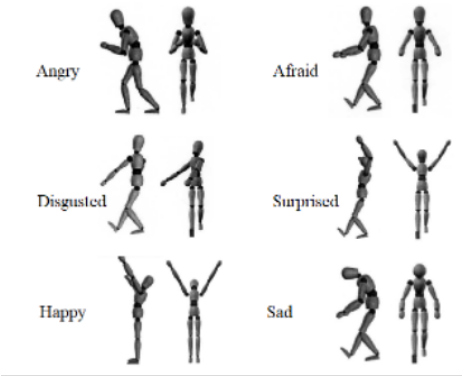

Figure 2.5: Ekman discrete display model based on body posture

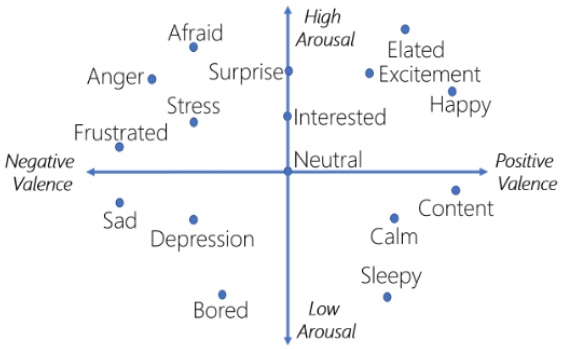

Figure 2.6: Valence–Arousal model of emotion

Other models of continuous emotion representation include Activation-Evaluation emotional space, Mehrabian and Russell's Pleasure-Arousal-Dominant (PAD) Theory and Appraisal theory. The continuous display method, in addition to being able to cover all human emotions, also leads to a higher detection rate. This approach can also be used in interactive games, and in some applications a display space with one or more limited dimensions can be used.

Other models of emotion expression include the semantic model, which is not the subject of this brief survey, is mostly theoretical and is not used in practical applications [12].

Most approaches examine categorized distinct emotions, and fewer studies refer to emotional dimensions-based recognition. Most researchers focus on a set of six main emotions (sadness, anger, fear, surprise, and disgust are present in most databases). However, many emotional states such as inattention, fear, shame and sensitivity are not found in that many databases [38]. The nature and characteristics of a generated or selected dataset largely depend on the application problem.

2.7 Evaluation (final model evaluation criteria)

One of the issues in the recognition process is how to evaluate the performance of the system. The performance of the system is evaluated based on its final output, the estimated emotion. Classification criteria can be used for discrete emotions that have been presented as a classification problem. One of these criteria is the accuracy or rate of correct recognition (Correct Classification Rate) which is equal to the number of correct recognition compared to the total test samples [32] [51]. Since the total accuracy criterion is not suitable for unbalanced data, in some research such as [12], the accuracy for each class is calculated and examined separately. For unbalanced classes, it is better to use other appropriate criteria for these data.

Another method of evaluating and displaying results is the use of the confusion matrix, from which more information can be obtained on how the classifier works. In many works with the recognition of emotion based on body posture, including [43], this method has been used. The confusion matrix is made up of a list of predicted emotions versus real emotions, and one of its advantages is its interpretability. Unfortunately, it is difficult to use this matrix to optimize a model.

Precision and recall are other classification criteria that are used in emotion recognition [51]. Precision is the number of correct predictions divided by the total number of all predictions. Recall is the number of correct predictions divided by the total number of samples in that class. These two criteria can be reported for both classes separately or as an average for all classes [4]. Averaging can be applied in two ways: micro averaging and macro averaging. In the macro mode, the desired criterion is calculated separately for each class and then mediated on all classes (all classes have the same effect on the evaluation). In micro mode, only one value is calculated and it summarizes the effect of all classes (suitable for unbalanced classes) [48].

Figure 2.7 presents computing different performance measures from Confusion Matrix. Precision and recall of each class are calculated [5].

The F1 score evaluation criterion is also one of the most important methods for evaluating the performance of a classifier, which is the weighted average of Recall and Precision. The highest value is equal to one and represents the best performance, and the lowest value is zero and represents the worst case classification efficiency. This criterion has also been used in some studies such as [51] [8] to measure the effectiveness of emotion recognition models.

This criterion is calculated as a value for two-class problems and separately for

Confusion Matrix		Predicted			False Negative (FN)	Recall
		Class 1	Class 2	Class 3		
Actual	Class 1	A	B	C	B + C	A/(A + B + C)
	Class 2	D	E	F	D + F	E/(D + E + F)
	Class 3	G	H	I	G + H	I/(G + H + I)
	False Positive (FP)	D + G	B+H	C+F	Overall Accuracy = A + E + I/(Sum of red and green squares)	
	Precision	A/(A + D + G)	E/(B + E + H)	I/(C + F + I)		

True positives. True Negatives. Misclassified cases. False Positives. False Negatives.

Figure 2.7: Different performance measures from confusion matrix

multi-class problems, and its formula is as follows:

$$F1Score = \frac{2 \times (Recall \times Precision)}{(Recall + Precision)}$$

Another evaluation criterion used in this field is the AUC, which means the area under the ROC curve. It shows the power of differentiation of the classification model in class separation. The higher the value, the higher the efficiency of the learning model. Some studies, such as [46], have used this criterion to evaluate the method of recognizing emotions based on body model.

Another criterion in evaluating and comparing methods of recognizing emotions is the time complexity of the recognition, which has been studied in some research [54]. The lower the time complexity or detection time, the better the model, and in real-world problems it is critical for real-time detection.

Researchers usually use separate training and test datasets, or data sharing methods such as leave one out cross validation (LOOCV), or k-fold cross validation. For example, in the field of emotion recognition, [3] used the LOOCV method and [26] used the 10-fold CV method for their experiments.

One of the challenges of this evaluation issue is the low efficiency of the methods introduced in uncontrolled environments. Since data is only available in controlled environments, their performance in applications such as video surveillance is challenged. Other challenges include the lack of a standard dataset, and standards to build models and compare method.

Chapter 3

Dataset

A proper dataset is essential in many research. However, there is no comprehensive and standard dataset available in recognition emotion from body motion. The existing datasets are often targeted for specific purpose and applications, and their production is time consuming and costly. Since there is no commonly benchmark, it is difficult for researchers to compare their models.

3.1 Data acquisition tools

Methods of data acquisition from the human body movement can generally be divided into perceptual and responsive. Responsive methods require special equipment on the human body and perceptual methods are more passive. Although responsive methods provide more accurate information compared to perceptual methods, they are not responsive to real-world applications.

Optical motion sensors can be placed on specific parts of the body and movement can then be tracked by a set of infrared or color (RGB) cameras [31]. Inertial motion capture systems use position, speed and acceleration sensors, bending sensors or pressure sensors on different parts of the body, thus the position of the joints and connections [43].

Another way to acquire body posture data is to use 2D and 3D cameras with image-based methods. Most of the studies in this review are based on cameras and image processing. One issue in using camera is the user's identity can easily be recognized. For privacy reasons, optical motion and inertial motion systems may be used.

3.2 Modality

Datasets can be classified into one-sided (face only or body gesture) or multifaceted (including face, voice and body movements). In the one-sided mode, emotion recognition is based on only one modality, while in the multi-sided mode, the combination of several modalities is used in the multi-side mode which often give more accurate results.

Numerous studies have been done in the recognition of emotions through human face and speech. On the other hand, little research and development has been carried in detection of emotions based on the human body. It has also been shown in some studies [55] that in a real environment the accuracy of facial modality is greater than that of speech and body.

Human body dataset for recognizing emotions may include only the upper body or the entire human body. Some focus on specific elements of the body, such as walking, to detect emotions [32].

3.3 Datasets

An ideal dataset should come from a real event or gathered in a real-world environment, be recorded covertly, and be emotion-oriented. Most of the existing datasets are not ideal. Many research work use self-generated data and are not available for review. Some multimodal datasets are not currently available [56] [24].

The most referenced datasets in the literature, for the recognition of emotion based on the state of the human body, is shown in table 3.1. The various data features, data accessibility and access links are also listed.

Name	Number of Sample	Year of generation	Number of emotion	Sensor	Multimodal (Modality)	Format	Access link
FABO [28]	206	2005	10	RGB camera	face and upper-body	video	FABO
HUMAINE [23]	240	2011	8	camera	face and body	video, sound	-
LIRIS-ACCEDE [11]	9800	2015	-	camera	face and upper-body	video, sound	LIRIS-ACCEDE
GEMEP [14]	$\geq$ 70000	2015	18	camera	face and body	video, sound	GEMEP
Emilya [27]	8026	2014	8	Xsens MVN motion capture	motion capture+ audio	video, sound	-
PACO [36]	4080	2006	4	motion capture	body	PTD-CSM	PACO
UCLIC [35]	183	2006	4	VICON (3D)	body	avatar	UCLIC
Emotional Body Motion Database [2]	1451	2014	11	camera-sensor	body	bvh-mvnx	-
OMG Emotion [10]	2400	2018	7	camera	face-body-sound	-	-
EmoPain [24]	35	2020	3	18 sensors	face and body	26 point in 3D-video	EmoPain
No Name [56]	118	2016	3	-	body	excel	Link
Action Database [34]	2783	2014	5	camera	body	HD video	Action Database
BEAST [21]	254	2011	4	camera-mat face	-	Image (bmp)	BEAST
No Name [44]	560	2018	6	kinect	posture, video, vocal	3D sckeleton (XEF)	-
EWalk [41]	1384	2019	4	camera	body	skeleton	-
IEMOCAP [13]	1384	2008	9	camera	face, upper-body, sound	video, sound, face, skeleton	IEMOCAP
MPI [52]	1447	2014	11	motion capture system	body	Bvh (3D)	MPI

Table 3.1: Datasets for emotion recognition from body motion

Chapter 4

Experiment

Interest in running high-quality CNN models under strict constraints on memory and computational budget have been raised. In this study an efficient architecture named MobileNet-V2, which is a convolutional neural network that is 53 layers deep is used [53]. Running deep networks on low compute devices such as mobile provide many advantages such as energy consumption, great user experience, and high accuracy performance. MobileNetV2 is an effective feature extractor for object detection. A real-time emotion detector is experimented using TensorFlow object detection API and python. These emotions are included excitement, sadness and neutral. This chapter presents the process of obtaining the object detection model:

- Data Collection and Labelling images;
- Train Object Detection Model;
- Real-time testing detection using webcam;

4.1 Object detection model

A transfer learning method is used against the TensorFlow Object Detection API to be able to train an object detector. In this study, the Single Shot Detector Mobilenet-V2 model is used. The ssd-mobilenet-v2-coco model is a Single-Shot multibox Detection (SSD) network intended to perform object detection. The model has been trained from the Common Objects in Context (COCO) image dataset [42]. This pre-trained model allows the function of transfer learning and trains the model faster. Single-shot multibox detector (SSD) is an object detector that is used to detect multiple objects

within a single image. The detection model is based on a feed-forward convolutional network that predicts the bounding boxes and confidence scores for each object [47]. Multiple feature layers of different sizes are used to predict the bounding box in an image. In this work, SSD is used as the object detector and MobileNetV2 as feature extractor. Figure 4.1 represents the architecture of SSD with MobileNetV2, which is used as the base network responsible for feature extraction.

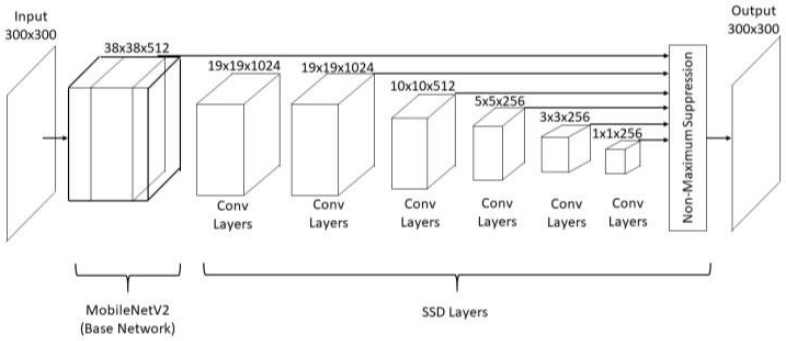

Figure 4.1: Architecture of SSD [47]

The progressive reduction in the size of feature layers allows the prediction of objects at multiple scales. The evaluation metric that is used is the loss function, which describes the confidence loss and the localization loss. Confidence loss describes the confidence of the object's prediction, while localization loss describes the offset of the default box from the center of the bounding box [47]. The pre-trained object detection model is trained using images of different poses as a training dataset.

Figure 4.2 represents three poses indicating an emotion expressed by the body. Around 400 sample image are collected for each emotion using webcam with OpenCV, which is a computer vision and machine learning library and can develop real-time computer vision applications, and python in the same condition, background, distance and space.These images have been labeled by the classes they belongs to in order to perform supervised learning and training.

4.1.1 LabelImg

Image labeling is done using the software LabelImg. FLabelImg is a graphical image annotation tool that is used to label images to prepare data for object detection. These labels are used to identify components in the image and to train the model. At this stage all the collected images (poses) are being labeled using LabelImg pack-

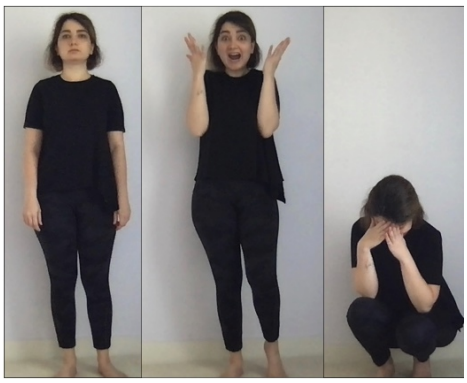

Figure 4.2: Three class of emotions (Neutral, Excitment, Sadness)

age.As figure 4.3 shows, detection boxes are drawn against different poses and the corresponding emotion is labeled.

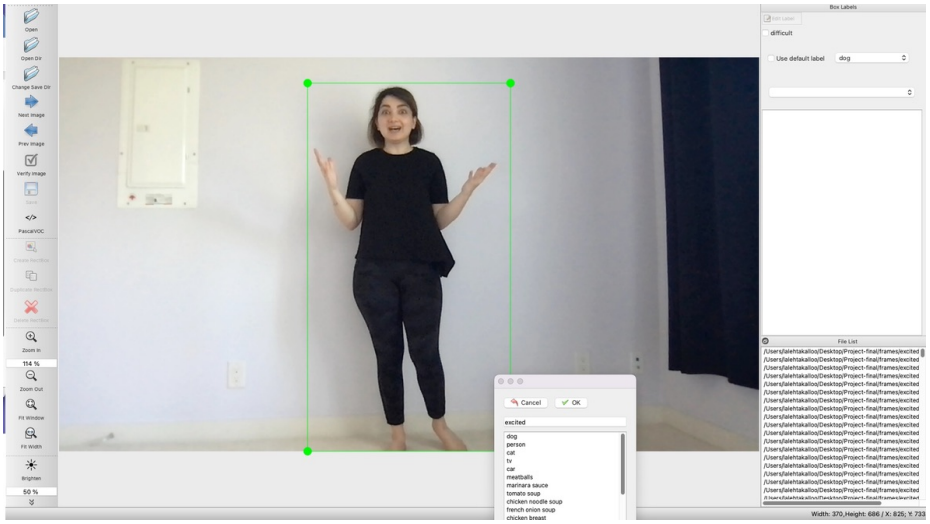

Figure 4.3: Labelling the image using labelImg

An xml file is created for each image, inside of these xml files any information needed to represent the objects are presented, as can be seen in figure 4.4 including the folder that the image is in, the file name, the path to that particular file, it's source, the size of that image (width, height, depth), the object, the selected label (where the bounding box is). The object detection model is to be trained based on these information.

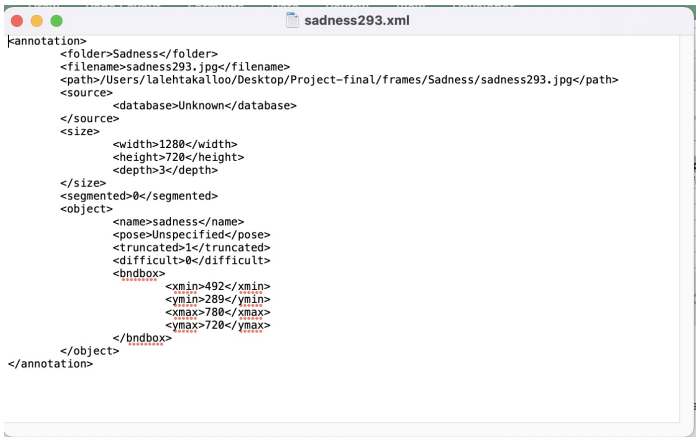

Figure 4.4: Produced xml file after labelling the image

4.1.2 Training

At this stage the images with their corresponded xml annotations should be split up into training and testing partitions so that it allows the model to be trained on a certain set of data and then being tested and evaluated on a separate partition. This will ideally help to reduce the chance of overfitting. This classification can be done by using a scientific method of sampling and selecting what is going to be in the training and testing partitions. However, in this study the selection have been done randomly, 70% of samples used for training and 30% for testing. Transfer learning is performed on a pre-trained model, SSD-MobileNet-V2, by retraining the model with the images. The training is performed until the classification-loss is 0.0909, localization-loss is 0.0158, regularization-loss is 0.1229, total-loss is 0.2296 and learning-rate is 0.07869, which took around 5000 steps and per-step time is around 0.200s.

4.1.3 Real-time detection

Figure 4.5 presents the classification of emotions from body movement using this emotion detection model.

Table 4.1 presents the performance and precision of the emotion detection model. According to analysis as shown in the table, the emotion detection model obtained an average precision of 74.34%, this can be improved by increasing the number of sample images.

Figure 4.5: Accuracy rate for emotion detection from body movement

Emotion	**First Precision**	**Second Precision**	**Third Precision**	**Average Precision**
Excitement	87%	79%	81%	82.34%
Sadness	86%	90%	89%	88.34%
Neutral	50%	54%	53%	52.34%

Table 4.1: Performance and precision of emotion detection

Chapter 5

Conclusions

In this report, recognizing emotions based on human body movement is presented and discussed. The basic concepts of body language, emotions, and their relationship are covered. The framework and components of an emotion recognition system are introduced. The datasets often cited in the literature are reviewed and the most referenced ones are presented with their characteristics. Also, TensorFlow-based object detection was used to perform emotion recognition and classification. The pre-trained object detection model, SSD-Mobilenet-V2, was able to obtain satisfactory precision. This model was able to detect emotions according to classes of excitement, sadness and neutral with an average accuracy of 74.34%.

5.1 Future work

Availability of dataset to the general public is a major issue in this field. There are also important parameters and characteristics for the dataset that have made this issue challenging. These characteristics include environmental conditions, gender, culture, nationality, and number of people, unrestricted or controlled environment, number of emotions. Also, there is no standard approach to labeling samples in such datasets, and there is no standardization, adaptation, or standardization between researchers. These heterogeneity problems make it difficult or impossible to compare the quality of datasets. The next step would be to create a large and available dataset with the correct labeling which covers every possibility and provides a high accuracy.

Bibliography

[1] An Overview of Human Pose Estimation with Deep Learning. https://beyondminds.ai/blog/an-overview-of-human-pose-estimation-with-deep-learning/.

[2] Emotional Body Motion Database. http://ebmdb.tuebingen.mpg.de/.

[3] Ferdous Ahmed, Brandon Sieu, and Marina Gavrilova. Score and rank-level fusion for emotion recognition using genetic algorithm. *IEEE 17th International Conference on Cognitive Informatics Cognitive Computing*, pages 46–53, 2018.

[4] Sharifa Alghowinem, Roland Goecke, Jeffrey Cohn, Michael Wagner, Gordon Parker, and Michael Breakspear. Cross-cultural detection of depression from nonverbal behaviour. *11th IEEE International conference and workshops on automatic face and gesture recognition*, 1:1–8, 2015.

[5] Muhammad Ali, Dae-Hee Son, Sang-Hee Kang, and Soon-Ryul Nam. An accurate ct saturation classification using a deep learning approach based on unsupervised feature extraction and supervised fine-tuning strategy. *Energies*, 10:1830, 2017.

[6] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. *IEEE conference on computer vision and pattern recognition*, pages 8387–8397, 2018.

[7] Mohd Ansari and Dushyant Kumar Singh. Human detection techniques for real time surveillance: A comprehensive survey. *Multimedia Tools and Applications 80*, pages 8759–8808, 2021.

[8] J Arunnehru and Kalaiselvi Geetha. Automatic human emotion recognition in surveillance video. *Intelligent Techniques in Signal Processing for Multimedia Security*, pages 321–342, 2017.

[9] Anthony Atkinson, Winand Dittrich, Andrew Gemmell, and Andrew Young. Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception*, 33:717–746, 2004.

[10] Pablo Barros, Nikhil Churamani, Egor Lakomkin, Henrique Siqueira, Alexander Sutherland, and Stefan Wermter. The omg-emotion behavior dataset. *IEEE International Joint Conference on Neural Networks*, pages 1–7, 2018.

[11] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6:43–55, 2015.

[12] D.r Bernhardt. Emotion inference from human body motion. Technical report, University of Cambridge, Computer Laboratory, 2010.

[13] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

[14] Tanja Bänziger, Marcello Mortillaro, and Klaus Scherer. Introducing the geneva multimodal expression corpus for experimental research on emotion perception. *Emotion*, 12:1161, 2012.

[15] James Charles, Tomas Pfister, Derek Magee, David Hogg, and Andrew Zisserman. Personalizing human video pose estimation. *IEEE conference on computer vision and pattern recognition*, pages 3063–3072, 2016.

[16] Yucheng Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, 192, 2020.

[17] Mark Coulson. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Nonverbal behavior*, 28:117–139, 2004.

[18] Elizabeth Crane and Melissa Gross. Motion capture and emotion: Affect detection in whole body movement. *International Conference on Affective Computing and Intelligent Interaction*, pages 95–101, 2007.

[19] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. *European conference on computer vision*, pages 428–441, 2006.

[20] B. De Gelder. Why bodies? twelve reasons for including bodily expressions in affective neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364:3475–3484, 2009.

[21] Beatrice De Gelder and Jan Van den Stock. The bodily expressive action stimulus test (beast). construction and validation of a stimulus basis for measuring perception of whole body expression of emotions. *Frontiers in psychology*, 2:181, 2011.

[22] P. Ravindra De Silva and Nadia Bianchi-Berthouze. Modeling human affective postures: an information theoretic characterization of posture features. *Computer animation and virtual worlds*, 15:269–276, 2004.

[23] Ellen Douglas-Cowie, Cate Cox, Jean-Claude Martin, Laurence Devillers, Roddy Cowie, Ian Sneddon, and Margaret McRorie. The humaine database. *Emotion-Oriented Systems*, pages 243–284, 2011.

[24] Joy Egede, Siyang Song, Temitayo Olugbade, Chongyang Wang, De Amanda, Hongying Meng, Min Aung, Nicholas Lane, Michel Valstar, and Nadia Bianchi-Berthouzed. Emopain challenge 2020: Multimodal pain evaluation from facial and bodily expressions. *15th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 849–856, 2020.

[25] J. L. Elman. *Encyclopedia of Language and Linguistics–2nd Edition*. Elsevier, 2005.

[26] Ahmed Ferdous, Bari ASM Hossain, and Gavrilova Marina. Emotion recognition from body movement. *IEEE Access*, 8:11761–11781, 2019.

[27] Nesrine Fourati and Catherine Pelachaud. Emilya: Emotional body expression in daily actions database. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 3486–3493, 2014.

[28] Hatice Gunes and Massimo Piccard. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. *18th IEEE International conference on pattern recognition*, 1:1148–115, 2006.

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

[30] M Kachouane, S Sahki, M Lakrouf, and N Ouadah. Hog based fast human detection. *24th International Conference on Microelectronics (ICM)*, pages 1–4, 2012.

[31] Asha Kapur, Ajay Kapur, Naznin Virji-Babul, George Tzanetakis, and Peter Driessen. Gesture-based affective computing on motion capture data. *International conference on affective computing and intelligent interaction*, pages 1–7, 2005.

[32] Michelle Karg, Kolja Kühnlenz, and Martin Buss. Recognition of affect based on gait patterns. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 40*, pages 1050–1061, 2010.

[33] Kyriaki Kaza, Athanasios Psaltis, Kiriakos Stefanidis, Konstantinos Apostolakis, Spyridon Thermos, Kosmas Dimitropoulos, and Petros Daras. Body motion analysis for emotion recognition in serious games. *International Conference on Universal Access in Human-Computer Interaction*, pages 33–42, 2016.

[34] Bruce Keefe, Matthias Villing, Chris Racey, Samantha Strong, Joanna Wincenciak, and Nick Barraclough. A database of whole-body action videos for the study of action, emotion, and untrustworthiness. *Behavior research methods*, 46:1042–1051, 2014.

[35] Andrea Kleinsmith, Ravindra De Silva, and Nadia Bianchi-Berthouze. Cross-cultural differences in recognizing affect from body posture. *Interacting with computers*, 18:1371–1389, 2006.

[36] Yingliang Ma, Helena Paterson, and Frank Pollick. A motion capture library for the study of identity, gender, and emotion perception from biological motion. *Behavior research methods*, 38:134–141, 2006.

[37] Yann Maret, Daniel Oberson, and Marina Gavrilova. Identifying an emotional state from body movements using genetic-based algorithms. *International Conference on Artificial Intelligence and Soft Computing*, pages 474–485, 2018.

[38] Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari. Survey on emotional body gesture recognition. *IEEE transactions on affective computing 12*, pages 505–523, 2018.

[39] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. *IEEE conference on computer vision and pattern recognition*, pages 4929–4937, 2016.

[40] Jonathan Posner, James Russell, and Bradley Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17:715–734, 2005.

[41] Tanmay Randhavane, Uttaran Bhattacharya, Kyra Kapsaskis, Kurt Gray, Aniket Bera, and Dinesh Manocha. Identifying emotions from walking using affective and deep features. *arXiv preprint arXiv:1906.11884*, 2019.

[42] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[43] Tomasz Sapiński, Dorota Kamińska, Adam Pelikant, and Gholamreza Anbarjafari. Emotion recognition from skeletal movement. *Entropy*, 21:646, 2019.

[44] Tomasz Sapiński, Dorota Kamińska, Adam Pelikant, Cagri Ozcinar, Egils Avots, and Gholamreza Anbarjafari. Multimodal database of emotional speech, video and gestures. *International Conference on Pattern Recognition*, pages 153–163, 2018.

[45] Konrad Schindler, Luc Van Gool, and Beatrice De Gelder. Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural networks*, 21:1238–1246, 2008.

[46] Min SH Sebastian Kaltwang, Bernardino Romera-Paredes, Brais Martinez, Aneesha Singh, Matteo Cella, and Michel Valstar. The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal emopain dataset. *IEEE transactions on affective computing*, 7:435–451, 2015.

[47] Teoh Sheng, Mohammad Shahidul, Norbahiah Misran, Hafiz Baharuddin, Haslina Arshad, Rashedul Islam, Muhammad Chowdhury, and Hatem Rmiliand Mohammad Islam. An internet of things based smart waste management system using lora and tensorflow deep learning model. *IEEE*, 8:148793–148811, 2020.

[48] Jiaqi Shi, Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro. Skeleton-based emotion recognition based on two-stream self-attention enhanced spatial-temporal graph convolutional network. *Sensors*, 21:205, 2021.

[49] Duc Thanh Nguyen, Wanqing Li, and Philip Ogunbona. Human detection from images and videos: A survey. *Pattern Recognition*, 51:148–175, 2016.

[50] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. *IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.

[51] Carla Viegas. Two stage emotion recognition using frame-level and video-level features. *15th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 912–915, 2020.

[52] Ekaterina Volkova, Stephan De La Rosa, Heinrich Bülthoff, and Betty Mohler. The mpi emotional body expressions database for narrative scenarios. *PloS one*, 9:e113647, 2014.

[53] Robert Wang, Xiang Li, and Charles Ling. Pelee: A real-time object detection system on mobile devices. *Advances in neural information processing systems*, 31, 2018.

[54] Weiyi Wang, Valentin Enescu, and Hichem Sahli. Adaptive real-time emotion recognition from body movements. *ACM Transactions on Interactive Intelligent Systems*, 5:1–21, 2015.

[55] Gou Wei, Li Jian, and Sun Mo. Multimodal (audio, facial and gesture) based emotion recognition challenge. *15th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 908–911, 2020.

[56] Xinhui Yuan and Marwa Mahmoud. Alanet: Autoencoder-lstm for pain and protective behaviour detection. *15th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 824–828, 2020.