



Visual Scene Context in Emotion Perception

DISSERTATION

Submitted to the Doctoral Programme in Network
and Information Technologies of the Universitat
Oberta de Catalunya in partial fulfillment of the
requirements for the degree of Doctor of Philosophy.

Author: Ronak Kosti

Thesis Director: Àgata Lapedriza

April 2019



© Copyright Ronak Kosti, 2019¹

¹all the ideas mentioned in this thesis are that of the author and his co-authors, except where the contribution of other research work is duly acknowledged

Acknowledgments

It would be dishonest to say that only countable people made this thesis possible for me. All of my interactions with my surroundings (social, cultural, academic, in the nature and otherwise) influenced this work in one way or the other. Having said that, there are people whose presence has been of utmost importance for me personally.

First and the foremost person who comes to mind is my supervisor, *Agata Lapedriza*. She has been the biggest influence in this journey. Her insights and inputs during the development of the whole thesis has been remarkable. I am forever grateful for her guidance in transforming myself from a *student* into a *researcher*. In addition, I would also like to thank *Jose M. Alvarez* for his technical insights, specifically in the development of code base for this thesis. I also thank my co-author *Adria Recasens* for his contributions.

The city of Barcelona with it's mountains, wonderful valleys, beaches and a unique cultural mix has provided an unparalleled environment to live and grow. I've made everlasting friendships and found amazing people whom I consider to be my family in this very city. I would specially like to thank Fernanda, Rosen, Samia, Pilar, Leila, Dani, Marta, Isuru, Greig, Nimesh, Manju, Waseem, Blanca, Joan, Raquel, Meritxell, Pedro, Amir, Negar, Eunice, Krizia, Eliza and Vilma.

An unflinching and continuous support to pursue this dream has come from my family. My parents and my brother are equal partners in the culmination of this dream of pursuing a doctorate. They have always encouraged and cared for me. I have their support in every stage of my life.

I thank all of my colleagues at UOC² whose help made me get things done faster. In the end, I want to thank my companion *Agata Bicz* for her understanding and continuous support. Her presence has made the future look more affirmative.

This thesis has been supported by the UOC Doctoral School Grant [November.2015 - October.2018]. This work has also been partially supported by the *Ministerio de Economía, Industria y Competitividad (Spain)*, under the Grant Ref. TIN2015-66951-C2-2-R. Thanks to NVIDIA as well for their generous hardware donations.

²Universitat Oberta de Catalunya, <https://www.uoc.edu/>

Our faculties of perception are limited even for simple things...

Stanisław Lem

Abstract

Recognizing emotion comes naturally to us. We are able to read people’s feelings quite well, from their facial expressions, behavior, body posture, appearance and social interaction with others. We often try to perceive or recognize, subconsciously and continuously, what people might be feeling or what are the emotional states of people in specific situations. This ability helps us understand what people are feeling and, depending on their emotional state, helps us respond appropriately. For example, if a person is sad and feeling unhappy, instinctively we are ready to offer our support and empathize with him. With a view to imparting such capability to machines, computer vision researchers have developed automatic emotion recognition techniques based on a person’s facial expressions and, in some cases, the body posture. Some of these methods work remarkably well in specific surroundings. However, their performance is limited in natural, unconstrained environments. Recent studies in psychology show that the scene context, in addition to facial expression and body posture, contributes essential information to our perception of people’s emotions. However, the processing of the context for automatic emotion recognition has not been explored in depth, partly due to the lack of proper data.

We present EMOTIons in Context (EMOTIC), a dataset of images of people in natural and diverse situations annotated with their apparent emotion. The EMOTIC database combines two different types of emotion representations: (1) 26 Emotion Categories, and (2) 3 Continuous Dimensions (*Valence, Arousal, and Dominance*). We present a detailed statistical and algorithmic analysis of the dataset along with the annotators’ agreement

analysis. We also develop a fusion Convolutional Neural Network (CNN) model which takes the person and his surrounding scene (context) as inputs to predict about his emotional state. We train this model on EMOTIC and analyse different configurations of the CNN model.

When the model is trained using both the format of emotion representation, the prediction performance is better when the context is present. Our results show that scene context contributes important information to automatically recognize emotional states and motivate further research in this direction.

List of contributions

Contributions of the Ph.D. research³

The validation of the research has been carried out with the publication of 3 original papers. All analysis and experimental or simulation results presented in the publications of this thesis have been produced by the author. The list of **published papers** is the following:

1. Ronak Kosti, Jose M. Alvarez, Adria Recasens, Agata Lapedriza. "Emotion Recognition in Context". *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1667-1675*
2. Ronak Kosti, Jose M. Alvarez, Adria Recasens, Agata Lapedriza. "EMOTIC: Emotions in Context Dataset". *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2017, pp. 61-69*
3. Ronak Kosti, Jose M. Alvarez, Adria Recasens, Agata Lapedriza. "Context Based Emotion Recognition using EMOTIC Dataset". *IEEE Transactions on Pattern Analysis and Machine Intelligence.* doi: 10.1109/TPAMI.2019.2916866

³Code, dataset and trained models available at: <https://github.com/rkosti/emotic>

Contents

| | |
|--|-----|
| Abstract | vii |
| List of contributions | ix |
| List of Figures | xv |
| List of Tables | xix |
| 1 Introduction | 1 |
| 1.1 Emotion Recognition | 1 |
| 1.1.1 Automatic Emotion Recognition and Applications | 3 |
| 1.2 Role of Context in Emotion Recognition | 5 |
| 1.2.1 Sources of Information for Emotion Recognition | 9 |
| 1.3 Motivation for Emotion Recognition | 12 |
| 1.4 Thesis Outline and Contributions | 13 |
| 2 Previous Research on Emotion Recognition from Images | 17 |
| 2.1 Emotion Representation Formats | 18 |
| 2.2 Emotion Recognition from Images | 20 |
| 2.2.1 Facial Expression Based Approaches | 20 |
| 2.2.2 Body Posture Based Approaches | 21 |
| 2.2.3 Group-level and whole-Image based Approaches | 23 |
| 2.3 Image-based Datasets for Emotion Recognition | 24 |
| 2.3.1 Shortcomings of the current Datasets | 26 |
| 3 EMOTIC Dataset | 29 |
| 3.1 EMOTIC Dataset Construction | 30 |
| 3.1.1 Image Data Collection | 31 |
| 3.1.2 Emotion Representation format for EMOTIC | 32 |
| 3.1.2.1 Continuous Dimensions (or Affect Dimensions) | 32 |

| | | |
|----------|---|-----------|
| 3.1.2.2 | Emotion Categories | 33 |
| 3.1.2.3 | Combined Emotion Representation for EMOTIC | 36 |
| 3.1.3 | Collecting Annotations | 36 |
| 3.1.3.1 | Interface Design | 37 |
| 3.1.3.2 | Annotation Quality Control Strategies | 46 |
| 3.2 | Analysis | 49 |
| 3.2.1 | Statistical Analysis | 51 |
| 3.2.2 | Annotator Agreement Analysis | 56 |
| 3.2.3 | Algorithmic Analysis | 60 |
| 4 | Modeling Emotion on EMOTIC dataset | 65 |
| 4.1 | Architectural Design | 65 |
| 4.1.1 | Person Features | 67 |
| 4.1.2 | Visual Scene-Context Features | 69 |
| 4.1.3 | EMOTIC Fusion Model | 70 |
| 4.2 | Multitask Learning (MTL) | 72 |
| 4.3 | Loss criterions and Evaluation metrics | 73 |
| 4.3.1 | Criterion for Emotion Categories (L_{disc}) | 73 |
| 4.3.2 | Criterions for Continuous Dimensions (L_{2cont} , SL_{1cont}) | 74 |
| 4.3.3 | Combined Criterions (L_{comb1} , L_{comb2}) | 75 |
| 4.3.4 | Performance Evaluation Metrics | 75 |
| 4.3.4.1 | Average Precision (AP) | 76 |
| 4.3.4.2 | Average Absolute Error (AE) | 76 |
| 5 | Experiments and Analysis | 79 |
| 5.1 | Experiments | 79 |
| 5.1.1 | B Model Experiments (person features) | 80 |
| 5.1.2 | I Model Experiments (visual scene features) | 80 |
| 5.1.3 | B + I Model Experiments (combined features) | 82 |
| 5.2 | Results' Analysis | 83 |
| 5.2.1 | B Model Analysis | 85 |
| 5.2.2 | I Model Analysis | 85 |
| 5.2.3 | B + I Model Analysis | 85 |
| 5.2.4 | Analysis Summary | 86 |
| 5.3 | Experiments with other architectures: SHG and Resnets | 86 |
| 5.4 | Discussions | 87 |
| 5.4.1 | Experiments with L_{comb2} (BEST EMOTIC MODEL) | 87 |

| | |
|---|------------|
| CONTENTS | xiii |
| 5.4.1.1 Quantitative Evaluation | 89 |
| 5.4.1.2 Qualitative Evaluation | 91 |
| 5.4.2 Sentibanks as Visual Context Features | 93 |
| 5.4.2.1 Context Features' Comparison | 95 |
| 6 Conclusions and Future Outlook | 97 |
| 6.1 Main Conclusions | 97 |
| 6.2 Future Outlook and Concluding Remarks | 99 |
| Appendices | 101 |
| A Comparison of Emotion Categories | 103 |
| B Facial Expressssions' Datasets | 107 |
| C Various sources of contextual information | 109 |
| Bibliography | 113 |

List of Figures

| | | |
|-----|--|----|
| 1.1 | Two people doing the same activity of reading a book in different surroundings. Depending on their surroundings, their perceived emotional states are different | 1 |
| 1.2 | Perceived emotion of the person changes from <i>face</i> → <i>visible body</i> → <i>whole image</i> | 4 |
| 1.3 | Examples showing people with different facial views, along with their body posture and the surrounding scene. What can be said about the emotional state of these people? | 6 |
| 1.4 | Emotion perception of two observers presented via <i>Emotion Categories</i> and <i>Natural Language</i> | 8 |
| 1.5 | Visual information for emotion recognition changes with the change in face poses, for a fixed facial expression (Ouamane [2015]) | 10 |
| 1.6 | Keeping the facial expression fixed, the body posture influences the perceived emotion of the person (Aviezer et al. [2017]) | 11 |
| 1.7 | Six different hand gestures suggesting distinct emotions (Link) | 11 |
| 1.8 | Type of the surrounding visual scene impacts the emotion perception. For example, the 2 people doing the same activity in different scenes are perceived to have distinct emotional states | 12 |
| 2.1 | Example for emotion perception for the person-in-context. The perceived emotion is using <i>Free Form</i> , <i>Affect Dimensions</i> and <i>Emotion Categories</i> . . | 19 |
| 2.2 | Different body postures that represent various emotion stimuli, while maintaining a uniform background (Adams Jr and Kleck [2005]) | 22 |
| 2.3 | Sample images from EMOTIC, CK (Kanade et al. [2000]) and EMO-TIONET (Fabian Benitez-Quiroz et al. [2016]) datasets in rows (a), (b) and (c) respectively | 27 |
| 3.1 | Example images from EMOTIC: The person-in-context is enclosed in a rectangular bounding-box | 31 |

| | |
|---|----|
| 3.2 Examples of annotated images in EMOTIC dataset for each of the 3 continuous dimensions viz. Valence, Arousal & Dominance. The person in the red bounding box has the corresponding value of the given dimension, mentioned at the top of each image | 33 |
| 3.3 Examples of annotated people in EMOTIC dataset for each of the 26 emotion categories (Table 3.1). The person in the red bounding box is annotated by the corresponding category. | 34 |
| 3.4 EQ task design: First Page showing disclaimers and instructions | 39 |
| 3.5 EQ task design: Main page of EQ task asking the general questions | 40 |
| 3.6 EQ task design: Warning message for EQ task if any question is missed (30th, in the above example) | 40 |
| 3.7 EC task design: First page showing disclaimer and instructions about the browser settings | 41 |
| 3.8 EC task design: Page showing instructions on how to attempt the task . . | 41 |
| 3.9 EC task design: Page showing the correct and incorrect ways of annotation | 42 |
| 3.10 CD task design: Page showing the instructions for CD task | 43 |
| 3.11 CD task design: Page showing the visual definition of Valence | 43 |
| 3.12 CD task design: Page showing the visual definition of Arousal | 44 |
| 3.13 CD task design: Page showing the visual definition of Dominance | 44 |
| 3.14 CD task design: Page showing an example on how to attempt the CD task | 45 |
| 3.15 CD task design: Page showing a correct way of annotating in CD task . . | 45 |
| 3.16 CD task design: Page showing an incorrect way of annotating in CD task . | 46 |
| 3.17 AMT interface designs | 47 |
| 3.18 Quality Control for EQ task by asking 2 trivial questions ((a) or (b)) and the warning sign (c) that doesn't allow to proceed if these questions have not been answered correctly | 48 |
| 3.19 Sample Images from EMOTIC dataset with their corresponding annotations in both the formats | 50 |
| 3.20 EMOTIC Statistics: Number of people annotated for each emotion category | 52 |
| 3.21 EMOTIC Statistics: Number of people annotated for every value of the three continuous dimensions (a,b,c) | 53 |
| 3.22 Co-variance between 26 emotion categories. Each row represents the occurrence probability of every other category given the category of that particular row. | 54 |

| | | |
|------|--|----|
| 3.23 | Distribution of continuous dimension values across emotion categories. Average value of a dimension is calculated for every category and then plotted in increasing order for every distribution. | 55 |
| 3.24 | Five different annotators for a given person in context | 56 |
| 3.25 | Representation of agreement between multiple annotators. Categories are sorted in decreasing order according to the average number of annotators that agreed for the category. | 58 |
| 3.26 | (a) Kappa values and (b) Standard deviation (Std), for each annotated person in validation set | 59 |
| 3.27 | Emotion distributions conditioned to a) image scene category, and b) image scene attribute | 61 |
| 3.28 | Summary of places and attributes with the highest and lowest values of Valence, Arousal and Dominance. | 62 |
| 3.29 | Illustration of 2 current scene-centric methods for extracting contextual features from the scene: AlexNet Places CNN outputs (place categories and attributes) and Sentibanks ANP outputs for three example images of the EMOTIC dataset. | 63 |
| 4.1 | Basic Alexnet (Krizhevsky et al. [2012]) with 5 Convolutional (Conv) layers for feature extraction and 3 Fully Connected (FC) layers for classification . | 68 |
| 4.2 | Person module based on Alexnet. $C*$ represent the combination of Conv + Recti-Linear Unit (ReLU) layers. $PL*$ are a combination of Pooling layer + Local Response Normalization layers | 68 |
| 4.3 | Filter weights of the Conv 1 layer of the basic Alexnet (Krizhevsky et al. [2012]), displaying various filter weights that help to extract low-level features | 69 |
| 4.4 | Scene module based on Alvarez and Petersson [2016]. $C*$ are Conv layers each followed by a ReLU layer. $CB*$ are a combination of Batch Normalization + ReLU layer | 70 |
| 4.5 | EMOTIC-CNN Fusion model trained with L_{comb1}/L_{comb2} criterion | 71 |
| 5.1 | B model configurations for different experiments (a, b, c) | 81 |
| 5.2 | I model configurations for different experiments (a, b, c) | 82 |
| 5.3 | B+I model configurations for different experiments (a, b) | 83 |
| 5.4 | JC and AE on the Test Set, along with comparisons for different models. The results are sorted with decreasing values of JC and AE | 90 |

| | | |
|-----|---|-----|
| 5.5 | Comparing predictions of $\mathbf{B}(L_{disc}, L_{comb1})$ and $\mathbf{B+I}(L_{comb2})$ models with <i>high</i> and <i>low</i> <i>JC</i> values. Red indicates incorrectly predicted emotion categories | 92 |
| 5.6 | Examples of images that evoke sentiments through their visual content | 94 |
| C.1 | Examples where part of the person’s body is not visible. However, due to prior knowledge, it is easy to predict what they are doing | 110 |
| C.2 | Frames from the movie Forrest Gump (Gump [1994]), showing the protagonist recounting a story from his past . His speech is transcribed into subtitles | 111 |
| C.3 | Emotional state of a person is influenced by the kind of activity being performed. For example, the perceived arousal level for the person doing stunts (a) is higher than the people playing chess (b), whereas it is the lowest for the sleeping girl (c) | 111 |
| C.4 | The gist of the social surrounding affects emotional states of the people in it (Dhall et al. [2017]). The happiness level of the people drinking beers (a) is higher than the family eating together (b), whereas that of people suffering (c) is the lowest | 112 |

List of Tables

| | | |
|-----|---|-----|
| 2.1 | Various publicly available facial expression datasets with their descriptions and data quantity | 25 |
| 3.1 | Proposed emotion categories with definitions. | 35 |
| 3.2 | Instruction summary for each HIT | 46 |
| 5.1 | Average Absolute Errors (<i>AE</i>) for various models, comparing performance of each with L_{2cont} and L_{comb1} criterions | 84 |
| 5.2 | Average Precision (<i>AP</i>) for various models, comparing performance of each with L_{disc} and L_{comb1} criterions | 84 |
| 5.3 | Average Precision (<i>AP</i>) obtained on test set per category. Comparing performance of B (L_{disc}) and B + I (L_{comb1}) models with their L_{comb2} counterparts | 88 |
| 5.4 | Average Absolute Error (<i>AE</i>) obtained on test set per each continuous dimension. Comparing performance of B (L_{disc}) and B + I (L_{comb1}) models with their L_{comb2} counterparts | 88 |
| A.1 | Emotion Categories' Comparisons between (A) and (B) | 104 |
| A.2 | Emotion Categories' Comparisons between (C) and (D) | 105 |
| B.1 | Various publicly available facial expression datasets with their references and Weblinks | 107 |

Chapter 1

Introduction

1.1 Emotion Recognition

In our daily lives, we are continuously engaged in trying to assess the emotions of people we interact with. We process the information about the person and his surroundings received through our sensory inputs like vision. While trying to determine the emotional state of the person, we often wonder why do people feel the way they do? Why is someone happy *and* why is anyone sad? We are affected by these feelings and quite often rely on them. More interestingly, what are the causes for feeling what we feel? Are feelings hard-coded inside us or do they change according to the situation and the environment we find ourselves in? For example in Figure 1.1 there are 2 people doing a common activity *i.e.* ‘reading a book’. When we pose the question: *what are they feeling or thinking?*, we



(a) Reading a book in the park



(b) Reading a book in the office

Figure 1.1: Two people doing the same activity of reading a book in different surroundings. Depending on their surroundings, their perceived emotional states are different

get different answers for each of the person. We perceive those differences due to their

surroundings. The person in Figure 1.1.a is reading in the park, and by the looks of his posture and the clothes he is wearing, he seems to be in a *relaxed* mood; whereas the person in Figure 1.1.b is reading in an office and his formal clothes along with his posture gives us the impression that he is in *work* mode. However, observe that both of them are *engaged* in their respective tasks. We see here that the difference of perception in their emotional state is caused by multiple reasons, but their surrounding environment plays an important part.

Complexity of our Emotion Perception The origins of emotions or rather the reasons behind elicitation of our feelings has been an interesting area of research. These basic questions have been investigated by many philosophers since the time of Plato and Aristotle (de Sousa [2017]). Later on, Descartes (Damasio [2002]) deconstructed the spectrum of human emotions into a few elemental components out of which all other emotions are synthesised. Although the inquiry was speculative and philosophical in nature, their attempts made way for future researchers to finding the underlying structure of human emotions. Darwin, a naturalist and a meticulous analyst (one of the most prominent figure in human history), was also puzzled by the richness of human feelings and their social constructs. He postulated that there might be universal basis for all emotions, that there are fixed set of emotions decipherable from facial expressions. He also investigated display of feelings in animals (Darwin [1998]).

During 1970's, psychologists Ekman and Friesen, inspired by Darwin's work, came up with a definitive coding of facial expressions into action units called *Facial Action Coding System* (FACS, Ekman and Friesen [1969]) which pushed forward emotion research. Their work on facial expression recognition helped in augmenting our understanding of human emotions and inspired the future research in human behaviour analysis. Recently, Hassin et al. [2013] and Aviezer et al. [2008a] studied the effect of body and context in the emotion perception. Their findings showed the importance and influence of multiple external sources for emotion perception.

Neuroscientific point of view, the researchers investigated the processes responsible for emotion elicitation in the brain. The limbic system in the brain, called the emotional centre of the brain, is made of *amygdala*, *hippocampus* and *hypothalamus* functional units (Stephani [2014]). *Amygdala* helps in processing the emotions like *Fear*, *Anxiety* and *Pleasure*; *Hippocampus* provides mechanisms for storing past experiences in the form of memories and; *Hypothalamus* controls the motor functions, including emotional responses. Together, they form the limbic functional system that helps us maintain our emotional health.

Computer scientists have also been engaged in emotional analysis, specifically com-

puting human emotions through expressions in face (Masuda et al. [2008]), voice (Aviezer et al. [2008b]), body pose (Bänziger et al. [2009a], Righart and De Gelder [2008]), EEG (Jirayucharoensak et al. [2014]) and text (Chen et al. [2014]). These recent advancements in computing technologies and the growing number of it's applications has made emotion recognition much more computationally viable than ever before. Many existing systems or applications exist that have need for better automatic emotion recognition.

Darwin (Darwin [1998]) posed a question: *Why do our emotions have variable forms of expression?* Emotional expressions once served particular functions (e.g. baring teeth in anger to prepare for attack), but now accompany particular emotions because of their usefulness in communicating one's feelings and thoughts. Taking inspiration from Darwin's work and building upon it, Paul Ekman (Ekman and Friesen [1969]) conducted a benchmark study which laid the foundation of popularly known basic emotions (*usually listed as happiness, sadness, fear, anger, surprise, and disgust*). However, this study excludes emotions that have higher cognitive processes like jealousy, envy, etc. These emotions show the sophisticated nature of our emotion perception (de Sousa [2017]).

During the early periods of human civilization, *Fear* played a central role in the survival of our species. It was essential to have non-cognitive reflexes for survival, and *Fear* helped elicit such reflexes. For example when a primitive man (defenseless) sees a lion approaching towards him, he knows immediately that he is in mortal danger and he takes action that can save his life. However, today, a similar effect of *Fear* is observed in different situations. A person afraid of cockroaches might go crazy, despite the small size and the harmlessness of the insect. These situations are very different from one another, however, the emotion *Fear* is observed in both the cases. Why should such different circumstances induce the same emotion? Is there any underlying factor that connects the two cases? It can be argued that the intensity of *Fear* in both the cases is variable. These cases show the sophisticated nature of our emotional structure.

1.1.1 Automatic Emotion Recognition and Applications

Recognizing emotion comes naturally to us. We are able to read people's feelings quite well, from their facial expressions, behaviour, body posture, appearance and social interaction with others. We do this recognition task subconsciously and continuously in our daily lives. It is needless, therefore, to emphasise the importance of recognizing emotional state of a person. It not only helps to understand what people are feeling in general, but also to respond appropriately. For example, if a person is sad and feeling unhappy, instinctively we are ready to offer our support and empathize with him.

When we look at a person it is very easy for us to put ourselves in his situation,

and even to *feel*, to some degree, things that this person appears to be feeling. We use frequently this exceptional ability of estimating how others feel in our everyday lives. Such *empathizing* capacity serves us to be more helpful, sensitive, sympathetic, affectionate and cordial in our social interactions. More generally, this capacity helps us understand other people, their motivations and goals behind their actions and to predict how they will react to different events. If a person is sad and feeling unhappy, instinctively we are ready to offer our support and empathize with him.

Figure 1.2 serves as a motivation for trying to unravel the functioning of our brains for such tasks. When we look at the images from left-to-right, the information content increases as one moves from *face* → *visible body* → *whole image*. Facial expression (Figure 1.2.a) suggest that the boy is annoyed, while the body posture (Figure 1.2.b) suggests that he is probably annoyed because of the apple, however, it is still not clear as his gaze is not in the direction of the apple. When we see the bigger visual context (Figure 1.2.c), it is revealed that he is indeed annoyed but not just with the apple. He is not happy about the girl eating the chocolate, almost feeling jealous about it. This analysis helps in understanding how our perception of the emotional state changes according to the different information available from the surroundings.

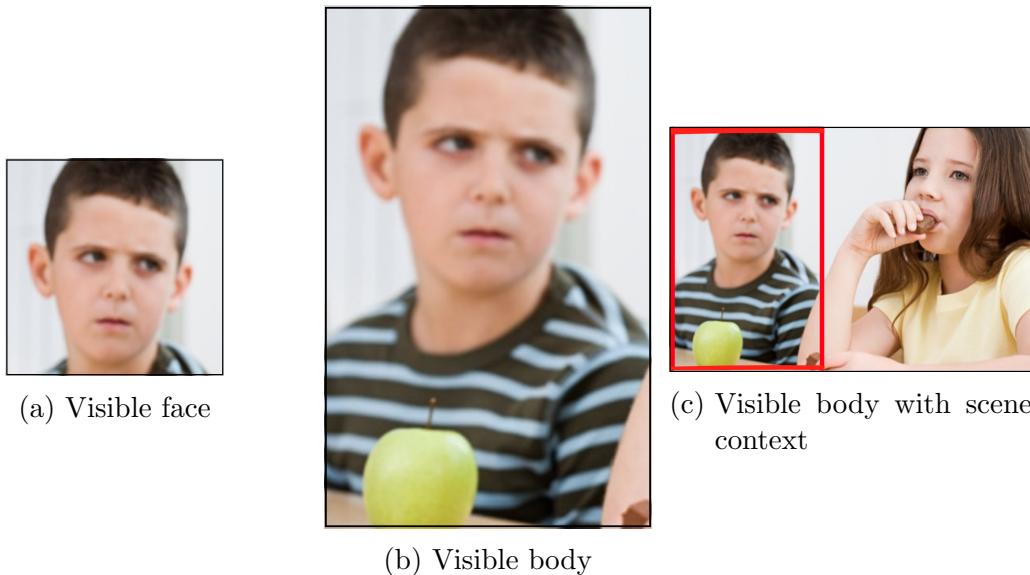


Figure 1.2: Perceived emotion of the person changes from *face* → *visible body* → *whole image*

Application point of view, there are many important scenarios where automatic emotion recognition is crucial. Human Computer Interaction (HCI) involves a verbal or non-verbal interchange between machine and human. There is a huge gap between how a machine works and the functioning of human beings. HCI techniques try to bridge

this gap. Emotion recognition is essential while designing such interfaces (Cowie et al. [2001]). Automatic emotion recognition techniques can come handy to improve the machine's limited capability to understand emotional response from humans. Another similar application is in the advanced driver assistance systems (or *ADAS*) technology. Human performance decreases during prolonged sleep deprivation (Posada-Quintero et al. [2018]). Due to this, the probability of accidents due to driver's drowsiness is increased. Automatic emotion recognition can help detect the emotions that are responsible for this kind of tiredness early on and help avoid road-accidents. Emotion recognition is also important for creating affective interactions for VR (Virtual Reality) applications. VR is used as a medium to elicit affective response while inside the virtual environments. Automatic emotion recognition is essential to create a feeling of *presence* inside the virtual environments (Riva et al. [2007]).

Apart from technological applications, emotion recognition is also essential from human behaviour analysis perspective as well. While interacting with technology, human emotions can be influenced and this in turn can affect the health of the individuals. Medical applications are increasingly taking this into consideration in their design (Luneski et al. [2008]). Another important application can be found in the field of mental health, where the experts try to improve the lives of people suffering from disabilities related to their mental health. Emotion regulation has shown good improvements in the treatment process (Berking and Wupperman [2012]). During emotion regulation, the researchers quite often take help of visual media that elicit affective reaction. The variability of the applications of emotion recognition has made the field important for the coming years.

1.2 Role of Context in Emotion Recognition

The place and/or the social situation that the person finds himself affects his emotional state and also influences the manner in which his feelings are perceived by an observer. The context of the situation is an important aspect while analysing people's feelings. Despite research focusing on facial expression and body pose, there is ample research work (Barrett et al. [2011], Aviezer et al. [2008a], Camras et al. [2006]) that asserts the importance of context in emotion perception. We try to understand the role of context through visual example shown in Figure 1.3.

There are 3 columns *viz.* *Face/Head*, *Body* and *Person in Context*. In the first column (Face/Head) of Figure 1.3.a, we see a female's face, she seems to be smiling suggesting that she might be feeling *happy*. We cannot be sure just by looking at the face. In second column (Body) of Figure 1.3.a, we see her body posture wearing a sports-wear.



(a) Full frontal view of the person's face is visible



(b) Facial profile of the person's face is visible



(c) Only the back of the person's face is visible

Figure 1.3: Examples showing people with different facial views, along with their body posture and the surrounding scene. What can be said about the emotional state of these people?

We can see that she is indeed happy about something, but we cannot yet be sure. One thing to note by her body posture is that something has grabbed her attention and she is looking towards that direction with some *anticipation*. In the third column (Person-in-Context) of Figure 1.3.a, we see that she is not only feeling happy, but also proud of her achievement. It is difficult to tell from the image which kind of sport it is, but using our common sense knowledge about such situations, we can say with certainty that she is the winner. This fact adds information to our previous knowledge about the face and her body posture and gives us more clues to understand what the person is feeling. The face and/or the body posture were not enough to understand her emotional state. The emotions suggested by facial expressions could change systematically and drastically depending on the intensity of the context in which the face of the person is embedded in (Aviezer et al. [2008a]). Figures 1.3.b & 1.3.c show examples of people whose faces and body are immersed into different situations. In Figure 1.3.b only the profile of the face of the person is visible, so it becomes complicated to anticipate his feelings. With his body posture visible (second column, Figure 1.3.b), a bit more information is revealed which indicates that the person is looking away toward something or someone - which apparently has his attention, but it is not enough to tell us what he is feeling. Only when we see the whole picture (third column, Figure 1.3.b) it becomes clearer that the person is in a meeting room and he is paying attention to a person talking, probably feeling *engaged* in the activity. Figure 1.3.c shows even more challenging situation. We just see the back of the person's head (first column) which does not give any information about the emotional state of the person. The body posture (second column) reveals part of the story, but it is only the whole image (third column) that portrays a bigger and more rich picture. We see that the boy is playing, so he is *engaged* in playing with other kids, and he is probably in a state of *anticipation*. We see that context is necessary to predict emotions when the face of the person is not visible or partially occluded. Even when it is completely visible (first column, Figure 1.3.a), the contextual cues present in the visual scene shapes our emotional perception very heavily. This is a clear example of how the context affects a person's emotional state and also how it is equally important when trying to estimate a person's emotional state. Context not only changes the perception of the emotional state, it also impacts what the person-in-context is actually feeling (Aviezer et al. [2008a]).

Psychological studies also discovered the effect of context on a person's emotional state. Barrett et al. [2011] contend that the perception of emotion is influenced by three different types of context which are based on *viz.* (i) *Stimulus*, (ii) *Perceiver* and (iii) *Culture*. The authors argue that *Stimulus* comes from the situational context that the subject is embedded in. Figure 1.2 is a representation of this argument. We can see

that despite having food (apple) in front, the boy gets annoyed because the girl gets to eat the chocolate. Her action of eating chocolate serves as a *stimulus* to make the boy feel annoyed. Various stimulus produces different kinds of emotional reactions - which in turn also affects the perceived emotion. *Perceiver* or *Observer* is another source of context according to the authors. It involves the use of verbal description by the perceivers. Emotion perception inherently is restricted by the limitation of the language. The authors affirm that quite often the association of the emotion words used to represent the actual facial expressions is obscure. An example of this is shown in Figure 1.4, where a single image was given to 2 observers and they were asked to recognize the emotion of the subject using emotion categories (Table 3.1) and natural language. When asked to choose the applicable emotions from a list, the observers concur on the categories. However we see that their emotion perception starts to differ when asked to respond using natural language. For *Observer 1*, the doctor doesn't like what he is doing, where as for *Observer 2* he is being sympathetic. This signifies that the facial expressions are not unambiguous, and that contextual information conditions the emotion perception. *Cultural* context plays a sociological role in emotion perception. Emotion perception changes across different cultures, even when the cultures share the same language. This indicates that the cultural context is not a cause of language differences between cultures, rather it is in itself an important contributor to the emotion perception. Culture dictates how and where a perceiver looks while making emotion judgements.



| Response Type → | Emotion Categories | Natural Language |
|-------------------|-------------------------------|--|
| Observer 1 | <i>Engagement, Excitement</i> | <i>To me it looks more like he does not like what he is doing, so I would say there is a small amount of displeasure involved.</i> |
| Observer 2 | <i>Engagement, Excitement</i> | <i>The doctor is engaged in his diagnostic and he sympathizes with the boy. The boy is not excited. He is in pain and looks sad.</i> |

Figure 1.4: Emotion perception of two observers presented via *Emotion Categories* and *Natural Language*

Another study by Aviezer et al. [2008c] provides evidences by experimentation that the facial expressions (and eventually the emotion of the person) is not invariant to the context. The authors found that the scene context and the body posture affect the emotion perception, even when the perception is done at early stages. It does not matter if the emotions are perceived through specific categories or affect dimensions, their experiments suggest that the context influences the emotion perception regardless of the representation

format of the emotions. Hence, context should be taken as an important modelling factor while designing emotion recognition systems.

Current systems which recognizing people's emotions through images are good in interpreting the emotions using face or body pose only, but they lack a holistic view and do not consider the context of the scene while making predictions. These systems rely heavily on facial features or work with only limited set of emotional states and do not work when the face is occluded. There is a need for a more robust system that can understand people's emotions while also taking context in consideration. There is a lack of research in this area mainly due to absence of a good labelled dataset which can overcome such limitations. In the following section we explore the various sources of context and try to understand their influence on the emotion perception.

1.2.1 Sources of Information for Emotion Recognition

We are surrounded by multitude of things in our day-to-day lives. Different inanimate objects (utensils, furnitures, machines, vehicles, infrastructure, nature, etc) surround us depending on the place and occasion. We live amongst (and are also surrounded by) people like our colleagues, friends and family - also depending on the circumstances, location and the time of the day. So the environment surrounding every person is unique and changes continuously. And these various surroundings can influence the emotional state of the person in many different ways. Due to these various situations, the perception (from an observer's point of view) of their emotional state is also heavily influenced. Here we list and discuss some such sources of context that are accessible to use through our vision capability.

1. **Face Pose:** Facial expressions have been and still are the primary source of emotion and also the main point of focus for research in emotion recognition (Ekman and Friesen [1969]; Essa and Pentland [1997]; Chanes et al. [2018]). Depending on the visibility of the face (or face pose), it contributes different kind of information for emotion recognition. For example in Figure 1.5, we see that a person with the same facial expression with 9 different face poses. When viewed independently, each image showcases distinct visual information.
2. **Body Posture:** One of the more effective way of communication includes the way one displays his body posture. For example, when the faces in Figure 1.6 are seen independent of their associated body pose, the emotion perceived is *Disgust*. However, depending on the different body postures, the perception of the emotion changes. When face with the same expression is put into different body context,



Figure 1.5: Visual information for emotion recognition changes with the change in face poses, for a fixed facial expression (Ouamane [2015])

emotions like *Anger* (Figure 1.6.b), *Sadness* (Figure 1.6.c) and *Fear* (Figure 1.6.d) are perceived. This has been experimentally demonstrated by Aviezer et al. [2017] (also, Martinez et al. [2016]) with ample more evidences (like Dael et al. [2012a]; Schindler et al. [2008]; Dael et al. [2012b]) that demonstrates the influence of body pose on the perception of emotion of the given person. These examples suggest that body posture is an important source of context for emotion perception and should be considered part of the emotion recognition process.

3. **Hand Gestures:** Hand gestures form an integral part of the body posture. Here, we discuss it's influence on emotion perception by keeping the facial expression as well as the body posture fixed. The gesture we use (or choose subconsciously - *learned through experience*) supplement the expression of our feelings. The gesture could be a simple movement of hand and can supplement our perception of the person's feelings. Similar to Figure 1.6, Figure 1.7 shows a person with the same facial expression and giving different hand gestures. We can see how emotion perception changes when we move from one gesture to the other. For instance, we see a huge contrast in gestures 2 and 5. Gesture 2 suggests that the person is approving positively where as gesture 5 is a complete opposite suggesting rejection or disapproval. Gesture is an important context for emotion perception. Mitra and Acharya [2007] is a good survey on different gesture recognition, specifically the ones with hand and faces.
4. **Visual Scene:** We travel to visit other places, cities or countries because we want to explore different towns, their characteristics, people, food and culture. While walking around a new village or neighbourhood, we see various things. We see



(a) Body posture depicting disgust



(b) Body posture depicting anger



(c) Body posture depicting sadness



(d) Body posture depicting fear

Figure 1.6: Keeping the facial expression fixed, the body posture influences the perceived emotion of the person (Aviezer et al. [2017])

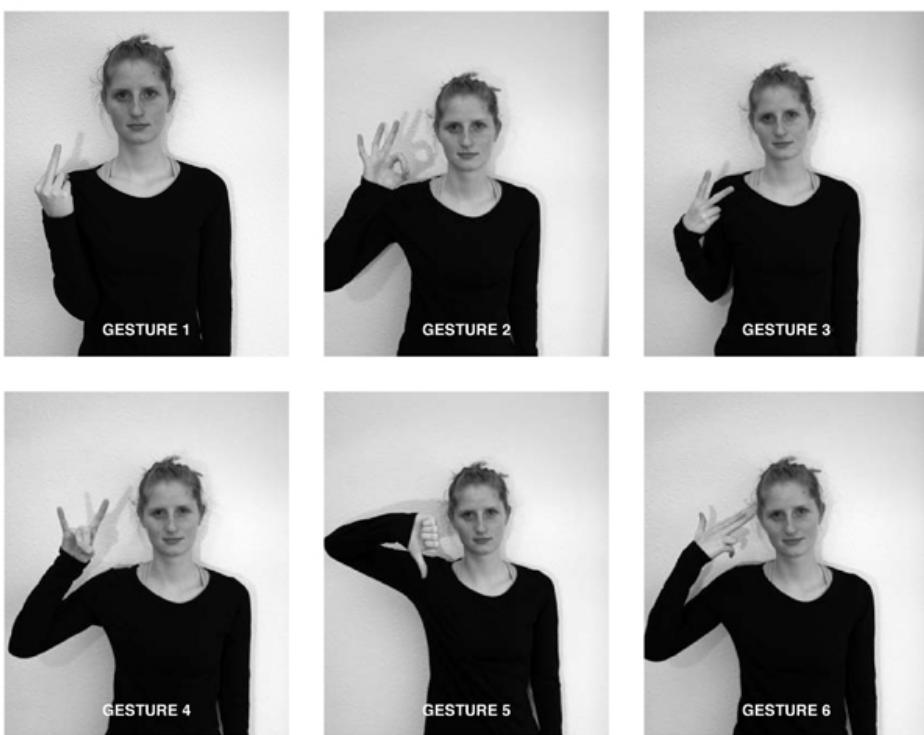


Figure 1.7: Six different hand gestures suggesting distinct emotions (Link)

a building that we have never seen before or have seen it but with a different architecture style. Sometimes there are streets or shops which reminds us of our own neighbourhood or something from our past. Such an experience gives us a new kind of feeling. We enjoy this state of being and are in general happy. But sometimes, there are situations or scenes that are disturbing or annoying, that makes us feel sad or anger or even fear. For example, while walking in an unknown town, we come across a deserted street in the middle of the night. It is very dark and there is no sound at all - this might be scary and might make you feel afraid. The visual scene described here is affecting the feelings of a person. Our perception of the surrounding scene has a direct influence to our emotional state. For example, when we see the two people in Figure 1.8 , we see that both are working but we perceive their emotional states to be different. We can interpret that the person in Figure 1.8.a is more relaxed as compared to the other, she is in her pyjamas which are more comfortable than wearing a suit and tie while working. Visual scene is a very important source of contextual information considering all the other sources. Visual scene or the immediate surrounding of the person contains more holistic information on the contexts affecting emotional state of the person.



Figure 1.8: Type of the surrounding visual scene impacts the emotion perception. For example, the 2 people doing the same activity in different scenes are perceived to have distinct emotional states

1.3 Motivation for Emotion Recognition

Human beings have different ways of sensing the physical world. We use tongue for taste, ears for listening, nose for smelling and eyes for vision. Each of these senses are unique and have their distinctive advantage over the other. However when we look at the different sources of context that can affect our emotions (discussed in subsection 1.2.1 and Appendix C), all of them except *audio* are visual contexts that are accessible (and eventually analyzable) through our vision capability. One can argue about the context

called *Prior Knowledge* as not being directly accessible through vision, however, once we have the visual input of the situation, then we can make a prediction about it which is based on our *prior knowledge* (or experience). This makes the visual input as a necessary requirement.

Computer vision being the primary field in understanding visual information, has a distinctive advantage to use state of the art algorithms developed for scene understanding (and visual information extraction) to understand emotions visually. In addition to computer vision, fields like psychology and cognitive neurosciences also helped us formulate our research problem. We divide our approach into 3 aspects. *First Step* is to find or extract the visual features responsible for the emotion; *Second Step* is to tag the extracted features with appropriate emotion labels so that they are distinguishable for further analysis; and *Third Step* is to develop a model that uses the visual features and their corresponding labels to make predictions about the perceived emotions.

Computer vision helps in visualising and perceiving the world and the people in it as is seen though human vision system. It also helps to capture various visual features which can be used to compute information about the world - type of objects, their position, relationship with other surrounding objects, etc (Lowe [1999]). Current research in computer vision has demonstrated that algorithms have surpassed human-level visual recognition (He et al. [2016]). Specifically, CNNs have helped achieve great performances on many vision-related problems. Using the robust power of CNNs, a machine can visualise a person's facial expressions, body features and the surroundings - crucial *first step* towards emotion recognition. Research related to emotion recognition in the field of psychology focuses on the cause, effect and response of emotions. This helped us understand the nature and different sources of emotion elicitation in humans. We learned how humans perceive emotions - *by forming categories of emotions and affect dimensions* (Russell [2003]) - which forms an important *second step* towards emotion recognition. Research in cognitive neuroscience (Stephani [2014]) showed us how different neuronal pathways in the human brain is responsible for elicitation, understanding and expression of emotions. This helps us understand how the emotions are perceived by the human brain, consequently aiding in development of a model for our emotion recognition pipeline - *third, and final step* towards emotion recognition.

1.4 Thesis Outline and Contributions

In this section we discuss the main scope of this thesis, briefly outline the content and coverage of each chapter and also summarize the main contributions of the thesis.

The principal goal of this thesis is to create the first model of automatic emotion recognition from a computer vision perspective that explicitly includes the visual scene context. We demonstrate the performance of the model through various empirical experiments and analysis. In working towards this goal, we constructed a novel dataset containing images of people taken in unconstrained environments - which provided sufficient sources of visual context for the purpose of emotion recognition. Then, we used crowd-sourcing techniques to generate emotion labels for people present in those images. The work was done by human workers, who were selected based on their qualification to being able to do such tasks. Then we used state of the art supervised learning algorithms in computer vision to design emotion recognition pipeline based on the collected data - which helped us demonstrate our principal goal.

Chapter 1 gives an introduction of the larger scope of the thesis and brief overview of the ideas being covered in this thesis. It introduces emotion recognition as the main problem domain of the thesis and discusses the relevance and importance in our daily lives with examples of applications. Then the chapter introduces the role and affects of visual context in emotion perception. The importance of context in emotion recognition is explained with examples detailing different sources of contexts that have direct influence on the emotion perception. It also expounds the importance and need for the development of an automatic emotion recognition system. Before closing, the chapter gives an outline of the whole document chapter-wise along with a brief summary of the main goals of the thesis and gives a brief description of the main contributions.

Chapter 2 covers the discussion about related research in the area of emotion recognition. The chapter starts with discussion of work related to emotion representation methods. Then an overview of current research in emotion recognition from facial expressions, body posture, group-level emotion recognition and related approaches is presented. Following this, a discussion on currently available datasets related to the task of emotion recognition is presented. The chapter closes with the comparison of contribution of this thesis with respect to the previous research and datasets.

Chapter 3 introduces our EMOTIC dataset for the first time. The chapter is divided into 2 main sections. The first section explains about the construction process of EMOTIC, detailing all the aspects in each step. The second section is devoted to the analysis of the generated annotations of the dataset. It presents a statistical and an algorithmic analysis along with the analysis of the agreement between workers for the

generated annotations. A discussion about common sense knowledge present in the visual scene ends the chapter.

Chapter 4 focuses on design of an appropriate architecture for the task of emotion recognition in context. It explains how state-of-art research in scene and object recognition helped design our CNN-based EMOTIC fusion model. The chapter closes by explaining the various criterion choices and the performance evaluation metrics used for training the model on EMOTIC dataset.

Chapter 5 details and lists all the empirical experiments and their respective analysis using the EMOTIC dataset and the fusion model. The chapter begins by explanation of baseline experiments and their analysis performed using different features. Then the chapter covers additional experiments done using more deeper networks with different types of architectures. It also describes how using a different loss criterion helped improve the performance. The chapter closes with a comparative analysis on using Sentibanks as visual context features for emotion recognition.

Chapter 6 is an account of conclusions from the experiments performed with respect to the current research paradigm. The chapter then explicitly describes the contributions of the thesis in the research domain. This final chapter end the dissertation by outlining different probable directions a future research can take based on the work presented in this thesis.

Main Contributions of the thesis are summarized below:

1. Generation of a novel dataset - EMOTIC - in the field of emotion recognition in context. Different scenes provide the necessary visual context. People in the images are annotated in their corresponding scene-contexts using 2 different emotion representations [Chapter 3]
2. A CNN fusion model designed in accordance with the characteristics of the EMOTIC dataset for automatic emotion recognition in visual scene context [Chapter 4]
3. Empirical experiments that demonstrate the influence of the visual scene contexts in emotion recognition. In addition, a comparative study of different sources of visual context features [Chapter 5]

Chapter 2

Previous Research on Emotion Recognition from Images

Emotion recognition is important for human behaviour analysis - which is an interesting area of research in many fields (psychology (Ekman and Friesen [1969], Izard [1971], Russell [2003])), cognitive neuroscience (Hassin et al. [2013], Aviezer et al. [2008a], Ritter et al. [2017])), computer science (Bänziger et al. [2009a])). As a consequence to the powerful techniques developed in the recent years after the success of deep learning, computer scientists are able to apply the new methods in multi-disciplinary fields. These new methods are well suited for complex problems like modelling human behaviour and emotions. By studying how humans interact with the world, we can model human behaviour. Emotion recognition forms an integral part of the human behaviour analysis.

Researchers have studied multiple cues that allow us to recognize a person's emotional state. For example, *viz.* face (Beristain and Graña [2009]), body (Dael et al. [2012a]), voice (Bänziger et al. [2009b]), gesture (Wan et al. [2016]), physiology (Bazgir et al. [2019]; Goshvarpour et al. [2018]), tactile (Gunes and Pantic [2010]), text (Ferreira et al. [2018]), brain-waves (Gunes and Pantic [2010]), etc. or a combination of these modalities (Bänziger et al. [2009b]; Gunes and Pantic [2010]) - which also constitute the various sources of emotion elicitation. Our vision helps us process the facial expressions, body postures, hand and body gestures and written text (in the form of literature, novels, etc. or any written material that can elicit emotions) for emotion recognition. Our audio capability allows us to listen to voices; and through our sense of touch we can sometimes feel the physiological changes in a person's body to understand what the person is feeling. Human beings use these modalities simultaneously to estimate a person's emotional state. Machines also have the capability to process these different modalities. The current research in emotion recognition is focused in trying to create automatic mechanisms that

can achieve commendable performance. We begin our review of related research by talking about various emotion representation formats present in the current research paradigm.

2.1 Emotion Representation Formats

Automatic emotion recognition algorithms need to capture the complex emotional states exhibited by humans. There are multiple ways for describing emotions. We mention the most commonly used representation formats:

1. **Free Form:** The emotions are described in the form of sentences and paragraphs using words and phrases used in our daily lives. This text carries information about the underlying affective states which is usually reflected in the usage of certain words or grammatical alternatives. There have been attempts in figuring out the best representation form for the emotions described in *Free Form* by building markup languages (Schröder et al. [2007]). The HUMAINE database is an example where the authors tried to bridge the gap between emotion elicitation and its annotation (Douglas-Cowie et al. [2011])
2. **Affect Dimensions:** The emotions are indicated using scales based on different affect terminologies. For example, Cowen and Keltner [2017] use 12 affect dimensions while gathering emotional experiences. Annotators (or workers who are paid to do such tasks) report their experiences on a given scale (an example of such a scale would be the *Likert* scale (Likert [1932])) for each of those 12 affect dimensions. Another, very popular approach is to use 3 independent affect dimensions (introduced by Mehrabian [1995]). The three dimensions are *viz.* *Valence, Arousal and Dominance*. *Valence* measures how positive or pleasant an emotion is, ranging from negative to positive; *Arousal* measures the agitation level of the person, ranging from non-active / calm to agitated / ready-to-act; and *Dominance* measures the control level of the situation by the person, ranging from submissive / non-control to dominant / in-control
3. **Emotion Categories:** The emotions are represented in discrete form using words that represent the characteristics pertaining to that emotion. Darwin [1998] was the first to suggest the permanence of human expressions. He suggested that the human emotions are universal and, based on the facial expressions, can be categorised into modular (or discrete) emotion categories like *Fear, Anger*, etc. Du et al. [2014] proposed a set of 21 facial emotion categories, defined as different combinations

of the basic emotions, like *happily surprised* or *happily disgusted*. With this categorization the authors were able to give a fine-grained detail about the expressed emotion. After multiple studies spread across many decades, modern psychologists and neuroscientists also concur the modal form of emotions (Ekman and Friesen [1969], Izard [1971])

Using these 3 methods of emotion description, the emotions of the person (enclosed in bounding-box) in Figure 2.1 can be described as follows:



Figure 2.1: Example for emotion perception for the person-in-context. The perceived emotion is using *Free Form*, *Affect Dimensions* and *Emotion Categories*

- (i) **Free Form** - *The person is lying down in pain and so multiple people are attending him because he is injured. The fact that he is lying down suggest that he is suffering in pain. He was involved in a sporting activity so maybe not sad or afraid but he might be annoyed at the inconvenience*
- (ii) **Affect Dimensions** - The range of each dimension is between 1 – 10, 1 being the lowest value and 10 the highest.
 - (a) *Valence = 2*: The person is injured and thus is in pain, so the valence is very low

- (b) *Arousal = 2*: Similarly, the person is not active any more due to the immobility caused by the accident
 - (c) *Dominance = 1*: He needs assistance to get better, clearly his ability to control himself is limited due to the injury, so dominance takes a very low value as well
- (iii) **Emotion Categories -** The following categories are relevant labels, with their conceptual descriptions
- (a) *Suffering* - psychological or emotional pain; distressed; anguished
 - (b) *Sadness* - feeling unhappy, sorrow, disappointed, or discouraged
 - (c) *Pain* - physical suffering
 - (d) *Fear* - feeling suspicious or afraid of danger, threat, evil or pain; horror

We mentioned 3 different formats of emotion representation with a corresponding example to have a broader understanding of the different formats available. Out of these formats, *Free Form* is the most flexible allowing one to convey the perceived emotion through verbal (or written) language. EMOTIC uses *Affect Dimensions* and *Emotion Categories* to label the perceived emotions of the people.

2.2 Emotion Recognition from Images

In this thesis, we focus on the importance of visual scene context on our emotion perception. We introduced and discussed a few sources of context that can affect the emotion recognition in section 1.3. Since our research is from the perspective of computer vision we, therefore, analyze the most prominent state-of-art studies for their contributions and shortcomings in this section.

2.2.1 Facial Expression Based Approaches

Most of the work has focused on the analysis of facial expression to predict emotions. The base of most of these methods is the Facial Action Coding System (FACS) (Friesen and Ekman [1978]), which encodes the facial expression using a set of specific localized movements of the face, called Action Units (AUs). These facial-based approaches usually use facial-geometry or appearance features to describe the face. The extracted features are then used to recognize AUs and infer the basic emotions from it. Ekman and Friesen (Ekman and Friesen [1971]) proposed the mapping of those action units into the following

six emotions: *anger, disgust, fear, happiness, sadness, and surprise*. Current state-of-the-art systems for emotion recognition from facial expression analysis use CNN to recognize emotions from AUs.

Essa and Pentland [1997] developed a computer vision system to annotate the dynamic movement of facial muscles using optical flow. The authors found that AUs of FACS are insufficient to encode the visual motion of muscles, primarily because they are based on the spatial pattern of the facial muscles. Their new proposed coding system, called FACS+, is able to encode the spatio-temporal functionality of the muscles (as opposed to spatial functionality in FACS) into their respective AUs. Pantic and Rothkrantz [2000] dived deeper into feature representation techniques specific to emotion recognition from face. Their system, called Integrated System for Facial Expression Recognition (ISFER), uses hybrid approach to facial feature extraction. Li et al. [2009] take advantage of the local features to model the facial expressions. They divide the face into sub-regions and extract it's local features like SIFT (Scale Invariant Feature Transform) and PHOG (Pyramid of Histogram of Oriented Gradients) that help in getting the texture and shape information of the sub-regions respectively.

Current state-of-the-art systems for emotion recognition from facial expression analysis use CNN to recognize emotions from AUs. The main advantages of CNN based algorithms is that they are able to highly reduce the dependency of the models on the physical properties of the face components and the pre-processing steps, and introduce an “end-to-end” learning system that is able to use the pixel values directly from the images. Breuer and Kimmel [2017] use CNNs for feature extraction and inference. Zhao et al. [2016] introduced a unified deep network, called Deep Region and Multi-label Learning (DRML), where a novel region layer induces important facial regions forcing the learned weights to capture structural information of the face. The complete network is end-to-end trainable, and automatically learns representations robust to variations inherent within a local region. Jung et al. [2015] trained 2 different CNNs for learning temporal appearance features and temporal geometry features respectively. A more intensive analysis on the different research work, based on deep learning and CNNs directed towards facial emotion recognition from images and video, can be found in Ko [2018].

2.2.2 Body Posture Based Approaches

Although the research in emotion recognition from a computer vision perspective is mainly focused in the analysis of the facial expressions (Beristain and Graña [2009]), recent research has also focused to consider other additional visual cues or multimodal approaches. A person’s body has lot of information that can help understand what that person is feel-

ing. This information could be in the form of gestures, body posture, facial expressions and head pose. Schindler et al. [2008] implement a computational model to understand body emotional language. Their main focus is studying the emotional language conveyed by the human body's posture when the subjects are shown stimuli to elicit emotional response. The faces of the subjects, although visible, are not very clear sometimes (as shown in Figure 2.2 Row-1 for *fearful* and *sad* emotions). The authors do not neglect this visual information, rather they include it in their model as a coarse representation. As seen in Figure 2.2, different body postures can still communicate the same emotion. This gave us a good motivation to consider the emotional language communicated by the human body posture as an important contextual cue for understanding emotional state in our model as well. Furthermore, Dael et al. [2012b] developed their Body Action and Posture (BAP) system to understand how the body posture conveys not only the emotion intensity but also some postures convey important information about particular emotions and augment the understanding of the overall emotional state of the person. In their attempt, they compared their findings with related theories and suggest that a more thorough design of experiments was needed.



Figure 2.2: Different body postures that represent various emotion stimuli, while maintaining a uniform background (Adams Jr and Kleck [2005])

In order to find out the effect of body posture on the perception of emotions, Aviezer et al. [2012] carried out an interesting experiment. They used a single face and superimposed it on different body poses (Figure 1.6). As the authors report, perception of the emotions changed from one posture to another.

The dynamic body movement is also an essential source for estimating emotion. Studies such as Kleinsmith and Bianchi-Berthouze [2007] & Kleinsmith et al. [2011] establish the relationship between affect and body posture using as ground truth the base-rate of human observers. The data consist of a spontaneous set of images acquired under a restrictive setting (people playing Wii games). Similarly, Dael et al. [2012a] also show that body posture conveys not only positive or negative emotions, but also specific emotions.

The research has also focused on using neural networks for modeling and training the body features to recognize emotions. Nicolaou et al. [2011] used the location of shoulders as additional information to the face features to recognize basic emotions. The authors compare the performance of traditional Support Vector Regressors (SVRs) against a bi-directional Long Short Term Memory (BLSTM). Due to the ability of LSTMs to model the past and future information, they could perform much better than SVRs. Schindler et al. [2008] demonstrate computationally that different body poses have different emotions. They use a biologically inspired design of neural network to model different body poses to recognize 6 basic emotions. They conducted experiments on a small dataset of non-spontaneous poses acquired under controlled conditions.

2.2.3 Group-level and whole-Image based Approaches

Apart from face and body, there are research work that have focused on more holistic approaches to emotion where the whole image is considered for extraction of relevant features. For example, Marchesotti et al. [2011] try to estimate the aesthetic aspects of an image computationally using image descriptors. Recent works mainly use the advantages offered by CNNs and deep learning to create better models and improve performance. Mou et al. [2015] presented a system of affect analysis in still images of groups of people, recognizing group-level arousal and valence from combining face, body and contextual information. Polanía and Barner [2017] use a hybrid network that uses global scene features, skeleton features of the group and the facial features to make predictions about the group-level emotion. Dolz and Pedersoli [2018] employ an attention mechanism in their CNN pipeline to incorporate group-level emotion information as part of their attention mechanism.

EmotiW (Emotion Recognition in the Wild) challenges (Dhall et al. [2017]) have provided researchers a platform where people can test their approaches to group-level emo-

tion recognition. Here, the researchers submit their solutions and are evaluated against a common test-set of samples. EmotiW host 3 databases: (1) The *AFEW* database (Dhall et al. [2012a]) focuses on emotion recognition from video frames taken from movies and TV shows, where the actions are annotated with attributes like name, age of actor, age of character, pose, gender, expression of person, the overall clip expression and the basic 6 emotions and a neutral category; (2) The *SFEW*, which is a subset of AFEW database containing images of face-frames annotated specifically with the 6 basic emotions and a neutral category; and (3) the *HAPPEI* database (Dhall et al. [2012b]), which addresses the problem of group level emotion estimation. In this work we can see an attempt to use context for the problem of predicting happiness in groups of people.

Image Sentiment Analysis deals with any type of image and the goal is to recognize the emotion an observer will have when looking at the image. This could be an interesting source of group-emotion features, when the image contains people in them. Chen et al. [2014] built a visual sentiment concepts called Adjective-Noun Pair (ANP) - which they discovered by mining millions of tags from web photos. ANPs can be used as features to find interesting information about the image.

2.3 Image-based Datasets for Emotion Recognition

The availability of appropriate data for visual recognition tasks is very important. Our goal in this thesis is to recognize emotions of people through images, so it is important to study the currently available datasets and analyze their feasibility for our purposes. Most of existing datasets for emotion recognition are centered in facial expression analysis. For example, the GENKI database (<http://mplab.ucsd.edu>) contains frontal face images of a single person with different illumination, geographic, personal and ethnic settings. Images in this dataset are labeled as *smiling* or *non-smiling*. Another common facial expression analysis dataset is the ICML Face-Expression Recognition dataset (Goodfellow et al. [2013]), that contains 28,000 images annotated with 6 basic emotions and a neutral category. On the other hand, the UCDSEE dataset (Tracy et al. [2009]) has a set of 9 emotion expressions acted by 4 persons. The lab setting is strictly kept the same in order to focus mainly on the facial expression of the person. In addition to these, there is a huge number of datasets that focus on facial expressions. Table 2.1 shows a list of publicly available datasets for face images. We mention short description about the kind of facial expression is present in the dataset and their corresponding quantity. Refer Table B.1 in Appendix B for their corresponding references and download links. Recently, Google launched a *Dataset Search* (Google)

| Dataset | Description | Data Info |
|---------------|---|---------------------------|
| CK+ | Posed and spontaneous facial expressions (123 Subjects) | 593 Video Sequences |
| CE | 22 Compound facial emotions (230 Subjects) | 5060 Images |
| DISFA+ | Posed and spontaneous facial expressions with 66 Facial Landmarks (27 Subjects) | 130000 Stereo Videos |
| Yale Face DB | 9 face poses and 64 illumination conditions (28 Subjects) | 16128 Images |
| MMI | Posed facial expressions - frontal and profile (75 Subjects) | 2900 Sequences and Images |
| KDEF | Posed facial expressions from 5 angles (70 Subjects) | 4900 Images |
| PubFig | Spontaneous face images (200 Subjects) | 58797 Images |
| ExpW | Spontaneous face images | 91793 Images |
| CASIA WebFace | Spontaneous face images from the Web (10575 Subjects) | 494414 Images |

Table 2.1: Various publicly available facial expression datasets with their descriptions and data quantity

Fabian Benitez-Quiroz et al. [2016] introduced a novel computer vision algorithm that can annotate millions of images of facial expressions in the wild. The algorithm can automatically detect the AUs and their respective intensities and map these to their respective emotion categories. This algorithm provides an efficient tool for generating large datasets.

There are datasets that also include the body of the person along with other related modalities. For example, the GEMEP database Bänziger et al. [2006] is multi-modal (audio and video) and has 10 actors playing 18 affective states. The dataset has videos of actors showing emotions through acting. Body pose and facial expression are combined along with multi-modal sources of producing output (audio and video). In another example, Dael et al. [2012b] developed a dataset that contains videos of actors showing emotions through acting - a combination of facial expressions and body pose.

The Looking at People (LAP) challenges and competitions (Escalera et al. [2017]) involve specialized datasets containing images, sequences of images and multi-modal data. The main focus of these datasets is the complexity and variability of human body configuration which include data related to personality traits (spontaneous), gesture recognition (acted), apparent age recognition (spontaneous), cultural event recognition (spontaneous), action/interaction recognition and human pose recognition (spontaneous).

COCO dataset (Lin et al. [2014]) is one of the most exhaustively annotated datasets available in computer vision community for visual recognition. COCO has been recently

annotated with object attributes (Patterson and Hays [2016]), including some emotion categories for people, such as *happy* and *curious*. These attributes show some overlap with the emotion categories that we define in this thesis (Table 3.1). However, COCO attributes are not intended to be exhaustive for emotion recognition, and not all the people in the dataset are annotated with affect attributes. The biggest advantage of this dataset is that it has the potential to be annotated with more exhaustive emotion attributes, given that it already has object annotation, instance segmentation, cations for the visual scene, people key-points, scene segmentations, stuff segmentation (which includes background and foreground) and some affect attributes. These visual recognition tasks can serve as good contextual cues for emotion recognition in context.

2.3.1 Shortcomings of the currrent Datasets

There has been a lot of research in emotion recognition from images due to which currently there are lot of datasets out there. The problem of emotional state recognition is extremely complex, but our hypothesis is that there are three important limitations in the current approaches and datasets that we have reviewed:

1. The datasets have been generated in lab environments or the images of the people are constrained (acted, posed or fixed)
2. The existing databases in emotion recognition (as discussed in the previous section) lack fine-grain labels of human emotions. Most studies classify emotions according to 6 categories, but this is far from the fine grain categorization that humans are capable of. In this work we introduce a more sophisticated set of 26 emotion categories and combine them with the common continuous dimensions (valence, arousal and dominance). This combination provides a rich description of the emotional state of a person
3. The visual scene context (the surroundings of the person) is an important source of information and has not been incorporated in previous studies

As an example, we compare the content of images present in EMOTIC with Cohn-Kanade (Lucey et al. [2010]) and EMOTIONET (Fabian Benitez-Quiroz et al. [2016]). EMOTIC includes the visual scene contextual features thereby enriching it with more information content. *For example:* We can see in Figure 2.3.(a,1) that the woman is in pain from some suffering and the other woman is trying to sympathize with her. Similarly we can see in Figure 2.3.(a,5) that an old man is completely engaged in his painting - these kind of information is absent in other datasets (Figure 2.3.b, Figure 2.3.c). EMOTIC also

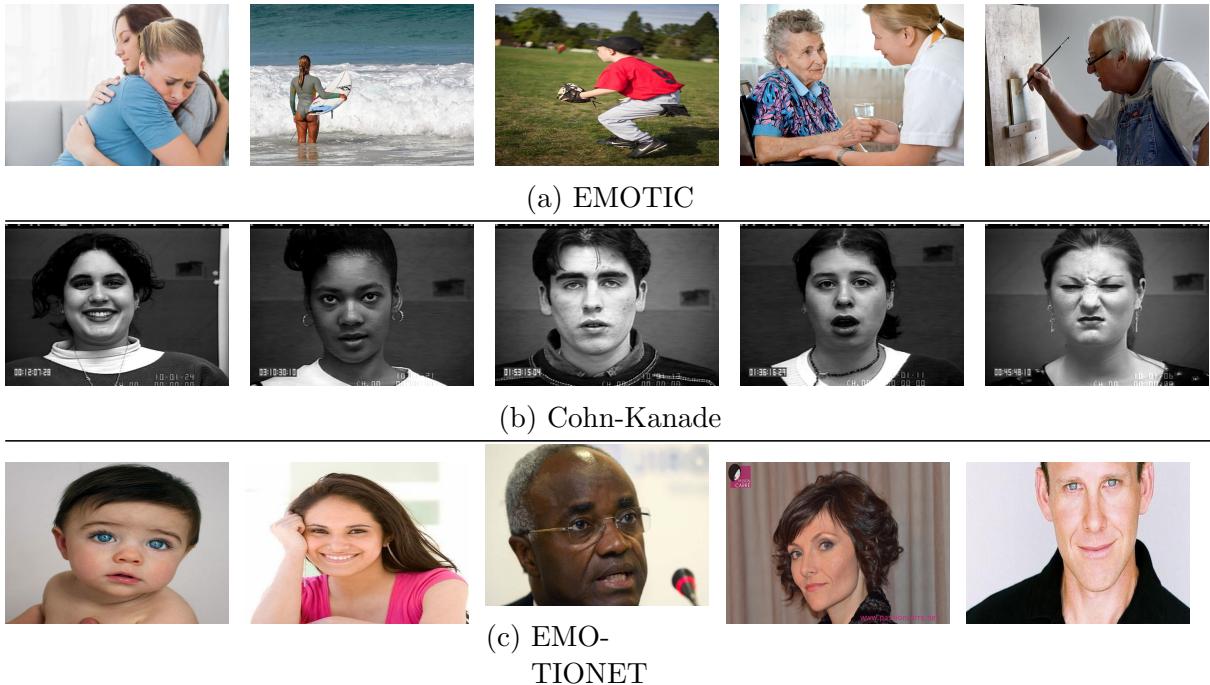


Figure 2.3: Sample images from EMOTIC, CK (Kanade et al. [2000]) and EMOTIONET (Fabian Benitez-Quiroz et al. [2016]) datasets in rows (a), (b) and (c) respectively

includes images with faces not clearly visible. *For example:* In Figure 2.3.(a,2) & 2.3.(a,3), the faces are obscured, however, their pose, attention, objects they hold and corresponding backgrounds provide more information about their emotional states.

Chapter 3

EMOTIC Dataset

Machine learning has become ubiquitous for the past few years. Due to the availability of fast, immense and cheap computing resources, machine learning algorithms are being used by increasing number of researchers, companies and various institutions to either improve state of the art or build smarter applications across disciplines. However, the performance of machine learning algorithms is limited by the amount of data available. Since there were enough computing resources, and not enough data, many new datasets have proliferated since. As a result there are many benchmark datasets in computer vision research available to work on. A good list of such datasets can be found on these links (1) Fisher [2018] and (2) Wikipedia [2018].

In previous chapter (chapter 2), we saw various emotion recognition datasets based on facial expressions, body pose and group-based emotion recognition. In all those datasets, none of them considered the visual context (Section 1.2.1) for emotion estimation. This fact, combined with the importance of context (Section 1.2) in emotion perception, gave motivation to create the **EMOTIC** dataset with essential characteristics absent in previous datasets.

Specifically, the EMOTIC dataset is characterised by:

1. **Appearance of Subjects:** The images of people in EMOTIC is natural and not acted (like in laboratory environments) and is not restricted by their facial expressions, head pose or the body postures.
2. **Presence of Context:** The background or the surrounding environment is present in the image and, also, not restricted to any particular location or setting. The images can show any place, view point or social situation. People can be doing different activities and have any object around them, including other people.

3. **Extensive Emotion labels:** The subjects are labelled with emotions that are very comprehensive. The emotion representations are unambiguous and encompasses widest range of human emotions. The apparent emotions of people are represented by a combination of 26 extensive discrete emotion categories and 3 continuous dimensions (Figure 3.19)

These attributes give a distinctive quality to the kind of data collected under the umbrella of EMOTIC. This differentiates EMOTIC from any other emotion-recognition dataset ever collected with regard to the type of images and their corresponding emotion labels.

3.1 EMOTIC Dataset Construction

The creation of the present EMOTIC dataset was divided into 2 releases:

1. In the first release, 18316 images were annotated overall. There are images with more than one person and, quite often, these people also have been annotated with emotion labels. So, overall, there are 23788 people, each with their own annotations. After dividing the dataset into *Train (70%), Test (20%) & Validation (10%)*, each person in the Test set was annotated by 2 additional distinct annotators. These supplemental annotations were carried out to have Test set of images with exhaustive labels from multiple different annotators. This created a bigger pool of labels for each person making it more efficient for testing the model after the training
2. In the second release, 44% more people were added to the previous collection making the final count of 34320 people in 23571 images. All the newly added instances in Test set were also annotated with 2 additional annotators similar to the previous release. For rigorous analysis of annotators' agreement, each annotation in the Validation set was annotated with 4 additional distinct annotators. Since Validation set has more images as compared to the Test set, it has more definitive content for doing agreement analysis amongst the annotators.

The creation of EMOTIC constitutes multiple stages for both the releases; with some of them iterative. It began with collection of images while creating the emotion representation formats simultaneously. Both these stages influence one another heavily. *For example:* while using the name of an emotion category (from the set of emotion categories compiled (Section 2.1) thus far) to look for related images, sometimes, we encountered

people whose emotional states were not represented in the set of emotion categories. During such scenarios, the newly found emotional state of the person was added in the set of emotion categories, and the image was also added to the main collection (Section 3.1.1).

3.1.1 Image Data Collection

EMOTIC is an image based dataset containing people as its main subject. So, it is only natural that we started looking at the current available datasets. We sought those datasets which have similarities to the characteristics that we desire in our dataset (page 29). Another important aspect is to only collect images that have the subjects' (people, in our case) location available in the image. Figure 3.1 gives an example of this aspect. While generating annotations (Section 3.1.3), the subject needs to be localised in the image so that it is distinguishable from the context to avoid ambiguity.

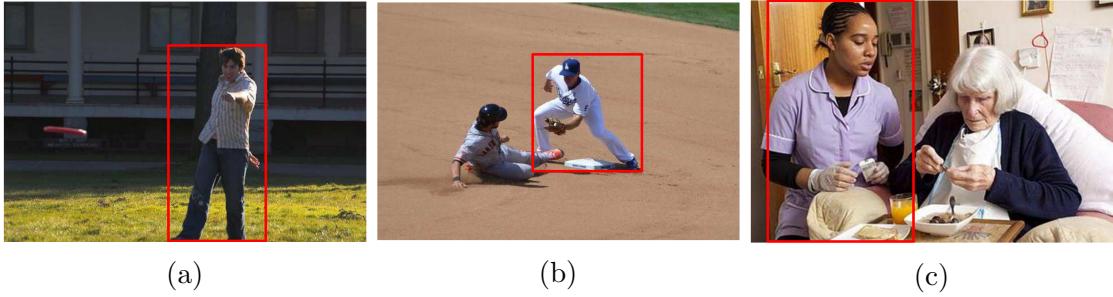


Figure 3.1: Example images from EMOTIC: The person-in-context is enclosed in a rectangular bounding-box

The main 3 sources for images in EMOTIC are the following:

1. **COCO (Lin et al. [2014]) (COCO)** - COCO is a large-scale object detection, segmentation, and captioning dataset. It also contains images with respective bounding-boxes for the objects (including people as shown in Figure 3.1) present in the images. There are 80 object categories that are annotated with their respective bounding-boxes and segmentation masks, including people. So, we parsed COCO for images containing only people and added those to the collection with their corresponding meta-data (including bounding-boxes)
2. **ADE20K (Scene Parsing Benchmark) (Zhou et al. [2017b])** - ADE20K is a scene parsing dataset with dense annotations of objects, including object parts. We parsed ADE20K and collected images containing people and added them to our collection along with their respective bounding-boxes

3. **Search Engines like Google** - We queried Google with words from the list of emotion categories (curated simultaneously (Section 2.1)), and collected images with desired characteristics (page 29). People in these images were manually localised with their respective bounding-boxes and then added in our collection

3.1.2 Emotion Representation format for EMOTIC

The emotion representation needs to be such that it can be handled computationally and is easy to compare for different images. Since there are numerous ways of ascribing emotional states using natural language, it becomes challenging to compare different annotations generated for the same image (section 2.1). In this thesis, we choose *viz.* *Affect Dimensions* and *Emotion Categories* as our emotion representation formats which we describe below.

3.1.2.1 Continuous Dimensions (or Affect Dimensions)

Affect Dimensions are very simple to understand and implement. Since the readings recorded from these dimensions could have real valued numbers, we call them *continuous dimensions*. The 3 continuous dimensions *Valence*, *Arousal* and *Dominance* were adopted as one of the formats for emotion representation for EMOTIC, Mehrabian [1995] called them Emotional State Model. In this model, emotions are represented as a tuple of (V, A, D) with values ranging from 1 to 10. Valence, Arousal and Dominance represent the axes of a 3D cartesian co-ordinate system. Valence represents the positiveness or pleasantness of an emotional state. A negative emotion has a lower Valence value while a positive emotion has a higher value (ref Figure 3.11). Similarly, Arousal represents the activeness of a person in a particular situation. If the person is calm then the Arousal values will be low, whereas for high activity the Arousal value will be high (ref Figure 3.12). The third dimension, Dominance, represents how much a person is in control of the situation. If a person is sad, in pain or is suffering for some reason, then she is not able to keep a check on herself under the circumstances. Lower Dominance value means that the person is not confident and unsure in the given situation. A high Dominance value means that the person comprehends the situation and is to some extent confident (ref Figure 3.13). Figure 3.2 shows examples of people annotated by their associated value of the given dimension.

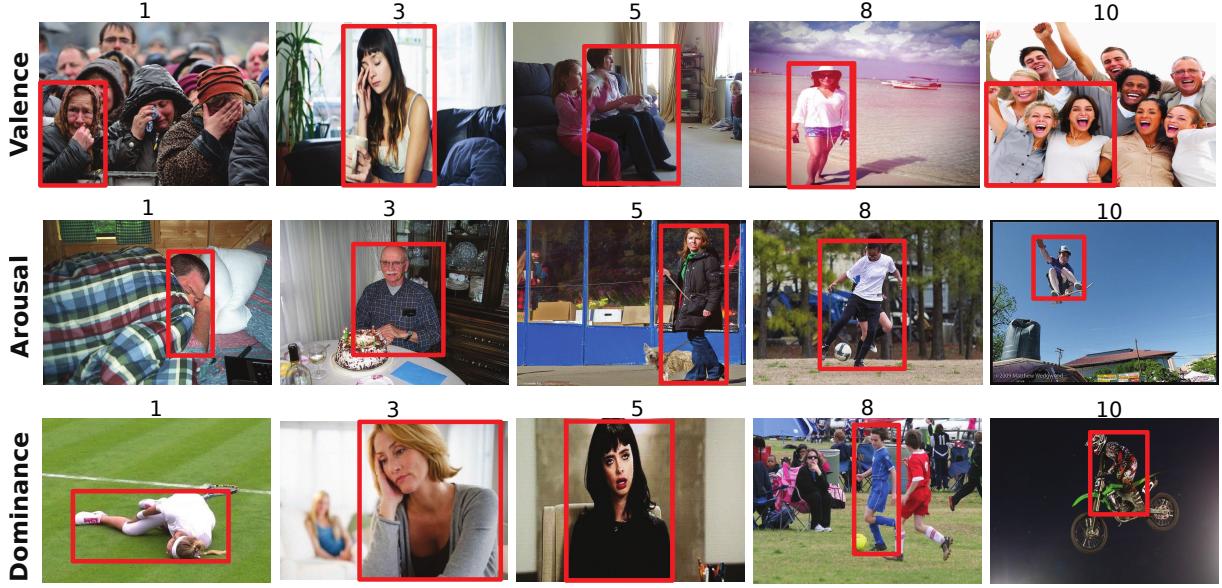


Figure 3.2: Examples of annotated images in the EMOTIC dataset for each of the 3 continuous dimensions viz. Valence, Arousal & Dominance. The person in the red bounding box has the corresponding value of the given dimension, mentioned at the top of each image

3.1.2.2 Emotion Categories

In the work by Cowen and Keltner [2017], the authors found through cross-validated regression that the affect dimensions fail to capture the whole spectrum of discrete emotion categories. In their regression analysis the authors found that their discrete categories (27 in all) were able to predict affect dimensions with 78% reliance, however, the affect dimensions could only capture 61% of the discrete categories - suggesting that the affect dimensions are not able to comprehensively represent the emotion space. Categorical emotions cover extensively affect dimensional space and not the other way around. *Continuous dimensions* alone aren't enough to represent the full breadth of emotional response. So we start building a list of emotion categories.

We collected an affect vocabulary from various resources like standard dictionaries (Dictionary [a], Dictionary [b]) and references on psychology (Picard [1997], Fernández-Abascal et al. [2010]). This vocabulary consists of a list of approximately 400 words representing a wide variety of emotional states. After a careful study of the definitions and the similarities amongst these definitions, we formed cluster of words with similar meanings. The clusters were formalized into 26 (Table 3.1) categories such that they were distinguishable from one another. For each category, it is possible to find an image of a person representing that emotion category. The final list of 26 affective categories (we call them *emotion categories*) were created taking into account the *Visual Separability*

criterion: words (that represent different affective states) that have similar definitions or meanings and not visually separable were grouped into one category. Those states for which it is not easy to find representative images were grouped into one category. For instance, *Anger* is defined by the words rage, furious and resentful. These affective states are different, but it is not always possible to separate them visually in a single image. Thus, the list of affective categories can be seen as a first level of a hierarchy, where each category has associated subcategories. It is interesting to note that the final list of emotion categories also includes the 6 basic emotions introduced by Ekman and Friesen [1969]. The emotion categories 2, 5, 16, 17, 21, 24 from Table 3.1 represent *Anger*, *Disgust*, *Fear*, *Happiness*, *Sadness*, *Surprise* respectively. However, there is one exception in that we used the more general term *Aversion* for the category *Disgust*. Thus, the category *Aversion* also includes the subcategories dislike, repulsion, and hate in addition to disgust. The list of the 26 emotional categories that represent various state of emotions and their corresponding definitions can be found in Table 3.1. Figure 3.3 gives 2 visual examples of each category. These images are selected from annotations of the EMOTIC dataset.



Figure 3.3: Examples of annotated people in EMOTIC dataset for each of the 26 emotion categories (Table 3.1). The person in the red bounding box is annotated by the corresponding category.

According to Keltner and Cordaro, the 26 number of emotion categories is not too large. They claim that the popularly known 6 basic emotions (Ekman and Friesen [1969]) are not exhaustive. In their work, they showed 2185 videos to their subjects. These videos elicit varied kinds of emotions. The people who watched the videos were asked to report their emotional experiences. The authors found out through an exhaustive analysis that there were at least 27 distinct emotional experiences reported. We found a significant

| |
|--|
| 1. Affection: fond feelings; love; tenderness |
| 2. Anger: intense displeasure or rage; furious; resentful |
| 3. Annoyance: bothered by something or someone; irritated; impatient; frustrated |
| 4. Anticipation: state of looking forward; hoping on or getting prepared for possible future events |
| 5. Aversion: feeling disgust, dislike, repulsion; feeling hate |
| 6. Confidence: feeling of being certain; conviction that an outcome will be favourable; encouraged; proud |
| 7. Disapproval: feeling that something is wrong or reprehensible; contempt; hostile |
| 8. Disconnection: feeling not interested in the main event of the surrounding; indifferent; bored; distracted |
| 9. Disquietment: nervous; worried; upset; anxious; tense; pressured; alarmed |
| 10. Doubt/Confusion: difficulty to understand or decide; thinking about different options |
| 11. Embarrassment: feeling ashamed or guilty |
| 12. Engagement: paying attention to something; absorbed into something; curious; interested |
| 13. Esteem: feelings of favorable opinion or judgment; respect; admiration; gratefulness |
| 14. Excitement: feeling enthusiasm; stimulated; energetic |
| 15. Fatigue: weariness; tiredness; sleepy |
| 16. Fear: feeling suspicious or afraid of danger, threat, evil or pain; horror |
| 17. Happiness: feeling delighted; feeling enjoyment or amusement |
| 18. Pain: physical suffering |
| 19. Peace: well being and relaxed; no worry; having positive thoughts or sensations; satisfied |
| 20. Pleasure: feeling of delight in the senses |
| 21. Sadness: feeling unhappy, sorrow, disappointed, or discouraged |
| 22. Sensitivity: feeling of being physically or emotionally wounded; feeling delicate or vulnerable |
| 23. Suffering: psychological or emotional pain; distressed; anguished |
| 24. Surprise: sudden discovery of something unexpected |
| 25. Sympathy: state of sharing others' emotions, goals or troubles; supportive; compassionate |
| 26. Yearning: strong desire to have something; jealous; envious; lust |

Table 3.1: Proposed emotion categories with definitions.

overlap between the 27 categories reported by the authors and the 26 emotion categories we defined in Table 3.1. The comparison between the 2 categories is reported in the Appendix A.

3.1.2.3 Combined Emotion Representation for EMOTIC

Studies conducted by various groups (Russell [2003]; Clore and Ortony [2013]; Scherer [2009]) have revealed the underlying *continuous dimensions* (*Valence*, *Arousal* and *Dominance*). In their work, Smith and Ellsworth [1985], review and ascertain the evidence of *Valence* and *Arousal* emotional dimensions. In addition, they discuss their new-found 6 appraisal dimensions affecting the emotional experience and showed how the dimensional approach influences the existing categorical approach. Multiple studies (Russell [1991]; Sabini and Silver [2005]) focused on the elicitation of emotional experiences that could be recorded in discrete forms of emotions like *anger* and *fear*. Russell [2003] also laid a descriptive ground-work to assist the combination of different modalities of emotion recognition, including *emotion categories* and *continuous dimensions*. The combination of *continuous dimensions* and *emotion categories* can be thought of a comprehensive tool for emotion annotation. Both capturing emotional states in different modalities. *Continuous Dimensions* use intensities across *Valence* and *Arousal* dimensions to capture the intensity of feeling; whereas *Emotion Categories* try to capture the essence of a specific emotion defined in a categorical fashion (Table 3.1). Each person in EMOTIC is annotated using both the formats of emotion representation to assist a deeper understanding of emotional state of people in different situations.

3.1.3 Collecting Annotations

After collecting images (Section 3.1.1) and building a comprehensive emotion representation format (Section 3.1.2), the next step is to get all the people in the images annotated. Since there are thousands of images, it is impossible to generate all the annotations in a lab or by a few people. A few hundred people are needed to generate all the required annotations. Also, it is not recommended to generate annotations by a fixed set of individuals since this might create invisible biases in the annotations. The emotion perception task is subjective in nature, so it is required that there are multiple different annotators.

Crowd-sourcing is a good method to gather annotations or labels in huge numbers with multiple different annotators. Amazon Mechanical Turk (AMT) is one such crowd-sourcing marketplace where there are annotators who can do such tasks anonymously. The annotators are called *Workers* and the people who launch their tasks for generating annotations are called *Requesters*. The tasks to be launched on AMT platform are called

Human Intelligence Task (HIT). The workers get paid accordingly for doing these HITs by the requesters. The identities of workers are hidden from requesters. Anyone residing in either of the 43 countries (Turk [a]) can sign-up on AMT to become a worker. For the first release of EMOTIC dataset, workers from 4 countries (Turk [b]) were allowed to become workers on the AMT platform, whereas for the second release of the EMOTIC dataset, workers from 43 countries (Turk [b]) were allowed to work on the AMT platform. AMT provides a huge workforce who have varied professional background, gender, age, demography, income and nationality (Ross et al. [2009]). According to the latest (2010 A.D.) study published on the background of the workers (Ross et al. [2010]), majority of the workers on AMT are young, highly-educated people with a good gender distribution - 48% males, 52% females. This ensures to certain extent that the responses will not be biased based on these criteria. AMT is used to collect the annotations for all the people in the EMOTIC dataset.

There are 3 HITs designed for generating all the annotations for EMOTIC. These are described as follows:

1. **Emotional Quotient (EQ) Task:** In this task, the workers are asked simple questions to gauge their emotional empathy skills. The questions are taken from a standard study done by Groen et al. [2015]. These questions are very general in nature asking about certain situations a person might face in real life and the decisions, thereby, he/she takes. These decisions (or responses) help us estimate their empathizing quotient. The worker needs to respond as if he is part of that situation
2. **Emotion Category (EC) Task:** The worker is shown the person-in-context along with all the 26 emotion categories. He has to put himself in that person's position, imagine how that person is feeling and select all the emotion categories that represent the emotional state of the person
3. **Continuous Dimension (CD) Task:** The worker is shown the person-in-context along with the 3 continuous dimensions. Again, he has to put himself in the place of person-in-context and choose the applicable levels for each of the continuous dimensions (Valence, Arousal and Dominance)

3.1.3.1 Interface Design

Three Human Intelligence Tasks (HITs), one for each of the 2 formats of emotion representation and one for EQ task are designed. The designing of the annotation interface

has two main focuses: (i) the task should be easy to understand interface-wise, and (ii) the interface fits the HIT in one screen which avoids scrolling.

EQ task design:

1. First page shown to the workers contains the *Disclaimer*, informing them that their responses are anonymous and will be used for research purposes only. The page briefly explains the tasks and also gives some general instructions about attempting the tasks. At the end of the page, they are notified about their browser settings. Figure 3.4 shows the first page of EQ task.
2. Figure 3.5 shows the main interface for EQ task. This page lists all the questions to be attempted
3. A warning message is displayed (as shown in Figure 3.6) if a worker does not attempt any of the questions or overlooks a question. The task is then not allowed to proceed until all the questions have been attempted
4. The next pages have one sample EC and CD tasks each along with their respective instructions. Their interface is exactly the same as the main tasks, shown in Figure 3.17

EC task design:

1. The first page (Figure 3.7) shows the disclaimer and instructions about the browser settings to the worker. It also mentions about the qualification requirement to attempt this particular task
2. Next pages show the instructions and an instance of how to annotate the images as shown in Figures 3.8, 3.9 respectively. It also shows the correct and incorrect ways of attempting the task as a guideline to the workers
3. Next, the worker is presented with the main task shown in Figure 3.17.a
4. If the worker misses by chance or tries to skip any question, then a warning sign similar to Figure 3.6 is shown and the worker is not allowed to proceed until the question is attempted
5. The instructions and examples (correct and incorrect), would be displayed at the bottom of each page during the main task, so that the worker can refer them in the same page without changing pages

CD task design:

1. First page of CD task is similar to that of EC task.

Analyzing how people feel

Disclaimer:

This is a Qualification task, after which you will be able to attempt HITs called 'Image Annotation Tasks' ([Link to the HITs](#)). Skip this Qualification task if you are able to open the link and attempt the HITs.

NOTE: This is a research study. Your anonymity is assured and your participation is voluntary. We do not receive or collect any personal information. You may decline further participation, at any time, without adverse consequences.

Instructions:

The goal of this study is to understand how people feel in different situations. You will be shown a pair of images containing people, and your task will be to describe how they are feeling.

The purpose of this qualification test is to ensure that you understand the task.

This qualification task is divided into 2 parts:

1. **First part:** You will be asked a set of questions (through multiple choices) to estimate your ability in interpreting other people's feelings. These questions come from a standardized test. The information gathered will not be linked to your ID or any information that might allow identifying you. Try to be accurate and honest in your responses.
2. **Second part:** You will be shown 2 images. Each image will have a form for you to fill in. The accuracy of your responses will be used to assign you a qualification.

Important:

1. Please turn off all website blockers that you might have installed in your browser, it might create problem to submit this HIT.
2. To navigate the previous or the next page, ONLY use the buttons at the bottom of the written text.
3. DO NOT use forward/backward button of the browser to navigate in this HIT, you will lose this HIT if you use browser buttons (read the above point again).

Click on the button 'Next' shown below to navigate to the next page.

Next

Figure 3.4: EQ task design: First Page showing disclaimers and instructions

Task: General Questions

Please, read the sentence carefully and select your level of agreement or disagreement from the 4 choices.

What is your Gender?

Male Female

1. I can easily tell if someone else wants to enter a conversation

Definitely Agree Slightly Agree Slightly Disagree Definitely Disagree

2. I find it difficult to explain to others things that I understand easily, when they don't understand it first time

Definitely Agree Slightly Agree Slightly Disagree Definitely Disagree

3. I really enjoy caring for other people

Definitely Agree Slightly Agree Slightly Disagree Definitely Disagree

4. I find it hard to know what to do in a social situation

Figure 3.5: EQ task design: Main page of EQ task asking the general questions

29. I am good at predicting what someone will do

Definitely Agree Slightly Agree Slightly Disagree Definitely Disagree

30. I tend to get emotionally involved with a friend's problems

Definitely Agree Slightly Agree Slightly Disagree Definitely Disagree

Back

Please attempt all the Questions

Continue

Figure 3.6: EQ task design: Warning message for EQ task if any question is missed (30th, in the above example)

Task: Image Annotation

IMPORTANT:

1. You must have the qualification called "Frames Labler val"
2. If you are unable to go 'Continue' using the button below, then turn off all website (or ad) blockers that you might have installed in your browser.
3. Navigate to the previous or the next page using the buttons shown at the bottom of the written text.
4. DO NOT use forward/backward button of the browser to navigate in this HIT, you will lose this HIT if you use browser buttons.

Click on the button 'Continue' below to navigate to the next page.

Figure 3.7: EC task design: First page showing disclaimer and instructions about the browser settings

Task: Image Annotation

NOTE: You are participating in a research study. Your anonymity is assured. The researchers who have requested your participation will not receive any personal information about you. Your participation in this research is voluntary. You may decline further participation, at any time, without adverse consequences.

Instructions (1/2)

- You will be shown an image with a yellow bounding box that will focus on a particular person.
- Your task is to think how that person feels given the situation in that image. Observe the whole image, not just the person and his/her face. Think also on how the surroundings affects the person.
- **Now consider each category separately. Look at the person and for each category think whether you see that particular emotion or state in this person or not.**
- And check all the relevant emotions that you think the person is feeling from the list of categories shown.

Figure 3.8: EC task design: Page showing instructions on how to attempt the task

2. Instruction page (similar to EC task) is shown in Figure 3.10
3. Next pages show the visual definition of each of the continuous dimensions. The levels (high and low) of each dimensions are shown with animated characters as well as sample example from the EMOTIC dataset. This visualization is targeted to help the worker for doing the CD task with visual aids. Figures 3.11, 3.12, 3.13 show the visual definitions of the dimensions
4. The next pages show (like in case of EC task), an example with correct and incorrect way of annotation. This is also serving as a guideline for the workers (Figures 3.14, 3.15, 3.16)
5. Next, the worker is presented with the main task shown in Figure 3.17.b
6. If the worker misses by chance or tries to skip any question, then a warning sign similar to Figure 3.6 is shown and the worker is not allowed to proceed until the question is attempted
7. The instructions and examples (correct and incorrect), would be displayed at the bottom of each page during the main task, so that the worker can refer them in the same page without changing pages

Instructions continued ... (2/2)

Example:

When I go through each of the emotion labels, I find that the person looks happy. He is engaged in conversation with people around him. Also, he feels excited to be amongst them. Hence, I should select all these emotions viz. Happiness, Engagement and Excitement. So, annotations like: Surprise, Disapproval or Annoyance would be considered incorrect.



| |
|---|
| <input type="checkbox"/> Peace (well being and relax) |
| <input type="checkbox"/> Affection (tend feelings/affection) |
| <input type="checkbox"/> Anticipation (state of expectation) |
| <input type="checkbox"/> Esteem (feelings of favor/esteem) |
| <input type="checkbox"/> Confidence (feeling of confidence) |
| <input checked="" type="checkbox"/> Engagement (paying attention/engaged) |
| <input type="checkbox"/> Pleasure (feeling of delight/pleasure) |
| <input type="checkbox"/> Happiness (feeling delight/happy) |
| <input checked="" type="checkbox"/> Excitement (feeling excited/excitement) |
| <input checked="" type="checkbox"/> Surprise (sudden discovery/surprised) |
| <input type="checkbox"/> Suffering (psychological suffering) |
| <input type="checkbox"/> Disapproval (feeling that something is wrong) |
| <input type="checkbox"/> Yearning (strong desire to have something) |
| <input type="checkbox"/> Fatigue (weariness/tiredness) |
| <input type="checkbox"/> Pain (physical suffering) |
| <input type="checkbox"/> Doubt/Confusion (doubt/confused) |
| <input type="checkbox"/> Fear (feeling suspicious or afraid) |
| <input type="checkbox"/> Sensitivity (feeling of being sensitive) |
| <input type="checkbox"/> Disturbance (upset/disturbed) |
| <input type="checkbox"/> Annoyance (bothered/irritated) |
| <input type="checkbox"/> Anger (intense displeasure) |
| <input type="checkbox"/> Disgust (feeling dislike or revulsion) |
| <input type="checkbox"/> Sadness (feeling unhappy/sad) |
| <input type="checkbox"/> Disconnection (not interacting) |
| <input type="checkbox"/> Embarrassment (feeling embarrassed) |
| <input type="checkbox"/> Sympathy (state of sharing) |



| |
|---|
| <input type="checkbox"/> Peace (well being and relax) |
| <input type="checkbox"/> Affection (tend feelings/affection) |
| <input type="checkbox"/> Anticipation (state of expectation) |
| <input type="checkbox"/> Esteem (feelings of favor/esteem) |
| <input type="checkbox"/> Confidence (feeling of confidence) |
| <input type="checkbox"/> Engagement (paying attention/engaged) |
| <input type="checkbox"/> Pleasure (feeling of delight/pleasure) |
| <input type="checkbox"/> Happiness (feeling delight/happy) |
| <input checked="" type="checkbox"/> Excitement (feeling excited/excitement) |
| <input checked="" type="checkbox"/> Surprise (sudden discovery/surprised) |
| <input type="checkbox"/> Suffering (psychological suffering) |
| <input checked="" type="checkbox"/> Disapproval (feeling that something is wrong) |
| <input type="checkbox"/> Yearning (strong desire to have something) |
| <input type="checkbox"/> Fatigue (weariness/tiredness) |
| <input type="checkbox"/> Pain (physical suffering) |
| <input type="checkbox"/> Doubt/Confusion (doubt/confused) |
| <input type="checkbox"/> Fear (feeling suspicious or afraid) |
| <input type="checkbox"/> Sensitivity (feeling of being sensitive) |
| <input type="checkbox"/> Disturbance (upset/disturbed) |
| <input checked="" type="checkbox"/> Annoyance (bothered/irritated) |
| <input type="checkbox"/> Anger (intense displeasure) |
| <input type="checkbox"/> Disgust (feeling dislike or revulsion) |
| <input type="checkbox"/> Sadness (feeling unhappy/sad) |
| <input type="checkbox"/> Disconnection (not interacting) |
| <input type="checkbox"/> Embarrassment (feeling embarrassed) |
| <input type="checkbox"/> Sympathy (state of sharing) |

[Back](#) [Continue to Images](#)

Figure 3.9: EC task design: Page showing the correct and incorrect ways of annotation

Task: Image Annotation

Instructions (1/2)

- You will be shown an image with a yellow bounding box that will focus on a particular person.
- Your task is to think how that person feels given the situation in that image. Observe the whole image, not just the person and his/her face. Think also on how the surroundings affect the person.
- ***Now consider each emotional dimension (viz. Valence, Arousal and Dominance) separately. Look at the person and for each of the three dimensions think what level does the person's feelings have on each of the dimensions.***
- Select that appropriate level for each of the dimensions
- Also, select the age and the gender of the person in the yellow box.

Figure 3.10: CD task design: Page showing the instructions for CD task

Definitions:

Valence: is the person having negative OR positive feelings?

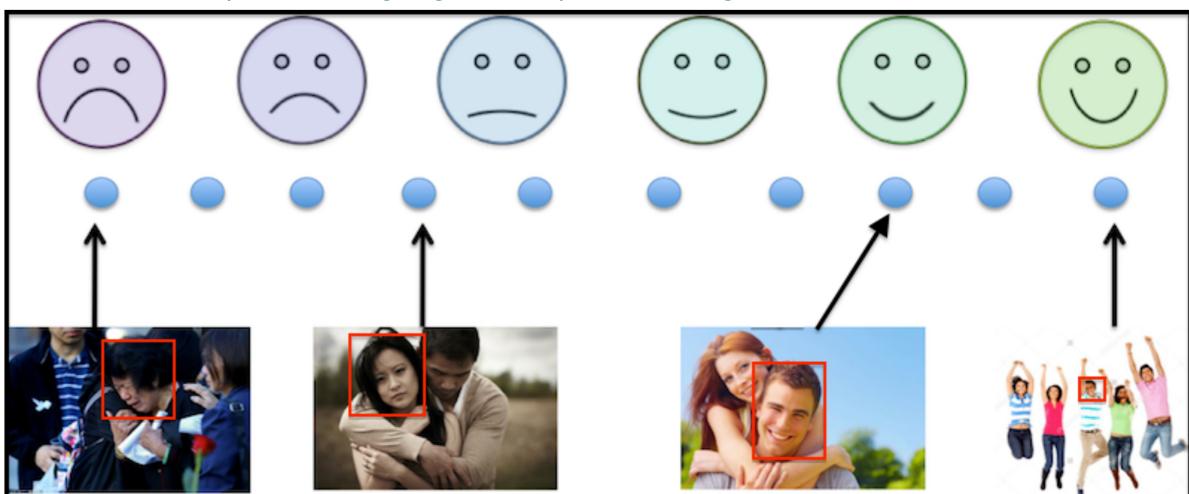


Figure 3.11: CD task design: Page showing the visual definition of Valence

Arousal: is the person calm (like sleepy) **OR** she feels energetic, agitated or excited and ready to act?

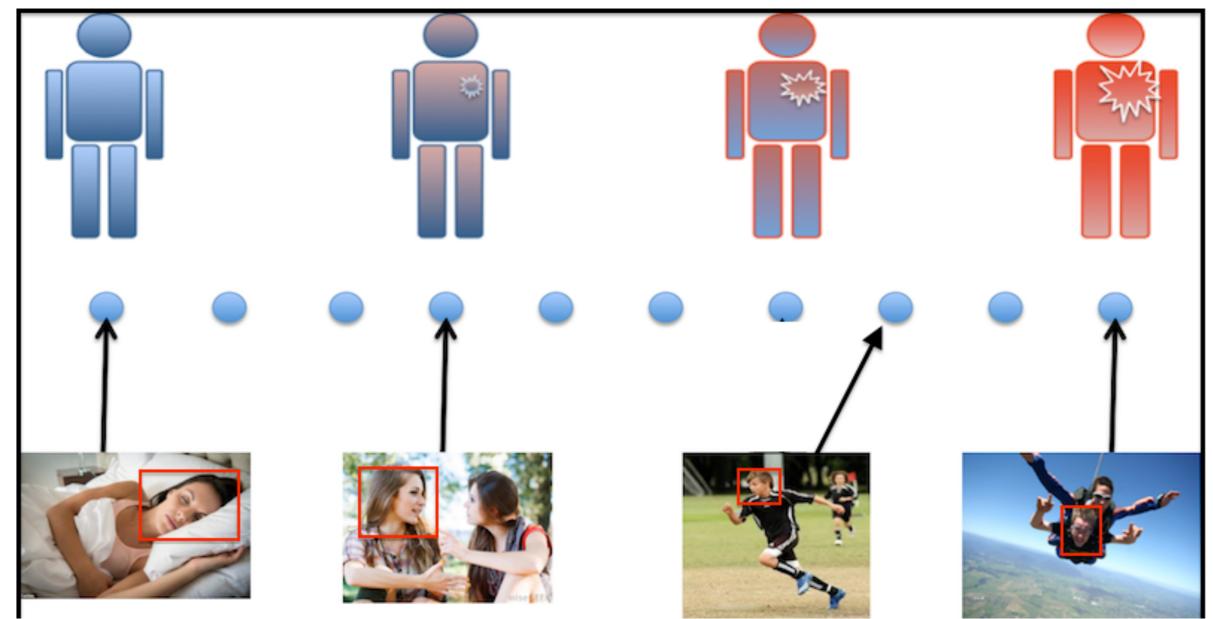


Figure 3.12: CD task design: Page showing the visual definition of Arousal

Dominance: does the person feel that the situation is not under her control **OR** she feels in control of the situation?



Figure 3.13: CD task design: Page showing the visual definition of Dominance

Instructions continued ... (2/2)

Example:

- We have to label each of the three emotional dimensions for this person.
- First, we observe the picture: the person is hugging a dog, she looks happy and relaxed. The dog is quiet and seems to enjoy this hugging as well. The environment looks quiet, we can see some vegetation and we notice it is sunny, the weather looks nice.
- Now let us analyze each one of the emotional dimensions:
 1. **Valence** (*is the person having positive OR negative feelings?*): In this case (image below), she seems to be experiencing a positive emotion. So, her valence emotional level is positive. If we can imagine situations where a person could experience even higher positive emotion we should rate this dimension with a high score, but not with the highest possible.
 2. **Arousal** (*is the person calm (like sleepy) OR she feels energetic, agitated or excited and ready to act?*): In this case (image below), She seems relaxed and quiet, her body pose indicates that she is not too active, so her arousal scale is towards calm. We can imagine, however, situations where a person can feel even more calm (for instance someone almost sleeping in the sofa). For this reason we score this dimension towards calm, but not the most calm possible.
 3. **Dominance** (*does the person feel that the situation is not under her control OR she feels in control of the situation?*): In this case (image below), the situation is not threatening at all, so the person seems to be in control. So, in the dominance scale, her emotional state suggests a high level of control of the situation.
- Also select the appropriate age and the gender of the person in yellow box.

Figure 3.14: CD task design: Page showing an example on how to attempt the CD task

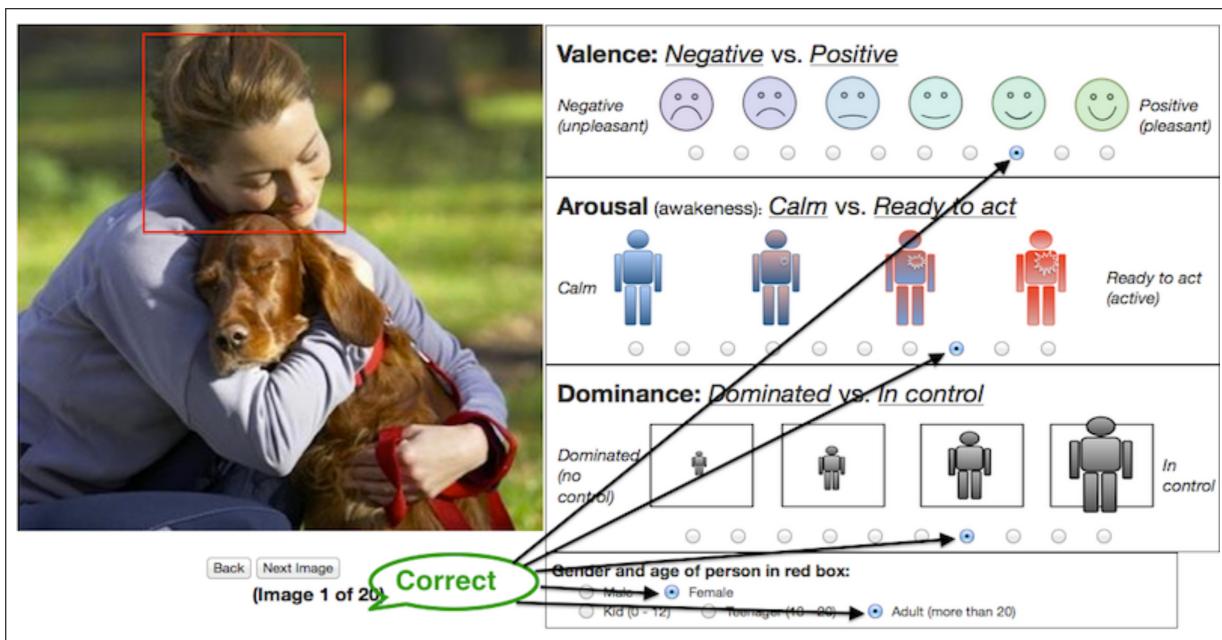


Figure 3.15: CD task design: Page showing a correct way of annotating in CD task



Figure 3.16: CD task design: Page showing an incorrect way of annotating in CD task

3.1.3.2 Annotation Quality Control Strategies

To ensure that the workers understand each task, we showed them how to annotate the images step-wise, by explaining two examples in detail. Also, instructions and examples (correct and incorrect both) were attached at the bottom on each page as a quick reference to the worker. Finally, a summary of the detailed instructions was shown at the top of each page (Table 3.2) so that the worker doesn't need to scroll to check the instructions each time.

Emotion Category

“Consider each emotion Category separately and, if it is applicable to the person in the given context, select that emotion category”

Continuous Dimension

“Consider each emotion dimension separately, observe what level is applicable to the person in the given context, and select that level”

Table 3.2: Instruction summary for each HIT

Quality control of EQ Task: In order to avoid random choice selection of the questions asked in the EQ task, 2 trivial questions (shown in Figure 3.18) were included whose correct and unambiguous response was known in advance. If the workers attempted these trivial questions incorrectly, they would be notified immediately that *one of their responses* was incorrect and that they need to change it response in order to proceed. This



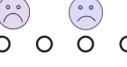
Peace (well being and relaxed/no worry/positive sensation/satisfied)
 Affection (fond feelings/tenderness/love/compassion)
 Expectation (state of anticipating/hoping on something or someone)
 Esteem (favorable opinion or judgment/gratefulness/admiration/respect)
 Confidence (feeling of being certain/proud/encouraged/optimistic)
 Engagement (occupied/absorbed/interested/paying attention to something)
 Pleasure (feeling of delight in the senses)
 Happiness (feeling delighted/enjoyment/amusement)
 Excitement (pleasant and excited state/stimulated/energetic/enthusiastic)
 Surprise (sudden discovery of something unexpected)
 Suffering (distressed/perturbed/anguished)
 Disapproval (think that something is wrong or reprehensible/contempt/hostile)
 Yearning (strong desire to have something/jealous/envious)
 Fatigue (weariness/tiredness/sleepy)
 Pain (physical suffering)
 Doubt/Confusion (difficulty to understand or decide/sceptical/lost)
 Fear (feeling afraid of danger/evil/pain/horror)
 Vulnerability (feeling of being physically or emotionally wounded)
 Disquietment (unpleasant restlessness/tense/worried/upset/stressed)
 Annoyance (bothered/irritated/impatient/troubled/frustrated)
 Anger (intense displeasure or rage/furious/resentful)
 Disgust (feeling dislike or repulsion/feeling hateful)
 Sadness (feeling unhappy/grief/disappointed/discouraged)
 Disconnection (not participating/indifferent/bored/distracted)
 Embarrassment (feeling ashamed or guilty)

[Back](#)
(Image 1 of 20)
[Go to Next Image](#)

(a) EC (Emotion Categories) task



Valence: Negative vs. Positive

| | | |
|--------------------------|---|------------------------|
| Negative (unpleasant) |  | Positive (pleasant) |
|--------------------------|---|------------------------|

Arousal (awakeness): Calm vs. Ready to act

| | | |
|------|---|--------------------------|
| Calm |  | Ready to act (active) |
|------|---|--------------------------|

Dominance: Dominated vs. In control

| | | |
|---------------------------|---|------------|
| Dominated (no control) |  | In control |
|---------------------------|---|------------|

Gender and age of the person in the yellow box

| | | |
|----------------------------------|--|--|
| <input type="radio"/> Male | <input type="radio"/> Female | |
| <input type="radio"/> Kid (0-12) | <input type="radio"/> Teenager (13-20) | <input type="radio"/> Adult (more than 20) |

[Back](#)
(Image 1 of 20)
[Go to Next Image](#)

(b) CD (Continuous Dimensions) task

Figure 3.17: AMT interface designs

helped control random choice selection by the workers. This made them aware that random responses will not be accepted, and they had to pay attention and attempt each question faithfully.

19. Snow is white in color

Definitely Agree Slightly Agree Slightly Disagree Definitely Disagree

(a) First trivial question asked on the EQ task

25. When you take two apples in left hand and three apples in right hand, in total you have five apples

Definitely Agree Slightly Agree Slightly Disagree Definitely Disagree

(b) Second trivial question asked on the EQ task

Your responses are not consistent, Please correct them

Back

Continue

(c) Warning message if either of the questions ((a) or (b)) is incorrect

Figure 3.18: Quality Control for EQ task by asking 2 trivial questions ((a) or (b)) and the warning sign (c) that doesn't allow to proceed if these questions have not been answered correctly

Quality control for EC and CD Tasks: Multiple strategies were adopted to have quality annotations in the EMOTIC dataset and avoid noise as much as possible without biasing the annotations.

1. A qualification task is conducted to shortlist viable workers who could understand and perform the main tasks (EC and CD) well. The qualification task has two parts: **(i)** The EQ task itself served as first part of the qualification task, and **(ii)** 2 sample image annotation tasks - one for each of our 2 emotion representations (emotion categories and continuous dimensions). The acceptable responses for the sample annotations were known in advance. The responses of the workers to this qualification task were evaluated and those who responded satisfactorily were chosen

as our main candidates. These workers were then allowed to do the EC and CD tasks to annotate images from EMOTIC dataset.

2. To avoid noisy annotations, 2 control images were randomly inserted in every annotation batch of 20 images. Again, the correct set of labels for the control images was known beforehand. Workers selecting incorrect labels on these control images were warned and if they still kept annotating incorrectly, they were not allowed to annotate further and their annotations were discarded.
3. According to Lerman and Hogg [2014], *random policy* is best for unbiased estimates of preferences. The authors mainly experiment on the influence of position bias on 4 different policies for presentation; where they conclude that *random policy* (in which the order of items shown to the participants is randomized) is best suited for unbiased estimates of recommendations. However sometimes it is not completely removed by simple rotation of the multiple choices (Blunch [1984]). But we overcome these limitations by randomizing the order of appearance of emotion labels in our data-generation steps. Following their research conclusion, the order of emotion categories shown to every worker was randomised. This helped us avoid *Position Bias*
4. EMOTIC is divided into three sub-sets *viz.* Train, Validation and Test. Individual annotations for Validation and Test were augmented by adding multiple worker responses. Particularly, each Validation set annotations was augmented by 4 more, resulting in 5 annotations for each person-in-context in the Validation set. Similarly, for Test set, in total there are 3 annotations for each person-in-context. In order to avoid same worker annotating same person-in-context again, those tasks were concealed from the workers who had already annotated those samples. In this manner, each of the multiple annotations in Validation and Test sets are by different workers. This helped bring variety in the annotations.

3.2 Analysis

We have the capacity of making reasonable guesses about other people's emotional state because of our capacity of being empathetic, putting ourselves into another's situation, and also because of our common sense knowledge and our ability for reasoning about visual information.

The highlight of EMOTIC is that it contains annotated images of people with the following characteristics:

1. Faces are not visible in entirety, quite often either the head (back-side) of the face or a profile from the side is visible (for example see Figures 3.19.b & 3.19.d). More than 25% of the people in EMOTIC have their faces partially occluded or with very low resolution, so we can not rely in facial expression analysis for recognizing their emotional state.
2. Body is not visible in entirety, quite often only the upper torso is visible (for example see Figures 3.19.b and 3.19.c)

EMOTIC presents a different task - to estimate people's emotions without directly using their facial expressions and body postures. Among the images of EMOTIC, a lot of them have significant partial occlusions in the face, or faces are shown in non-frontal views. For this reason, the task of estimating person's emotional state can not be approached with facial expression analysis only, presenting us with a new challenging task.

Figure 3.19 shows sample annotated images in the EMOTIC dataset. Figure 3.19.a shows that the person is performing an activity that needs attention to predict the upcoming curve on the road, so he is feeling *anticipation*. Since he is doing a thrilling activity, he feels *excited* about it and he is *engaged* or focused in this activity to avoid unexpected injuries. This explains the emotion categories annotated. Also, in terms of continuous dimensions, since he is engaged in a precarious activity and it seems that he likes it, the valence value is 6 - signifying a little positivity in his feelings. The high Arousal value (= 9) clearly shows that he is involved in an intense activity. When one looks at his posture, it is apparent that he is in control and is confident - this explains the high Dominance value (= 10). Similar interpretations can be made about other annotations.

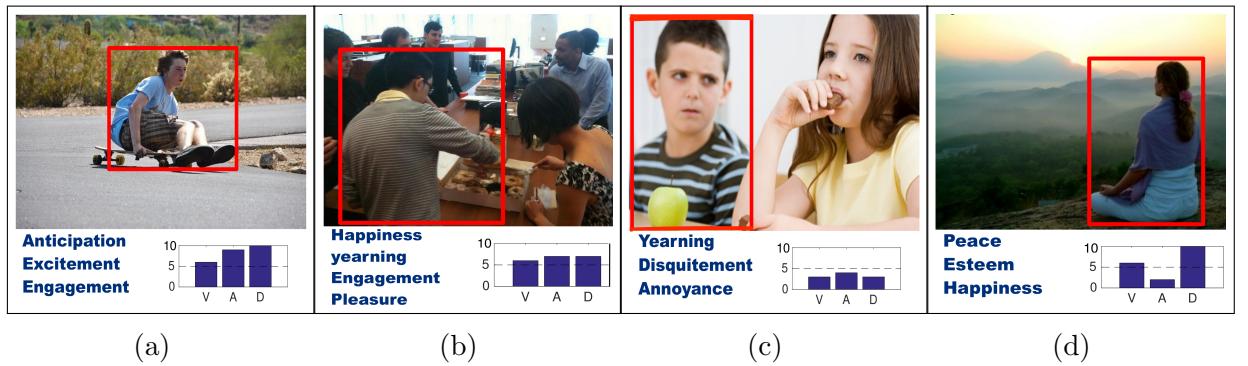


Figure 3.19: Sample Images from EMOTIC dataset with their corresponding annotations in both the formats

In Figure 3.19.c, the kid feels a strong desire (*yearning*) for eating the chocolate instead of the apple. From his facial expression we can see that he is a bit *disquiet* and *annoyed*

about it. Apparently, he feels disappointed with the situation, hence the low Valence (= 3) and Dominance (= 3) values. Figures 3.19.b & 3.19.d shows people whose faces are not visible, however, they have been annotated as well. In fact, if we consider the context, we can make reasonable estimation (like in real world) about emotional states even when the face of the person not visible (as illustrated in Figures 3.19.b & 3.19.d) The person in bounding-box of Figure 3.19.b is picking a doughnut and he probably *yearns* for it. He is participating in a social event with his colleagues, showing *engagement*. We can also say that he is also feeling *pleasure* eating the doughnuts; and is also possibly *happy* for the relaxed break along with other people. In Figure 3.19.d, the person is admiring the beautiful landscape with *esteem*. She seems to be enjoying the moment (*happiness*), and she seems calmed and relaxed (*peace*). We do not know exactly what is on the people's minds, but we are able to reasonably extract relevant affective information just by looking at them in their situations.

After the first phase of annotations (1 annotation per person-in-context), the images were divided into three sets: **Train** (70%), **Validation** (10%), and **Test** (20%) sets maintaining a similar affective category distribution across the different sets. After that, Test set was annotated by 2 additional distinct annotators to analyse the annotation agreements amongst the annotators. In second phase, all the images in Validation set were annotated by an additional 4 distinct annotators per annotation. The Validation set annotations (being much higher in numbers) were used to study the consistency of the annotations across different annotators (more in Section 3.2.2). The dataset statistics and algorithmic analysis on the EMOTIC dataset are detailed in Sections 3.2.1 and 3.2.3 respectively.

3.2.1 Statistical Analysis

EMOTIC dataset is a collection of images of people in unconstrained environments annotated according to their apparent emotional states. The dataset contains 23,571 images and 34,320 annotated people. Overall, the images show a wide diversity of contexts, containing people in different places, social settings, and engaged in diverse activities. The posture of people is not limited either. As shown in Figure 3.19, sometimes only the face or upper body part is visible (Figure 3.19.c), sometimes we see the whole body but the face is not visible (Figure 3.19.d) and sometimes only the upper body part is visible while the face is occluded.

The last release of the EMOTIC dataset contains 34,320 annotated people, where 66% of them are males and 34% of them are females. There are 10% children, 7% teenagers and 83% adults amongst them. Figure 3.20 shows the number of annotated people for

each of the 26 emotion categories, sorted by decreasing order. Notice that the data is unbalanced amongst the categories, which makes the dataset particularly challenging. An interesting observation is that there are more examples for categories associated to positive emotions, like *Happiness* or *Pleasure*, than for categories associated with negative emotions, like *Pain* or *Embarrassment*. The category with most examples is *Engagement*. This is because in most of the images people are doing something or are involved in some activity, showing some degree of engagement. Figures 3.21.a, 3.21.b and 3.21.c show the number of annotated people for each value of the 3 continuous dimensions. In this case as well unbalanced data is observed, but it is fairly distributed across the 3 dimensions which is good for modelling.

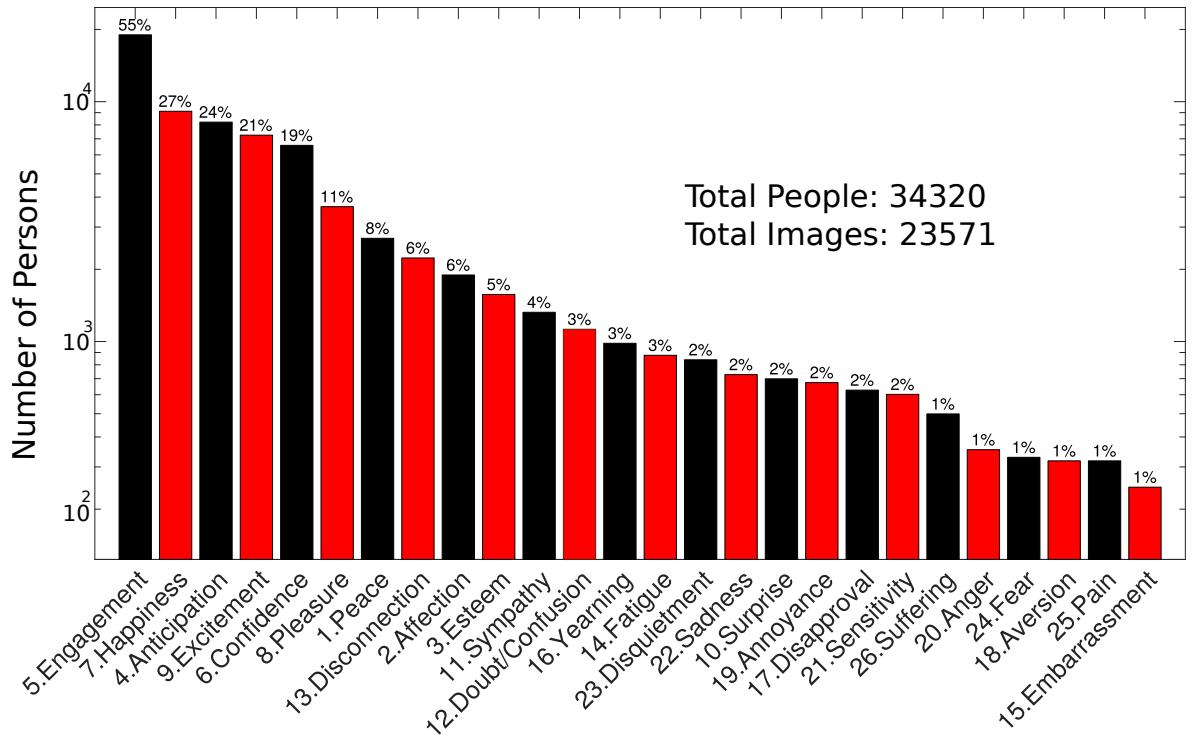


Figure 3.20: EMOTIC Statistics: Number of people annotated for each emotion category

EMOTIC shows interesting patterns of category co-occurrences. For example, after computing conditional probabilities, Figure 3.22 shows the co-occurrence rates of any two categories. Every value in the matrix (r, c) (r represents the row category and c column category) is a co-occurrence probability (in %) of category r if the annotation also contains the category c , that is, $P(r|c)$. It is observed, for instance, that when a person is labelled with the category *Annoyance*, then there is 46.05% probability that this person is also annotated by the category *Anger*. This means that when a person seems to be feeling *Annoyance* it is likely (by 46.05%) that this person might also be feeling *Anger*. We also used a k-means (Kanungo et al. [2002]) clustering on the category annotations

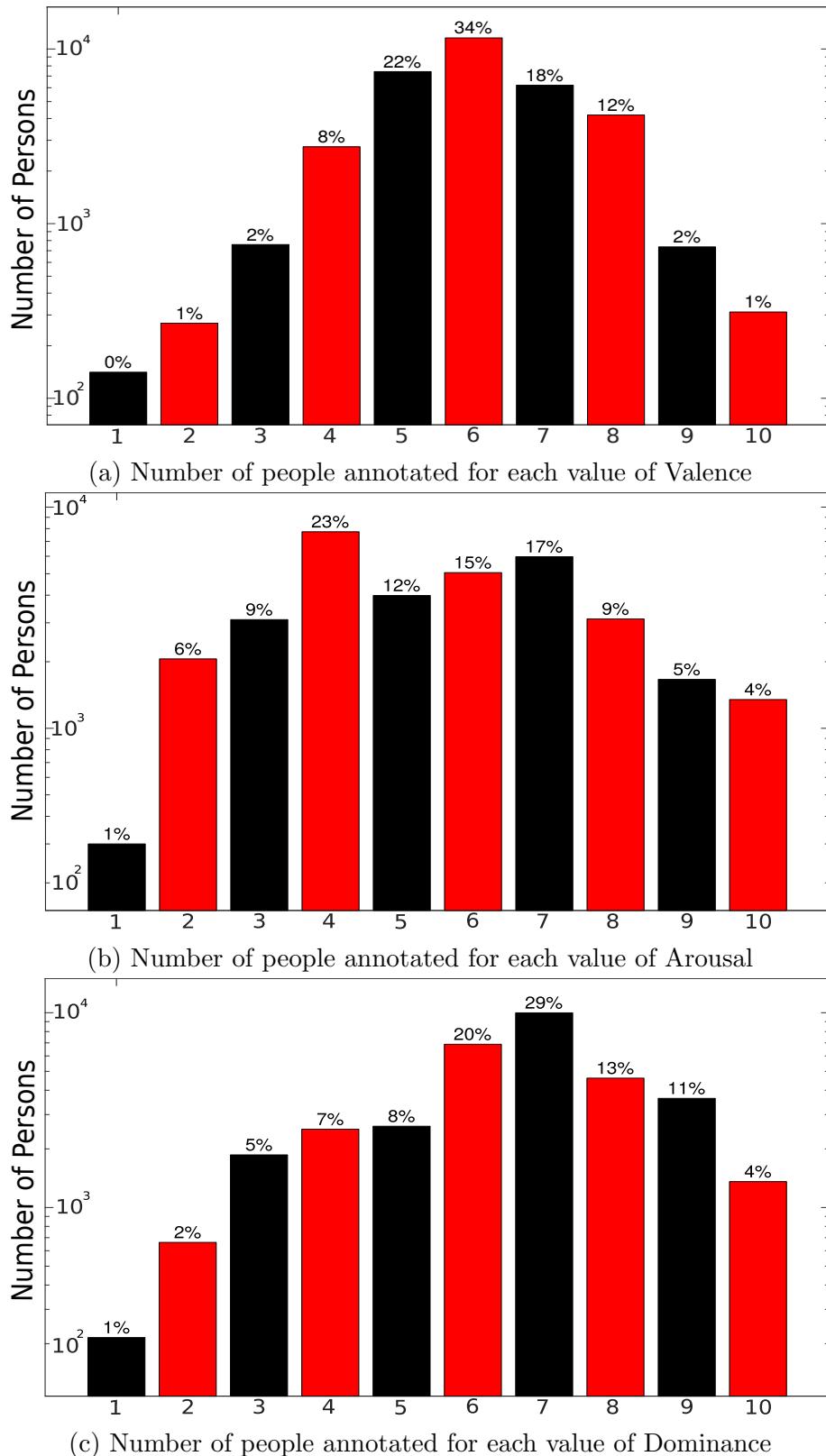


Figure 3.21: EMOTIC Statistics: Number of people annotated for every value of the three continuous dimensions (a,b,c)

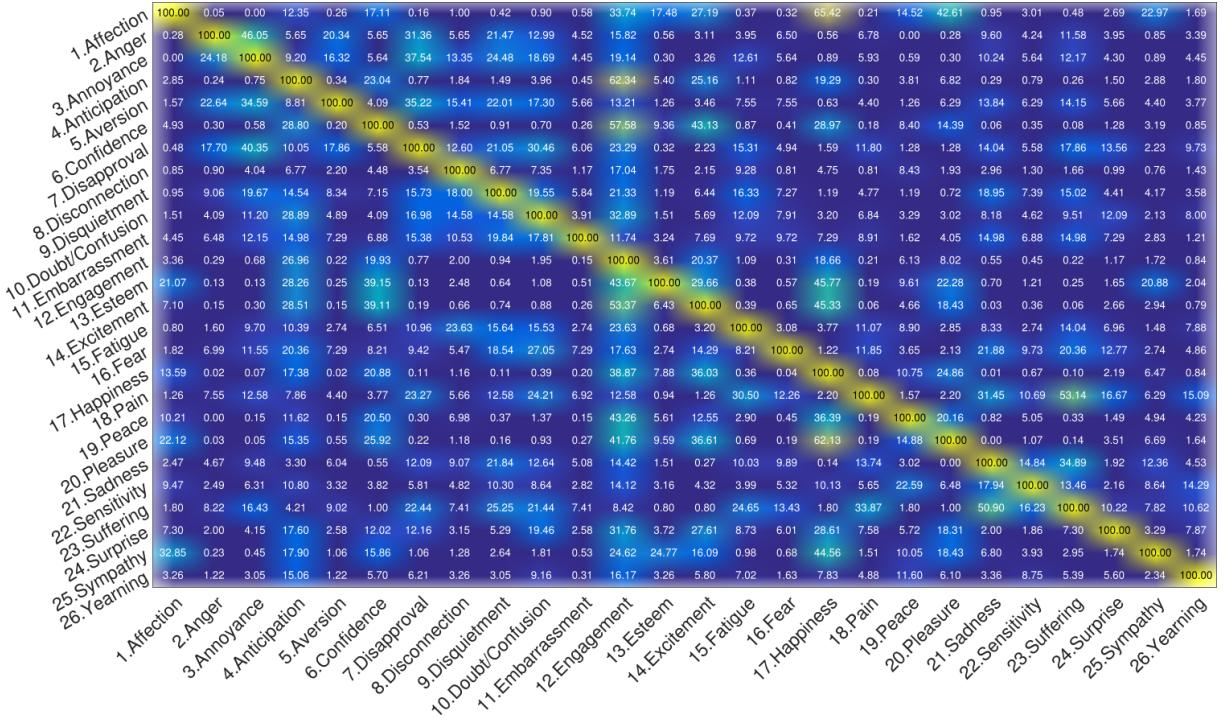


Figure 3.22: Co-variance between 26 emotion categories. Each row represents the occurrence probability of every other category given the category of that particular row.

to find groups of categories that occur frequently. It was found, for example, that these category groups are common in the EMOTIC annotations: $\{Anticipation, Engagement, Confidence\}$, $\{Affection, Happiness, Pleasure\}$, $\{Doubt/Confusion, Disapproval, Annoyance\}$, $\{Yearning, Annoyance, Disquietment\}$.

Figure 3.23 shows the distribution of each continuous dimension across the different emotion categories. For each plot, categories are arranged in increasing order of their average values of the given dimension (calculated for all the instances containing that particular category). Thus, it is observed from Figure 3.23.a that emotion categories like *Suffering*, *Annoyance*, *Pain* correlate with low Valence values (feeling less positive) in average whereas emotion categories like *Pleasure*, *Happiness*, *Affection* correlate with higher Valence values (feeling more positive). Also interesting is to note that a category like *Disconnection* lies in the mid-range of Valence value which makes sense. When we observe Figure 3.23.b, it is easy to follow that emotional categories like *Disconnection*, *Fatigue*, *Sadness* show low Arousal values and we see high activeness for emotion categories like *Anticipation*, *Confidence*, *Excitement*. Finally, Figure 3.23.c shows that people are not in control when they show emotion like *Suffering*, *Pain*, *Sadness* whereas when the Dominance is high, emotion categories like *Esteem*, *Excitement*, *Confidence* occur more

3.2. ANALYSIS

55

often.

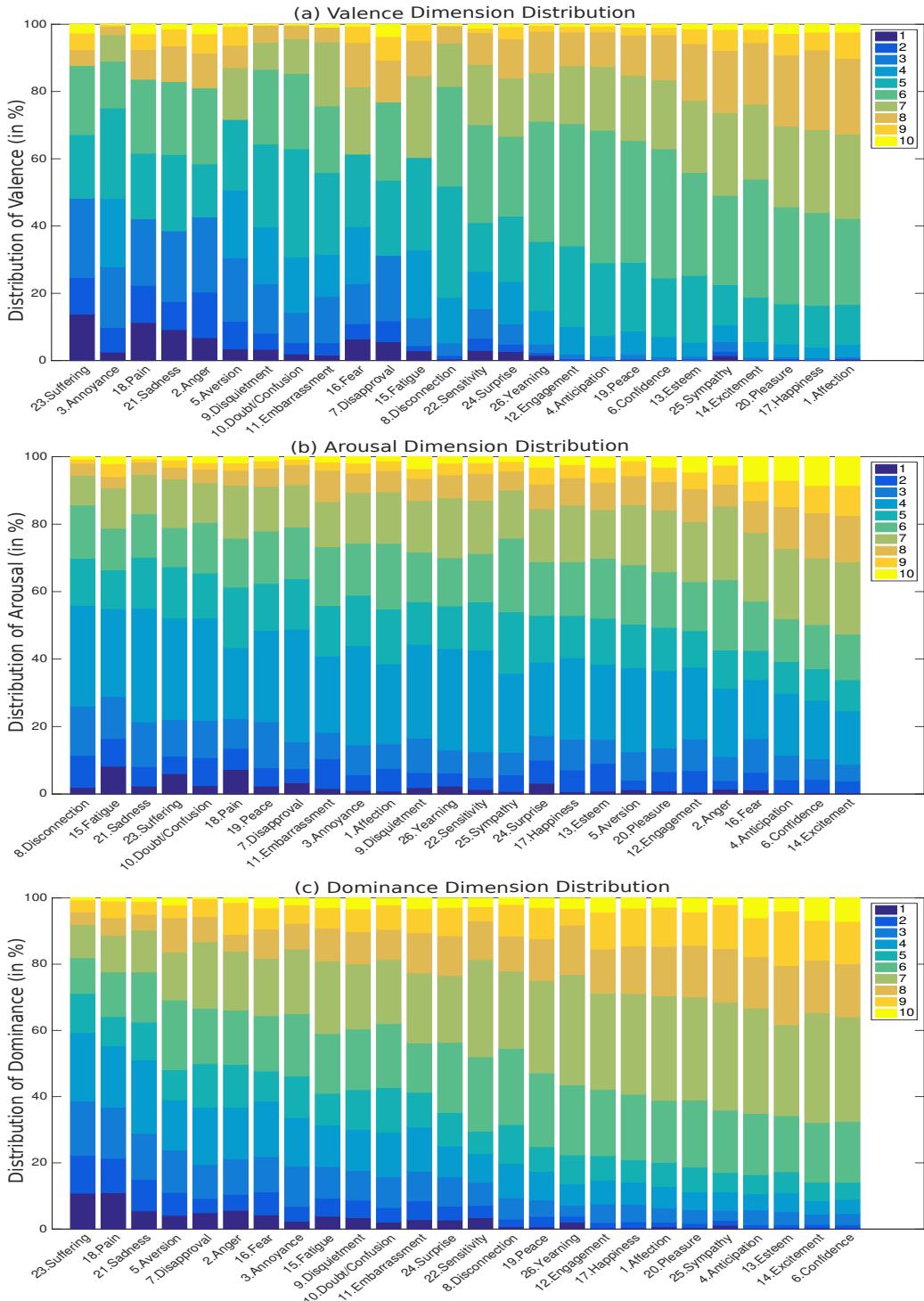


Figure 3.23: Distribution of continuous dimension values across emotion categories. Average value of a dimension is calculated for every category and then plotted in increasing order for every distribution.

3.2.2 Annotator Agreement Analysis

There is no direct measure to find the agreement between the annotators given the subjective nature of the annotation task. In this section a study on the annotators' agreement level using the images in the Validation set is presented. 2 methods are used to measure the annotation consistency. Since emotion perception is a subjective task, each perceiver can recognise different emotions after seeing the same image. For example in both Figure 3.24.a and 3.24.b, the person in bounding-box seems to feel *Affection*, *Happiness* and *Pleasure* and the annotators have annotated with these categories with consistency. However, not everyone has selected all these emotion categories. Also, it is seen that annotators do not agree in the emotions *Excitement* and *Engagement*. However, these categories are reasonable in this situation. Another example is that of *Roger Federer* hitting a tennis ball in Figure 3.24.c. He is seen predicting the ball (or *Anticipating*) and clearly looks *Engaged* in the activity. He also seems *Confident* in getting the ball. In spite of the annotation process being subjective in nature and not all annotators agreeing on every annotation, their responses have good quality and subtlety.

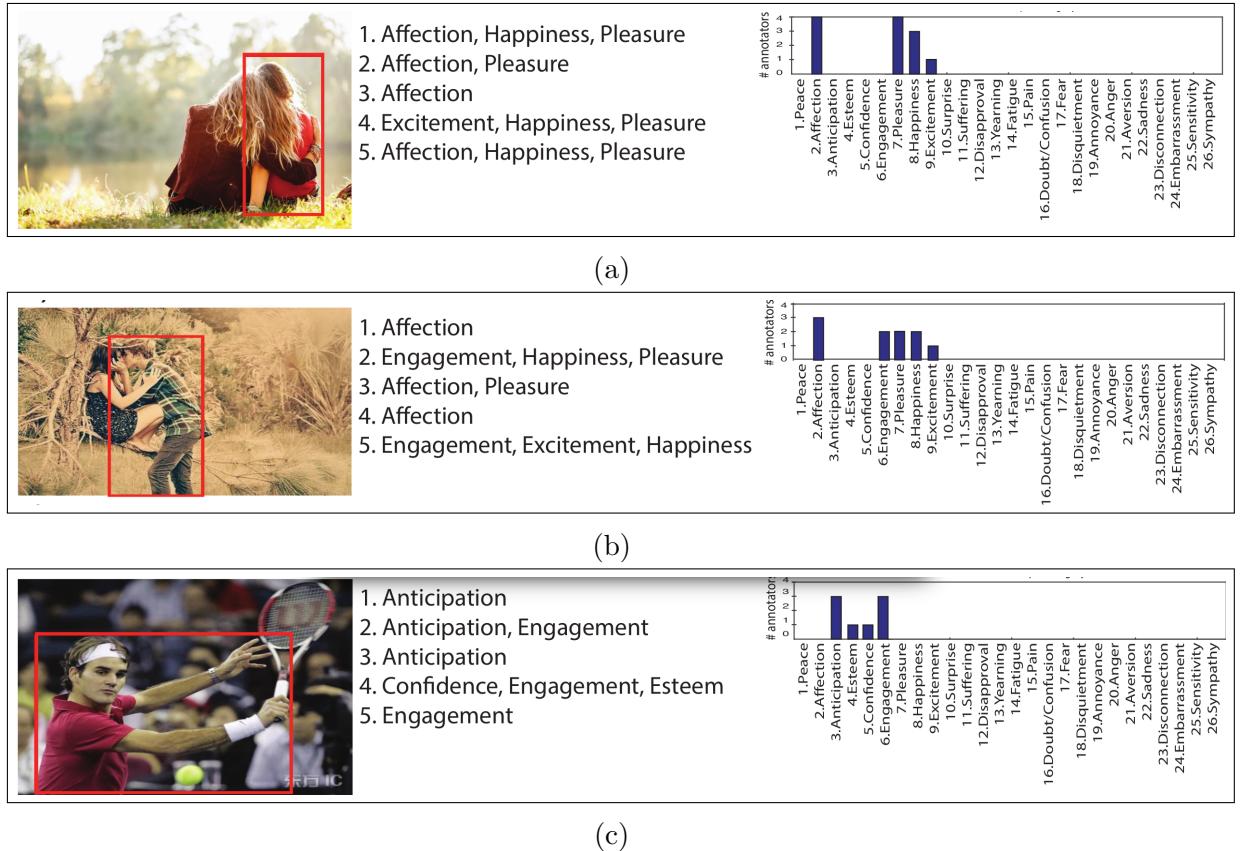


Figure 3.24: Five different annotators for a given person in context

After these observations and finding interesting co-occurrences amongst the categories

in the statistical analysis (Section 3.2.1), different quantitative analysis on the annotation agreement were conducted. First focus was on analysing the agreement level in the category annotation. Given a category annotated (or assigned) to a person in an image, the number of annotators agreeing for that particular category is considered as an agreement measure. Accordingly, it was calculated, for each category and for each annotation in the validation set, the agreement amongst the annotators and those values were sorted across categories. Figure 3.25 shows the distribution on the percentage of annotators agreeing for an annotated category across the validation set.

There seems a need to find a criteria with which we could compare annotator-agreement analysis amongst the discrete categories. Normalize each category with the number of people annotated for that category, then empirically weigh the number of people annotated by 5, 4, 3, 2, 1 annotators and quantify in the form of a rank (an annotation agreed upon by 5 annotators has the highest importance and is given the highest weight). This rank ranges, in case of EMOTIC dataset, between [1.04, 2.87]. Practical values of this rank have the limits $[0, N]$ - where N is the number of annotators for each annotation. Accordingly, the categories are sorted based on this rank and plotted in decreasing order of *annotator-agreement* in Fig. 3.25. We observe that *Engagement* has the highest *annotator-agreement* which means that for each instance that *Engagement* is annotated, 62% of times 3 or more annotators (out of 5) agree. Similarly, for *Pain*, of all the instances where it is annotated, there are 2 or more annotators who agree 26% of times.

The agreement between all the annotators for a given person using *Fleiss' Kappa* (κ) was also computed. *Fleiss' Kappa* is a common measure to evaluate the agreement level among a fixed number of annotators when assigning categories to data. In general, for the validation set, if an annotator selects an emotion category, the probability that he is in agreement with at least one of the four other annotators in selecting this category is 50%. In case of EMOTIC, given a person to annotate, there is a subset of 26 categories. If we have N annotators per image, that means that each of the 26 categories can be selected by n annotators, where $0 \leq n \leq N$. Given an image we compute the Fleiss' Kappa per each emotion category first, and then the general agreement level on this image is computed as the average of these Fleiss' Kappa values across the different emotion categories. We obtained that more than 50% of the images have $\kappa > 0.30$. Figure 3.26.a shows the distribution of kappa values across the validation set for all the annotated people in the validation set, sorted in decreasing order.

Keeping the annotations' parameters constant, we tried to find a *random* agreement between the annotators. This random agreement value, over 1000 iterations for EMOTIC is $\kappa \approx 0.15$. Notice that total disagreement gives $\kappa = 0$. The random kappa value

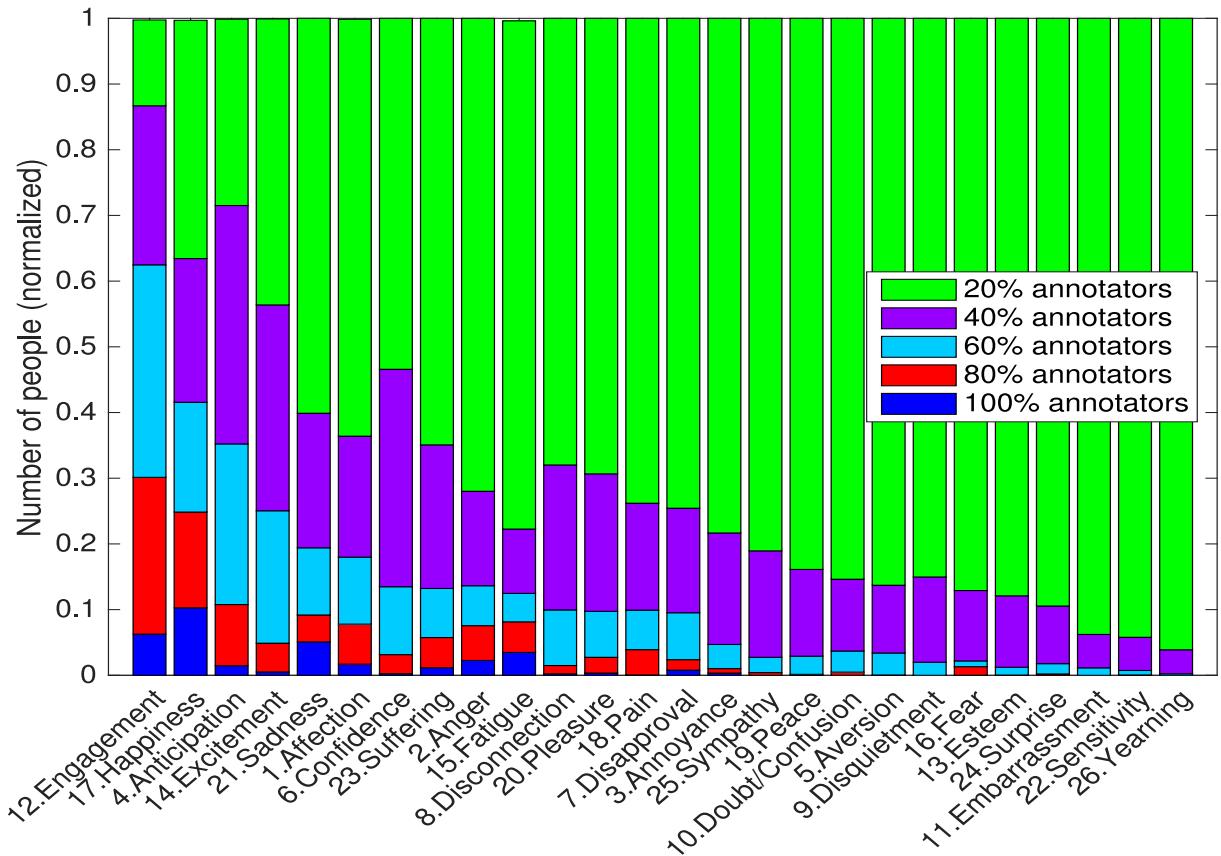


Figure 3.25: Representation of agreement between multiple annotators. Categories are sorted in decreasing order according to the average number of annotators that agreed for the category.

$(\kappa \approx 0.15)$ in comparison to the actual value ($\kappa > 0.30$) indicates that there is a significant agreement level even though the task of emotion recognition is subjective.

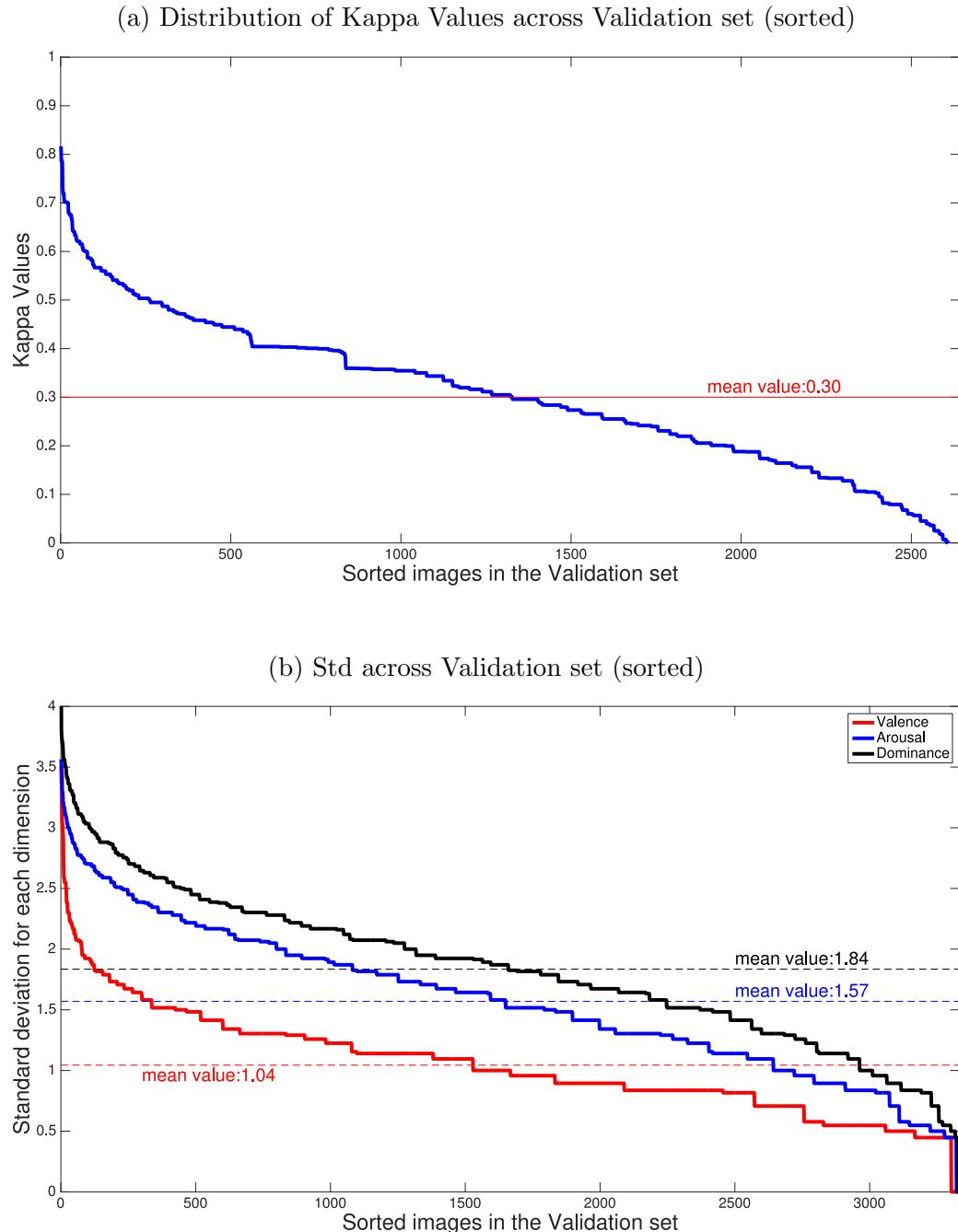


Figure 3.26: (a) Kappa values and (b) Standard deviation (Std), for each annotated person in validation set

Regarding to the continuous dimensions, the agreement is measured by the standard deviation (SD) of the different annotations. In general, the average SD across the Validation set is 1.04, 1.57 and 1.84 for Valence, Arousal and Dominance respectively - indicating

that Dominance has higher (± 1.84) dispersion than the other dimensions. It reflects that annotators disagree more often for Dominance than for the other dimensions which is understandable since Dominance is more difficult to interpret than Valence or Arousal [1995]. As a summary, Figure 3.26.b shows the standard deviations of all the images in the validation set for all the 3 dimensions, sorted in decreasing order.

An important aspect of doing agreement analysis is the tool or method used. For example, the agreement between the annotators decreases if the scales for capturing the responses is increased (Whitehill et al. [2014]). In general, random agreement between annotators is higher for a binary scale (1 or 0) as compared to when there are n options to choose from ($n > 2$). We did a similar agreement analysis for continuous dimension's representation for EMOTIC. Reducing the scale from [1 – 10] to [1 – 5], we re-calculated the average SD across the Validation set and found that it decreases, suggesting higher agreement. The *new* average SD across the Validation set in contrast to the previous values are **(0.54, 1.04)**, **(0.82, 1.57)**, **(0.94, 1.84)** for *Valence*, *Arousal* and *Dominance* respectively. Similar interpretations can be made for the new values, however, the important point to note is that the SD decreases when we reduce the scales. Clearly, lower SD indicates better agreements, depending on the scale used.

The average values of each dimension for a given category is also a good characterization of annotation agreement. For example, *Affection* has $(V, A, D) = (6.8, 5.3, 6.6)$ - suggesting high positiveness, medium activeness and high control. This interpretation makes sense when we see *Affection* in Figure 3.3(2). Similarly, for *Suffering*, $(V, A, D) = (3.7, 4.7, 4.3)$ - low positiveness (or high negativity), medium-low activeness and low control. Again, when we observe a person who is *Suffering* (example: Figure 3.3(26)), we see that he is feeling negative emotions, is not too aroused and is not in control. Such comparisons are consistent across categories indicating good agreement amongst annotators.

3.2.3 Algorithmic Analysis

EMOTIC Dataset contributes to the research community with rich data to understand people's emotions in various contexts. In this section state-of-the-art scene recognition systems by Zhou et al. [2017a] are used to observe interesting patterns in the distribution of emotions shown in different places or environments.

A CNN trained on Places dataset (Zhou et al. [2017a]) is used to predict the scene-category and scene-attributes for all the images in EMOTIC. With this information, the dataset analysis for EMOTIC is divided per place category to show additional statistics of the dataset. Figure 3.27.a shows the probability of each emotion conditioned to a particular place. This is, $P(\text{emo}|\text{places}) = N_{\text{emo}}/N_{\text{people}}$ where N_{people} is the total number

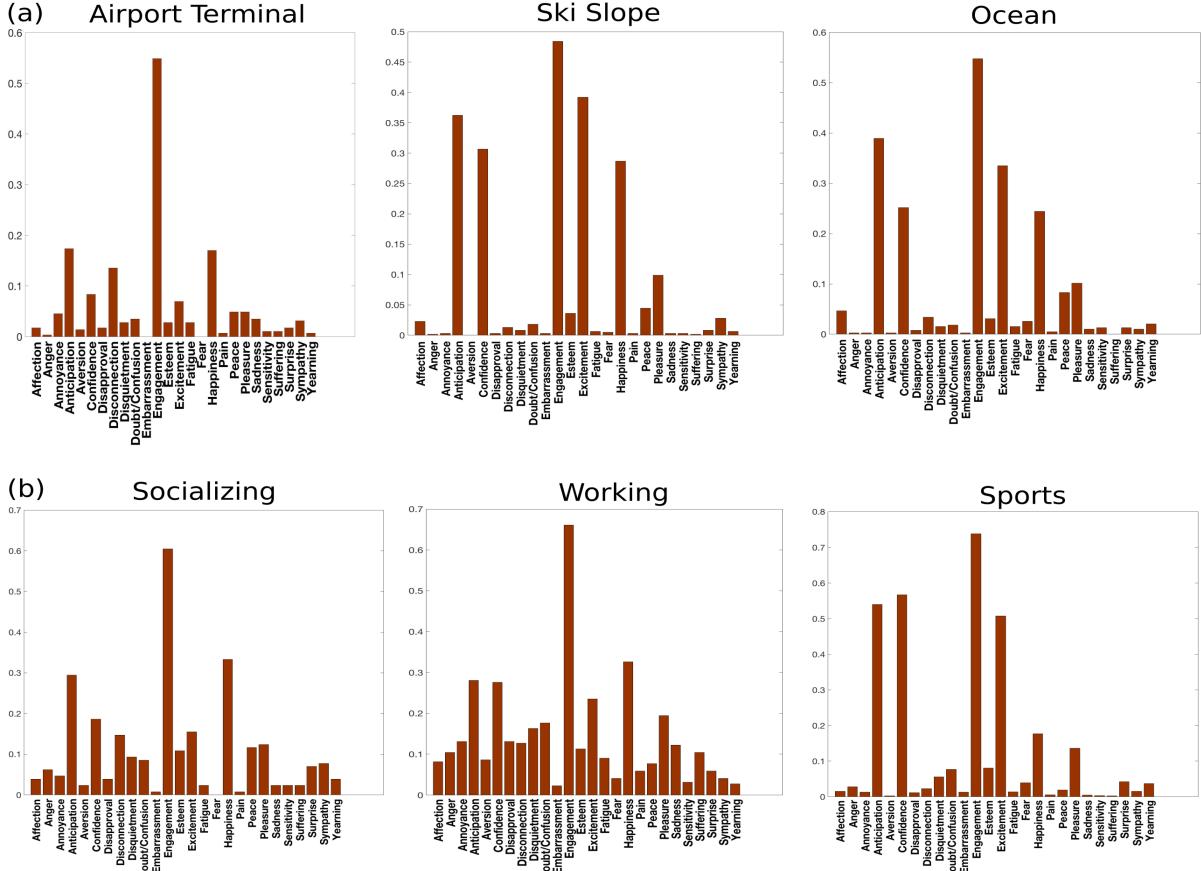


Figure 3.27: Emotion distributions conditioned to a) image scene category, and b) image scene attribute

of people present in the images of that particular scene category and N_{emo} is the number of those people labeled with this particular emotion.

Representative distributions are shown in Fig. 3.27. As can be seen, the distribution of emotions varies among different place categories (Fig. 3.27.a). For example, we see that people in the *Ski Slope* frequently experience *Anticipation* or *Excitement*, which are associated to the activities that usually happen in this place category. Compare sport-related images and working-environment related images (Fig. 3.27.b), we see that people in sport-related images usually show *Excitement*, *Anticipation* and *Confidence*, but they show *Sadness* or *Annoyance* less frequently. Interestingly, these two categories appear with higher frequency in working environments than in social gatherings. In places such as the *Airport Terminal*, only a few emotions dominate the distribution where as in places like *Oceans* & *Ski Slopes*, range of activities are comparatively more.

| | Valence | Arousal | Dominance |
|---------|--|---|---|
| Highest | Orchard, Field Wild Ski Slope, Ski Resort Restaurant | Baseball Field Baseball Stadium Outdoor Track, Ski Slope Fairway | Ski Slope, Outdoor Track Baseball Stadium Baseball Field Race Course |
| Lowest | Foliage, Snow Sailing/boating Cold , Moist | Sailing/Boating Competing, Sports Moist, Cold | Sailing/Boating Snow, Competing Cold, Sports |
| | Auditorium, Boxing Ring, Assembly Line, Shower, Beauty Salon | Restaurant Patio Lobby, Hospital Room Waiting Room, Cockpit | Jail Cell, Shower Waiting Room Forest Path, Beauty Salon |
| | Electric Lighting Shopping, Driving Congregating, Working | Enclosed Area, Wood Electric Lighting Socializing, Shopping | Enclosed Area, Working, Electric Lighting Socializing, Congregating |

Figure 3.28: Summary of places and attributes with the highest and lowest values of *Valence*, *Arousal* and *Dominance*.

Fig. 3.28 summarizes places and attributes with an overall highest and lowest value of *Valence*, *Arousal* and *Dominance*, the three continuous variables used in the dataset. We observe some interesting patterns here as well. For instance, places with the highest *Dominance* value are sport-related places and sport-related attributes. Furthermore, low *Dominance* categories contain places like *Shower* or *Jail Cell* or attributes like *Enclosed Area* or *Working*, where the freedom of movement is reduced. Finally, places and attributes with the highest value of *Valence* are usually related to pleasure or enjoyment. However, categories with low *Valence* values are related to high stress situations such as *Driving* or *Working*. Overall, these observations suggest that some common sense knowledge patterns related with emotions and context could be extracted from the data.

Common Sense Knowledge in Visual Scenes We illustrate how current scene-centric systems can be used to extract contextual information that can be potentially useful for emotion recognition. In particular, we illustrate this idea with a CNN trained on Places dataset (Zhou et al. [2017a]) and with the Sentibanks Adjective-Noun Pair (ANP) detectors (Jou et al. [2015]; Chen et al. [2014]), a Visual Sentiment Ontology for image sentiment analysis. As a reference, Figure 3.29 shows Places and ANP outputs for

sample images of the EMOTIC dataset.

| Places CNN output | Sentibanks ANP (score). Top 8. |
|---|--|
|  <p>Place Category: kindergarten_classroom, classroom Attributes: no_horizon, enclosed_area, man-made, working, cloth, wood, socializing, plastic, congregating.</p> | early_childhood (0.203) early_education (0.087) early_learning (0.047) elementary_schools (0.045) elementary_education (0.041) creative_kids (0.040) final_exam (0.024) young_child (0.019) |
|  <p>Place Category: landfill Attributes: natural_light, open_area, dirt, sunny, no_horizon, rugged_scene, dry, foliage, trees.</p> | long_range (0.024) outdoor_adventure (0.022) outdoor_education (0.020) hard_work (0.016) healthy_lifestyle (0.007) environmental_portrait (0.007) active_volcano (0.007) big_bear (0.007) |

Figure 3.29: Illustration of 2 current scene-centric methods for extracting contextual features from the scene: AlexNet Places CNN outputs (place categories and attributes) and Sentibanks ANP outputs for three example images of the EMOTIC dataset.

We computed the ANP for each image in EMOTIC. The positive sentiment scores denote the presence of that particular ANP. These ANPs describe the apparent sentiment conveyed by the image. The detected ANP with their respective labelled emotion categories, we found interesting patterns. For example, in images with people labeled with *Affection*, the most frequent ANP is *young_couple*, while in images with people labeled with *Excitement* we found frequently the ANPs *last_game* and *playing_field*. Also, we observe a high correlation between images with *Peace* and ANP like *old_couple* and *domestic_scenes*, and between *Happiness* and the ANPs *outdoor_wedding*, *outdoor_activities*, *happy_family* or *happy_couple*.

Overall, these observations suggest that some common sense knowledge patterns related with emotions and context could be potentially extracted, automatically, from the data.

Chapter 4

Modeling Emotion on EMOTIC dataset

We, as human beings, perceive emotions of other people on daily basis which constitutes an important part of our social skills. Over time, we have become more adept in being empathetic towards a fellow human being. We have learned to recognize different emotional states of people depending on the visual information available to us from the surroundings. Such information could be present in the immediate environments surrounding us in various forms. Many of such sources have been explored in section 1.2.1. These sources are very important aspect to understanding people's emotions. It is, therefore, essential to take into account all such sources. Machines are not as sophisticated as humans in making estimations about people's emotions. There are many machine learning algorithms which use facial features, and sometimes body postures, to estimate a person's emotional state. Such algorithms (face based- Beristain and Graña [2009], body based - Schindler et al. [2008]) are modeled on that specific feature of the person, while disregarding the other visual contexts present in the image.

4.1 Architectural Design

In section 1.2.1 we described various sources of context (including visual ones) that influence emotional states of the people embedded in those situations. The context present in the immediate surroundings affect and can modulate the emotions of the people present in those settings. So, in order to describe the emotional state we integrate the features that contain such information in the modeling process. It is very difficult to extract such specific visual features (like edges, shape, color, texture, shapes, objects, etc) from each image and model them; mainly because they change in size, orientation and shape. It is

also possible that a particular feature can have more impact than the other. This kind of information is very difficult to hand-craft. Traditional computer vision algorithms rely on such techniques. But since the importance of a given feature cannot be defined for all the images, it is impractical to use such methods. CNNs are a good way of extracting such features. They are variant of Multi Layer Perceptron (MLP), with a specific connectivity pattern. Inspired by neurons in the visual cortex of animals (Fukushima [1988]; Hubel and Wiesel [1968]; Riesenhuber and Poggio [1999]), they have multiple overlapping receptive fields (filters) that capture the global information in the visual field while simultaneously preserving the local features. This helps in modeling the prior knowledge present in the form of visual features automatically without hand-crafting them.

In this chapter, CNN architecture is introduced for modeling emotion recognition based on EMOTIC dataset. **CNNs have the following desirable properties** that would help us create a model to predict the emotion of a person:

- Locality preserving feature extraction or translation invariance - If a feature is present in a different location, or is displaced, the CNN is still able to capture those patterns. Pooling (Pooling) is the operation that accomplishes this feature for the CNNs.
- Hierarchical feature extraction - The initial filters of the CNN extract low-level features (edges, corners) and later layers become highly specific and extract more abstract shapes and objects (Zhou et al. [2015]).
- Multi-grained feature extraction from fine to coarse granularity - The different filter sizes in association with their strides accomplish this aspect of the CNNs. Smaller filters would be able to extract finer features, whereas the bigger filters are able to view and extract coarser features.
- Low pre-processing as compared to the MLPs - The number of parameters to learn for CNN is drastically lower than that of a corresponding MLP designed for the same task. This is because the CNNs rely on filter weights which extract local features depending on the size of their receptive fields (filter size). The number of filter weights are much lower in comparison to the parameters of an MLP. This provides faster computation and low training time for CNNs.
- Can be trained end-to-end - The CNN does not need different training schemes for various structural parts. All the components of the CNN work together and learn in conjunction with one another. This end-to-end learning is not only convenient but

also gives a holistic approach to training. The CNN is able to learn highly complex tasks like image recognition, machine translation and BAP recognition.

The CNN-based model is expected to learn recognition of the person's emotional state using all the contextual features present in the image. CNNs are chosen because they have properties that suit the purpose of emotion recognition. They can capture the context present in the image along with the features of the person. As discussed in section 1.2.1, the surrounding visual scene and the body posture of the subject constitute one of the main sources of visual context. CNNs have the capability to learn scene-specific and person-specific features automatically.

4.1.1 Person Features

All the research attempts in developing a model for emotion recognition using facial expressions (section 2.2.1) and the emotional language communicated by the body (section 2.2.2) make a compelling case to explicitly model person features for understanding the emotional state of the person. As discussed in Chapter 3, all the annotated people in EMOTIC also have their respective bounding boxes (example Figure 3.1). So, we use the CNNs capability to extract all of the person features. In EMOTIC we have images with the person embedded in the surrounding context. We use the bounding box of that person to extract person features using a CNN that was pre-trained on object recognition. Specifically, we use a standard object recognition CNN to extract the features of the person's body. We employ a pre-trained Alexnet (Krizhevsky et al. [2012]) which was trained on Imagenet (Deng et al. [2009]) to extract person features with some modifications. Figure 4.1 shows the basic structure of Alexnet, containing 5 Conv layers and 3 FC layers. The shown network doesn't include intermediate layers that are not trainable (like ReLU). Because it was pre-trained on objects (including person), this Alexnet provides us with features that represent the person.

For modeling emotion perception based on EMOTIC, we remove the FC layers and add Spatial Average Pooling layer, followed by a layer that flattens all the values into a single dimensional vector representation - called **nn.View** - part of Torch7 (Collobert et al. [2011]) library. A detailed view of the network is shown in Figure 4.2. Each of the Conv layer is followed by a ReLU layer - which rectifies negative values. C^* refers to the combination of CNN + ReLU. After each of the first 2 Conv layers there is a combination of *Max Pooling layer + Local Response Normalization* (Krizhevsky et al. [2012]) layers, *PL1* and *PL2* in Figure 4.2. Without further fine-tuning or transfer learning, this network generates features related to the object it is presented in the form of an image. In our case, the input is the body of the person so it generates features related to that person.

Since our task is that of emotion recognition from images, quite different from that of object recognition or localization, we need to retrain all these layers on our dataset for the specific task of emotion recognition. The first filters of a CNN normally learn low-level image features like edges, corners, etc (Zeiler and Fergus [2014]). Figure 4.3 shows all the weights of all 96 filters corresponding to the Conv 1 layer of the pre-trained Alexnet model. These filters capture the low-level features present in the image. Further layer filters progressively learn higher object abstractions. Zeiler and Fergus [2014] give a deeper visual understanding of which part of the image do these filters activate. Depending on the layer, these filters will be activated to different pattern in an image.

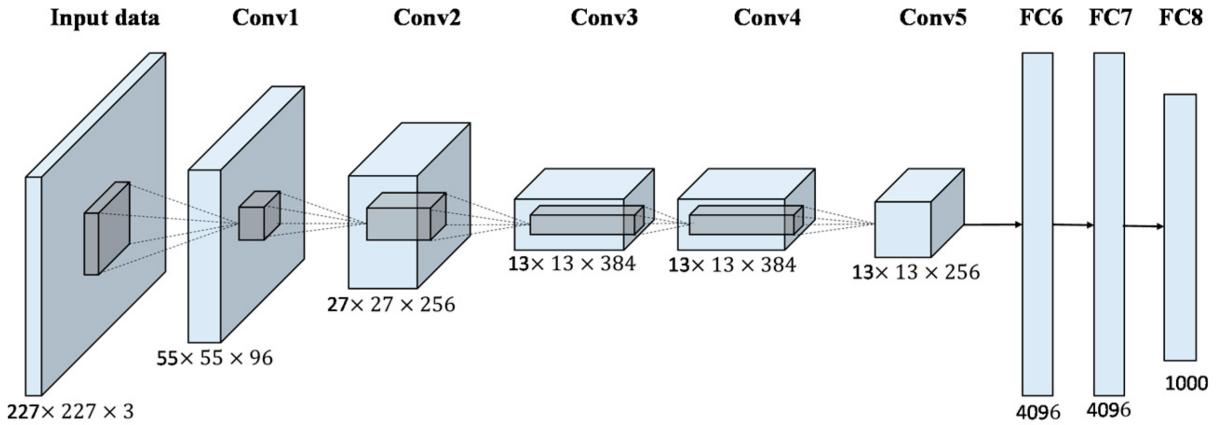


Figure 4.1: Basic Alexnet (Krizhevsky et al. [2012]) with 5 Conv layers for feature extraction and 3 FC layers for classification

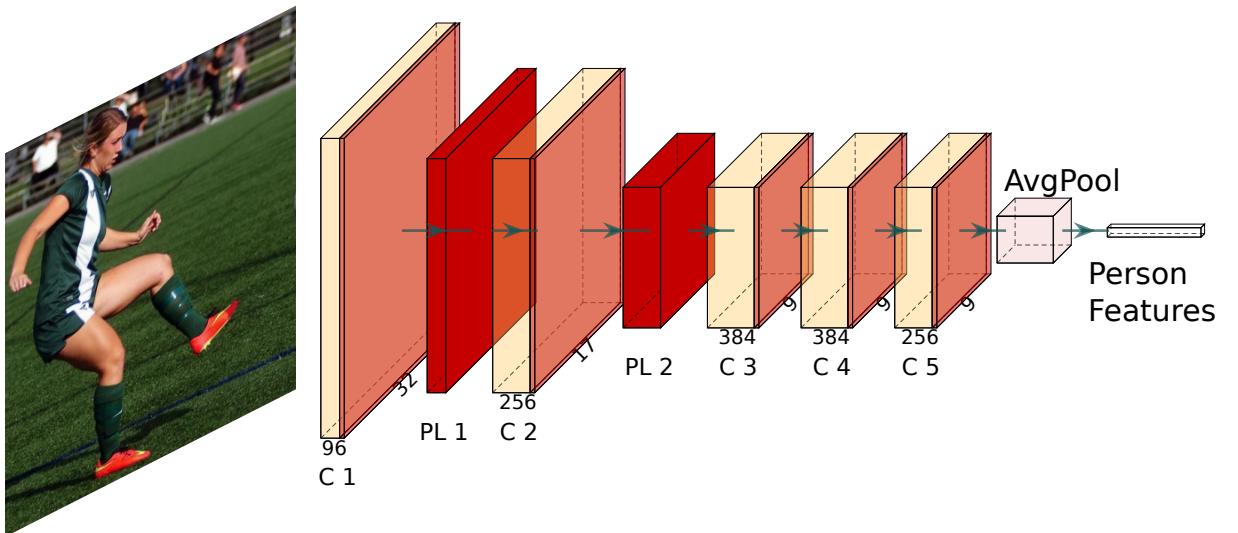


Figure 4.2: Person module based on Alexnet. $C*$ represent the combination of Conv + ReLU layers. $PL*$ are a combination of Pooling layer + Local Response Normalization layers

4.1.2 Visual Scene-Context Features

Visual context includes the scene category, its attributes, the dynamics between other objects present in the scene. Everything that can directly and indirectly affect the subject (person) under consideration has contextual influence over the subject. Aviezer et al. [2008a] gave an experimental evidence on how the context affects the emotion perception from faces only. In a more recent cognitive-neuroscientific study, the authors (Hassin et al. [2013]) show how the visual context not only influences the intensity of emotion perception but also changes the categorical perception of emotion. Martinez et al. [2016] showed how important the context (the surrounding scene and objects) plays in determining the emotional state of the person from his body posture and gives us insight to look for more information present in the surrounding visual scene to understand the emotional state of the person. We also explored various visual sources of context in section 1.2.1. These arguments gave us motivation to model the visual scene as one of the main sources of context in our emotion recognition model.

In order to capture these aspects, we need a network that can extract holistic features from the whole image. A network that is trained to capture the scene specific features is a good choice. For this, we fine-tune a pre-trained network called PlacesCNN (Zhou et al. [2017a] - based on Alexnet Deng et al. [2009]). This network was previously trained on a scene-specific database called Places2 (Zhou et al. [2017a]) of 10 million images to classify scenes. Alvarez and Petersson [2016] introduced a technique in the structure of the PlacesCNN called *DecomposeMe*, where they use 1D convolutions on PlacesCNN to reduce the computations of the first layers of the network (most of the computations in a CNN occur in the first Conv layers - Denton et al. [2014]) and also reduce the memory footprint,

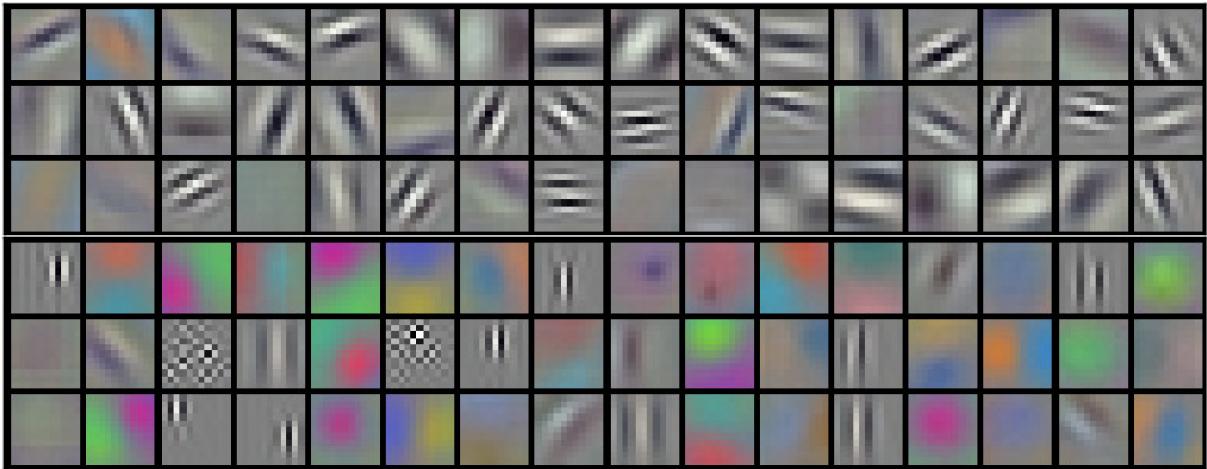


Figure 4.3: Filter weights of the Conv 1 layer of the basic Alexnet (Krizhevsky et al. [2012]), displaying various filter weights that help to extract low-level features

while simultaneously increasing classification accuracy. These CNN networks provide competitive performance while the number of parameters is low. The proposed CNN network to capture scene-specific features (Alvarez and Petersson [2016]) consists of 16 Conv layers with 1-dimensional kernels, effectively modeling 8 layers using 2-dimensional kernels. We use this modified PlacesCNN model, pre-trained on Places2 database using the *DecomposeMe* technique, to capture scene specific features as contextual cues for our emotion recognition process. In fact, Zhou et al. [2015] show that intermediate CNN layers learn object level features representations when the network is trained for scene recognition. The network has the capability to learn object level features which is apt for our purposes.

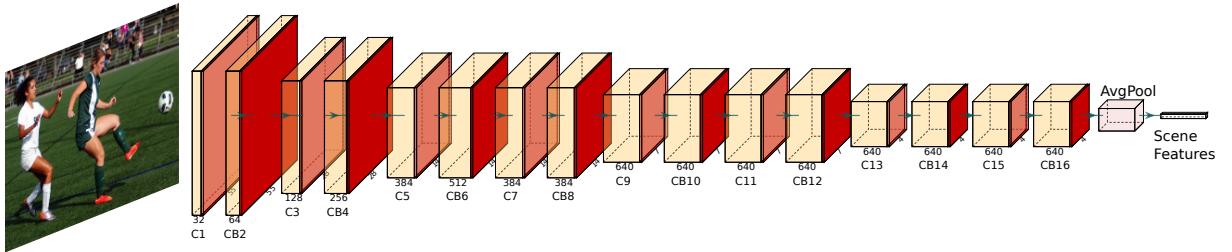


Figure 4.4: Scene module based on Alvarez and Petersson [2016]. $C*$ are Conv layers each followed by a ReLU layer. $CB*$ are a combination of Batch Normalization + ReLU layer

Similar to the network for extracting person features (section 4.1.1), in this network when we extract visual context features, we get rid of the final FC layers and add a Spatial Average Pooling layer followed by a layer that flattens all the values into a single dimensional vector representation - called **nn.View** - part of Torch7 (Collobert et al. [2011]) library. Figure 4.4 shows the basic structure of the network used for extracting visual context features. $C*$ represents Conv layer followed by a ReLU layer, whereas $CB*$ represents Conv layer followed by a Batch Normalization (Ioffe and Szegedy [2015]) and a ReLU layer. All the Conv filters are one-dimensional filters (more details in Alvarez and Petersson [2016]).

4.1.3 EMOTIC Fusion Model

Visual scene-context features (section 4.1.2) and person features (section 4.1.1) are essential to make predictions about the emotional state of the person embedded in the given situation. We want to take into consideration both these features while training the network on EMOTIC dataset. We want to show the network the scene as well as the person features so that while training (and while making predictions) the network is able to visually interpret the features required for emotion estimation. For this, we designed a fusion

model. It has three modules. The network architecture of our EMOTIC-CNN Fusion model is shown in Figure 4.5. The first two modules are feature extractors, taken directly from person module (Figure 4.2) and scene module (Figure 4.4). Accordingly, these two modules do the heavy lifting of generating features from the images. The *first module* (Visual Context Features) takes the whole image as the input and generates scene-related features. The *second module* (Person Features) takes the visible body of the person and generates features related to the body. The *third module*, called the *fusion module*, works as a fusion layer and concatenates features generated by the previous two modules. The combined features are then passed through tuple of FC layers for training.

Fusion module uses three FC layers. First FC layer (FC 1) reduces the dimensionality of the concatenated features (person features (256) + scene features (640) = 896) to 256 and then, to learn independent representations for each task, the reduced features are passed through two FC layers (FC 2 and FC 3) (Caruana [1997]), each for 3 Continuous Dimensions and 26 Emotion Categories respectively. The output of FC 1 layer is first passed through a dropout layer which helps avoid over-fitting the model (Srivastava et al. [2014]). The output FC 2 gives the regressed values of the 3 continuous dimensions, whereas FC 3 gives a probability distribution. This distribution from validation set of images is used to calculate the thresholds for each category and find the ones that activated for each test images.

Our EMOTIC-CNN is designed in a way that we are able to maintain the localization of different parts of the image while achieving a design fitting our emotion recognition system.

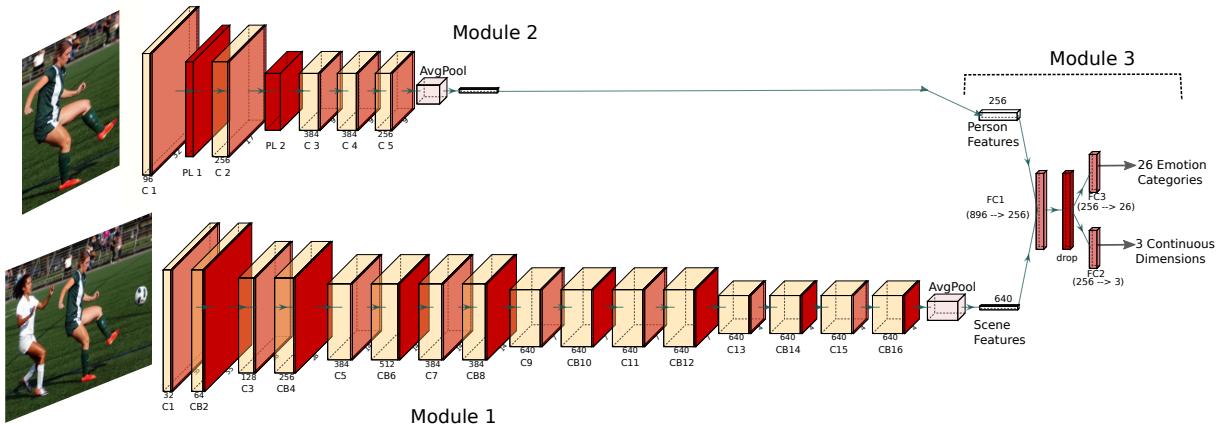


Figure 4.5: EMOTIC-CNN Fusion model trained with L_{comb1}/L_{comb2} criterion

4.2 Multitask Learning (MTL)

Multitask Learning (MTL) (Caruana [1997]) in neural network is an approach to inductive transfer learning. It uses the information content in the training signals of a related task to improve the generalization performance of the network. For example, let's say that the task is to detect faces of people from an image. We know, through the human body configuration, that the relation of face to the rest of the body is hierarchical. So, if we know one part it is easy to infer the other part. So the task of body detection is a related task to face detection. While training for body detection, the model will also embed, inductively, the information of face. The content of training signals for body detection also has face information, so training for body detection will also help to improve the generalization performance of the face detection task.

In non-neural models like kernel methods, bayesian algorithms and linear models, normally the generalization has been improved by enforcing regularization (Yuan and Lin [2006]). Block-sparse method imposes l_1/l_q norm regularization on the jointly learned parameters. In another method for learning task relationship, the technique involves using constraints that would enforce clustering of tasks as a regularization. In learning task relationships, bayesian based methods use gaussian as a prior distribution. Most of these methods would take some characteristic of the features as a regularizer to constrain the model to work better for the given task. In multitask neural models that are trained end-to-end, the task relatedness itself imposes the regularization which helps to improve the generalization of the model. The need to exclusively define a regularization is avoided in neural networks.

Normally, the main learning task is divided into smaller sub-tasks. These sub-tasks are trained on different models independently. The performance of the main task is then calculated by combining the performance of models of all such sub-tasks. MTL technique takes advantage of task-relatedness to learn a common model representation for the related tasks.

MTL technique is inherent to our task of emotion recognition as well, while training our model end-to-end. EMOTIC contains people annotated with 2 different formats of emotion representation *viz.* Emotion Categories and Continuous Dimensions (section 3.1.2). Both these representations have different ways of representing the perceived emotion of the person, therefore each of these representations constitute tasks that are related to one another. Emotion Categories use 26 emotions (Table 3.1) to do that, whereas Continuous Dimensions use 3 dimensions (*viz.* Valence, Arousal and Dominance). So for every annotated person in EMOTIC, their emotional state can be described by these 2 different approaches. We take advantage of this aspect of our dataset in designing our fusion model

(section 4.1.3), where we have a single network that learns both representations simultaneously. We conducted experiments with individual representations and with combined representations. Our results show that the performance of both the individual tasks in a combined form of learning has improved as compared to learning with individual tasks alone.

4.3 Loss criterions and Evaluation metrics

The first two feature extraction modules are initialized with weights from models pre-trained on two different large-scale classification datasets called ImageNet (Deng et al. [2009]) and Places (Zhou et al. [2017a]). ImageNet consists of data belonging to broad-ranging classes including *person*. This motivates us to use the network called Alexnet (pre-trained on Imagenet) to extract features related to the target person. Places dataset is used for understanding high-level visual recognition tasks such as recognizing scene categories. The network, called PlacesCNN [based on DecomposeMe (Alvarez and Petersson [2016]), pre-trained on Places], helps to model scene related features into our emotion recognition system.

We train our recognition system end-to-end, learning the parameters jointly using stochastic gradient descent with momentum and weight decay. As mentioned, the first two modules are initialized using pre-trained models (Places (Zhou et al. [2017a]) and Imagenet (Deng et al. [2009])) while the fusion module is trained from scratch. The batch size is set to 52 - twice the size of the emotion categories. We found empirically after testing multiple batch sizes (including multiples of 26 like 26, 52, 78, 108) that batch-size of 52 gives the best performance.

In this section, we introduce and describe the different losses used for training the models. We discuss the basis for choosing them and their relevance for our training goals. Specifically, we use weighted euclidean loss for emotion categories, margin euclidean loss and smooth L_1 (Girshick [2015]) for continuous dimensions and a weighted combined loss for joint training of emotion categories and continuous dimension. We discuss each in the following sections.

4.3.1 Criterion for Emotion Categories (L_{disc})

In our emotion recognition problem, we have formalised 26 different emotion categories (Table 3.1). Each person in EMOTIC can be labeled with multiple emotion categories (sample example in Figure 3.19). We see that there are multiple possible emotion classes (26 different emotion categories) to choose from and each different input can be labeled

with multiple labels making this a multiclass-multilabel problem. In addition, there is an inherent class imbalance as the number of training examples is not the same for each emotion category (Figure 3.20) nor for each score of the continuous dimensions (Figure 3.21.a,b,c). Therefore, we use a weighted euclidean loss to overcome this issue. We found empirically that this loss is more effective than using Kullback–Leibler divergence or a multi-label multi-classification hinge loss. The weighted euclidean loss for our emotion categories is defined as follows:

$$L_{disc} = \frac{1}{N} \sum_{i=1}^N w_i (\hat{y}_i^{disc} - y_i^{disc})^2 \quad (4.1)$$

where N is the number of categories (26 in this case), \hat{y}_i^{disc} is the prediction and y_i^{disc} is the ground-truth label for the i^{th} category. The parameter w_i is the weight assigned to each category. Weight values are defined as $w_i = \frac{1}{\ln(c+p_i)}$, where p_i is the probability of the i^{th} category and c is a parameter to control the range of valid values for w_i . Using this weighting scheme the values of w_i are bounded as the number of instances of a category approach to 0. This is particularly relevant in our case as we set the weights globally based on the occurrence of each category for the entire dataset. Experimentally, we obtained better results using this approach compared to setting the weights batch-wise.

4.3.2 Criterions for Continuous Dimensions (L_{2cont} , SL_{1cont})

In this learning task, there are 3 dimensions whose values are learned. Each dimension has values in the range from 1 to 10 so we model this task as a regression problem. There are multiple annotators for each annotation in validation and test sets and since the annotation is a subjective evaluation, we compare the performance using two different robust losses : (1) a margin Euclidean loss L_{2cont} , and (2) the Smooth L_1 SL_{1cont} . The former defines a margin of error (v_k) when computing the loss for which the error is not considered. The margin Euclidean loss for continuous dimension is defined as:

$$L_{2cont} = \frac{1}{\#\mathcal{C}} \sum_{k \in \mathcal{C}} v_k (\hat{y}_k^{cont} - y_k^{cont})^2 \quad (4.2)$$

where $\mathcal{C} = \{Valence, Arousal, Dominance\}$ and $\#\mathcal{C} = 3$, \hat{y}_k^{cont} and y_k^{cont} are the prediction and the normalized ground-truth for the k^{th} dimension and v_k ($= 0, 1$) is a binary weight to represent the error margin:

$$\begin{aligned} v_k &= 0, & \text{if } |\hat{y}_k^{cont} - y_k^{cont}| < \epsilon \\ &= 1, & \text{otherwise} \end{aligned} \quad (4.3)$$

If the predictions are within the error margin, *i.e.* error is smaller than ϵ , then these predictions do not contribute to updating the weights of the network during back propagation.

The Smooth L_1 loss refers to the absolute error using the squared error if the error is less than a threshold (set to 1 in our experiments). This loss has been widely used for object detection (Girshick [2015]) and, in our experiments, has been shown to be less sensitive to outliers. Precisely, the Smooth L_1 loss is defined as follows

$$SL_{1cont} = \sum_{k=1}^3 v_k \begin{cases} 0.5x_k^2, & \text{if } |x_k| < 1 \\ |x_k| - 0.5, & \text{otherwise} \end{cases} \quad (4.4)$$

where $x_k = (\hat{y}_k^{cont} - y_k^{cont})$, and v_k is a weight assigned to each of the continuous dimensions and it is set to 1 in our experiments.

4.3.3 Combined Criterions (L_{comb1}, L_{comb2})

We define the combined loss function as a weighted combination of two separate losses as explained above:

$$L_{comb1} = \lambda_{disc} L_{disc} + \lambda_{cont} L_{2cont} \quad (4.5)$$

$$L_{comb2} = \lambda_{disc} L_{disc} + \lambda_{cont} SL_{1cont} \quad (4.6)$$

The parameters $\lambda_{(disc, cont)}$ weigh the importance of each loss and are set empirically using the validation set. After various combinations, we found that using $\lambda_{disc} = \lambda_{cont} = 0.5$ gives the best performance. Using equal weights gives equal importance to both the criterions which prevents inducing bias for a particular task. L_{disc} and L_{cont} (L_{2cont} , SL_{1cont}) represent the losses corresponding to learning the emotion categories (section 3.1.2.2) and the continuous dimensions (section 3.1.2.1) respectively. This combined loss criterion is used to enforce the multi-task training in the model computationally. It calculates the loss from both the tasks and back-propagates it through the model while training. This enables the model to generalize its performance by learning the parameters in a shared manner.

4.3.4 Performance Evaluation Metrics

The task for learning emotion categories is modeled as multi-class classification and the task for learning continuous dimensions is modeled as regression. So we measure the performance of our fusion model using 2 different evaluation metrics.

4.3.4.1 Average Precision (AP)

For this evaluation metric, first we find a threshold for each emotion category from the validation set predictions. After setting the model in evaluation mode, forward pass of a sample from validation set generates 26 predictions - probabilities for the 26 emotion categories. We collect all such predicted probability values for all the validation set samples. Next, for each of the emotion category, their predictions for all the samples together with the labels are used to plot Precision-Recall curves (Powers [2011]). For each of these 26 curves, we find a point where *Precision=Recall* which is the threshold for that particular category. We now have 26 thresholds for the 26 emotion categories. These thresholds are then applied to the predicted probabilities of the model for the test set samples. For example, when we forward pass a test set sample, the model predicts 26 values for the emotion categories. If a predicted value is above it's respective threshold, this indicates that the model triggers that emotion category with higher probability. Similarly we apply the threshold to all the test set samples to find the predicted emotion categories for each of them. Apart from this, in order to measure the performance of the model in a holistic manner, we take the predicted probabilities for the test samples together with their respective labels and we find the precision-recall curve for each emotion category. The area enclosed by these curves (for each of the emotion category) constitutes what is called as *Average Precision*. This metric represents the average performance of the model for emotion categories. Average Precision can have values ranging from 0 to 100, where 0 implies that precision and recall both were *nil* and 100 indicates that there was a perfect recall and perfect precision.

4.3.4.2 Average Absolute Error (AE)

The continuous dimension task is a regression and we use Average Absolute Error (AE) to measure it's performance. An error for a given continuous dimension is the absolute value of the difference between it's predicted value and the target value. An error is considered for the calculation only if it is lower than the predefined error margin. Such errors are averaged over all the inputs for each of the continuous dimensions. This performance measure is similar to the criterion for continuous dimension (subsection 4.3.2), except that instead of a square we consider the absolute values. Thus, the AE is calculated as:

$$AE = \frac{1}{|\mathcal{C}|} \sum_{k \in \mathcal{C}} v_k |\hat{y}_k^{cont} - y_k^{cont}| \quad (4.7)$$

where $\mathcal{C} = \{Valence, Arousal, Dominance\}$, \hat{y}_k^{cont} and y_k^{cont} are the prediction and the normalized ground-truth for the k^{th} dimension, m is the number of samples and $v_k (=0,1)$

is a binary weight to represent the error margin. This error is also weighted in the same manner as the loss criterion (equation 4.3). This AE vector is the metric to represent the average performance of the model for the continuous dimensions. Lower the AE, better the performance.

Chapter 5

Experiments and Analysis

This chapter discusses all the experiments conducted on variations of the EMOTIC fusion model (introduced in Chapter 4, section 4.1.3). All these experiments use the EMOTIC dataset (Chapter 3). We establish baselines through experiments that take into account only the surrounding scene features or only the person features to learn about the perceived emotions of the person-in-context. Then, we combine both these features and train the fusion model (section 4.1.3) end-to-end. We observe that the performance is better with the combined features than the individual features. Thus demonstrating the influence and importance of the surrounding visual scene in the perception of emotion of a person embedded in that scene, in addition to the features extracted from the person alone.

In this section, we present all our experiments on the EMOTIC dataset. The goals of these experiments are two fold:

1. Demonstrate the influence of visual context information on the perception of emotion (in addition to the body of the person)
2. Using multitask learning for emotion perception, the two emotion representations help one another to improve the generalization performance of the model. Ultimately achieving better results for the combined training of the tasks as compared to the individual training

5.1 Experiments

EMOTIC being the first dataset of it's kind, we set up baseline performance for our CNN fusion model. We want to observe the influence of different features on the task of emotion perception for which we designed our fusion model. To this end, we design

the experiments into 3 parts - each part with a different input given to the network. We denote the experiments according to the following:

1. CNN models trained with *only* the body of the person as input *as* \mathbf{B} ; with focus on person features
2. CNN models trained with *only* the whole image as input *as* \mathbf{I} ; with focus on visual scene features
3. CNN models trained with body of the person *and* the whole image as inputs *as* $\mathbf{B} + \mathbf{I}$; with focus on the combined features

Each of the above parts constitute 3 different experiments. There are 2 main tasks *viz.* (1) learning 26 emotion categories and (2) learning 3 continuous dimensions. The first two experiments focus on learning these tasks independently. In the third experiment, we do a combined training for both the tasks. For each part we modify the architectures depending on the task. The details for each part is discussed in the following sections.

5.1.1 B Model Experiments (person features)

The focus of this set of experiments is to train a CNN model that can extract person related features and use this to make emotion predictions. We do a supervised training of the person module (Figure 4.2) with input as the visible body of the person with their respective labels. The purpose of training this model is to understand and analyze the network behavior when only the body of the person is presented (or fed) to the network as an input. We perform the experiments where the network learns the person features, and trains in 3 different experiments and compare their performance for emotion recognition.

Depending on the task (learning emotion categories, continuous dimensions or their combined form), the model configurations with their respective criterions (L_{disc} , L_{2cont} and L_{comb1}) are shown in Figure 5.1.a,b,c respectively. While training with criterion L_{comb1} , the **B** model shares the parameters between both the sub-tasks.

5.1.2 I Model Experiments (visual scene features)

In these experiments, we train a CNN model on visual scene related features for emotion prediction and observe how the network behaves when only the image of the scene is given as a training signal. We train visual scene model using only the image (visual scene) as our input (Figure 4.4) with their respective labels. 3 different experiments use 3 different criterions *viz.* L_{disc} , L_{2cont} and L_{comb1} for their respective tasks. Figure 5.2 shows the

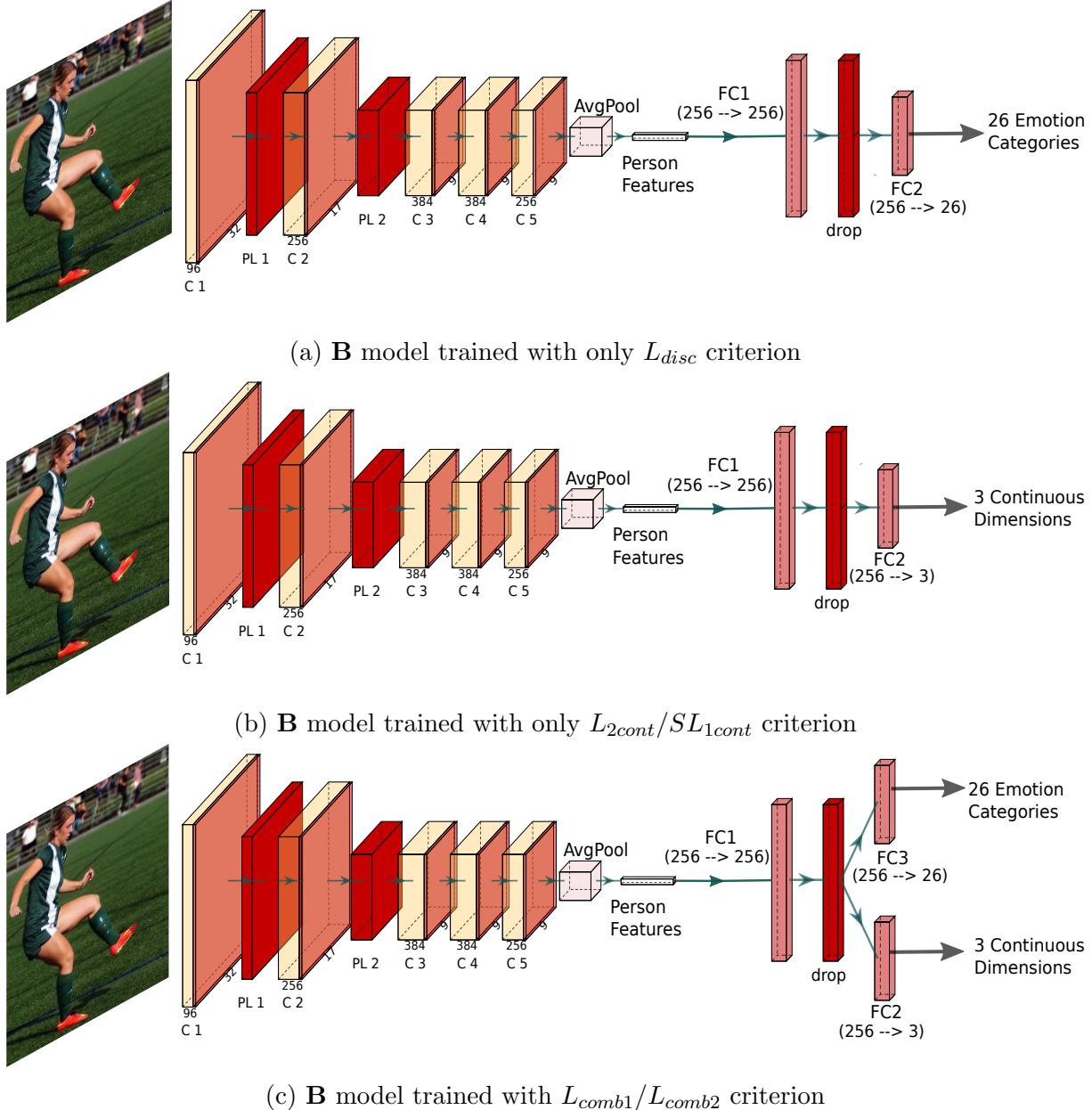


Figure 5.1: **B** model configurations for different experiments (a, b, c)

different CNN configurations trained for the **I** model. Similar to the body model, while training with criterion L_{comb1} , the **I** model also shares the parameters between both the sub-tasks.

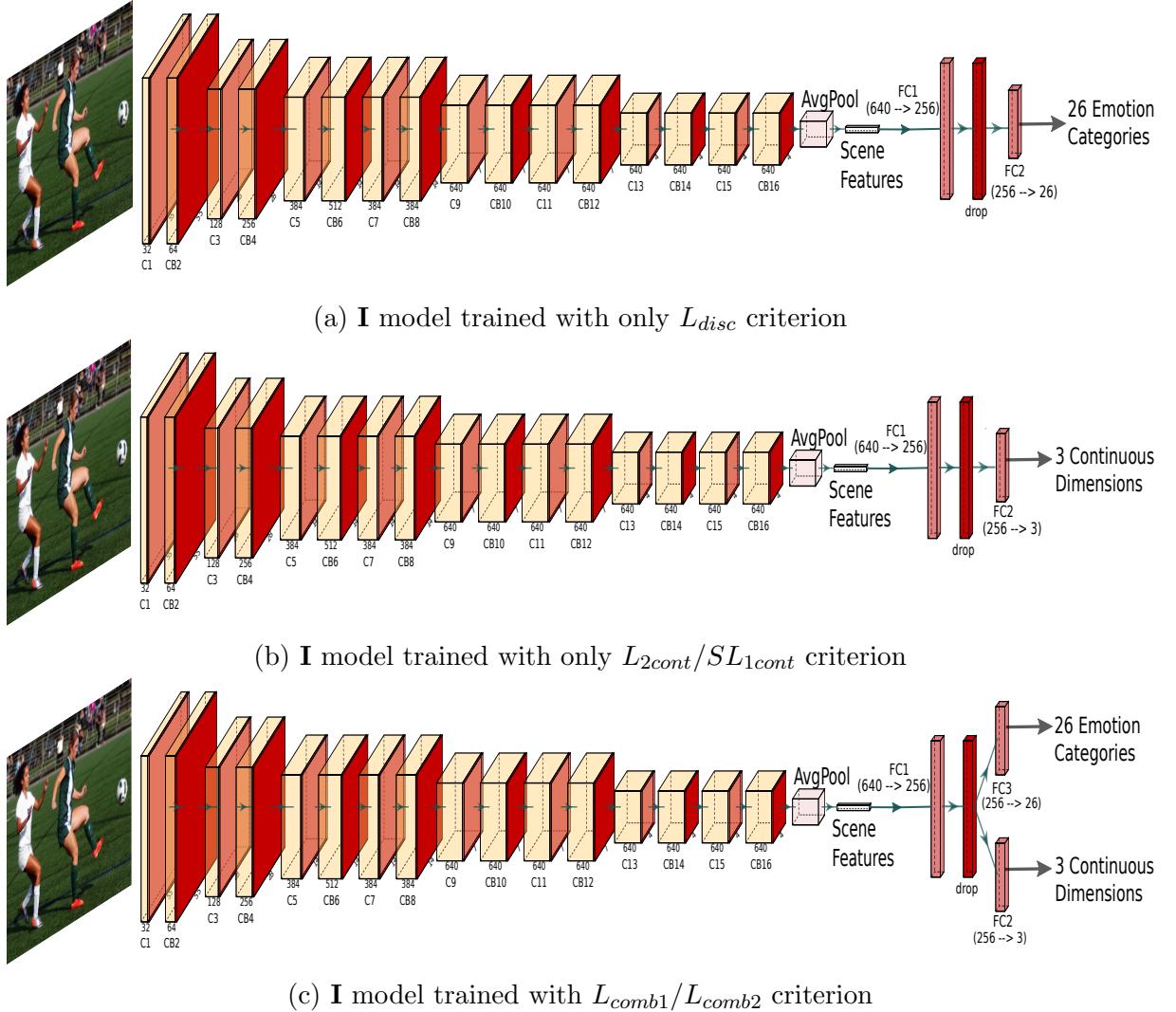


Figure 5.2: **I** model configurations for different experiments (a, b, c)

5.1.3 B + I Model Experiments (combined features)

In this set of experiments, we use the visible body along with the visual scene as our inputs for our fusion model (Figure 4.5), and train with their respective labels. Through these experiments we would be able to observe the behaviour of the CNN when it trains with both, *the person features & the visual scene features* for emotion recognition. Similar to previous experiments the fusion model shares the parameters for both the sub-tasks when trained with the combined criterion L_{comb1} . Figures 4.5,5.3.a,b show the different

CNN configurations trained for the **B+I** model for the different tasks.

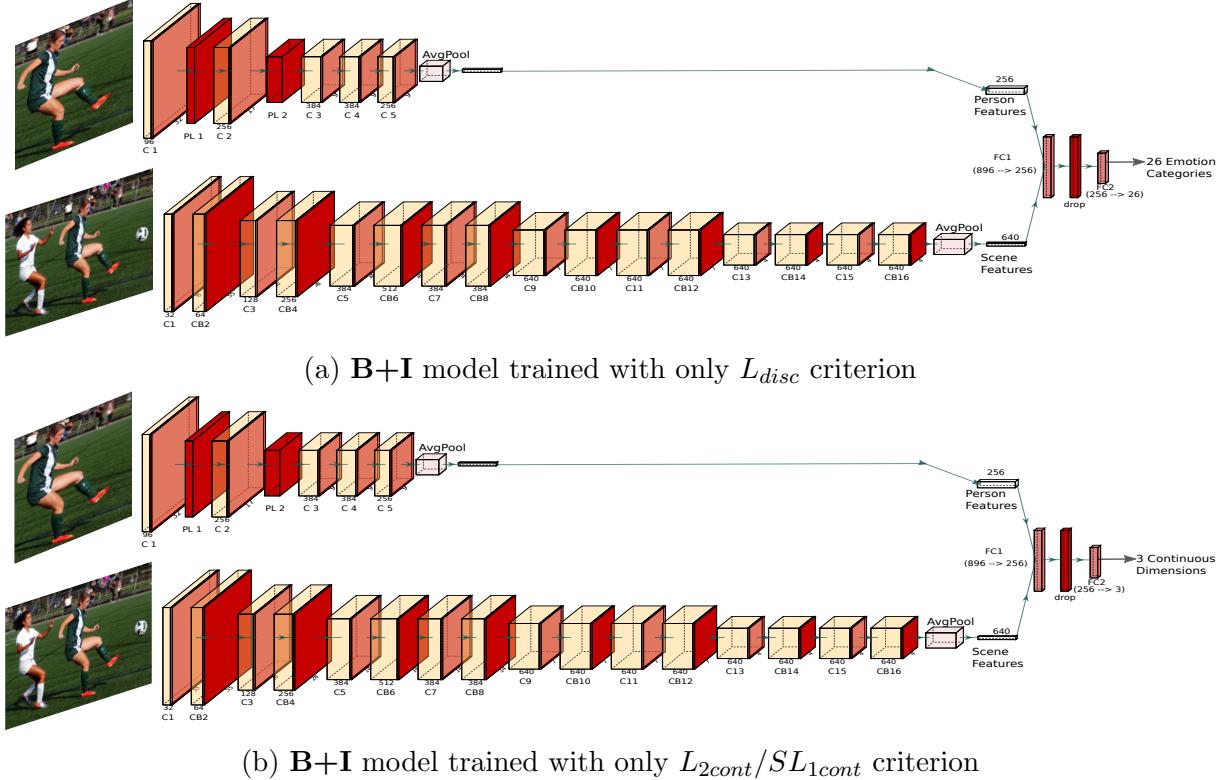


Figure 5.3: **B+I** model configurations for different experiments (a, b)

5.2 Results' Analysis

We trained different models on the EMOTIC dataset with various input configurations and loss functions. In this section we observe and compare how the 2 emotion representations (section 3.1.2) are learned independently using different criterions (L_{2cont} & L_{disc}) and compare them with the combined criterion (L_{comb1}), for each of the 3 experiments (person features (**B**), visual scene features (**I**) and their combination (**B + I**)). Table 5.2 summarizes the Average Precision (**AP**) of different model configurations comparing the performance of combined criterion (L_{comb1}) with discrete criterion (L_{disc}); and Table 5.1 summarizes the Average Absolute Error (**AE**) of different model configurations comparing the performances of the combined criterion (L_{comb1}) with the continuous criterion (L_{2cont}). The performances are compared on the images from the testing set of EMOTIC.

| Continuous Dimensions | Input to network | | | | | |
|-----------------------|------------------|---------------|-------------|-------------|-------------|-------------|
| | B | | I | | B + I | |
| | L_{2cont} | L_{comb1} | L_{2cont} | L_{comb1} | L_{2cont} | L_{comb1} |
| Valence | 0.0543 | 0.0537 | 0.0600 | 0.0541 | 0.0613 | 0.0546 |
| Arousal | 0.0661 | 0.0600 | 0.0620 | 0.1060 | 0.0622 | 0.0648 |
| Dominance | 0.0589 | 0.0570 | 0.0604 | 0.0687 | 0.0604 | 0.0573 |
| Mean | 0.0597 | 0.0569 | 0.0608 | 0.0763 | 0.0613 | 0.0589 |

Table 5.1: Average Absolute Errors (AE) for various models, comparing performance of each with L_{2cont} and L_{comb1} criterions

| Emotion Categories | Input to the network | | | | | |
|---------------------|----------------------|--------------|--------------|-------------|--------------|--------------|
| | B | | I | | B + I | |
| | L_{disc} | L_{comb1} | L_{disc} | L_{comb1} | L_{disc} | L_{comb1} |
| 1. Affection | 22.89 | 21.80 | 19.51 | 21.03 | 19.46 | 21.16 |
| 2. Anger | 07.55 | 06.45 | 06.88 | 05.64 | 08.10 | 06.45 |
| 3. Annoyance | 10.72 | 07.82 | 05.73 | 08.49 | 09.79 | 11.18 |
| 4. Anticipation | 57.68 | 58.61 | 53.75 | 57.12 | 52.27 | 58.61 |
| 5. Aversion | 06.04 | 05.08 | 05.37 | 05.33 | 05.58 | 06.45 |
| 6. Confidence | 70.15 | 73.79 | 65.76 | 68.23 | 60.59 | 77.97 |
| 7. Disapproval | 09.62 | 07.63 | 09.01 | 07.84 | 08.10 | 11.00 |
| 8. Disconnection | 19.42 | 20.78 | 16.89 | 17.72 | 20.79 | 20.37 |
| 9. Disquietment | 14.83 | 14.32 | 13.75 | 14.08 | 14.66 | 15.54 |
| 10. Doubt/Confusion | 28.17 | 29.19 | 28.39 | 28.11 | 28.47 | 28.15 |
| 11. Embarrassment | 02.72 | 02.38 | 02.60 | 02.52 | 02.58 | 02.44 |
| 12. Engagement | 85.08 | 84.00 | 81.33 | 84.72 | 81.72 | 86.24 |
| 13. Esteem | 17.05 | 18.36 | 17.90 | 17.26 | 17.54 | 17.35 |
| 14. Excitement | 69.22 | 73.73 | 65.98 | 68.09 | 65.20 | 76.96 |
| 15. Fatigue | 08.62 | 07.85 | 06.63 | 07.55 | 07.87 | 08.87 |
| 16. Fear | 11.55 | 12.85 | 13.01 | 11.54 | 12.11 | 12.34 |
| 17. Happiness | 57.07 | 58.71 | 55.64 | 56.04 | 54.55 | 60.69 |
| 18. Pain | 05.78 | 03.65 | 02.77 | 03.88 | 04.79 | 04.42 |
| 19. Peace | 20.31 | 17.85 | 17.35 | 18.79 | 17.69 | 19.43 |
| 20. Pleasure | 44.36 | 42.58 | 42.17 | 42.86 | 42.34 | 42.12 |
| 21. Sadness | 10.01 | 08.13 | 11.72 | 06.14 | 09.11 | 10.36 |
| 22. Sensitivity | 05.43 | 04.23 | 04.76 | 04.24 | 04.09 | 04.82 |
| 23. Suffering | 09.33 | 04.90 | 08.34 | 06.01 | 07.41 | 07.65 |
| 24. Surprise | 17.53 | 17.20 | 16.05 | 16.26 | 16.77 | 16.42 |
| 25. Sympathy | 11.20 | 10.66 | 11.60 | 11.29 | 10.52 | 11.44 |
| 26. Yearning | 07.76 | 07.82 | 08.28 | 07.93 | 07.64 | 08.34 |
| Mean | 24.23 | 23.86 | 22.74 | 23.03 | 22.68 | 24.88 |

Table 5.2: Average Precision (AP) for various models, comparing performance of each with L_{disc} and L_{comb1} criterions

5.2.1 B Model Analysis

Research in emotion perception has mainly focused on the facial expressions (section 2.2.1) and body pose (section 2.2.2) of the person. This motivated us to create experiments with features focused on the body of the person as the network input. Person features are extracted from the visible body of the person so they include the facial features (if the face is visible) as well as features off the body of the person. Since body of a person (which also includes faces) is the main source for emotion perception, we consider model **B** as the benchmark with which we compare the performances of other models. In columns 2 & 3 of Tables 5.2 & 5.1 each, we provide results for the experiments focused only on the person features. For emotion categories, **B** model with L_{disc} has better *AP* as compared to L_{comb1} . For continuous dimensions, the *AE* performance of **B**(L_{comb1}) model is better by 0.0028 points as compared to **B**(L_{2cont}) model (lower *AE* is better). So, **B**(L_{disc}) and **B**(L_{comb1}) model register best performance for emotion categories and continuous dimensions respectively for this set of experiments.

5.2.2 I Model Analysis

Affect analysis of images has inspired research in sentiment and emotion analysis (Soleymani et al. [2017]). This inspired us to use the whole image and train different models on it for emotion perception. These set of experiments focuses on the visual features extracted from images as a whole for training and prediction. Columns 4 and 5 of Tables 5.2 & 5.1 furnish the results of the experiments for emotion categories and continuous dimensions respectively. As shown, none of the models is able to perform better than their counterpart of **B** models. The best **I**(L_{comb1}) model gives an *AP* lower by 1.2 points for emotion categories as compared to **B**(L_{disc}) model. When compared for continuous dimensions, the **I**(L_{2cont}) model gives lower performance as compared to **B**(L_{comb1}) model by 0.0039 *AE* points.

5.2.3 B + I Model Analysis

Taking cues from the above 2 experiments, we want to observe the model performance when both the features (person and visual scene context) are used in conjunction as the input to the model. Columns 6 and 7 of Tables 5.2 & 5.1 furnish the results of the experiments for emotion categories and continuous dimensions respectively. The model which considers person features as well as visual contextual information (**B+I** model) outperforms the **B** model for one task and is similar in performance for the other. For emotion categories, **B+I**(L_{comb1}) model gives 0.65 points higher *AP* than **B**(L_{disc}) model.

This could be attributed to the fact that **B+I** model is able to learn additional features supplied by the input image. For continuous dimensions, **B+I**(L_{comb1}) model gives lower *AE* performance by 0.002 points as compared to the **B**(L_{comb1}) model, which is not very different as the **B+I**(L_{comb1}) model is the second best by performance for continuous dimensions.

5.2.4 Analysis Summary

After reviewing the above experiments, it is clear to see how person features together with the visual scene context features learn a better representation for emotion perception. Neither **B** model nor **I** model could capture independently what the **B + I** model captured by joining both the features. This result clearly suggests that multiple tasks learned together helps to improve the predictions. Jointly learning both emotion representations is more effective and improves the performance. In addition, as shown column 7 of Table 5.2, the model achieves better performance when visual contextual information (**I** models) is taken into account in addition to person features (**B** models). For **B** model, the input feature space contains only the visible body part whereas the input feature space for **B+I** model has a wider feature space covering, in addition to person features, the visual scene (the surrounding context).

Note that **B** model as well as **I** model have 2 different models that give better performance for emotion categories and continuous dimensions each. However, quite interestingly a single **B + I**(L_{comb1}) model gives better performance for both.

It is now evident that the **B+I**(L_{comb1}) model gives better performances for emotion categories as well as continuous dimensions in comparison to the **B** and **I** models.

5.3 Experiments with other architectures: SHG and Resnets

We conducted more experiments on our proposed fusion CNN baseline model. Concretely, we experimented with different state-of-art CNN architectures for the feature extraction modules of our model. In particular, we used ResNets (He et al. [2016]) which won the ILSVRC 2015 image recognition challenge (Deng et al. [2009]) as visual scene-context feature extractor. ResNets have been trained on ImageNet 2012 dataset and achieved a top-5 error rate of 3.57% on the test set of ImageNet with ensembles, and a top-5 error rate of 4.49% on the validation set with a single model. The previous models like VGG (Simonyan and Zisserman [2014]) and GoogLeNet (Szegedy et al. [2015]) achieved

top-5 test error rates of 7.32 and 6.66. The performance improvement in ResNets for the task of image recognition over previous models is attributed to the fact that they address the degradation problem (He and Sun [2015]; Srivastava et al. [2015]) using deep residual learning. ResNets achieve more depth and complexity in their architecture which motivated us to use them for a more complex visual scene-context feature extractor. A state-of-the-art Stacked Hour Glass (SHG, Newell et al. [2016]) network was used as body feature extractor. SHG has a unique architecture that allows it to learn features from the body of the person for the task of pose estimation. We get rid of the last layer from SHG and use it as our person feature extractor. We trained these networks with different parameters (batch sizes, learning rates) and found that their performances were lower as compared to using less deeper models or less complex models like AlexNet. In particular, the best obtained result in the discrete categories was $AP = 22.38$, which is lower by 2.5 points as compared to our baseline $\mathbf{B} + \mathbf{I}(L_{comb1})$ model. We observed that the performances do not improve, showing that increasing the complexity of the CNN is not helping the model to improve the emotion recognition task.

5.4 Discussions

5.4.1 Experiments with L_{comb2} (BEST EMOTIC MODEL)

L_{comb2} (Equation 4.6) is a loss that combines Smooth L_1 (SL_{1cont}) and weighted euclidean loss (L_{disc}). The Smooth L_1 loss (Equation 4.4) refers to the absolute error using the squared error if the error is less than a threshold (set to 0.1 in our experiments). This loss has been widely used for object detection (Girshick [2015]) and, in our experiments, has been shown to be less sensitive to outliers. Precisely, the Smooth L_1 loss is a robust $L1$ loss that is less sensitive to outliers than the $L2$ loss used in R-CNN (Girshick et al. [2014]) and SPPnet (He et al. [2015]). Normally, when the regression labels are unbounded, training with $L2$ loss (Equation 4.2) can require careful tuning of learning rates in order to prevent exploding gradients.

We re-trained our models with L_{comb2} . Results for emotion categories in the form of AP are summarized in Table 5.3, whereas for continuous dimensions in the form of AE are summarized in Table 5.4. Notice that the $\mathbf{B+I}$ models outperform the \mathbf{B} models in all categories except *Esteem*. Clearly, the combination of person features and visual scene-context features ($\mathbf{B+I}$ model) is better than the person features (\mathbf{B} model). For continuous dimensions we can observe from the Table 5.4 that the performance increases for $\mathbf{B} + \mathbf{I}$ model but not for \mathbf{B} model.

Another noteworthy thing is that $\mathbf{B} + \mathbf{I}(L_{comb2})$ model has AP higher by 2.5 points as

| Emotion Categories | Input to the network | | | |
|----------------------------|----------------------|--------------|--------------|--------------|
| | B | | B + I | |
| | L_{disc} | L_{comb2} | L_{comb1} | L_{comb2} |
| 1. Affection | 21.80 | 16.55 | 21.16 | 27.85 |
| 2. Anger | 06.45 | 04.67 | 06.45 | 09.49 |
| 3. Annoyance | 07.82 | 05.54 | 11.18 | 14.06 |
| 4. Anticipation | 58.61 | 56.61 | 58.61 | 58.64 |
| 5. Aversion | 05.08 | 03.64 | 06.45 | 07.48 |
| 6. Confidence | 73.79 | 72.57 | 77.97 | 78.35 |
| 7. Disapproval | 07.63 | 05.50 | 11.00 | 14.97 |
| 8. Disconnection | 20.78 | 16.12 | 20.37 | 21.32 |
| 9. Disquietment | 14.32 | 13.99 | 15.54 | 16.89 |
| 10. Doubt/Confusion | 29.19 | 28.35 | 28.15 | 29.63 |
| 11. Embarrassment | 02.38 | 02.15 | 02.44 | 03.18 |
| 12. Engagement | 84.00 | 84.59 | 86.24 | 87.53 |
| 13. Esteem | 18.36 | 19.48 | 17.35 | 17.73 |
| 14. Excitement | 73.73 | 71.80 | 76.96 | 77.16 |
| 15. Fatigue | 07.85 | 06.55 | 08.87 | 09.70 |
| 16. Fear | 12.85 | 12.94 | 12.34 | 14.14 |
| 17. Happiness | 58.71 | 51.56 | 60.69 | 58.26 |
| 18. Pain | 03.65 | 02.71 | 04.42 | 08.94 |
| 19. Peace | 17.85 | 17.09 | 19.43 | 21.56 |
| 20. Pleasure | 42.58 | 40.98 | 42.12 | 45.46 |
| 21. Sadness | 08.13 | 06.19 | 10.36 | 19.66 |
| 22. Sensitivity | 04.23 | 03.60 | 04.82 | 09.28 |
| 23. Suffering | 04.90 | 04.38 | 07.65 | 18.84 |
| 24. Surprise | 17.20 | 17.03 | 16.42 | 18.81 |
| 25. Sympathy | 10.66 | 09.35 | 11.44 | 14.71 |
| 26. Yearning | 07.82 | 07.40 | 08.34 | 08.34 |
| Mean | 23.86 | 22.36 | 24.88 | 27.38 |

Table 5.3: Average Precision (AP) obtained on test set per category. Comparing performance of **B**(L_{disc}) and **B + I**(L_{comb1}) models with their L_{comb2} counterparts

| Continuous Dimensions | Input to network | | | |
|-----------------------|------------------|---------------|-------------|---------------|
| | B | | B + I | |
| | L_{comb1} | L_{comb2} | L_{comb1} | L_{comb2} |
| Valence | 0.0537 | 0.0545 | 0.0546 | 0.0528 |
| Arousal | 0.0600 | 0.0630 | 0.0648 | 0.0611 |
| Dominance | 0.0570 | 0.0567 | 0.0573 | 0.0579 |
| Mean | 0.0569 | 0.0581 | 0.0589 | 0.0573 |

Table 5.4: Average Absolute Error (AE) obtained on test set per each continuous dimension. Comparing performance of **B**(L_{disc}) and **B + I**(L_{comb1}) models with their L_{comb2} counterparts

compared to $\mathbf{B} + \mathbf{I}(L_{comb1})$ model. This shows that using SL_{1cont} (L_{comb2}) loss criterion in place of L_{2cont} (L_{comb1}) for continuous dimension boosts the performance of the model in general. We can see this fact reflected in columns 4 and 5 of Table 5.3. $\mathbf{B} + \mathbf{I}(L_{comb2})$ model outperforms $\mathbf{B} + \mathbf{I}(L_{comb1})$ model in all but one emotion category (*Happiness*, the *AP* for both these models is same for *Yearning*). In addition to emotion categories, the performance in continuous dimensions of $\mathbf{B} + \mathbf{I}(L_{comb2})$ model is better than $\mathbf{B} + \mathbf{I}(L_{comb1})$ model as shown in Table 5.4 (Lower *AE* is better). Also note that $\mathbf{B} + \mathbf{I}(L_{comb2})$ model has better performance by *AP* of 3.52. This improvement is a good sign of how the visual scene affects the emotion perception as compared to only person features.

These experiments with L_{comb2} show the importance of using a loss that is less sensitive to outliers. These experiments gave us $\mathbf{B} + \mathbf{I}(L_{comb2})$ model which outperforms all the other models for EMOTIC dataset.

5.4.1.1 Quantitative Evaluation

We focus now on a detailed quantitative analysis of our results. We keep our discussion limited to \mathbf{B} and $\mathbf{B} + \mathbf{I}$ models. We compare the performance of these models on our 2 principal forms of emotion representations (emotion categories and continuous dimensions).

Performance analysis with Emotion Categories: We focus now on analyzing the performance of the models trained on emotion categories. To this end, we analyze precision-recall curves and use the validation set to find, for each emotion category, the points where *Precision* = *Recall*. With these values (thresholds) we find the Jaccard coefficient (*JC*) for all the samples in the test set. The *JC* coefficient is computed as follows: per each category we use as threshold for the detection of the category the value where *Precision* = *Recall*. Higher values of *JC* are better, with a maximum value of 1, where the detected categories and the ground truth categories are exactly the same. Then, the *JC* coefficient is computed as:

$$JC = \frac{\text{Intersection}(\text{detected categories, ground-truth categories})}{\text{Union}(\text{detected categories, ground-truth categories})} \quad (5.1)$$

The summary of the results obtained per each instance in the testing set can be found in Figure 5.4. Specifically, Figure 5.4.a shows *JC* for all the samples in the test set for 4 different models. The examples are sorted in decreasing order of the *JC* coefficient. We can observe that for $\mathbf{B} + \mathbf{I}(L_{comb2})$ model, 65% of people have a $JC \geq 0.3$ suggesting good retrieval rates for emotion categories. In contrast the best model for $\mathbf{B}(L_{disc})$, has 59% of people with $JC \geq 0.3$. This difference in retrieval rates is a good measure to realize

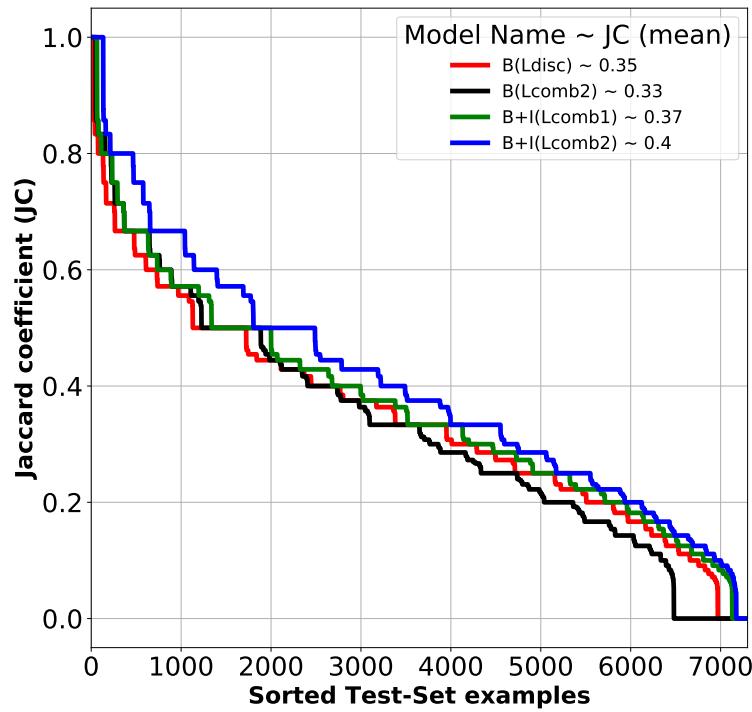
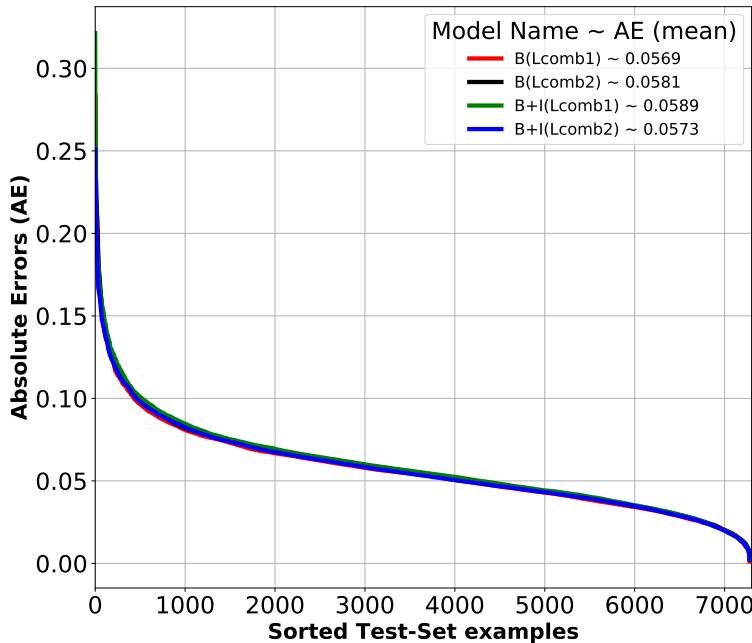
(a) Jaccard Coefficient (JC) of the predicted Emotion Categories(b) AE in the estimation of three Continuous Dimensions

Figure 5.4: JC and AE on the Test Set, along with comparisons for different models.
The results are sorted with decreasing values of JC and AE

the importance of using the combined inputs (image and body) to predict the correct categories. It can be said that the visual scene features are helping the person features to improve the performance, indicating the importance of visual scene context in emotion perception. This is a good evidence to change our understanding of emotion perception in general. Instead of focusing only on the features related to the person (face, body pose, or gestures), it is essential to look at the surrounding scene context.

Performance analysis with Continuous Dimensions: Results for continuous dimensions in the form of mean value of AE (the lower, the better) for the 3 dimensions are plotted in Figure 5.4.b. Samples are sorted by increasing order of AE . Again, these results are consistent with the observations made from Table 5.4. We can see that the models are almost equivalent. The standard deviation (std) of the AE is 0.00077 which is lower by 3rd order of magnitude from the range of their associated label values. The low std is also reflected in Figure 5.4.b by the thick plot. All the compared models give very similar results. From these results, we can conclude that combining contextual features from the visual scene with person features (**B + I** models) and jointly training them improves the AP (Table 5.3) compared to other model configurations. The network takes advantage of the multiple emotion representation to improve its performance.

5.4.1.2 Qualitative Evaluation

Quantitative analysis give a computational insight into the performance of the models. However, due to the inherent subjectivity of the task of emotion perception, it is important to look at the qualitative examples. Figure 5.5 shows qualitative predictions for the best **B** and **B+I** models. Categories in *Red font* indicate incorrectly predicted emotion categories. For **B** model, the predictions for emotion categories are taken from **B** (L_{disc}) model, whereas the predictions of continuous dimensions are taken from **B** (L_{comb1}) model. These examples were randomly selected among samples with high $JC \geq 0.8$ and a low $AE < 0.2$ value (Figure 5.5.a) and samples with low $JC \geq 0.4$ and $AE < 0.3$ value (Figure 5.5.b). The graphic compares the **B** and **B + I** models' predictions with their corresponding ground truths. As shown, in general, **B+I** model outperforms **B**. These examples show how the evaluation measures good predictions.

As we can see, the predictions of the continuous dimensions do not vary much when comparing **B** and **B+I** models. However, we observe significant differences in the case of the emotion categories. For example row 1 in Figure 5.5.a shows that **B + I** model predicts all the emotion categories faithfully, whereas the **B** model predicts 3 incorrect categories, meanwhile their continuous dimension predictions do not differ by huge margins. Another



| Ground Truth | | B (Ldisc, Lcomb1) | B+I (Lcomb2) |
|---------------------|---------|-----------------------------------|----------------------|
| Anticipation | V: 0,57 | Anticipation V: 0,61 | Anticipation V: 0,62 |
| Confidence | A: 0,83 | Confidence A: 0,61 | Confidence A: 0,70 |
| Engagement | D: 0,67 | Engagement D: 0,67 | Engagement D: 0,66 |
| Excitement | | Happiness Surprise Sympathy | Excitement |
| | | JC: 0,57 | JC: 1,00 |
| Anticipation | V: 0,50 | Anticipation V: 0,50 | Anticipation V: 0,62 |
| Confidence | A: 0,63 | Confidence A: 0,54 | Confidence A: 0,56 |
| Engagement | D: 0,67 | Engagement Esteem | Engagement D: 0,61 |
| Excitement | | Excitement Happiness | Excitement Happiness |
| Happiness | | Peace | |
| | | JC: 0,71 | JC: 1,00 |
| Anticipation | V: 0,60 | Anticipation V: 0,59 | Anticipation V: 0,63 |
| Confidence | A: 0,33 | Confidence A: 0,52 | Confidence A: 0,56 |
| Engagement | D: 0,63 | Disquietment D: 0,61 | Engagement D: 0,63 |
| Doubt/Confusion | | Doubt/Confusion Engagement | |
| Excitement | | Excitement Happiness | Excitement Happiness |
| Happiness | | Happiness, Surprise | |
| | | JC: 0,75 | JC: 0,71 |
| Anticipation | V: 0,60 | Anticipation V: 0,59 | Anticipation V: 0,64 |
| Confidence | A: 0,53 | Aversion A: 0,52 | Confidence A: 0,54 |
| Excitement | D: 0,73 | Engagement D: 0,58 | Engagement D: 0,62 |
| Happiness | | Happiness Peace | |
| Peace | | | |
| Pleasure | | | |
| | | JC: 0,38 | JC: 0,71 |

(a) Predictions with *high JC* values

| Ground Truth | | B (Ldisc, Lcomb1) | B+I (Lcomb2) |
|---------------------|---------|----------------------------------|-----------------------|
| Affection | V: 0,67 | Anticipation V: 0,59 | Anticipation V: 0,62 |
| Anticipation | A: 0,43 | Confidence A: 0,52 | Confidence A: 0,52 |
| Engagement | D: 0,83 | Doubt/Confusion D: 0,61 | Disconnection D: 0,62 |
| Esteem | | Engagement | Engagement |
| Happiness | | Pain | Excitement |
| Peace | | Pleasure | Happiness |
| Pleasure | | | Pleasure |
| | | JC: 0,30 | JC: 0,40 |
| Affection | V: 0,53 | Anticipation V: 0,60 | Affection V: 0,62 |
| Anticipation | A: 0,70 | Engagement A: 0,56 | Anticipation A: 0,58 |
| Disquietment | D: 0,63 | Happiness D: 0,63 | Confidence D: 0,63 |
| Engagement | | Peace | Engagement |
| Fear | | Pleasure | Excitement, Happiness |
| Sympathy | | | Pleasure, Sympathy |
| | | JC: 0,22 | JC: 0,40 |
| Annoyance | V: 0,40 | Affection, Anticipation, V: 0,59 | Affection V: 0,62 |
| Engagement | A: 0,33 | Aversion A: 0,52 | Anticipation A: 0,50 |
| Excitement | D: 0,63 | Confidence D: 0,61 | Disconnection D: 0,59 |
| Fatigue | | Disapproval, Doubt/Confusion, | Doubt/Confusion |
| Doubt/Confusion | | Embarassment, Engagement | Engagement, Happiness |
| Fear | | Esteem, Fatigue, Happiness, | Peace, Pleasure |
| Surprise | | Peace, Pleasure, Sympathy | Surprise |
| | | JC: 0,17 | JC: 0,23 |
| Anger | V: 0,50 | Anticipation V: 0,60 | Affection V: 0,64 |
| Annoyance | A: 0,33 | Confidence A: 0,50 | Anticipation A: 0,54 |
| Aversion | D: 0,67 | Disconnection D: 0,63 | Disquietment D: 0,62 |
| Doubt/Confusion | | Engagement | Doubt/Confusion |
| Sadness | | Happiness | Engagement |
| Surprise | | Pain | Happiness |
| | | JC: 0,00 | JC: 0,08 |

(b) Predictions with *low JC* valuesFigure 5.5: Comparing predictions of **B**(L_{disc} , L_{comb1}) and **B+I**(L_{comb2}) models with *high* and *low JC* values. Red indicates incorrectly predicted emotion categories

example is of row 3 in Figure 5.5.b. Here we see that both the models predict 3 correct emotion categories each, however the **B** model has lower JC because it is predicting 11 incorrect categories in comparison to the 6 incorrectly predicted by the **B + I** model.

Best prediction of emotion categories ($JC = 1$) is by the **B+I** model (Figure 5.5.a, row 1) and the worst ($JC = 0$) is by the **B** model (Figure 5.5.b, row 4). For the worst prediction, the continuous dimension predictions have a mean AE of 0.103 whereas the best has a mean AE of 0.063. From Figure 5.4.b we see that the mean AE for all the models is ≈ 0.06 . So the model that gives best prediction for emotion categories also predicts the continuous dimensions within the mean AE ; whereas the model that predicts worse for emotion categories also predicts bad for their continuous dimension counterpart. This exaggerates that our best EMOTIC fusion model (**B + I** (L_{comb2}) model) is able to perform better for both the emotion representation.

5.4.2 Sentibanks as Visual Context Features

Social media has become ubiquitous. Sharing of text, gifs, images, videos, and other forms of media has become very commonplace. Quite often people use these mediums to convey a specific message, a feeling or some other point of view. These medium very often transgress language limitations due to their visual appeal. The visual content itself is very intense to invoke specific sentiments. For example, the samples in Figure 5.6 convey lot of information and sentiments through their visual content. In Figure 5.6.a, we see a popular image of people bringing down the famous Berlin wall. This image arouses intense sentiments and serves as a symbol to demolishing the turbulent history. In another example shown in Figure 5.6.d, we see a beautiful sunset at a beach lined with palm trees and light waves. This image evokes a feeling of admiration to the nature's beauty.

In computer vision it is feasible and computationally easy to detect and recognize objects from an image. However, to recognize the reason that evoke sentiments is still an open area of research. In Figure 5.6.c it is easy to find the position and recognize the players, their poses, the ball, the net and the grass. From a football fan perspective, this is an amazing goal. However it is difficult to compute the sentiment this image conveys. The authors (Chen et al. [2014]) try to address this problem called the *affective gap* in their work. They pose 2 interesting questions: (1) how are images in various languages used to express affective visual concepts, e.g. beautiful place or delicious food? And (2) how are such affective visual concepts used to convey different emotions and sentiment across languages? They have build a visual sentiment concepts called ANP - which they discovered by mining millions of tags from web photos. This ANP is considered to fill the



Figure 5.6: Examples of images that evoke sentiments through their visual content

affective gap between recognition and sentiment.

These ANP detectors extract sentiments conveyed by the image. This motivated us to use these features as a visual context feature for our emotion recognition system. First, we analyse these features and then compare the performance of our model with these ANP detectors.

5.4.2.1 Context Features' Comparison

The goal of this section is to compare different context features for the problem of emotion recognition in context. A key aspect for incorporating the context in an emotion recognition model is to be able to obtain information from the context that is actually relevant for emotion recognition. Since the context information extraction is a scene-centric task, the information extracted from the context should be based in a scene-centric feature extraction system. That is why our baseline model uses a Places CNN for the context feature extraction module. However, recent works in sentiment analysis (detecting the emotion of a person when he/she observes an image) also provide a system for scene feature extraction that can be used for encoding the relevant contextual information for emotion recognition.

To compute body features, denoted by \mathbf{B}_f , we fine tune an AlexNet ImageNet CNN with EMOTIC database, and use the average pooling of the last convolutional layer as features. For the context (image), we compare two different feature types, which are denoted by \mathbf{I}_f and \mathbf{I}_S . \mathbf{I}_f are obtained by fine tuning an AlexNet Places CNN with EMOTIC database, and taking the average pooling of the last convolutional layer as features (similarly to \mathbf{B}_f), while \mathbf{I}_S is a feature vector composed of the sentiment scores for the ANP detectors from the implementation of Chen et al. [2014].

To fairly compare the contribution of the different context features, we train Logistic Regressors for the following features and combination of features: (1) \mathbf{B}_f , (2) $\mathbf{B}_f + \mathbf{I}_f$, and (3) $\mathbf{B}_f + \mathbf{I}_S$. For the discrete categories we obtain mean average precisions as $AP = 23.00$, $AP = 27.70$, and $AP = 29.45$, respectively. For the continuous dimensions, we obtain AE as 0.0704 , 0.0643 , and 0.0713 respectively. We observe that, for the discrete categories, both \mathbf{I}_f and \mathbf{I}_S contribute relevant information to the emotion recognition in context. Interestingly, \mathbf{I}_S performs better than \mathbf{I}_f , even though these features have not been trained using EMOTIC. However, these features are smartly designed for sentiment analysis, which is a problem closely related to extracting relevant contextual information for emotion recognition, and are trained with a large dataset of images.

Chapter 6

Conclusions and Future Outlook

In this thesis we focused on the importance of considering the visual scene context in the problem of automatic emotion recognition. We began by introducing the problem of emotion recognition (section 1.1), then we discussed the role of contexts in emotion perception (section 1.2). With the focus on the visual scene context, we discussed the relevant related research in the domain of emotion recognition through images (section 2.2). We also explored various datasets that are publicly available for the same (section 2.3). This discourse led to the observation of their shortcomings with respect to our goal of emotion recognition in context.

We presented the EMOTIC database (chapter 3), a dataset of 23,571 natural unconstrained images with 34,320 people labeled according to their apparent emotions with two different emotion representation formats (section 3.1.2). We also provided different statistics and algorithmic analysis on the EMOTIC database (section 3.2). Then, we proposed a baseline fusion CNN model for emotion recognition in scene context (section 4.1.3) and their related baseline experiments (section 5.1). We also compare different feature types for encoding the contextual information (section 5.4.2). The obtained results show the relevance of using visual scene context to recognize emotions and, in conjunction with the EMOTIC dataset, motivate further research in this direction¹.

6.1 Main Conclusions

The following summary of observations are made based on the challenges faced and tackled while actively conducting research for the principal goals of the thesis:

1. There was a lack of proper dataset of images for emotion recognition in context. EMOTIC dataset, presented in this thesis, is our attempt towards bridging this gap

¹Dataset and trained models are available on the project site: <http://sunai.uoc.edu/emotic/>

and trying to make emotion research from images more accessible. EMOTIC is one of it's kind in the field of emotion recognition through images. It is a collection of images of people annotated according to their apparent emotional states. Images are spontaneous and unconstrained, showing people doing various activities in different situations. Figure 3.19 shows some examples of images in the EMOTIC database along with their corresponding annotations.

2. EMOTIC Fusion CNN model (section 4.1.3) is designed based on the EMOTIC dataset. The model is constructed in a way that it implements the use of the visual scene and the body of the person as it's inputs. This strategy helps the network to see the whole image along with the focus on the person while training. This network scheme helps to observe the effect of visual scene in emotion recognition. The network uses state-of-art pretrained modules to ease the process of transfer learning to achieve the empirical results (Tables 5.3, 5.4); while training end-to-end.
3. Baseline experiments (Table 5.2) show that using visual scene features, in addition to the person features, improves the performance of the fusion model as compared to using individual features. This experiment shows that the visual scene influences and is an important cue for emotion recognition.
4. The fusion model is trained for multiple tasks in a joint manner, using a combination of two separate loss functions (section 4.3.3), one each for emotion categories and continuous dimensions. A single fusion model gives the best performance (Table 5.3), where as the models with single inputs (either Image (**I**) or Person (**B**)) couldn't improve their performance when trained for both the tasks jointly.
5. Smooth L_1 loss (SL_{1cont}) is designed to be less sensitive to the outliers while training. When applied to our model we observe that it improves the performance of the fusion model (Table 5.3).
6. The low performance of our model ($AP = 27.38$) cannot be attributed to the low capacity of the network. While training deeper networks like SHG and Resnets (section 5.3), we could achieve $AP = 21.46$ as compared to our current $AP = 27.38$. So, using deeper networks does not necessarily mean that the performance will improve .
7. Emotion Recognition in Scene Context and Image Sentiment Analysis are different problems that share some common characteristics. While Emotion Recognition aims to identify the emotions of a person depicted in an image, Image Sentiment

Analysis consists of predicting what a person will feel when observing a picture (does not necessarily contain a person). We found that features from a model trained on Image sentiment features and Scene context features are both good sources of visual context information for the task of emotion recognition in context (section 5.4.2).

6.2 Future Outlook and Concluding Remarks

This thesis experimentally showed the importance of visual scene context for emotion perception. We demonstrated, with empirical experiments, that visual scene context plays an important role in emotion perception. This is an interesting and challenging stage for the future research in this direction. Below we preview some of the probable directions that the current research could potentially take.

Probable lines of research

- **Data Augmentation:** The recent work by Azulay and Weiss [2018] states that data augmentation cannot be used for all the networks. Modern CNNs are not invariant to image transformations. Due to the subsampling introduced by the pooling layers, and due to the *photographer's bias*, the networks don't learn data invariance. If all the sources of invariances were considered and data augmented based on those then the data size increases exponentially which might not be a good idea. The authors suggest (by virtue of Sampling Theorem) to always have a *stride + pooling* combination while designing networks - which is supposed to introduce invariability in the networks. This suggestion could be used to design a completely new network model for emotion recognition using EMOTIC dataset.
- **Quality of EMOTIC:** The extended dataset contains 4 additional annotations for the validation set which improves the quality of the labels. This is important since the model continuously validates itself during training. Future work can focus more on improving the qualitative aspects of the dataset. Maybe include more annotations for training set of images for a more deterministic training.
- **Quantity of EMOTIC:** If a dataset has bias (depends on whether it is high or low), increasing the data quantity depends on it. More data doesn't help with high bias, however with high variance in the data, adding more samples might just help. The EMOTIC fusion CNN model does not have a high performance ($AP = 27.38$) and with deeper networks, its performance is poorer ($AP = 21.46$). Due to this it would be very difficult to improve performance by extending the dataset quantity.

However, a different training strategy could help improve the performance. The inherent dataset bias could be used as a starting point to design new strategies for training (He and Garcia [2009]).

- **Emotion Captioning or using Free Form for representing Emotions:** Captioning generates a natural language description of the image contents. Current captioning systems can even generate semantic description of the image contents (You et al. [2016]). However, for emotion captioning, one needs to consider the affective content of the image and the situation of the person whose emotion caption needs to be acquired. An example of *emotion captioning* is shown in Figure 1.4. We see the difference of opinion of both the observers. These words can be transformed to vectors using word embedding methods like **word2vec** (Mikolov et al. [2013]). This would make it easier to compute and quantify language. However, there are two bigger challenges from computational perspective *viz.* **1.** The choice of words and phrases by the observer is highly subjective and varies so it becomes difficult to compute semantic similarity of 2 different samples of *Free Form*, and **2.** Once the *Free Form* responses are generated, it becomes difficult to compare between different responses due to their variable lengths and dissimilarities in the semantic content. Without comparison, it becomes difficult to understand the similarities and differences in the perceived emotional states, which inevitably restricts further useful analysis. Computing the gist (text summarization) of the emotional content of a given *Free Form* expression is also challenging. These challenges pose an entirely different research direction.

Concluding Remarks: The paradigm of research in emotion recognition in context is new and challenging. This thesis has been an attempt in observing the influence and importance of visual scene in emotion recognition. Overall our results are far from the accuracies obtained in other visual recognition problems, showing that the EMOTIC dataset and the problem of emotion recognition in context is a challenging area of research. We hope that this work serves as landmark for the future research.

Appendices

Appendix A

Comparison of Emotion Categories

Keltner and Cordaro found that the subjects of their experiments reported multiple emotion categories. After intensive analysis they found that there were 27 distinct emotion categories. We compare these with the ones that are presented in this thesis (Table 3.1, Section 3.1.2.2).

The following 2 tables (A.1 and A.2) summarizes the semantic overlap present in the emotion categories. It is interesting to see that majority of the categories have similarities in both works. This is a good example of how 2 different approaches converged to similar conclusions about the emotion categories. The ones presented in Table 3.1 were designed through meticulous analysis of previous works, dictionaries, including literature on psychology and clustering techniques of semantically similar word meanings (Section 3.1.2.2). On the other hand, the authors (Keltner and Cordaro) discovered their categories in the experiments they conducted for their research. They showed 2185 videos to their subjects and those people generated self-reports after watching the videos. Their analysis of these observations helped them reveal the 27 distinct emotion categories.

Keys for Table A.1:

A - Emotion Categories reported by Keltner and Cordaro

B - EMOTIC Emotion Categories (Table 3.1) that overlap with Keltner and Cordaro

Keys for Table A.2:

C - EMOTIC Emotion Categories (Table 3.1)

D - Emotion Categories reported by Keltner and Cordaro that overlap with EMOTIC Emotion Categories (Table 3.1)

NA - No semantic overlap was found

| # | A | Definitions of A | B |
|----|------------------------|---|----------|
| 1 | Admiration | feeling impressed, pride, amazement | 6,13 |
| 2 | Adoration | love, adoration, happiness | 1,17 |
| 3 | Aesthetic appreciation | awe, calmness, wonder | 19,24 |
| 4 | Amusement | amusement, laughter, humor | 17 |
| 5 | Anger | anger, angry disgust, boiling with anger | 2,5 |
| 6 | Anxiety | anxiety, fear, nervousness | 9,16 |
| 7 | Awe | awe, amazement, feeling impressed | 24 |
| 8 | Awkwardness | awkwardness, amused embarrassment, embrassment | 11,17 |
| 9 | Boredom | boredom, annoyance, interest | 3,8,12 |
| 10 | Calmness | calmness, peacefulness, serenity | 19 |
| 11 | Confusion | confusion, curiosity, interested confusion | 10,12 |
| 12 | Craving | hunger, desire, satiation of hunger | 26 |
| 13 | Disgust | disgust, feeling grossed out, extreme disgust | 5 |
| 14 | Empathic Pain | pain, emapathic pain, shock | 18 |
| 15 | Entrancement | interest, amazement, feeling intrigued | 12 |
| 16 | Excitement | excitement, adrenaline rush, awe | 14 |
| 17 | Fear | fear, feeling scared, extreme fear | 16 |
| 18 | Horror | shock, horror, feeling scared | 16 |
| 19 | Interest | interest, amazement, feeling intrigued | 12 |
| 20 | Joy | happiness, extreme happiness, love | 1,17 |
| 21 | Nostalgia | nostalgia, boredom, reminescence | 8 |
| 22 | Relief | relief, deep relief, sense of narrow escape | 19 |
| 23 | Romance | love, romantic love, romance | 1 |
| 24 | Sadness | sadness, extreme sadness, sympathy | 21,25 |
| 25 | Satisfaction | feeling impressed, satisfaction, awestruck surprise | 6,13,24 |
| 26 | Sexual Desire | sexual arousal, feeling horny, sexual desire | 26 |
| 27 | Surprise | surprise, shock, amazement | 12,24 |

Table A.1: Emotion Categories' Comparisons between (A) and (B)

| # | C | Definitions of C | D |
|----|-----------------|---|---------------|
| 1 | Affection | fond feelings; love; tenderness | 2,20,23 |
| 2 | Anger | intense displeasure or rage; furious; resentful | 5 |
| 3 | Annoyance | bothered by something or someone; irritated; impatient; frustrated | 14 |
| 4 | Anticipation | state of looking forward; hoping on or getting prepared for possible future events | 12 |
| 5 | Aversion | feeling disgust, dislike, repulsion; feeling hate | 5,13 |
| 6 | Confidence | feeling of being certain; conviction that an outcome will be favorable; encouraged; proud | 1 |
| 7 | Disapproval | feeling that something is wrong or reprehensible; contempt; hostile | NA |
| 8 | Disconnection | feeling not interested in the main event of the surrounding; indifferent; bored; distracted | 9,21 |
| 9 | Disquietment | nervous; worried; upset; anxious; tense; pressured; alarmed | 6 |
| 10 | Doubt/Confusion | difficulty to understand or decide; thinking about different options | 11 |
| 11 | Embarrassment | feeling ashamed or guilty | 8 |
| 12 | Engagement | paying attention to something; absorbed into something; curious; interested | 11,15,19 |
| 13 | Esteem | feelings of favorable opinion or judgment; respect; admiration; gratefulness | 1 |
| 14 | Excitement | feeling enthusiasm; stimulated; energetic | 16 |
| 15 | Fatigue | weariness; tiredness; sleepy | NA |
| 16 | Fear | feeling suspicious or afraid of danger, threat, evil or pain; horror | 22,23 |
| 17 | Happiness | feeling delighted; feeling enjoyment or amusement | 2,4,20 |
| 18 | Pain | physical suffering | 14 |
| 19 | Peace | well being and relaxed; no worry; having positive thoughts or sensations; satisfied | 10,22,25 |
| 20 | Pleasure | feeling of delight in the senses | 26 |
| 21 | Sadness | feeling unhappy, sorrow, disappointed, or discouraged | 24 |
| 22 | Sensitivity | feeling of being physically or emotionally wounded; feeling delicate or vulnerable | 14 |
| 23 | Suffering | psychological or emotional pain; distressed; anguished | 14 |
| 24 | Surprise | sudden discovery of something unexpected | 7,15,19,25,27 |
| 25 | Sympathy | state of sharing others' emotions, goals or troubles; supportive; compassionate | 24 |
| 26 | Yearning | strong desire to have something; jealous; envious; lust | 12 |

Table A.2: Emotion Categories' Comparisons between (C) and (D)

Appendix B

Facial Expressions' Datasets

| Dataset | Reference | WebLink |
|---------------|---------------------------|---|
| CK+ | Kanade et al. [2000] | http://www.consortium.rim.cmu.edu/ckagree/ |
| CE | Du et al. [2014] | http://cbcsl.ece.ohio-state.edu/dbform_compound.html |
| DISFA+ | Mavadati et al. [2013] | http://mohammadmahoor.com/disfa/ |
| Yale Face DB | Georghiades et al. [2001] | http://vision.ucsd.edu/~iskwak/ExtYaleDatabase/ExtYaleB.html |
| MMI | Pantic et al. [2005] | https://mmifacedb.eu/ |
| KDEF | Lundqvist et al. [1998] | http://kdef.se/ |
| PubFig | Kumar et al. [2009] | http://www.cs.columbia.edu/CAVE/databases/pubfig/ |
| ExpW | Zhang et al. [2015] | http://mmlab.ie.cuhk.edu.hk/projects/socialrelation/index.html |
| CASIA WebFace | Yi et al. [2014] | http://www.voidcn.com/article/p-tnjoaphx-bqo.html |

Table B.1: Various publicly available facial expression datasets with their references and Weblinks

Appendix C

Various sources of contextual information

Here, we list some other related sources of context that can affect the perception of emotions. Since this thesis does not cover the implementation or deep discussion of these sources, we add them here so that it can serve as a reference for a curious reader.

1. **Prior Knowledge:** Our brain uses prior learned knowledge to predict when there is gap in understanding or missing information (Barrett [2017]). For example, the facial expression of the boy in Figure 1.2.a are ambiguous. His expressions suggest that he is annoyed (and/or a bit angry) at something, but we are unsure of the reason. This can happen due to several factors. Until we know the object (or the reason) that incited this feeling in him, we cannot be sure about our perception of his feelings. Maybe the object is not in the image (i.e. *Out-of-Sight*) or maybe it is occluded. From Figures 1.2.b, 1.2.c we can deduce that he is annoyed because of the girl eating a chocolate and probably refusing to share it with him. So when presented with an image of a person, our brains try to fill in the missing contextual information to make sense. Due to occlusion, vital knowledge that might help reveal the apparent emotion state is concealed. Figure C.1.a shows that a person is on a surf board. We can deduce safely that he is doing tricks on the surfboard, even though we cannot see major parts of his body. Similarly we can see that the tennis player is focused while tying her lace (Figure C.1.b). Her face is occluded but we can safely assume that she is engaged in her task. It is easy for us due to our past experience and knowledge to predict what the person might be doing or feeling. Without this knowledge we are at a loss to describe what might be behind that surfboard. Prior knowledge could be useful as an important source of contextual information as well.



(a) Surfer performing a trick



(b) Tennis player tying her shoe lace

Figure C.1: Examples where part of the person’s body is not visible. However, due to prior knowledge, it is easy to predict what they are doing

2. **Audio:** We not only listen to music for the sake of entertainment, but also enjoy and appreciate the audio quality. We are able to recognize the change in tone and pitch of singer’s voice (Bazgir et al. [2019]). And we use this skill to effectively communicate more subtle expression of thought. Humour, for instance, is a good example where the performer uses various combinations of words, gestures and voice intonation to bring out the comical aspect of his act. It is very difficult to watch a movie without sound. Clearly, audio serves as a major source of context to understand what the person is feeling. For example in Figure C.2 try to see the frames without reading the subtitles. Here we see that the protagonist (Tom Hanks) is talking. Unless we listen to him, it is difficult to realise why his expressions are changing over the frames. Once we listen (in this case, read the subtitles) to what he says, then it makes sense.
3. **Activity being conducted:** There are different kinds of objects present in our immediate surroundings. We interact with them depending on the activity. Different activities can stir distinct emotional reactions. The body movement, gaze, facial expressions are different for different activities. The person performing high intensity sports (for example, a boy showing off his skills on a skater-board in Figure C.3.a) usually has high engagement (i.e. focussed) and high arousal with extreme physical activeness, where as a person playing a board game (Figure C.3.b) is not active physically, but is using a lot of mental ability. Another example is that of a girl sleeping (Figure C.3.c) with a very low physical and mental activity.
4. **Social Surrounding:** When a group of people come together for a common purpose, such a situation is considered a social surrounding. The social context of the environment where the person is located can influence the emotional state of the person. The surrounding could be either of a formal (or informal) party, cele-



Figure C.2: Frames from the movie Forrest Gump (Gump [1994]), showing the protagonist recounting a story from his past . His speech is transcribed into subtitles



Figure C.3: Emotional state of a person is influenced by the kind of activity being performed. For example, the perceived arousal level for the person doing stunts (a) is higher than the people playing chess (b), whereas it is the lowest for the sleeping girl (c)

bration, festival, community gathering, seminar, workshop, presentation, meeting, demonstration, etc. One can find different types of apparent emotions in people within these distinct situations. The people involved in any kind of social activity (being part of the social surrounding) also influence one another's feelings. More specifically, together they show group emotion (Dhall et al. [2017]). Depending on the kind of gathering, the apparent group emotion can be different. Figure C.4 gives few examples of people showing different types of emotions depending on the gathering. Figure C.4.a shows people are enjoying a few drinks, they seem happy. Figure C.4.b shows a small group of children supervised by an adult and they are eating from a common plate, they seem calm. And Figure C.4.c shows that the people are sad and suffering.



(a) People having drinks (b) People eating together (c) Sad group of people

Figure C.4: The gist of the social surrounding affects emotional states of the people in it (Dhall et al. [2017]). The happiness level of the people drinking beers (a) is higher than the family eating together (b), whereas that of people suffering (c) is the lowest

Bibliography

Adams Jr, R. B. and Kleck, R. E. (2005). Effects of direct and averted gaze on the perception of facially communicated emotion. *Emotion*, 5(1):3. (pages xv, 22).

Alvarez, J. and Petersson, L. (2016). DecomposeMe: Simplifying ConvNets for End-to-End Learning. (pages xvii, 69, 70, 70, 70, 73).

Aviezer, H., Ensenberg, N., and Hassin, R. R. (2017). The inherently contextualized nature of facial emotion perception. *Current Opinion in Psychology*, 17:47–54. (pages xv, 10, 11).

Aviezer, H., Hassin, R., Bentin, S., and Trope, Y. (2008a). Putting facial expressions back in context. *First impressions*, pages 255–286. (pages 2, 5, 7, 7, 17, 69).

Aviezer, H., Hassin, R. R., Ryan, J., Grady, C., Susskind, J., Anderson, A., Moscovitch, M., and Bentin, S. (2008b). Angry, disgusted, or afraid? studies on the malleability of emotion perception. *Psychological science*, 19(7):724–732. (page 3).

Aviezer, H., Hassin, R. R., Ryan, J., Grady, C., Susskind, J., Anderson, A., Moscovitch, M., and Bentin, S. (2008c). Angry, disgusted, or afraid? studies on the malleability of emotion perception. *Psychological science*, 19(7):724–732. (page 8).

Aviezer, H., Trope, Y., and Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, 338(6111):1225–1229. (page 23).

Azulay, A. and Weiss, Y. (2018). Why do deep convolutional networks generalize so poorly to small image transformations? *CoRR*, abs/1805.12177. (page 99).

Bänziger, T., Grandjean, D., and Scherer, K. R. (2009a). Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (mert). *Emotion*, 9(5):691. (pages 3, 17).

- Bänziger, T., Grandjean, D., and Scherer, K. R. (2009b). Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (mert). *Emotion*, 9(5):691. (page 17, 17).
- Bänziger, T., Pirker, H., and Scherer, K. (2006). Gemep-geneva multimodal emotion portrayals: A corpus for the study of multimodal emotional expressions. In *Proceedings of LREC*, volume 6, pages 15–019. (page 25).
- Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt. (page 109).
- Barrett, L. F., Mesquita, B., and Gendron, M. (2011). Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290. (pages 5, 7).
- Bazgir, O., Mohammadi, Z., and Habibi, S. A. H. (2019). Emotion Recognition with Machine Learning Using EEG Signals. (pages 17, 110).
- Beristain, A. and Graña, M. (2009). Emotion recognition based on the analysis of facial expresions: a survey. *New Mathematics and Natural Computation*, 5(02):513–534. (pages 17, 21, 65).
- Berking, M. and Wupperman, P. (2012). Emotion regulation and mental health: recent findings, current challenges, and future directions. *Current opinion in psychiatry*, 25(2):128–134. (page 5).
- Blunch, N. J. (1984). Position bias in multiple-choice questions. *Journal of Marketing Research*, 21(2):216–220. (page 49).
- Breuer, R. and Kimmel, R. (2017). A deep learning perspective on the origin of facial expressions. *arXiv preprint arXiv:1705.01842*. (page 21).
- Camras, L. A., Bakeman, R., Chen, Y., Norris, K., and Cain, T. R. (2006). Culture, ethnicity, and children’s facial expressions: A study of european american, mainland chinese, chinese american, and adopted chinese girls. *Emotion*, 6(1):103. (page 5).
- Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75. (pages 71, 72).
- Chanes, L., Wormwood, J. B., Betz, N., and Barrett, L. F. (2018). Facial expression predictions as drivers of social perception. *Journal of Personality and Social Psychology*, 114(3):380–396. (page 9).

- Chen, T., Borth, D., Darrell, T., and Chang, S.-F. (2014). Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*. (pages 3, 24, 62, 93, 95).
- Clore, G. L. and Ortony, A. (2013). Psychological construction in the occ model of emotion. *Emotion Review*, 5(4):335–343. (page 36).
- Collobert, R., Kavukcuoglu, K., and Farabet, C. (2011). Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS workshop*, number EPFL-CONF-192376. (pages 67, 70).
- Cowen, A. S. and Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, page 201702247. (pages 18, 33).
- Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80. (page 5).
- Dael, N., Mortillaro, M., and Scherer, K. R. (2012a). Emotion expression in body action and posture. *Emotion*, 12(5):1085. (pages 10, 17, 23).
- Dael, N., Mortillaro, M., and Scherer, K. R. (2012b). Emotion expression in body action and posture. *Emotion*, 12(5):1085. (pages 10, 22, 25).
- Damasio, A. R. (2002). Descartes' error: Emotion, reason and the human brain. *Bulletin of the American Meteorological Society*, 83(5):742. (page 2).
- Darwin, C. (1872/1998). *The expression of the emotions in man and animals*. Oxford University Press, USA. (pages 2, 3, 18).
- de Sousa, R. (2017). Emotion. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2017 edition. (pages 2, 3).
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee. (pages 67, 69, 73, 73, 86).

- Denton, E. L., Zaremba, W., Bruna, J., LeCun, Y., and Fergus, R. (2014). Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in neural information processing systems*, pages 1269–1277. (page 69).
- Dhall, A. et al. (2012a). Collecting large, richly annotated facial-expression databases from movies. (page 24).
- Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., and Gedeon, T. (2017). From individual to group-level emotion recognition: EmotiW 5.0. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 524–528. ACM. (pages xviii, 23, 112, 112).
- Dhall, A., Joshi, J., Radwan, I., and Goecke, R. (2012b). Finding happiest moments in a social context. In *Asian Conference on Computer Vision*, pages 613–626. Springer. (page 24).
- Dictionary, M.-W. Merriam-webster online english dictionary. <https://www.merriam-webster.com>. Accessed: 2017-21-06. (page 33).
- Dictionary, O. Oxford english dictionary. <http://www.oed.com>. Accessed: 2018-21-06. (page 33).
- Dolz, J. and Pedersoli, M. (2018). An Attention Model for group-level emotion recognition. (page 23).
- Douglas-Cowie, E., Cox, C., Martin, J.-C., Devillers, L., Cowie, R., Sneddon, I., McRorie, M., Pelachaud, C., Peters, C., Lowry, O., Batliner, A., and Hoenig, F. (2011). *The HUMAINE Database*, pages 243–284. (page 18).
- Du, S., Tao, Y., and Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462. (pages 18, 107).
- Ekman, P. and Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica*, 1(1):49–98. (pages 2, 3, 9, 17, 19, 34, 34).
- Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124. (page 20).
- Escalera, S., Baró, X., Escalante, H. J., and Guyon, I. (2017). Chalearn looking at people: Events and resources. *CoRR*, abs/1701.02664. (page 25).

- Essa, I. A. and Pentland, A. P. (1997). Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763. (pages 9, 21).
- Fabian Benitez-Quiroz, C., Srinivasan, R., and Martinez, A. M. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5562–5570. (pages xv, 24, 26, 27).
- Fernández-Abascal, E. G., García, B., Jiménez, M., Martín, M., and Domínguez, F. (2010). *Psicología de la emoción*. Editorial Universitaria Ramón Areces. (page 33).
- Ferreira, P. M., Pernes, D., Fernandes, K., Rebelo, A., and Cardoso, J. S. (2018). Dimensional emotion recognition using visual and textual cues. *arXiv preprint arXiv:1805.01416*. (page 17).
- Fisher, C. O. B. (2018). Computer Vision Online. <http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm>. [Online; accessed 2-October-2018]. (page 29).
- Friesen, E. and Ekman, P. (1978). Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*. (page 20).
- Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1(2):119–130. (page 66).
- Georghiades, A., Belhumeur, P., and Kriegman, D. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660. (page 107).
- Girshick, R. (2015). Fast R-CNN. *IEEE International Conference on Computer Vision (ICCV 2015)*, pages 1440–1448. (pages 73, 75, 87).
- Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587. (page 87).
- Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukier-ski, W., Tang, Y., Thaler, D., Lee, D.-H., et al. (2013). Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer. (page 24).

- Google. Dataset search - google. <https://toolbox.google.com/datasetsearch>. (page 24).
- Goshvarpour, A., Abbasi, A., and Goshvarpour, A. (2018). An accurate emotion recognition system using ECG and GSR. *Biomedical Journal*, (2017). (page 17).
- Groen, Y., Fuermaier, A. B. M., Den Heijer, A. E., Tucha, O., and Althaus, M. (2015). The empathy and systemizing quotient: The psychometric properties of the dutch version and a review of the cross-cultural stability. *Journal of Autism and Developmental Disorders*, 45(9):2848–2864. (page 37).
- Gump, F. (1994). Robert zemeckis. *Paramount Pictures*. (pages xviii, 111).
- Gunes, H. and Pantic, M. (2010). Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, 1(1):68–99. (page 17, 17, 17).
- Hassin, R. R., Aviezer, H., and Bentin, S. (2013). Inherently ambiguous: Facial expressions of emotions, in context. *Emotion Review*, 5(1):60–65. (pages 2, 17, 69).
- He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284. (page 100).
- He, K. and Sun, J. (2015). Convolutional neural networks at constrained time cost. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5353–5360. (page 87).
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916. (page 87).
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (pages 13, 86).
- Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243. (page 66).
- Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*. (page 70).
- Izard, C. E. (1971). The face of emotion. (pages 17, 19).

- Jirayucharoensak, S., Pan-Ngum, S., and Israsena, P. (2014). Eeg-based emotion recognition using deep learning network with principal component based covariate shift adaptation. *The Scientific World Journal*, 2014. (page 3).
- Jou, B., Chen, T., Pappas, N., Redi, M., Topkara, M., and Chang, S.-F. (2015). Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 159–168. ACM. (page 62).
- Jung, H., Lee, S., Yim, J., Park, S., and Kim, J. (2015). Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2983–2991. (page 21).
- Kanade, T., Cohn, J. F., and Tian, Y. (2000). Comprehensive database for facial expression analysis. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 46–53. IEEE. (pages xv, 27, 107).
- Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7):881–892. (page 52).
- Keltner, D. and Cordaro, D. T. Understanding multimodal emotional expressions: Recent advances in basic emotion theory. *ISRE's Sourcebook for Research on Emotion and Affect, Andrea Scarantino (Ed.)*. Accessed: 2018-21-06. (pages 34, 103, 103, 103, 103).
- Kleinsmith, A. and Bianchi-Berthouze, N. (2007). Recognizing affective dimensions from body posture. In *Proceedings of the 2Nd International Conference on Affective Computing and Intelligent Interaction, ACII '07*, pages 48–58, Berlin, Heidelberg. Springer-Verlag. (page 23).
- Kleinsmith, A., Bianchi-Berthouze, N., and Steed, A. (2011). Automatic recognition of non-acted affective postures. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(4):1027–1038. (page 23).
- Ko, B. (2018). A brief review of facial emotion recognition based on visual information. *sensors*, 18(2):401. (page 21).

- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105. (pages xvii, xvii, 67, 67, 68, 69).
- Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. (2009). Attribute and simile classifiers for face verification. In *2009 IEEE 12th International Conference on Computer Vision*, pages 365–372. IEEE. (page 107).
- Lerman, K. and Hogg, T. (2014). Leveraging position bias to improve peer recommendation. *PLOS ONE*, 9(6):1–8. (page 49).
- Li, Z., Imai, J.-i., and Kaneko, M. (2009). Facial-component-based bag of words and phog descriptor for facial expression recognition. In *SMC*, pages 1353–1358. (page 21).
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*. (page 18).
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer. (pages 25, 31).
- Link, W. Hand gesture meanings. <https://airfreshener.club/quotes/meaning-gestures-hand-italian.html>. Accessed: 2019-03-04. (pages xv, 11).
- Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee. (page 13).
- Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE. (page 26).
- Lundqvist, D., Flykt, A., and Öhman, A. (1998). The karolinska directed emotional faces (kdef). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, 91:630. (page 107).
- Luneski, A., Bamidis, P. D., and Hitoglou-Antoniadou, M. (2008). Affective computing and medical informatics: state of the art in emotion-aware medical applications. *Studies in health technology and informatics*, 136:517. (page 5).

- Marchesotti, L., Perronnin, F., Larlus, D., and Csurka, G. (2011). Assessing the aesthetic quality of photographs using generic image descriptors. In *2011 International Conference on Computer Vision*, pages 1784–1791. IEEE. (page 23).
- Martinez, L., Falvello, V. B., Aviezer, H., and Todorov, A. (2016). Contributions of facial expressions and body language to the rapid perception of dynamic emotions. *Cognition and Emotion*, 30(5):939–952. PMID: 25964985. (pages 10, 69).
- Masuda, T., Ellsworth, P. C., Mesquita, B., Leu, J., Tanida, S., and Van de Veerdonk, E. (2008). Placing the face in context: cultural differences in the perception of facial emotion. *Journal of personality and social psychology*, 94(3):365. (page 3).
- Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., and Cohn, J. F. (2013). Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160. (page 107).
- Mehrabian, A. (1995). Framework for a comprehensive description and measurement of emotional states. *Genetic, social, and general psychology monographs*. (pages 18, 32, 60).
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. (page 100).
- Mitra, S. and Acharya, T. (2007). Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(3):311–324. (page 10).
- Mou, W., Celiktutan, O., and Gunes, H. (2015). Group-level arousal and valence recognition in static images: Face, body and context. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 5, pages 1–6. IEEE. (page 23).
- Newell, A., Yang, K., and Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer. (page 87).
- Nicolaou, M. A., Gunes, H., and Pantic, M. (2011). Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105. (page 23).
- Ouamane, A. (2015). *Reconnaissance Biométrique par Fusion Multimodale du Visage 2D et 3D*. PhD thesis. (pages xv, 10).

- Pantic, M. and Rothkrantz, L. J. (2000). Expert system for automatic analysis of facial expressions. *Image and Vision Computing*, 18(11):881–905. (page 21).
- Pantic, M., Valstar, M., Rademaker, R., and Maat, L. (2005). Web-based database for facial expression analysis. In *2005 IEEE International Conference on Multimedia and Expo*, pages 5 pp.–. (page 107).
- Patterson, G. and Hays, J. (2016). Coco attributes: Attributes for people, animals, and objects. In *European Conference on Computer Vision*, pages 85–100. Springer. (page 26).
- Picard, R. W. (1997). *Affective computing*, volume 252. MIT press Cambridge. (page 33).
- Polanía, L. F. and Barner, K. E. (2017). Group-Level Emotion Recognition using Deep Models on Image Scene , Faces , and Skeletons. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. (page 23).
- Posada-Quintero, H. F., Bolkhovsky, J. B., Qin, M., and Chon, K. H. (2018). Human performance deterioration due to prolonged wakefulness can be accurately detected using time-varying spectral analysis of electrodermal activity. *Human factors*, page 0018720818781196. (page 5).
- Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. (page 76).
- Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019. (page 66).
- Righart, R. and De Gelder, B. (2008). Rapid influence of emotional scenes on encoding of facial expressions: an erp study. *Social cognitive and affective neuroscience*, 3(3):270–278. (page 3).
- Ritter, S., Barrett, D. G., Santoro, A., and Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. *arXiv preprint arXiv:1706.08606*. (page 17).
- Riva, G., Mantovani, F., Capideville, C. S., Preziosa, A., Morganti, F., Villani, D., Gaggioli, A., Botella, C., and Alcañiz, M. (2007). Affective interactions using virtual reality: the link between presence and emotions. *CyberPsychology & Behavior*, 10(1):45–56. (page 5).

- Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., and Tomlinson, B. (2010). Who are the crowdworkers?: Shifting demographics in mechanical turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, pages 2863–2872, New York, NY, USA. ACM. (page 37).
- Ross, J., Zaldivar, A., Irani, L., and Tomlinson, B. (2009). Who are the turkers? worker demographics in amazon mechanical turk. *Department of Informatics, University of California, Irvine, USA, Tech. Rep.* (page 37).
- Russell, J. A. (1991). Culture and the categorization of emotions. *Psychological bulletin*, 110(3):426. (page 36).
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145. (pages 13, 17, 36, 36).
- Sabini, J. and Silver, M. (2005). Why emotion names and experiences don't neatly pair. *Psychological inquiry*, 16(1):1–10. (page 36).
- Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition and emotion*, 23(7):1307–1351. (page 36).
- Schindler, K., Van Gool, L., and de Gelder, B. (2008). Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural networks*, 21(9):1238–1246. (pages 10, 22, 23, 65).
- Schröder, M., Devillers, L., Karpouzis, K., Martin, J.-C., Pelachaud, C., Peter, C., Pirker, H., Schuller, B., Tao, J., and Wilson, I. (2007). What should a generic emotion markup language be able to represent? In Paiva, A. C. R., Prada, R., and Picard, R. W., editors, *Affective Computing and Intelligent Interaction*, pages 440–451, Berlin, Heidelberg. Springer Berlin Heidelberg. (page 18).
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. (page 86).
- Smith, C. A. and Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, 48(4):813. (page 36).
- Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., and Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14. (page 85).

- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958. (page 71).
- Srivastava, R. K., Greff, K., and Schmidhuber, J. (2015). Training very deep networks. In *Advances in neural information processing systems*, pages 2377–2385. (page 87).
- Stephani, C. (2014). Limbic system. In Aminoff, M. J. and Daroff, R. B., editors, *Encyclopedia of the Neurological Sciences (Second Edition)*, pages 897 – 900. Academic Press, Oxford, second edition edition. (pages 2, 13).
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9. (page 86).
- Tracy, J. L., Robins, R. W., and Schriber, R. A. (2009). Development of a faces-verified set of basic and self-conscious emotion expressions. *Emotion*, 9(4):554. (page 24).
- Turk, A. M. Mturk is now available to requesters from 43 countries. <https://blog.mturk.com/mturk-is-now-available-to-requesters-from-43-countries-77d16e6a164e>. Published: 2017-6-06. (page 37).
- Turk, A. M. We've made it easier for more requesters to use amazon mechanical turk. <https://blog.mturk.com/weve-made-it-easier-for-more-requesters-to-use-amazon-mechanical-turk-ab2ae649c5>. Published: 2016-13-10. (page 37, 37).
- <http://mplab.ucsd.edu>. The MPLab GENKI Database. (page 24).
- Wan, J., Zhao, Y., Zhou, S., Guyon, I., Escalera, S., and Li, S. Z. (2016). Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 56–64. (page 17).
- Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., and Movellan, J. R. (2014). The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98. (page 60).
- Wikipedia (2018). Machine Learning datasets. https://en.wikipedia.org/wiki/List_of_datasets_for_machine_learning_research. [Online; accessed 2-October-2018]. (page 29).

- Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014). Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*. (page 107).
- You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. (2016). Image captioning with semantic attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (page 100).
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67. (page 72).
- Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833, Cham. Springer International Publishing. (page 68, 68).
- Zhang, Z., Luo, P., Loy, C.-C., and Tang, X. (2015). Learning social relation traits from face images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3631–3639. (page 107).
- Zhao, K., Chu, W.-S., and Zhang, H. (2016). Deep region and multi-label learning for facial action unit detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (page 21).
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2015). Object detectors emerge in deep scene cnns. *International Conference on Learning Representations*. (pages 66, 70).
- Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017a). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*. (pages 60, 60, 62, 69, 69, 73, 73).
- Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. (2017b). Scene parsing through ade20k dataset. In *Proc. CVPR*. (page 31).