2020

# Learning Context-sensitive Human Emotions in Categorical and Dimensional Domains

Pooyan Balouchian

*University of Central Florida*

## STARS Citation

Balouchian, Pooyan, "Learning Context-sensitive Human Emotions in Categorical and Dimensional Domains" (2020). *Electronic Theses and Dissertations, 2020-*. 325.
https://stars.library.ucf.edu/etd2020/325

# EMOTION RECOGNITION: FROM CONTEXT-SENSITIVE DEEP LEARNING TO UNSUPERVISED RANKING OF CONTINUOUS EMOTIONS

by

POOYAN BALOUCHIAN
M.Sc. Eastern Mediterranean University, 2009

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
in the Department of Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Fall Term
2020

Major Professor: Hassan Foroosh

# ABSTRACT

Still image *emotion recognition* (*ER*) has been receiving increasing attention in recent years due to the tremendous amount of social media content on the Web. Many works offer both *categorical* and *dimensional* methods to detect image sentiments, while others focus on extracting the true social signals, such as *happiness* and *anger*. Deep learning architectures have delivered great success, however, their dependency on large-scale datasets labeled with (1) *emotion*, and (2) *valence*, *arousal* and *dominance*, in *categorical* and *dimensional* domains respectively, introduce challenges the community tries to tackle. Emotions offer dissimilar semantics when aroused in different *contexts*, however *"context-sensitive" ER* has been by and large discarded in the literature so far. Moreover, while *dimensional* methods deliver higher accuracy, they have been less attended due to (1) lack of reliable large-scale labeled datasets, and (2) challenges involved in architecting unsupervised solutions to the problem. Owing to the success offered by multi-modal *ER*, still image *ER* in the single-modal domain; *i.e*. using only still images, remains less resorted to. In this work, (1) we first architect a novel fully automated dataset collection pipeline, equipped with a built-in semantic sanitizer, (2) we then build *UCF-ER* with $50K$ images, and *LUCFER*, the largest labeled *ER* dataset in the literature with more than $3.6M$ images, both datasets labeled with *emotion* and *context*, (3) next, we build a single-modal *context-sensitive ER* CNN model, fine-tuned on *UCF-ER* and *LUCFER*, (4) we then claim and show empirically that infusing *context* to the unified training process helps achieve a more balanced *precision* and *recall*, while boosting performance, yielding an overall classification accuracy of 73.12% compared to the state of the art 58.3%, (5) next, we propose an unsupervised approach for ranking of continuous emotions in images using *canonical polyadic (CP) decomposition*, providing theoretical proof that rank-1 CP decomposition can be used as a ranking machine, (6) finally, we provide empirical proof that our method generates a Pearson Correlation Coefficient, outperforming the state of the art by a large margin; *i.e*. 65.13%

(difference) in one experiment and 104.08% (difference) in another, when applied to *valence rank estimation*.

I would like to wholeheartedly dedicate this dissertation to my amazing father who is the main reason and motivation behind all and any progress I ever made in my life, including but not limited to learning the philosophy of life, the true and practical meaning of human values, patience, honesty, sense of humor, good thoughts, good words and good deeds, friendship, and the list goes on and on, and last but not least learning English. I won’t be able to appreciate enough for your honest efforts in teaching me how to live life. I LOVE you beyond imagination and I can’t wait to regroup with you in a better world again. I would like to also dedicate this dissertation to my incredible mother who taught me the true meaning of unconditional love. Your patience, perseverance, encouragement and support from day one has been key in my life. I LOVE you so much.

# ACKNOWLEDGMENTS

I would like to hereby acknowledge the efforts, patience and guidance delivered by my knowledgeable adviser, Dr. Hassan Foroosh, for his tireless and generous assistance during the course of my PhD studies. Needless to mention that without his guidance, this dissertation would not carry the same weight. Moreover, I would like to acknowledge the valuable contributions of the honorable committee members, Dr. Charles Hughes, Dr. Ulas Bagci and Dr. Valerie Sims, for their constructive feedback and motivation. The generous and accurate feedback delivered during the proposal process contributed to the higher quality and validity of this dissertation and I would like to express my sincere appreciation for that. Last, but not least, I would like to extend my thanks to my wife, Marjaneh, for her continuous and reassuring encouragement, and for volunteering to help when needed the most during the course of my PhD.

Three previous works published by the author of this dissertation have been used in this manuscript. The copyright permissions for using these publications are attached in appendices B, C and D.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1: INTRODUCTION

In this chapter[1], we first provide a brief history of *Emotion Recognition*, discussing the core principles in the field. Next, *Affective Computing*, as an interdisciplinary field spanning computer science, psychology, and cognitive science is further introduced. We then pinpoint the motivation behind this work as well as the existing gaps we attempt to tackle accordingly. The chapter is concluded by touching on the main contributions we have made to the area of *Affective Computing*.

## Emotion Recognition: A brief history

Many philosophers, since the dawn of civilization, have reflected on the nature of emotions. Aristotle, one of the greatest philosophers of all time distinguished four humors. The Enlightenment philosophers attempted to identify human emotions and moods. The development of the field of psychology in the twentieth century, however, enabled a thorough analysis of human emotions [13]. The term "emotion" constitutes a hypothetical construct; *i.e.* a conceptual and operational definition of an underlying phenomenon that constitutes the object of theory and research. While the term "emotion" is used interchangeably with terms such as *affect*, *feeling*, *sentiment* or *mood*, psychologists define the construct as a process of changes in different components rather than a homogeneous state. Moreover, the differentiation of the emotions (*e.g.* fear, anger, joy, etc) is based on specific configurations of changes in the components. The question of *how many emotions are*

[1] This chapter includes excerpts from three works previously published by the author of this dissertation:
(1) "Context-Sensitive Single-Modality Image Emotion Analysis: A Unified Architecture from Dataset Construction to CNN Classification", Pooyan Balouchian, Hassan Foroosh, 2018 25th IEEE International Conference on Image Processing (ICIP), 1932-1936
(2) "LUCFER: A large-scale context-sensitive image dataset for deep learning of visual emotions", Pooyan Balouchian, Safaei M., Foroosh H., 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), 1645-1654
(3) "An Unsupervised Subspace Ranking Method for Continuous Emotions in Face Images", Pooyan Balouchian, Safaei M., Cao X., Foroosh H., 2019 30th British Machine Vision Conference

*there* and *what are they* plays a crucial role in the design of *human affects* systems. This question is answered very differently depending on the theoretical stance adopted [14]. Next, the major theoretical models of *human affect* is briefly described [15].

### *Dimensional Models*

Wundt [16] was the first to explicitly formulate the first *dimensional model* of *human affect*. He expressed emotions vary with respect to three independent dimensions: *i.e.* pleasantness-unpleasantness (*valence*), rest-activation (*arousal*), and relaxation-attention (*dominance*). Osgood suggested that all emotions are perceived in terms of *valence* and *arousal* [17]. Since then, many dimensional theorists hypothesized that the classification of emotions is described by one or both of these central dimensions.

### *Discrete Emotion Models*

Discrete emotion theorists define a limited number of basic emotions, such as fear, anger, joy, sadness, disgust and etc. as the core classes every emotion is characterized by. Some models suggest that the number of core emotions is determined by evolved neural circuits, such as circuit models [18], or by phylogenetically continuous classes of motivation such as aggression, as in motivational models; *e.g.* Plutchik [11]. These core emotions, or a combination thereof, explain a variety of emotion-descriptive verbal labels in different languages and cultures.

### *Meaning-oriented Models*

Semantic constructs in the emotion vocabularies are used to discuss the structure of different emotions. This is based on the assumption that there exists a mapping between the semantic and psychological structures of different emotions. A similar theory claims that emotions are constructed by socio-culturally determined behavior, suggesting that emotions cannot be simply reduced to basic psycho-biological patterns and that an emotion needs to be defined based on its semantics in the culture it's being contextualized in [15].

### *Adaptational Models*

Theorists proponent of the adaptational models suggest that the evolution of mankind has equipped us with organisms that are capable of reacting automatically to certain events. Such models are referred to as *biological preparedness* models of *human affect*. An example includes the work carried on by Oahman [19], suggesting that humans are equipped with automatic detection mechanisms when exposed to certain threat stimuli, such as snakes and spiders.

### *Componential Models*

Componential emotion theorists believe that emotions are aroused by means of a cognitive appraisal of events, suggesting that different reactions in response to certain situations is produced by the outcome of this evaluation process [14].

## Affective Computing

*Affective computing* revolves around the creation of and interaction with machine systems, capable of sensing, recognizing, responding to, and influencing emotions [20]. *Affective computing* is a multi-disciplinary field that spans across multiple, and somewhat fundamentally disparate, disciplines such as psychology, sociology, engineering, computer science, linguistics and physiology.

Such disparate and wide range of disciplines that *affective computing* relates to, suggests the immense complexity in understanding of human emotions and its importance in today's world.

In this section, we further discuss the modern science of *affective computing*, by first introducing the concepts of *affect sensing* and *affect generation*. Moreover, the application areas of *affective computing* are touched on and the ethical issues arising from building of affective systems are further outlined.

### *Affect Sensing*

The first step towards building an *affective computing* system is recognizing emotions. The term *affect sensing* refers to computer systems, capable of recognizing human emotions by means of receiving data through signals and patterns [20].

Facial expressions are informative due to their visibility and omnipresence [21]. Face gestures such as smiling and frowning convey valuable information with the emotion being aroused. Basic emotion categories, as they relate to the *discrete* (*categorical*) emotion model, can be identified using methods such as HMM, optical flow, active appearance model, and neural networks [22]. These systems may be used in combination relying on *early fusion* or *late fusion* techniques. A widely used system in *emotion classification* is the Facial Action Coding System (FACS), created by Paul Ekman, Wallace V. Friesen, and Joseph C. Hager [23]. FACS identifies facial expressions as well as the muscles producing these expressions. In addition to FACS, the Emotion Facial Action Coding System (EMFACS) is a similar system used to score the facial actions, applicable to sensing emotions [23].

Other systems contributing to *affect sensing* include posture and gesture recognition, which deals with identifying different body states, when emotions are aroused as a result of being exposed to certain visual stimuli. These systems help recognize gestures, resulting from the movement of different body parts, and postures as they relate to the position of the body of the person subject of showing emotions.

Furthermore, the vocal expressions are considered as valuable cues in determining human emotions. These include cues suggesting the most important content in the message, and cues arising from the speaker's affective state [24]. The common vocal cues with emotion categories are speech rate, pitch average, pitch range, intensity, voice quality, pitch changes and articulation [25].

Finally, another modality that has recently attracted a lot of attention in the *affective computing* community is text. With recent advancements in the field of Natural Language Processing (NLP) techniques, sentiment analysis tools benefit from computational linguistics and text analysis to boost the performance in recognizing emotions. Dominating tools and technologies used for this purpose include WordNet-Affect [26], SenticNet [27], and SentiWordNet [28].

### *Affect Generation*

Another sub-area of *affective computing* involves designing systems and robotic agents capable of conveying intention in a human-perceptible fashion. Optimal human-computer interaction require computer agents to possess the following attributes: embodiment in the physical environment, quick reactions to unseen events, computational power to meet goals [29]. There exist numerous use cases that humans benefit from by having intelligent robots, capable of understanding as well as conveying emotions in a human-like fashion in the context of office, medicine, hotel use, cooking, marketing, entertainment, recreation, therapy and rehabilitation, among others [30].

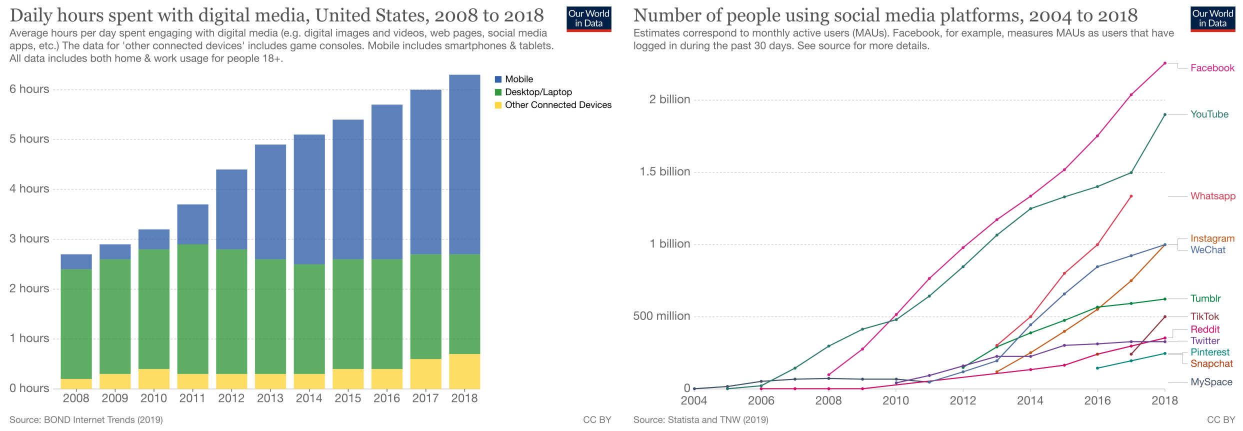
### *Applications of Affecting Computing*

*Affective computing* has a wide range of application areas. Picard [20] classifies *affective computing* application areas into three main categories; *i.e.* (1) systems that detect human emotions, (2) systems capable of expressing human-like emotions, and (3) systems capable of “feeling” an emotion. Healthcare systems benefit from *affective computing* in different ways. Research has indicated that patients dealing with Asperger syndrome (AS) or high functioning autism (HFA) experience difficulty expressing their emotions compared to those without these syndromes. Such traits affect their relationship with others [31]. Tools have been developed to track the behavioral patterns, moods and triggers for these patients in an attempt to equip their therapists with better insight into the patients emotional outbursts.

Furthermore, in the context of education system, *affective computing* systems help teachers to better understand the engagement of students during lectures. This is done by monitoring conversational cues, body language, as well as facial expression recognition using variety of sensors [32].

## Motivation

The popularity of social networks has contributed to rapid growth of multimedia content; *i.e.* image, text, audio and video, on the web. Figure 1.1 depicts statistics portraying the daily hours spent with digital media in the United States from 2008 to 2018, as well as the trend of the number of people using social media platforms from 2004 to 2018 [4]. The user engagement on popular social media such as Facebook, Instagram, Snapchat, YouTube, among others, is mainly characterized by interaction with multimedia content in the form of images and videos. Users post more than 100 million posts on Instagram on a daily basis, 5 billion videos get viewed on YouTube, with more



(a) Daily hours spent with digital media, United States, 2008 to 2018 [4]

(b) Number of people using social media platforms, 2004 to 2018 [4]

Figure 1.1: Social Media Usage Statistics [4]

than 60 billion messages sent on WhatsApp. SnapChat is host to more than 109 million users per day with more than 3 billion daily snaps and 10 billion daily video views. The tremendous number of multimedia content uploaded on these platforms has resulted in tremendous demand for data retrieval and understanding.

*Affective computing* as the interdisciplinary field spanning computer science, psychology, and cognitive science helps in the study and development of systems and devices that can recognize, interpret, process, and simulate human affects [33]. *Sentiment analysis* and *emotion recognition* are among the main areas leveraging *affective computing* techniques. Image *sentiment analysis* is a coarse-grained approach that deals mainly with the polarity of the image, detecting if the still image is categorized as positive, negative or neutral. On the other hand, image *emotion recognition*, entails a fine-grained, deep dive into the themes associated with each emotion, dealing with recognizing the exact emotion aroused when exposed to certain visual stimuli. While tremendous amount of research has been carried on in the area of *sentiment analysis*, still image *emotion*

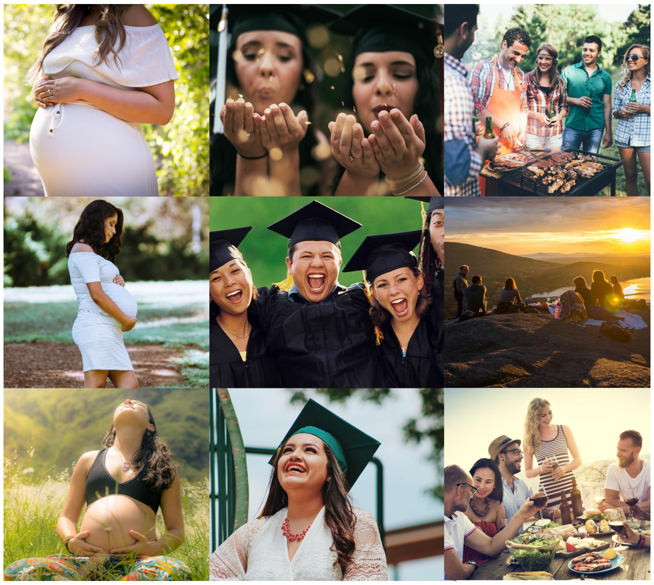


Figure 1.2: Excerpt from noise-reduced *LUCFER* labeled by AMT workers depicting the *happiness emotion* in 3 different contexts (from left to right): (1) *pregnancy*, (2) *graduation* and (3) *picnic*.

*recognition* remains to be less attended [33].

While in a face-to-face interaction, humans detect and interpret interactive signals of their communicator with little effort, design and development of an automated system that accomplishes the same purpose is rather difficult [34]. Moreover, detecting the *context* the emotion conveys exacerbates the challenging nature of *emotion recognition*, where *context* is defined as the whole set of secondary characteristics of a situation or secondary properties of a cognitive or motivational state of an individual which may modify the effect of an effective stimulation (stimulus) or an oriented activity [35].

## Main Contributions

Deep learning has recently enabled robust feature learning, yielding promising results in a variety of computer vision and multimedia tasks such as image classification and scene detection. The challenge, however, is the demanding nature of these systems for large-scale datasets required for training. Lack of such large-scale datasets urged us to build *UCF ER*[2], a *context-sensitive emotion recognition* dataset containing 50,000 images, labeled with *emotion* and *context*. We then built *LUCFER*[3], a dataset containing $3.6M$ still images labeled with *emotion* and *context*, along with a rich set of metadata including objects, bounding boxes, related searches, related images, objects and etc. Both datasets are publicly available for research purposes under the *Creative Commons Attribution 4.0 International* license. *LUCFER* is 156 times larger than the largest dataset of the kind currently available; *i.e*. Flickr-Instagram dataset [12]. To build *LUCFER*, we first architected a fully-automated dataset collection pipeline, equipped with a semantic sanitizer component. Chapter 3 further provides details on the architecture of the designed pipeline, providing statistics on *LUCFER*.

We further train a *context-sensitive* classifier to classify images based on both *emotion* and *context*, pioneering the first single-modal *context-sensitive emotion recognition* CNN model. Using an empirical approach, we claim and show that embedding *context* as part of a unified training process not only helps boost performance, but also helps deliver a more balanced *precision* and *recall*. We draw the conclusion that CNNs are better fit to learn emotion prediction models by running experiments on fine-grained (context-sensitive) compared to coarse-grained (context-free) datasets, running extensive experiments supporting the claim.

Having approached *emotion recognition* from a *categorical* angle leveraging deep learning, we

[2] UCF ER: https://cil.cs.ucf.edu/dataset-2/ucf-er/

[3] LUCFER: https://cil.cs.ucf.edu/dataset-2/labeled-ucf-emotion-recognition/

explored the problem in the *dimensional* domain. Continuous *dimensional* models of human affect have shown to offer a higher accuracy in identifying a broad range of emotions compared to the discrete *categorical* approaches dealing only with emotion categories such as *joy*, *sadness*, *anger*, etc. Unlike the majority of existing works benefiting from *dimensional* models of human affect (VAD; *i.e.* Valence, Arousal and Dominance) that mainly rely on training-based (supervised) approaches, here we propose a fully unsupervised novel method for ranking of continuous emotions in images using *canonical polyadic decomposition*. To better portray the efficacy of our proposed approach, we provide theoretical and empirical proof that our system is capable of generating a Pearson Correlation Coefficient (PCC) that outperforms the state of the art by a large margin; *i.e.* 65.13% (*i.e.* difference in PCC) in one experiment and 104.08% (*i.e.* difference in PCC) in another, when applied to *valence rank estimation*. Towards this aim, we run experiments on four major *emotion recognition* datasets; *i.e.* *CK+*, *AFEW-VA*, *SEMAINE* and *AffectNet*, and provide comprehensive analysis on the observed results accordingly. Our datasets are selected in a way to include images collected under controlled environments such as a laboratory setting; *e.g.* *CK+* and *SEMAINE*, semi-controlled environments; *e.g.* *AFEW-VA*, and uncontrolled environments (from the *wild*); *e.g.* AffectNet.

We further performed extensive ablation studies to monitor the performance of our designed ranking machine. Our ablation studies are designed to measure the fault-tolerance of the proposed method. Details on these studies are discussed in chapter 4, section 4.

In chapter 2, we will dig into more details, summarizing the current state of the art in the area of *emotion recognition*. The sub-areas and open problems we have tackled in this work will be further pointed out in the same chapter, specifying similarities and differences between our work and those proposed by state of the art.

In chapter 3, the architecture of the designed dataset collection pipeline gets further discussed and comprehensive statistics on *UCF ER* and *LUCFER* will be provided accordingly.

Chapter 4 further digs into details of our proposed approach to the problem of *emotion recognition*, portraying empirical and theoretical proofs, when applicable.

Finally, in chapter 6, we summarize the work we are presenting by pinpointing the gaps we have filled, lessons learned during the whole process, doors closed, while pointing out doors opened as a result of our effort in this work, concluding by providing some insight on the future directions the community is currently exploring.

# CHAPTER 2: LITERATURE REVIEW

In this chapter[1], we further review the main building blocks of *emotion recognition* in both *categorical* and *dimensional* domains. The widely-used emotion models and the state of the art in *emotion recognition* will further be discussed in this chapter. We further pinpoint how our work is similar or different compared to state of the art, touching on the sub-areas we address.

## Emotion Models in Psychology

There exist mainly two governing emotion representation models deployed in the field of psychology: *categorical* and *dimensional*. The *categorical* models classify human emotions into a number emotion classes, *e.g.* happiness, anger and etc. Some of the widely used and dominant *categorical* models the *affective computing* community benefits from include Mikels' eight emotions [36], Ekman's six basic emotions [37] and Plutchik's wheel of emotions [38].

An emotion is referred to as *sentiment* when classified into positive, neutral, or negative polarities. Human emotions, however, are better modeled as continuous coordinate points in a 3D or 2D Cartesian space; *i.e.* *valence*, *arousal* and *dominance* (VAD) [39]. *VAD* is the most widely used *dimensional* model of human affect, where *valence* represents the pleasantness ranging from positive to negative, *arousal* represents the intensity of emotion ranging from excited to calm, and *dominance* represents the degree of control ranging from controlled to in control. *Dominance* is

[1] This chapter includes excerpts from three works previously published by the author of this dissertation:
(1) "Context-Sensitive Single-Modality Image Emotion Analysis: A Unified Architecture from Dataset Construction to CNN Classification", Pooyan Balouchian, Hassan Foroosh, 2018 25th IEEE International Conference on Image Processing (ICIP), 1932-1936
(2) "LUCFER: A large-scale context-sensitive image dataset for deep learning of visual emotions", Pooyan Balouchian, Safaei M., Foroosh H., 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), 1645-1654
(3) "An Unsupervised Subspace Ranking Method for Continuous Emotions in Face Images", Pooyan Balouchian, Safaei M., Cao X., Foroosh H., 2019 30th British Machine Vision Conference

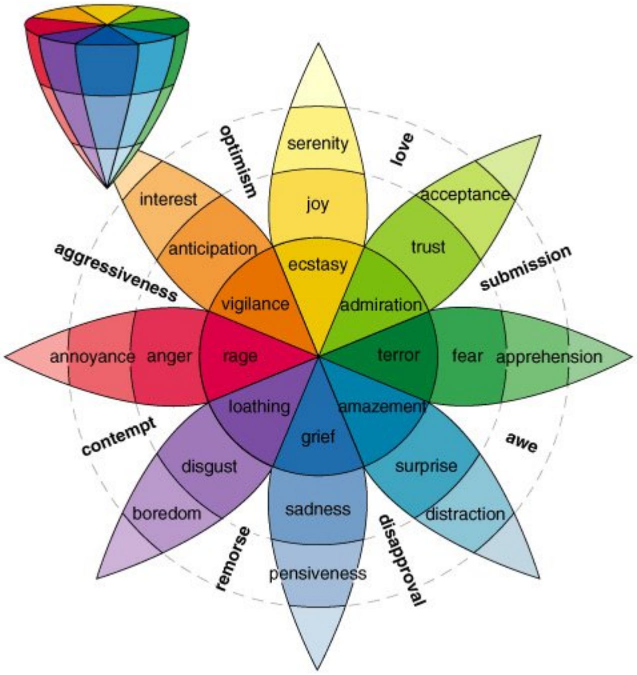


Figure 2.1: Plutchik's Wheel of Emotions

difficult to measure and is often omitted, leading to the commonly used two dimensional *VA* space [40].

In our work and as part of proposing a deep learning solution in the *categorical* domain, we adopt the *Plutchik*'s wheel of emotions model depicted in figure 2.1 to build *UCF ER* (UCF Emotion Recognition) and *LUCFER* datasets. This choice was motivated by the depth the *Plutchik*'s wheel of emotions represents compared to other emotion models. In 1980, Robert Plutchik constructed a wheel-like diagram of emotions, depicted in figure 2.1, visualizing eight basic emotions: *joy*, *trust*, *fear*, *surprise*, *sadness*, *disgust*, *anger* and *anticipation* [11]. Plutchik's three-dimensional model describes the relations among emotions, which is extremely helpful in understanding how complex emotions interact and change over time, hence embedding *valence* as part of its emotion definition system. The eight sectors are designed to indicate that there are eight primary emotion dimensions. The cone's vertical dimension represents intensity - emotions intensify as they move

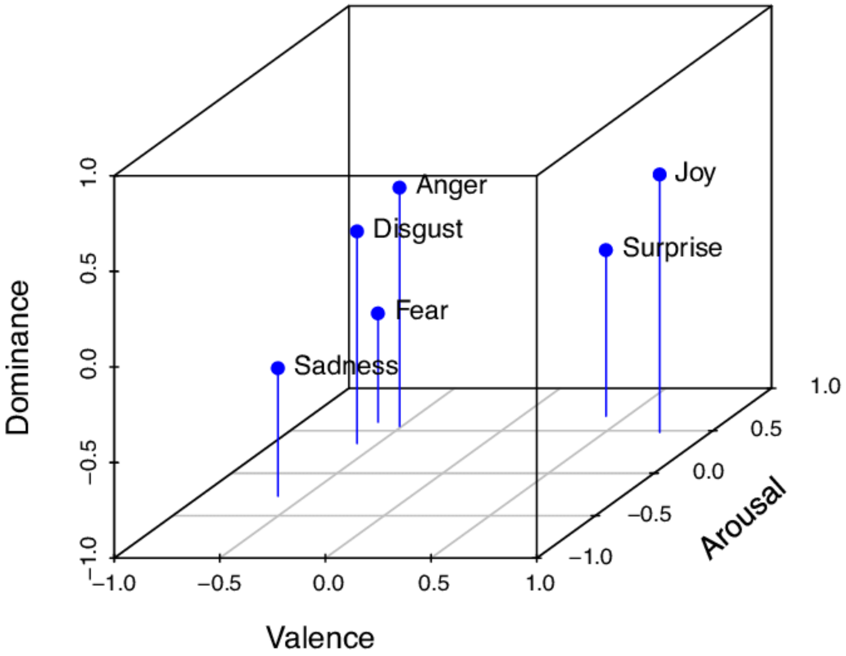


Figure 2.2: VAD Model of *affective computing*, depicting *valence*, *arousal* and *dominance*

from the outside to the center of the wheel.

Furthermore, with respect to our proposed method in the *dimensional* space, we employ the *valence* dimension of the widely-used *VAD* model of *affective computing*. Figure 2.2 provides a visual depiction of the 3-dimensional *VAD* model in a -1 to +1 rating scale, along with a number of sample emotion categories in this space.

## Image Datasets for Affective Computing

In the early phases of *affective computing*, before the social networks were widely used by users to share multimedia content on, the image datasets built and adopted were small in size. **IAPS** (International Affective Picture System), a widely used emotion dataset offers complex scenes, with each image associated with the mean and standard deviation (STD) of *VAD* ratings in a 9-point scale

Table 2.1: Publicly accessible *emotion* datasets showing (1) number of images, (2) emotion model adopted, (3) labeling method used, (4) whether or not *context-sensitive* labels exist, and (5) the label type; *i.e. dimensional*, *categorical* or hybrid. *VAD* stands for Valence, Arousal, Dominance.

| **Dataset** | **# Images** | **Emotion Model** | **Labeling Approach** | **Context-Sensitive Labels?** | **Label Type** |
|---|---:|:---:|:---:|:---:|:---:|
| IAPS [41] | 1,182 | VAD | Human Judges | No | Dimensional |
| ArtPhoto [42] | 806 | Mikels | Human Judges | No | Categorical |
| Flickr-Instagram [12] | 23,308 | Mikels | Amazon Mech. Turk | No | Categorical |
| Emotion6 [43] | 1,980 | Ekman+neutral | Amazon Mech. Turk | No | Categorical |
| EMOTIC [44] | 18,316 | 26 hand-picked categories | Amazon Mech. Turk | No | Hybrid |
| FlickrLDL [45] | 10,700 | Mikels | Human Judges | No | Categorical |
| TwitterLDL [45] | 10,045 | Mikels | Human Judges | No | Categorical |
| IAPSa [36] | 246 | Mikels | Human Judges | No | Dimensional |
| GAPED [46] | 730 | Sentiment, VA | Human Judges | No | Hybrid |
| MART [47] | 500 | Sentiment | Relative Score Method | No | Categorical |
| devArt [47] | 500 | Sentiment | Relative Score Method | No | Categorical |
| Tweet [48] | 603 | Sentiment | Amazon Mech. Turk | No | Categorical |
| FlickrCC [48] | ˜500,000 | Plutchik's wheel of emotions | Amazon Mech. Turk | No | Categorical |
| Flickr [49] | ˜300,000 | Ekman | Keyword Matching | No | Categorical |
| IESN [50] | 1,012,901 | Mikels, VAD | Human Judges | No | Hybrid |
| **UCF ER** [1] | **50,000** | Plutchik's wheel of emotions | Human Judges | **Yes** | Categorical |
| **LUCFER** [2] | **3.6M** | Plutchik's wheel of emotions | Amazon Mech. Turk | **Yes** | Hybrid |

[41]. This dataset was manually built and labeled by 100 students. The **IAPSa** dataset is a subset of **IAPS**, composed of 246 images, annotated by 20 students [36]. **ArtPhoto**, which includes a set of 806 artistic photographs is annotated by the artist uploading the photos to a photo sharing website [42]. **GAPED** (The Geneva Affective Picture Database) is comprised of 520 negative, 121 positive, and 89 neutral images [46], tagged with *valence* and *arousal* in a 0-100 rating scale. **MART** and **devArt** datasets both contain 500 abstract paintings each. **MART** contains paintings by professional artists and **devArt** consists of paintings by amateur artists, collected from the Museum of Modern and Contemporary Art of Trento and Rovereto [47], and the "DevianArt" social network [47], respectively.

With the advent of social networks, the volume of multimedia content on the web started piling up fast. This led to creation of large-scale datasets constructed from these sites. The Tweet sentiment dataset includes 470 and 113 positive and negative sentiments respectively [48]. The **FlickrCC** dataset is constructed based on 1,553 adjective noun pairs (ANPs) to generate a total of $500K$ Flickr creative common (CC) images [48]. In this dataset, the images are mapped to the *Plutchik*'s wheel of emotions. The **Flickr** dataset contains a set of $300K$ images [49]. In this dataset, the emotion category associated to an image is defined via synsets (list of synonymous words) to the adjective words found in an image's tags and comments. A widely used emotion recognition dataset, *i.e*. **Flickr-Instagram** [12], consists of 23,308 images collected from Flickr and Instagram. **Flickr-Instagram** uses a keyword-based search approach to crawl Flickr and Instagram. The collected images are then labeled by 225 Amazon Mechanical Turk (MTurk) workers in an attempt to sanitize the collected dataset. The **Emotion6** dataset [43] is a well-balanced dataset with 330 images representing each of the Ekman's emotion categories. 15 AMT Workers verified each image, offering high confidence in the validity of each image assigned to an emotion category. The **IESN** (Image-Emotion-Social-Net) dataset [50] contains around $1M$ images collected from Flickr, used for personalized emotion prediction. **FlickrLDL** and **TwitterLDL** datasets [45] are constructed for discrete emotion distribution learning. The former one is a subset of **FlickrCC**, which are labeled by 11 viewers. The latter one consists of 10,045 images which are collected by searching various sentiment keywords from Twitter and labeled by 8 viewers.

Table 2.1 lists popular *emotion recognition* datasets including *UCF ER* and *LUCFER* along with relevant statistics for each. *LUCFER* differs from the similar datasets mainly in terms of its *volume*, *rich metadata*, *multi-dimensional* labels; *i.e*. *<emotion-context>* it is tagged with, and the fact that it has *valence* self-embedded. *UCF ER* and *LUCFER* contain 50,000 and $3.6M$ still images respectively, tagged with *emotion* and *context* categories. Our datasets do not merely rely on human judges or AMT workers, but the images are also validated by a semantic sanitizer, taking advantage of a WordNet-based semantic similarity component. This resulted in a higher confidence in the quality of the constructed dataset. The sanitized set is further enhanced in size through integration with the *Bing's Cognitive Services API*. The architecture of our dataset construction pipeline is further discussed in chapter 3 comprehensively.

## Deep Learning Methods for Still Image Emotion Recognition

Research in the area of *emotion recognition* is mainly focused on inferring emotions via multimodal approaches, taking advantage of text, still images, audio and video data, relying on deep learning frameworks. Mainly due to the their capacity that can be controlled by varying their depth and breadth [51], CNNs have helped many works taking advantage of transfer learning. In a recent survey on the topic of emotion and social signals [52], it is argued that aside from textual and written information provided by users, multimodal human behavioral information that is present within the media provides a vast source of information, referred to as *affective and social content*. The majority of these methods model emotions using *categorical* (using the emotion category as labels) or *dimensional* approaches (*valence*, *arousal* and *dominance*) studied by emotion theorists [53]. Most of the models surveyed in [54] and [55] adopt the *categorical* approach.

Some of the recent works involving *categorical* approaches include [56], [57], [49], [12], [42], [58], [59], [60], [61], while [62], [63], [64] and [65] propose *dimensional* methods, with some employing hybrid models leveraging both methods [44].

In [59] and [60], efforts are made to address the problem of visual sentiment analysis based on CNNs, where the sentiments are predicted using multiple affective cues. It is argued that providing the localized information of the affective images in addition to the holistic representations, helps boost performance when experimented on six benchmark datasets; i.e. *Flickr-Instagram* (FI), *Flickr*, *Instagram*, *EmotionROI*, *Twitter I* and *Twitter II*. In another work, [61], a multi-task deep framework is developed to leverage the ambiguity and relationship between emotional categories

Table 2.2: Related works on deep learning based *affective computing* of still images. Under **Task**, *cla* stands for *classification*, *ret* stands for *retrieval*, and *dis_d* stands for *discrete distribution.*

| **Base Network** | **Pre-trained** | **Context** | **Datasets** | **Dataset size** | **Task** | **Result** |
|---|---|---|---|---|---|---|
| Custom [68] | no | no | FlickrCC | ˜500K | cla | 0.781 |
| AlexNet [12] | yes | no | FI | 23,308 | cla | 0.583 |
| Custom [69] | no | no | FI, IAPSa, ArtPhoto | 23,308, 246, 806 | cla | 0.730, 0.902, 0.855 |
| GoogleNet-Inception [70] | yes | no | Flickr-Instagram, IAPSa, Abstract, ArtPhoto | 23,308, 246, 279, 806 | cla<br>ret | 0.676, 0.442, 0.382, 0.400<br>0.780, 0.819, 0.788, 0.704 |
| AlexNet [43] | yes | no | Emotion6 | 1,980 | dis_d | 0.480 |
| VGG16 [61] | yes | no | Emotion6, FlickrLDL, TwitterLDL | 1,980, 10,700, 10,045 | dis_d | 0.420, 0.530, 0.530 |
| **VGG16 (our work)** | **yes** | **yes** | **UCF ER** | **50,000** | **cla** | **0.711** |
| **ResNet-50 (our work)** | **yes** | **yes** | **UCF ER** | **50,000** | **cla** | **0.766** |
| **VGG16 (our work)** | **yes** | **yes** | **LUCFER** | **3.6M** | **cla** | **0.731** |

for visual sentiment prediction, showing that the proposed method performs favorably against state of the art.

Some other works utilize *user demographics* to infer the emotion of an image ([66] and [67]). These demographics include gender, age, social and political views of a user on an array of social networks. These efforts borrow the theory that a correlation exists between the user's behavioral patterns and the demographics.

In the closest efforts in the literature, [12] and [44] have made valuable contributions in the area of *emotion recognition* dataset construction and building prediction models, forming a baseline for other works to be compared against. [12] introduced an *emotion recognition* dataset of 23,308 strongly labeled images to address the challenge posed by unavailability of a large-scale well labeled dataset specifically for the task of *emotion recognition.* Meanwhile, [12] evaluated the deep visual features extracted from differently trained neural network models, suggesting the deep CNN features outperform the state-of-the-art hand-tuned features for visual emotion analysis. Their results demonstrate a classification accuracy of 58.3% on their fine-tuned CNN, forming a baseline for the community to work with. They choose eight emotion categories derived from a psychological study in [36]; *i.e*. Mikels' eight emotions.

In an effort to alleviate the challenges imposed by weakly labeled datasets, [68] employs CNNs. They follow a three-step process. They first design a CNN to perform sentiment analysis, collecting roughly $500K$ training samples, employing a baseline sentiment algorithm in order to label Flickr images. Next, a progressive strategy is employed in order to fine-tune the deep network. Finally, the performance on Twitter images is boosted by inducing domain transfer with a small number of manually labeled Twitter images.

More recently, [70] explored *deep metric learning* to observe the correlation of emotional labels with the same polarity, and further proposed a multi-task deep framework in an effort to optimize retrieval and classification tasks. Taking into consideration the relations among emotional categories in the Mikels' wheel, they jointly optimized a novel sentiment constraint with the cross-entropy loss. Extending triplet constraints to a hierarchical structure, the sentiment constraint employs a sentiment vector based on the texture information from the convolutional layer to measure the difference between affective images.

The deep methods discussed so far mainly focus on the dominant emotion prediction. There exist, however, some other methods proposed on emotion distribution learning. A work pioneering this methodology is a mixed bag of emotions, in which a deep CNN regressor (CNNR) is trained for each emotion category in Emotion6 [43] dataset based on AlexNet. The number of output nodes is changed to 1, facilitating prediction of a real value for each emotion category. The Softmax loss is further replaced with Euclidean loss. The probabilities of all emotion categories are normalized to enforce the sum of different probabilities to be 1. CNNR poses some limitations. Firstly, the predicted probability is not guaranteed to be a non-negative value. Secondly, the correlation among different emotion categories are ignored as the regressor for each emotion category is trained independently.

A key area that remains less attended is ignoring the *context* in which the emotion is conveyed.

In [44], effort is made to address the problem of emotion state recognition in context. [44] introduces the EMOTIC database, containing 18,316 images, collected in non-controlled environments containing people in context, combining two different types of annotations; *i.e*. 26 emotional categories, and 3 continuous emotional dimensions in the VAD (*Valence*, *Arousal*, and *Dominance*) space.

The dominant deep learning based methods discussed are summarized in table 2.2 along with unique features associated with each method. Our work, compared to [12] and [44], is different in that *LUCFER* is labeled with 275 (as opposed to 8) different *emotion-context* categories, containing more than $3.6M$ labeled images, hence constructing a dataset, 156 times larger than the largest dataset of the kind currently available in the community; *i.e*. 23,308 from [12]. Moreover, *LUCFER* is not only enriched with emotion categories, but also the *context* in which the emotion is triggered under, forming a 2-dimensional label. The term *context* in [44] is defined by the visual features defining the background of the main subject in the image, treated as helpers to better define the emotion in question. On the contrary, *context* in our work is defined as the circumstances that form the setting for an event, in terms of which it can be fully understood and assessed, hence treating it as a semantic, rather than a visual, element in the image. Moreover, borrowing from the *Pluthik*'s wheel of emotions [11], we quantized the values 0 to 10 and assigned them to the relevant emotion categories based on where on the *Plutchik*'s wheel each emotion category appears.

## Machine Learning for Ranking

Ranking data is an important problem of machine learning, mainly approached as a supervised, semi-supervised or reinforcement learning technique. Its application areas include information retrieval, document retrieval, collaborative filtering, sentiment analysis and online advertising [71]. Chapelle and Chang (2011), among other researchers in the field, argue that state-of-the-art learning to rank models can be categorized into three types [72].

Pointwise methods, such as decision tree models and linear regression, directly learn the relevance score of each instance. The final ranking is achieved by simply sorting the result list by these document scores. Ordinal regression and classification algorithms could benefit from this approach when used to predict the score of a single query-document pair.

Pairwise methods, such as rankSVM [73] learn to classify preference pairs by learning a binary classifier that prefers one document over the other given pair of documents. The goal for the ranker is to minimize the number of inversions in the ranking; *i.e*. cases where the pair of results are in the wrong order relative to the ground truth. One could argue that pairwise approaches might have an edge compared to pointwise approaches considering the fact that predicting relative ordering is closer to the nature of ranking as opposed to predicting class label or relevance score. A number of popular *learning to rank* approaches take advantage of pairwise techniques, namely RankNet [74], LamdaRank and LamdaMART [75].

Listwise methods, such as LambdaMART [75] tend to directly optimize the measurement for evaluating the entire ranking list. Moreover, methods have been proposed that combine more than one of these categories, *e.g*. GBRank [76], which proposes a combination of pointwise decision tree models and pairwise loss. A widely used method utilized as a pairwise approach is rankSVM [73], which is considered an extension to standard Support Vector Machines (SVM) by Boser *et al*. (1992) [77] and Cortes and Vapnik (1995) [78].

Moreover, there are works on unsupervised ranking of images leveraging a probabilistic approach, such as one by Horster *et al*. (2009) [79]. Horster *et al*. (2009) [79] hypothesizes the photos at the peaks of a distribution are the most likely photos for any given category, making such images the most representative. In contrast to such probabilistic methods, ours does not require a large number of images in order to derive a high quality ranking.

### Works on Valence Estimation

While deep architectures have been proven to yield promising results in different computer vision tasks, specifically classification of multimedia content, they have not been able to offer similar robust results when applied on continuous *valence* and *arousal* estimation. There exist a large number of works in the domain of *valence* and *arousal* estimation, with the majority of these methods focusing on multimodal audio/video-based estimation as opposed to single modality approaches in the domain of still images. *Valence estimation* was mainly approached using coarse-level methods, posing the problem as a classification problem (*e.g.* positive vs. negative). Later, researchers in the field started treating the problem in the continuous domain [80], [81], [82]. The majority of these approaches mainly rely on metadata derived from other modalities such as text, audio and video to perform *valence estimation* [80].

The continuous audio/visual emotion challenge [83], AVEC in short, started in 2011, aims to bring together researchers from the audio and video analysis communities around *emotion recognition* with the goal being to recognize the four continuously valued affective dimensions; *i.e. valence*, *arousal*, *dominance* and *expectancy*. Initially starting with a subset of SEMAINE dataset [83] and later switching to RECOLA [84], computer vision researchers proposed solutions to tackle the problem of audio/visual *emotion recognition* in the continuous domain, reporting performance using the Pearson Correlation Coefficient (PCC).

Several methods have been proposed to address *valence estimation* in the continuous domain, the summary of which is presented here in table 2.3.

Numerous existing approaches to *valence estimation* use static regression [82], [106], [107], [108]. These methods that are mainly used as baseline methods range from linear regression, partial least squares regression to kernel-based methods such as Nadaraya-Watson kernel regression [86].

Table 2.3: Related works on continuous *valence estimation*. Modalities are abbreviated as V: video, AV: Audio/Video, AVM: Audio/Video/Meta-data and AVP: Audio/Video/Physiological. PCC stands for Pearson Correlation Coefficient.

| **Method** | **Modality** | **Dataset** | **Valence (Average PCC)** |
|---|---|---|---:|
| OA BLSTM-NN [85] | AV | Semaine subset | 0.796 |
| N-W kernel regression [86] | AV | Semaine (AVEC 12) | 0.341 |
| Fuzzy inference system [87] | AVM | Semaine (AVEC 12) | 0.42 |
| CSR [88] | AV | Semaine subset | 0.21 |
| SVR and CCRF [89] | V | Semaine (AVEC 12) | 0.343 |
| RF [90] | V | Semaine (AVEC 12) | 0.454 |
| Doubly sparse RVM [91] | V | Semaine (AVEC 12) | 0.31 |
| Time-delay NN [92] | V | Semaine (AVEC 12) | 0.308 |
| Time-delay NN [92] | V | AVEC13 | 0.127 |
| SVR [93] | AV | AVEC13 | 0.135 |
| SVR [94] | AVM | AVEC14 | 0.587 |
| CCA [95] | V | AVEC14 | 0.381 |
| LR [96] | AV | AVEC14 | 0.493 |
| Deep belief network [97] | AV | AVEC14 | 0.528 |
| OA RVM [98] | AVP | AVEC15 | 0.588 |
| LR + boosted regression trees [99] | AV | AVEC15 | 0.501 |
| RF + gradient boosting + SGD [100] | AVP | AVEC15 | 0.490 |
| RNN [101] | AVP | AVEC15 | 0.590 |
| LSTM-RNN [102] | AVP | AVEC15 | 0.627 |
| Deep BLSTM-RNN [103] | AVP | AVEC15 | 0.616 |
| LSTM - kalman filter [104] | AVP | AVEC16 | 0.689 |
| RF + LR [105] | AVP | AVEC16 | 0.634 |

Moreover, different types of fusion methods, including early fusion and late fusion are proposed with the former offering methods that combine geometric and appearance features before training, and the latter combining estimations resulted from different modalities and later fusing them together forming a uniform estimation.

In the next two chapters, *i.e.* 3 and 4, we provide details on the architecture of our proposed dataset construction pipeline, followed by our proposed methods to the problem of *emotion recognition* in context, tackling the problem from both *categorical* as well as *dimensional* angles, highlighting the differences between our methods vs. state of the art.

# CHAPTER 3: DATASET CONSTRUCTION

With reference to figure 1.1, the volume of multimedia content on the web is increasing at a rapid pace. On the other hand, due to their demanding nature for large-scale datasets, deep learning architectures designed to tackle problems in various domains have urgent need for large-scale structured datasets that are labeled and available for use. This necessity urged us to architect and develop a reusable dataset collection pipeline built for large-scale datasets. We further used the pipeline to collect *UCF ER* and *LUCFER*, containing 50,000 and $3.6M$ images respectively and used both datasets to benchmark our proposed methods.

In this chapter[1], we first provide details on the emotion definition system used. Next, we elaborate on the method developed to collect images from the wild, and the labeling process adopted. We then discuss the approach employed to enhance *LUCFER*'s size using *Bing's Cognitive Services API* [2]. Furthermore we provide detailed statistics on both *UCF ER* and *LUCFER*.

## *Emotion Definition System*

There exist a number of different emotion definition systems in the context of cognitive sciences. These include the work by [36] (adopted in [12]), collecting descriptive emotional category data on subsets of the *International Affective Picture System* (IAPS) to identify images that elicit one

[1] This chapter includes excerpts from three works previously published by the author of this dissertation:
(1) "Context-Sensitive Single-Modality Image Emotion Analysis: A Unified Architecture from Dataset Construction to CNN Classification", Pooyan Balouchian, Hassan Foroosh, 2018 25th IEEE International Conference on Image Processing (ICIP), 1932-1936
(2) "LUCFER: A large-scale context-sensitive image dataset for deep learning of visual emotions", Pooyan Balouchian, Safaei M., Foroosh H., 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), 1645-1654
(3) "An Unsupervised Subspace Ranking Method for Continuous Emotions in Face Images", Pooyan Balouchian, Safaei M., Cao X., Foroosh H., 2019 30th British Machine Vision Conference

[2] https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api/

discrete emotion more than others. A tree-structured list of emotions was described in [109], and later featured in [110]. In [37], it is demonstrated that there exists a high agreement across members of different cultures on selecting emotional labels that fit facial expressions. These include *happiness*, *surprise*, *anger*, *disgust*, *fear* and *sadness*.

In 1980, Robert Plutchik constructed a wheel-like diagram of emotions portrayed in figure 2.1, adopted here in this work, visualizing eight basic emotions: *joy*, *trust*, *fear*, *surprise*, *sadness*, *disgust*, *anger* and *anticipation*. *Plutchik*'s three-dimensional model [11], describes the relations among emotions, which is extremely helpful in understanding how complex emotions interact and change over time, hence embedding *valence* as part of its emotion definition system. The eight sectors are designed to indicate that there are eight primary emotion dimensions. The cone's vertical dimension represents intensity - emotions intensify as they move from the outside to the center of the wheel.

### *Collecting Images from the Wild*

One of the challenges posed by architectures employing deep learning frameworks is the need for large-scale datasets. This problem has been tackled in some domains, but not others. Despite efforts made in the area of *emotion recognition*, this problem remains to be a challenge. In a valuable recent effort [12], a dataset of 23,308 images is constructed by querying Flickr and Instagram using a similar approach in [111]. In [111], Flickr is queried using Kobayashi's 16 affective categories as keywords. If an image's labels or the author's comments contain one affective category, they consider the image associated with the affect.

In another recent contribution, [44] introduces the EMOTIC database, composed of images from MSCOCO [112], Ade20k [113] and images that were manually downloaded from Google search engine. Their database contains 18,316 images, combining two different representation formats:

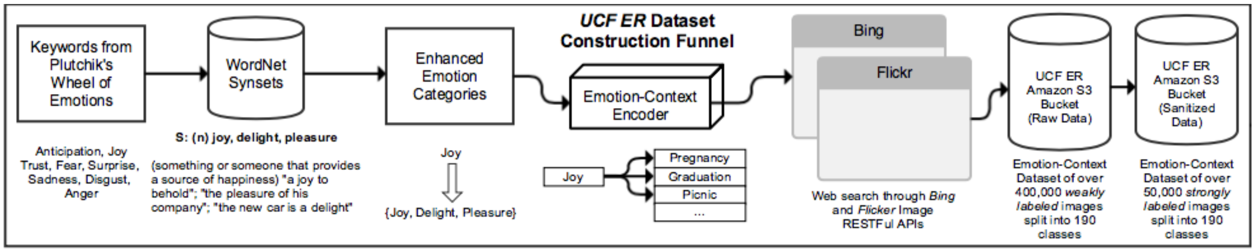

Keywords from Plutchik's Wheel of Emotions
WordNet Synsets
Enhanced Emotion Categories
UCF ER Dataset Construction Funnel
Emotion-Context Encoder
Bing
Flickr
UCF ER Amazon S3 Bucket (Raw Data)
UCF ER Amazon S3 Bucket (Sanitized Data)
Anticipation, Joy Trust, Fear, Surprise, Sadness, Disgust, Anger
S: (n) joy, delight, pleasure
(something or someone that provides a source of happiness) "a joy to behold"; "the pleasure of his company"; "the new car is a delight"
Joy
{Joy, Delight, Pleasure}
Joy
Pregnancy
Graduation
Picnic
...
Web search through Bing and Flicker Image RESTFul APIs
Emotion-Context Dataset of over 400,000 weakly labeled images split into 190 classes
Emotion-Context Dataset of over 50,000 strongly labeled images split into 190 classes

Figure 3.1: *UCF ER* dataset construction pipeline.

*i.e. discrete categories* and *continuous dimensions*.

In this work, we propose a novel approach to develop a scalable, configurable and re-usable system to collect images from the wild. We first designed a large-scale dataset collection pipeline "LDAC 1.0", used to build *UCF ER* [1]. We later developed "LDAC 2.0" [2], which is an enhanced version of "LDAC 1.0", used to build *LUCFER*. The two systems are further elaborated on in the following sections.

### *LDAC 1.0 and UCF ER*

Figure 3.1 depicts *LDAC 1.0*'s dataset construction pipeline. To build *UCF ER*, we enriched the set of emotion keywords, adopted from the *Plutchik*'s wheel of emotions, by querying *WordNet* synsets to append the synonymous words to each emotion category. Moreover, we formed a set of *contexts* each emotion category relates to, based on the frequency of different contexts each of the eight emotions appear in. For instance, the emotion *happiness* is defined in contexts, including but not limited to *graduation*, *birthday party*, *pregnancy*, *wedding ceremony* and *sport event*, among others. Using this technique, we formed a matrix of 190 *emotion-context* pairs.

Next, we crawled the Web, utilizing the RESTFul APIs offered generously by Flickr [3] and Bing

[3] Flickr API https://www.flickr.com/services/api/

[4], collecting a dataset of over 400,000 images. We then eliminated the duplicates using fdupes [114]. Moreover, we employed 10 image processing and computer vision experts to evaluate the relevancy of the images with respect to the labels; *i.e. emotions* and *contexts*. The experts were instructed to eliminate photos including text cues. We ended up with 50,000 strongly labeled, noise-reduced images labeled with *emotion* and *context*. To the best of our knowledge, this is the first emotion recognition dataset that enriches the label with *context* in addition to *emotion*. The current largest dataset, not benefiting from *context*, refers to the dataset provided generously by [12], containing 23,308 images.

A key observation is that [12] initially collects over $3M$ raw images, and finally ends up with a sanitized dataset of 23,308 images, indicating 99.3% noise in the initial dataset collected. In our work, following the approach depicted in figure 3.1, our raw dataset of 400,000 images was reduced to 50,000, resulting in 87.50% noise. [12] uses an approach similar to [111] to collect images from Flickr and Instagram, treating an image associated with the affect if an image's labels or the author's comments contain one affective category. The strategy adopted in our web crawling, depicted in 3.1, plays a crucial role in reducing noise available in our raw collected dataset, resulting in a strongly labeled dataset more than double the size of the current largest dataset of the like; *i.e.* [12].

*LDAC 2.0 and LUCFER*

The enhanced version of *LDAC 1.0*, *i.e. LDAC 2.0* [2] includes two additional components compared to *LDAC 1.0*. *LDAC 2.0* is integrated to Amazon Mechanical Turk as well as *Bing's Cognitive Services API*. The functionality of these newly added components are further discussed in this section.

[4] https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api/

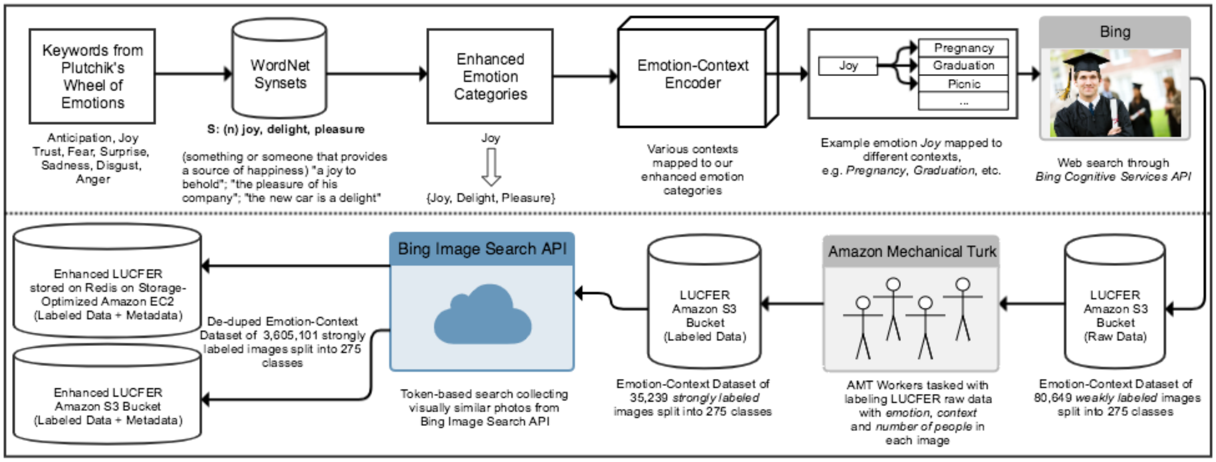
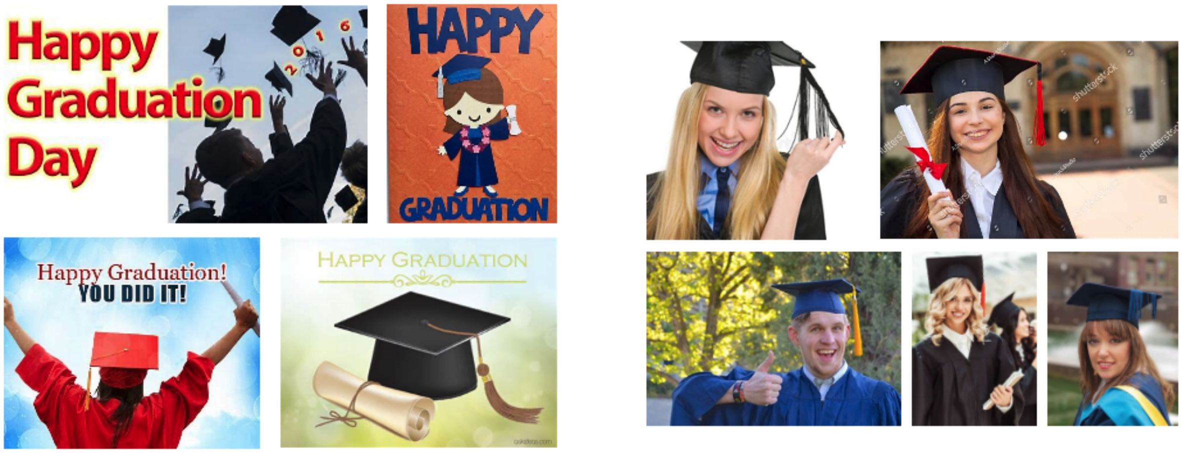

Keywords from Plutchik's Wheel of Emotions
Anticipation, Joy Trust, Fear, Surprise, Sadness, Disgust, Anger
WordNet Synsets
S: (n) joy, delight, pleasure
(something or someone that provides a source of happiness) "a joy to behold"; "the pleasure of his company"; "the new car is a delight"
Enhanced Emotion Categories
Joy
{Joy, Delight, Pleasure}
Emotion-Context Encoder
Various contexts mapped to our enhanced emotion categories
Joy
Pregnancy
Graduation
Picnic
...
Example emotion Joy mapped to different contexts, e.g. Pregnancy, Graduation, etc.
Bing
Web search through Bing Cognitive Services API
LUCFER Amazon S3 Bucket (Raw Data)
Emotion-Context Dataset of 80,649 weakly labeled images split into 275 classes
Amazon Mechanical Turk
AMT Workers tasked with labeling LUCFER raw data with emotion, context and number of people in each image
LUCFER Amazon S3 Bucket (Labeled Data)
Emotion-Context Dataset of 35,239 strongly labeled images split into 275 classes
Bing Image Search API
Token-based search collecting visually similar photos from Bing Image Search API
De-duped Emotion-Context Dataset of 3,605,101 strongly labeled images split into 275 classes
Enhanced LUCFER stored on Redis on Storage-Optimized Amazon EC2 (Labeled Data + Metadata)
Enhanced LUCFER Amazon S3 Bucket (Labeled Data + Metadata)

Figure 3.2: *LUCFER* dataset construction pipeline.

To build *LUCFER*, we start from 24 emotion categories; *i.e.* 8 basic emotion categories plus an additional set of 16 emotion categories available in the *Plutchik*'s wheel of emotions, covering *valence* [11]. Next, we queried *WordNet* [115] to extract terms synonymous to our initial list of 24 emotions at hand. We then combined 42 *contexts* with these emotions to form a 2-dimensional *emotion-context* matrix containing 275 *emotion-context* pairs. This helps us further analyze the effect of infusing *context* into the training process. Figure 3.2 further depicts the architecture of our dataset collection pipeline.

To crawl the Web for images, we used *Bing's Cognitive Services API*. To take advantage of Bing's filters including *face-only*, *include body parts* and etc., we flagged the *emotion-context* pairs with the relevant filters when applicable, in order to enforce them during search. Figure 3.3 portrays a side-by-side comparison between the same search ran "with" and "without" Bing's head & shoulder filter applied. Using this approach, we collected 80,649 images from the *wild*. Next, we employed Amazon Mechanical Turk workers to label the weakly-labeled images resulted from our initial search. AMT workers answered an array of questions on each image. The questions were designed to validate the weakly-labeled (1) *emotion* and (2) *context*, while also capturing the



(a) Keyword "happiness graduation" searched on Bing Images with no filter applied

(b) Keyword "happiness graduation" searched on Bing Images with head & shoulder filter applied

Figure 3.3: Keyword "happiness graduation" searched on Bing Images "with" and "without" *head & shoulder* filter applied

(3) number of humans in the image and (4) whether or not the image is a drawing, synthesized, cartoon-based or real. The workers flagged 43.60% of the images as correctly labeled, resulting in a validated 35,239 noise-reduced images strongly-labeled with *emotion* and *context*. Table 3.1 shows the proportion of noise vs. correctly labeled images in *Flickr-Instagram* [12], *UCF ER* [1] and *LUCFER* [2]. The noticeable difference in the noise percentage in our method as opposed to state of the art, prior to the validation step, is attributed to the (1) way we construct our *emotion-context* pair, and (2) the use of filters in our search strategy, as well as (3) the semantic sanitization we perform as a final step to sanitize our dataset. Excerpt from the noise-reduced *LUCFER* is depicted in figure 3.4. Next, we will discuss the dataset enhancement method we employed in [2].

### *Enhancing LUCFER Size*

To enhance the size of our dataset, we took advantage of *Bing*'s "visually similar images" feature available in its *Cognitive Services API*. This enabled us to query for images similar to the images

Table 3.1: Noise percentage in the raw dataset collected by different dataset collection strategies

| Dataset | Raw Size | Sanitized Size | Noise % | Context | Filter-based Search |
|---|---:|---:|---:|:---:|:---:|
| Flickr-Instagram [12] | ˜3M | 23,308 (strongly labeled) | 99.3% | No | No |
| IESN [50] | ˜21M | 1M (weakly labeled) | 95% | No | No |
| UCF ER [1] | ˜400,000 | 50,000 (strongly labeled) | 87.5% | Yes | No |
| **LUCFER** [2] | ˜80,000 | 35,239 (strongly labeled)<br>3.6M (weakly labeled) | **56.4%**<br>N/A | Yes | **Yes** |

collected during the first step of the process. It is worth mentioning that the dataset at hand during this phase includes only noise-reduced strongly labeled images labeled by AMT workers. Taking advantage of this feature, our system re-crawled the Web, collecting 8,498,660 images. Next we (i) de-duped the dataset using *fdupes* [114], and (ii) minimized the noise by comparing the labels of the AMT labeled images against the labels on the automatically captured images, eliminating those violating a predetermined semantic similarity measure. This led to creation of *LUCFER*, containing a total of 3,605,101 unique images. To avoid the *class imbalance* problem, elaborated on in section 4.4, we sub-sampled the dataset to have an equal number of images per class.

We hope that *UCF ER* and *LUCFER* serve the *Multimedia* and *Computer Vision* communities, enabling both single-modal and multi-modal methods run experiments on our datasets, specifically taking advantage of the rich metadata *LUCFER* is equipped with. Figure 3.4 depicts excerpts from *LUCFER*, portraying samples from each basic emotion category, along with the *context* embedded in the photo. This figure displays samples from the strongly labeled set validated by AMT workers as well as the visually similar images collected using *Bing's Cognitive Services API*.

### *Dataset Statistics*

In this section, statistics for both *UCF ER* and *LUCFER* are shared to enable us do a comparative analysis between our datasets vs. those of the state of the art.

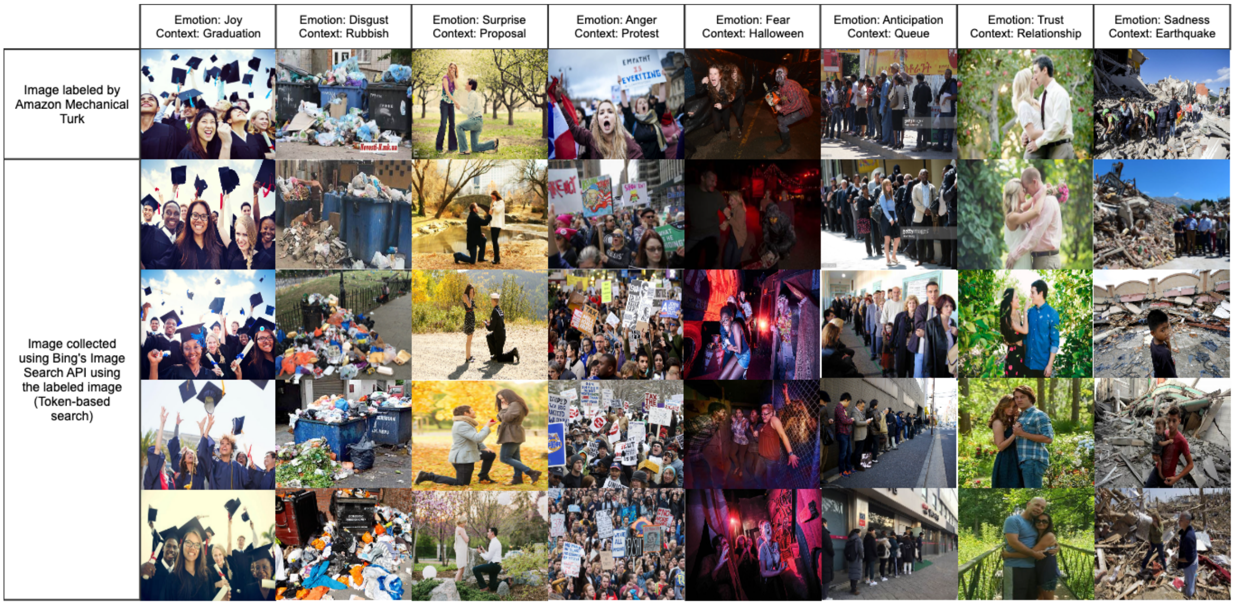


Figure 3.4: Excerpts from *LUCFER* displaying samples from the (1) strongly labeled set validated by AMT workers, and (2) images pulled using *Bing's Cognitive Services API*.

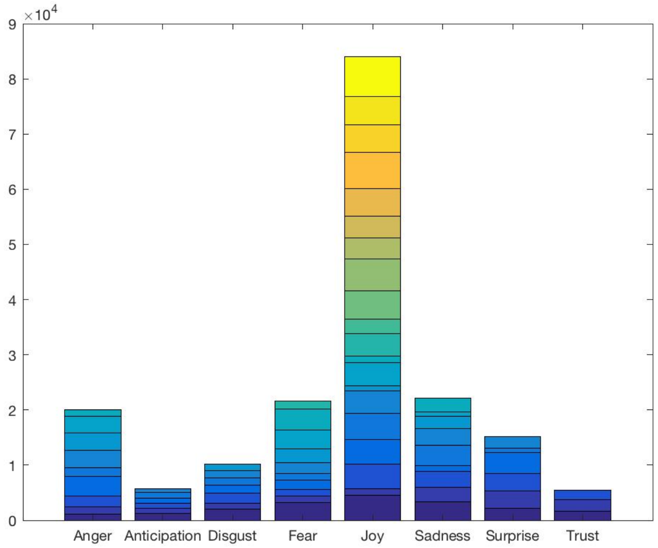


Figure 3.5: Distribution of images per *emotion* category. Stacked bars show *context* distribution across each *emotion* category.

Table 3.2: Excerpts from *UCF ER emotion-context* matrix

| | **Emotion** |
|---|---:|
| **Context** | *Joy* |
| Baby | 4,558 |
| Pregnancy | 5,074 |
| Game/Sport | 9,777 |
| Family/Friends | 9,017 |
| Gathering/Reunion | 8,089 |
| Graduation | 4,119 |
| Job/Work/Business | 4,001 |
| Outdoor | 5,064 |
| Party/Event | 5,839 |
| Peace/Relaxation | 10,413 |
| Picnic | 3,874 |
| Travel/Adventure | 7,227 |

| | **Emotion** |
|---|---:|
| **Context** | *Anger* |
| Argument/Quarrel | 2,018 |
| Protest/Demonstration | 6,370 |
| Failure | 1,639 |
| Fight | 3,108 |
| Police Encounter | 3,245 |

| | **Emotion** |
|---|---:|
| **Context** | *Fear* |
| Adventure | 2,326 |
| Police Encounter | 3,760 |
| Watching Movie | 3,507 |
| Burglary | 2,117 |

*UCF ER*

Figure 3.5 depicts the distribution of images per *emotion* category. Stacked bars show *context* distribution across each *emotion* category. As shown in this figure, the number of images across different *emotion* categories is imbalanced. *Class imbalance* is a classical problem associated to dataset sizes. We approach this problem by making adjustments to the class weights in our CNN configuration, shown in figure 3.1 to make our predictions unbiased.

Table 3.2 shows excerpts from *UCF ER*, reporting on the number of images per *emotion-context* pair.

*LUCFER*

Table 3.3 lists an array of useful statistics on *LUCFER* showing the basic emotions, synonymous emotions covering different degrees of *valence*, relevant *contexts* paired with different emotion categories, along with the number of images collected for each *emotion-context* pair. It is worth noting that even though a number of *emotion-context* pairs are semantically similar; such as {*violence-demonstration*} compared to {*outburst-demonstration*}, the dataset is eventually de-duped, leaving only unique images used in the training phase. *LUCFER* also benefits from a rich and structured set of metadata suitable for *computer vision* tasks involving *emotion recognition* ranging from still image single-modality to multi-modal approaches. *LUCFER*'s images are accompanied by captions, recognized entities (people) in the image, person's gender, bounding boxes, types of clothes the person is wearing, image resolution and dimensions, and tags among other useful data points.

Table 3.4 shows some commonly-used datasets in the area of *emotion recognition*.

Table 3.3: Statistics on *LUCFER* dataset showing (1) basic emotions from *Plutchik*'s wheel of emotions [11] and the total number of images collected for each basic emotion, (2) synonymous emotions covering different degrees of *valence*, (3) *context*s paired with each emotion set (basic emotion + context, or synonymous emotion + context), and (4) the total number of *emotion-context* pairs formed as a result of combining basic emotions and synonyms with their relevant contexts.

| Basic Emotions (# of Images) | Synonymous emotions covering different degrees of *valence* | Contexts (# of Images) | Emotion-Context Pairs |
|---|---|---|---|
| Anger (611,031) | Violence, Resentment, Outburst, Rage, Indignation, Unhappiness, Frustration, Discontent, Annoyance, Outrage, Displeasure, Animosity | Argument (66,866), Demonstration (95,125), Fight (67,651), Police (80,467), Protest (191,098), Sports (83,095), Work-related (26,729) | 62 |
| Anticipation (89,083) | Enthusiasm, Vigilance, Expectation | Pregnancy (42,160), Standing in Queue (14,422), Sports (19,178), Work-related (13,323) | 6 |
| Disgust (242,095) | Fatigue, Monotony, Lethargy, Indifference, Apathy, Boredom, Contempt, Dislike | Food (20,510), People (14,597), Relationships (4,368), Rubbish (19,272), Sports (30,912), Studying (70,159), Watching TV (31,521), Work-related (50,756) | 18 |
| Fear (196,501) | Shock, Intimidation, Dread, Stress, Anxiety, Concern, Despair, Doubt, Horror, Panic, Worry, Unease, Scare, Apprehension, Disquiet, Mistrust, Suspicion, Terror, Awe | Adventure (11,984), School Exam (60,962), Halloween (29,560), Police (50,522), Relationships (21,049), Surgery (22,424) | 25 |
| Joy (1,208,429) | Happiness, Rapture, Pleasure, Delight, Gladness, Cheer, Amusement, Serenity, Calmness, Tranquility, Euphoria, Elation, Bliss, Ecstasy, Peace | Babies (69,558), Birthday (42,334), Graduation (103,159), Group Event (178,116), Party (121,566), Picnic (100,285), Pregnancy (61,466), Romance (261,372), Sports (66,737), Traveling (69,421), Wedding (134,415) | 85 |
| Sadness (840,691) | Heartache, Melancholy, Bummer, Pensiveness, Grief, Agony, Discomfort, Mourning, Remorse, Gloom, Distress, Depression, Sorrow, Misery, Heartbreak, Anguish, Hopeless | Earthquake (295,202), Funeral (50,637), Hurricane (183,135), Natural Disaster (233,474), Police (60,164), Romance (18,079) | 53 |
| Surprise (164,051) | Disturbance, Astonishment, Bewilderment, Amazement, Interruption, Interference, Distraction | Babies (48,182), Conversations (12,668), Driving (29,689), Engagement/Proposal (17,287), Gifts (56,225) | 9 |
| Trust (278,250) | Faith, Confidence, Acceptance, Admiration, Adoration, Affection, Applause, Assurance, Praise, Appreciation, Esteem | Babies (14,159), Business Relationship (34,616), Events (18,610), Nature (54,117), Relationships (55,852), Religion (35,141), Spirituality (36,747), Sports (29,008) | 17 |
| **Total** | | | **275** |

Table 3.4: Statistics on four existing *emotion recognition* datasets, sharing similar emotion categories, showing the class imbalance problem. Excerpt from *You* et al. 2016. [12]

| Dataset | Amusement | Anger | Awe | Contetment | Disgust | Excitement | Fear | Sadness |
|---|---|---|---|---|---|---|---|---|
| IAPS-Subset | 37 | 8 | 54 | 63 | 74 | 55 | 42 | 62 |
| ArtPhoto | 101 | 77 | 102 | 70 | 70 | 105 | 115 | 166 |
| Abstract Paintings | 25 | 3 | 15 | 63 | 18 | 36 | 36 | 32 |
| You *et al.* | 4,942 | 1,266 | 3,151 | 5,374 | 1,658 | 2,963 | 1,032 | 2,922 |

Table 3.5: Total number of images per emotion category for *UCF ER* and *LUCFER*

| Dataset | Anger | Anticipation | Disgust | Fear | Joy | Sadness | Surprise | Trust | **Total** |
|---|---|---|---|---|---|---|---|---|---|
| **UCF ER** | 5,788 | 1,320 | 3,250 | 5,550 | 22,859 | 6,200 | 3,825 | 1,208 | **50,000** |
| **LUCFER** | 611,031 | 89,083 | 242,095 | 196,501 | 1,208,429 | 840,691 | 164,051 | 278,250 | **3.6M** |

# CHAPTER 4: PROPOSED METHODS

In this chapter[1], we discuss the methods we propose for *emotion recognition*. As previously discussed, we approach this problem from two different angles; *i.e. categorical* and *dimensional*. We first approach the problem from a *categorical* perspective, elaborating on (1) the CNN architectures we design, (2) fine-tuning our CNNs, and (3) the class de-contextualization step added to our CNNs. To provide empirical proof that our system is capable of delivering superior performance compared to state of the art, we benchmark our methods on the *Flickr-Instagram* [12], *UCF ER* [1] and *LUCFER* [2] datasets, and further analyze the result of experiments.

Next, we tackle the problem of *emotion recognition* from a *dimensional* perspective based on the *VAD* model of *human affect*. To this end, we develop an unsupervised subspace ranking method for continuous emotions; we formulate the problem and provide details on *rank-1 cp-decomposition*, employed as an unsupervised ranking machine here in this work. Finally, we deliver theoretical proof on our proposed unsupervised ranking method.

[1] This chapter includes excerpts from three works previously published by the author of this dissertation:
(1) "Context-Sensitive Single-Modality Image Emotion Analysis: A Unified Architecture from Dataset Construction to CNN Classification", Pooyan Balouchian, Hassan Foroosh, 2018 25th IEEE International Conference on Image Processing (ICIP), 1932-1936
(2) "LUCFER: A large-scale context-sensitive image dataset for deep learning of visual emotions", Pooyan Balouchian, Safaei M., Foroosh H., 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), 1645-1654
(3) "An Unsupervised Subspace Ranking Method for Continuous Emotions in Face Images", Pooyan Balouchian, Safaei M., Cao X., Foroosh H., 2019 30th British Machine Vision Conference

## CNN $_{\text{UCF ER}}$ - Categorical Approach to Emotion Recognition

### *CNN $_{UCF\ ER}$ fine-tuning*

In this section, we provide details regarding the approach proposed for training the CNNs used to run benchmarks on *UCF ER*. We first chose *VGGNet 16* as shown in figure 4.1. As shown in this figure, the strongly labeled images from *UCF ER* are fed into VGGNet 16. We resized the images to *224x224*, the default input size for VGGNet 16. We modified the last layer from 1,000 to 190 classes and left the rest of the fully-connected layers intact, tasked with learning a possibly non-linear function in the invariant feature space, provided by the convolutional layers. The CNN is tasked with learning a function to classify images into one of the 190 classes derived from our *emotion-context* matrix. With respect to the fact that, to the best of our knowledge, no other work has provided a *context-sensitive emotion recognition* dataset, our results inferred from such classification could not be initially compared against state of the art in a fair fashion. This urged us to add a new layer on top of the fully-connected layer to perform de-contextualization on the predicted classes by means of mapping the images from a *context-sensitive* domain onto a *context-free* domain accordingly.

We created the new de-contextualization layer by starting from the modified version of softmax derived from the previous step. The new layer gets as input a $K$-dimensional vector of probabilities, where $K$ represents the number of classes; *i.e*. 190. It then converts this representation to a $J$-dimensional vector of probabilities, where $J$ represents the number of de-contextualized classes; *i.e*. 8, with reference to figure 4.1. The experimental setup will be discussed in section 4 accordingly.

Moreover, as depicted in figure 4.2, we also designed the CNN using *ResNet 50*, resized images to *299x299*, the default input size for ResNet 50, modified the last layer from 1,000 to 190 and left

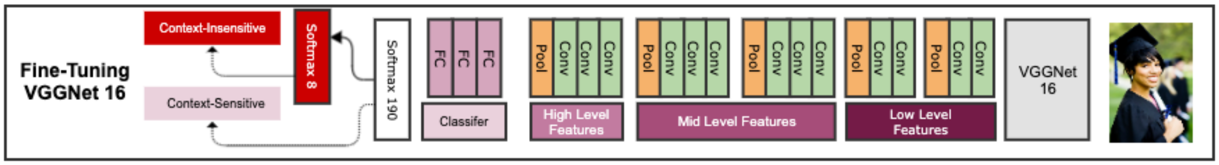


Figure 4.1: CNN Fine-tuning - VGGNet 16

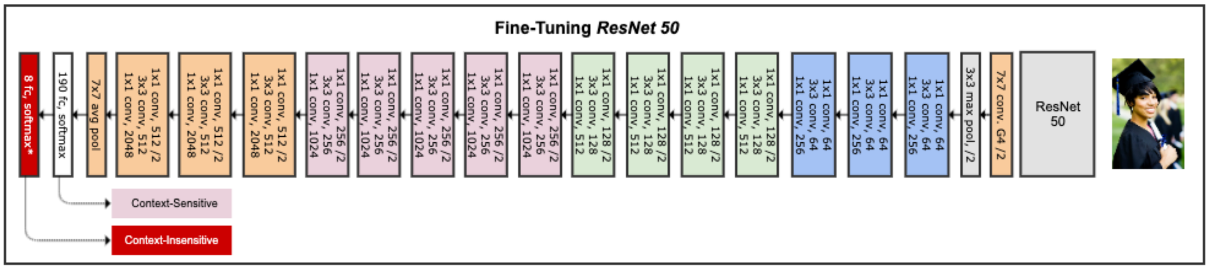


Figure 4.2: CNN Fine-tuning - ResNet 50

the rest of the fully-connected layers intact.

*Experiments on UCF ER*

We first split our dataset of 50,000 strongly labeled images into training, test and validation sets, 80%, 15% and 5% respectively. Table 4.1 shows the experimental setup. *UCF ER* contains a merged set of images collected from *Flickr* and *Bing*. This promotes an unbiased evaluation towards only one source otherwise.

Table 3.5 shows the number of instances in each emotion category. As shown in figure 3.5, the number of instances are imbalanced across different emotion categories. This is a classical problem, referred to as the *class imbalance* problem, and we approach it by making adjustments to the class weights in our CNN configuration to make the predictions unbiased.

Next, we configured the CNNs with a batch size of 8, max iteration of 312,500, resulting in 50 epochs for our experiment on the strongly labeled *UCF ER*. We first fine-tuned the pre-trained

Table 4.1: Training, test and validation experimental setup on multiple model-dataset pairs

| **Model-Dataset Pair** | Training ˜ | Testing ˜ | Validation ˜ | # Epochs |
|---|---:|---:|---:|:---:|
| VGGNet 16 - Weakly Labeled *UCF ER* | 320,000 | 60,000 | 20,000 | 50 |
| ResNet 50 - Weakly Labeled *UCF ER* | 320,000 | 60,000 | 20,000 | 50 |
| VGGNet 16 - Strongly Labeled *UCF ER* | 42,370 | 4,995 | 2,500 | 50 |
| ResNet 50 - Strongly Labeled *UCF ER* | 42,370 | 4,995 | 2,500 | 50 |
| VGGNet 16 - *Flickr-Instagram* [12] | - | - | 23,000 | - |
| ResNet 50 - *Flickr-Instagram* [12] | - | - | 23,000 | - |

VGG 16 and ResNet 50 models on the weakly labeled *UCF ER*, serving as our baseline. We then further fine-tuned them on the strongly labeled *UCF ER*. Finally, we used the latter model to evaluate the accuracy of our predictions on both *UCF ER* as well as *Flickr-Instagram* [12] datasets. These experiments are carried on using *Caffe* on 2 GPU machines; *i.e.* a GeForce GTX TITAN X with 15 GB of memory and a P2 xlarge Amazon instance with 4 virtual CPUs and 61 GiB of memory.

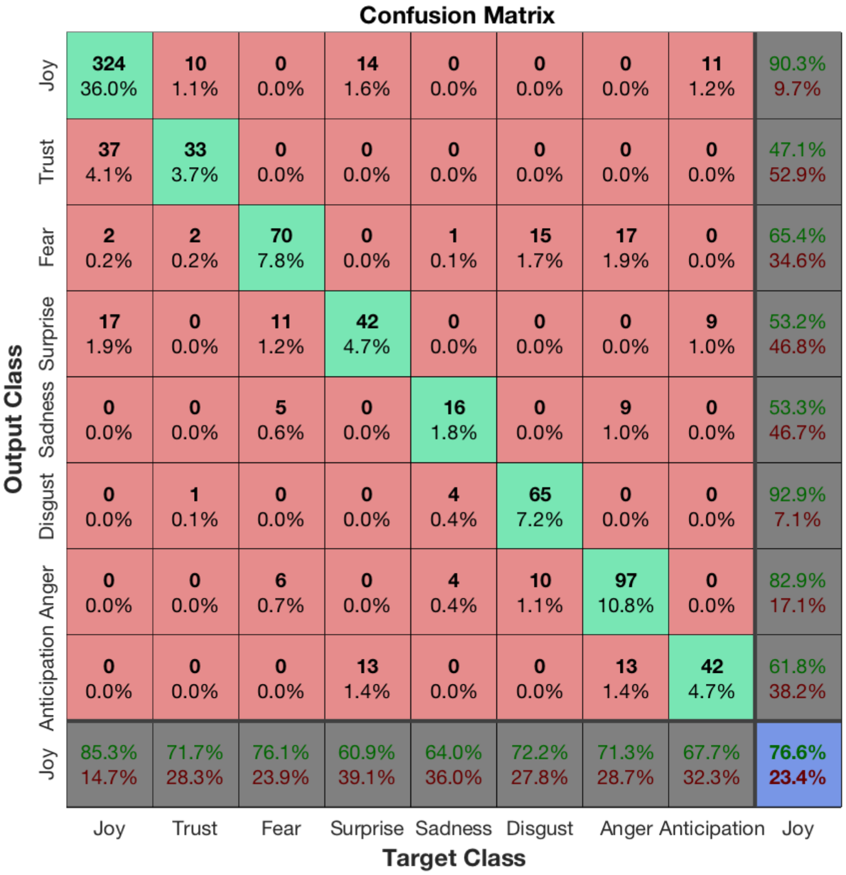
### *Discussion*

Table 4.2 reports multiple performance metrics; *i.e. overall precision*, *overall recall*, *overall F1 score* and *overall accuracy*.

With refernece to table 4.2, we also include *overall F1 score* for the eager reader considering that

Table 4.2: Performance on context-sensitive classification

| **Model-Dataset Pair** | | **Performance Metrics %** | | | |
|---|---|---|---|---|---|
| Model | Dataset | *Overall Precision* | *Overall Recall* | *Overall F1 Score* | *Overall Accuracy* |
| *Fine-tuned VGGNet 16* | *Weakly Labeled UCF ER* | 22.74 | 23.02 | 20.01 | **23.68** |
| *Fine-tuned ResNet 50* | *Weakly Labeled UCF ER* | 24.51 | 25.43 | 22.33 | **24.95** |
| *Fine-tuned VGGNet 16* | *Strongly Labeled UCF ER* | 65.26 | 65.64 | 63.97 | **71.1** |
| *Fine-tuned ResNet 50* | *Strongly Labeled UCF ER* | 69.97 | 71.53 | 69.62 | **76.6** |
| *Fine-tuned VGGNet 16* | *You et al.* [12] | 65.42 | 66.01 | 62.33 | **67.01** |
| *Fine-tuned ResNet 50* | *You et al.* [12] | 66.47 | 66.59 | 63.17 | **67.91** |



| | Joy | Trust | Fear | Surprise | Sadness | Disgust | Anger | Anticipation | Joy |
|---|---|---|---|---|---|---|---|---|---|
| Joy | 324<br>36.0% | 10<br>1.1% | 0<br>0.0% | 14<br>1.6% | 0<br>0.0% | 0<br>0.0% | 0<br>0.0% | 11<br>1.2% | 90.3%<br>9.7% |
| Trust | 37<br>4.1% | 33<br>3.7% | 0<br>0.0% | 0<br>0.0% | 0<br>0.0% | 0<br>0.0% | 0<br>0.0% | 0<br>0.0% | 47.1%<br>52.9% |
| Fear | 2<br>0.2% | 2<br>0.2% | 70<br>7.8% | 0<br>0.0% | 1<br>0.1% | 15<br>1.7% | 17<br>1.9% | 0<br>0.0% | 65.4%<br>34.6% |
| Surprise | 17<br>1.9% | 0<br>0.0% | 11<br>1.2% | 42<br>4.7% | 0<br>0.0% | 0<br>0.0% | 0<br>0.0% | 9<br>1.0% | 53.2%<br>46.8% |
| Sadness | 0<br>0.0% | 0<br>0.0% | 5<br>0.6% | 0<br>0.0% | 16<br>1.8% | 0<br>0.0% | 9<br>1.0% | 0<br>0.0% | 53.3%<br>46.7% |
| Disgust | 0<br>0.0% | 1<br>0.1% | 0<br>0.0% | 0<br>0.0% | 4<br>0.4% | 65<br>7.2% | 0<br>0.0% | 0<br>0.0% | 92.9%<br>7.1% |
| Anger | 0<br>0.0% | 0<br>0.0% | 6<br>0.7% | 0<br>0.0% | 4<br>0.4% | 10<br>1.1% | 97<br>10.8% | 0<br>0.0% | 82.9%<br>17.1% |
| Anticipation | 0<br>0.0% | 0<br>0.0% | 0<br>0.0% | 13<br>1.4% | 0<br>0.0% | 0<br>0.0% | 13<br>1.4% | 42<br>4.7% | 61.8%<br>38.2% |
| Joy | 85.3%<br>14.7% | 71.7%<br>28.3% | 76.1%<br>23.9% | 60.9%<br>39.1% | 64.0%<br>36.0% | 72.2%<br>27.8% | 71.3%<br>28.7% | 67.7%<br>32.3% | 76.6%<br>23.4% |

Figure 4.3: Confusion Matrix for *UCF ER* on ResNet 50

in cases when class distribution is imbalanced, *F1 score* offers useful information when presented alongside *accuracy* [116]. *F1 score* is calculated as 2*(*Recall* * *Precision*) / (*Recall* + *Precision*). We compute *overall F1 score*, with reference to table 4.2, by averaging individual *F1 scores* for each emotion category.

We first ran experiments on *UCF ER* weakly labeled noisy dataset of 400,000 images. *Accuracies* of 23.68% and 24.95% are reported by VGGNet 16 and ResNet 50 respectively. For this experiment, we configured the CNNs to include an additional catch-all class for noisy images. We then ran experiments on the strongly labeled *UCF ER*, resulting in *accuracies* of 71.1% and

76.6%, by VGGNet 16 and ResNet 50 respectively. The confusion matrix for this experiment is depicted in figure 4.3. This suggests a noticeable boost in performance compared to state of the art; *i.e.* accuracy of 58.3% by [12]. The injection of *context* as a dependency to the fine-tuning process from one side, and our proposed Web crawling strategy from the other side, contribute to such performance boost.

Finally, we ran another experiment on *Flickr-Instagram* [12], treating the entire dataset as the validation set since our pre-trained models on *UCF ER* are utilized. 6 categories have overlapping classes between *UCF ER* and *Flickr-Instagram* [12]. This experiment, therefore, was run on the VGGNet 16 and ResNet 50, trained on the 6 shared classes from *UCF ER*. Results yield accuracies of 67.01% and 67.91%, respectively. The relatively high accuracy on *Flickr-Instagram* dataset [12] demonstrates the efficacy of adding *context* to our unified training approach; *i.e.* higher prediction accuracy compared to the 58% reported by [12], run on their fine-tuned CNN.

### CNN $_{\text{LUCFER}}$ - Categorical Approach to Emotion Recognition

Having collected *LUCFER*, containing more than $3.6M$ images, labeled with *emotion* and *context*, we designed and implemented a CNN architecture to learn an *emotion recognition* model. We further utilized this model to empirically observe the effect of using a large-scale *context-sensitive* dataset for this task. To this end, in this section, we elaborate on the approach adopted to learn a *context-sensitive emotion recognition* classifier by providing details on the (1) fine-tuning process and the (2) dimensionality reduction method adopted in this work. The latter is accomplished via adding a new layer on top of the last fully connected layer of the network; *i.e.* fc8, tasked with mapping the *context-sensitive* onto the *context-free* domain.

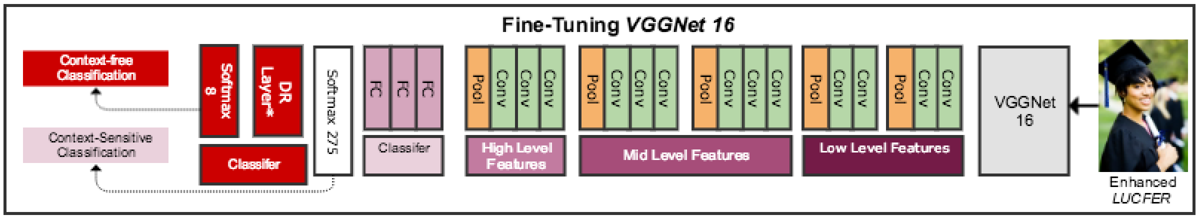


Figure 4.4: *LUCFER*'s CNN Fine-tuning. *DR stands for Dimensionality Reduction.

### *$CNN_{LUCFER}$ fine-tuning*

A classical problem in machine learning is training classifiers with *imbalanced datasets*. This leads to suboptimal classification performance, with the standard classifiers being overwhelmed by the large ones, while ignoring the small ones [117]. In many works, the ratio of the small to large classes is as drastic as 1 to 1,000 or even more. Re-sampling and combination methods are used to alleviate the *imbalance* problem at the cost of having a smaller training set.

Given the large-scale nature of *LUCFER* and the web crawling strategy adopted here, we are less challenged with this problem. Other *emotion recognition* datasets, however, suffer from this problem to some extent. These include, but are not limited to the works mentioned in table 2.1. Some of these works ([12] and [44]) make improvements compared to other ones by collecting relatively larger datasets. Table 3.4 shows the distribution of images across the emotion categories adopted by the mentioned related works. Table 3.3 shows *LUCFER*'s class distribution across eight basic emotions from *Plutchik*'s wheel of emotions [11]. The reader's attention is attracted to the significant difference in the size of each class in *LUCFER* compared to related efforts.

We first resized the images to *224x224*, *VGGNet 16*'s default input size. The CNN is tasked with learning a function to classify images into one of the 275 classes, mentioned in chapter 3. We started from the *VGGNet 16* model, pre-trained on *ImageNet*. We modified the last fully connected layer; *i.e.* fc8, from 1,000 to 275 classes and experimented with different learning rates. We left the rest of the fully-connected layers intact to learn a non-linear function in the invariant feature space, provided by the convolutional layers. Figure 4.4 depicts the CNN architecture discussed.

### *Class De-contextualization*

In the closest works to ours, [12] and [44], classification is performed on 8 and 26 emotion categories respectively. However, here, due to adding *context* into the mix, our method performs classification on 275 categories; *i.e. emotion-context* pairs. This initially prevented us from being able to perform a fair comparison against state of the art. To enable a fair comparison, we mapped our *context-sensitive* model onto a *context-free* domain; *i.e.* migrating from 275 to 8 classes. This was done via a *Dimensionality Reduction* approach, referred to as *de-contextualization* in this work, further explained in this section.

We perform *de-contextualization* by mapping high dimensional input data; *i.e.* 275, onto a low dimensional space; *i.e.* 8, such that in the target space, neighborhood points from the input domain are mapped to one another, forming a new datapoint with partially common characteristics. To better formulate the problem:

$$Let\ \ \mathcal{S} = \{\overrightarrow{I}_i, ..., \overrightarrow{I}_n\}\ \ be\ the\ set\ of\ input\ vectors$$
$$where\ \overrightarrow{I}_i \in \mathcal{D}^D\ for\ all\ values\ of\ i$$

The goal is to find a parametric function:

$$\mathcal{F}_W : \mathcal{D}^D \longrightarrow \mathcal{D}^d$$
$$where\ d\ is\ of\ a\ lower\ dimension\ compared\ to\ D$$

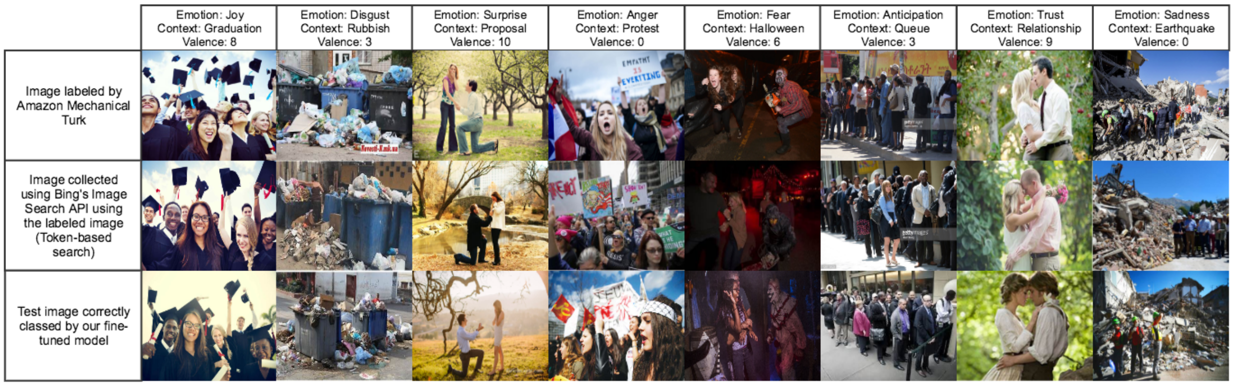


Figure 4.5: Excerpt from *LUCFER* with all eight basic emotion categories, portraying (1) images labeled by AMT workers, (2) visually similar photos collected from the wild using *Bing Image Search API* Token-based search, and (3) test images correctly classified using our model.

We implemented the *de-contextualization* in the CNN level by adding a new layer on top of the last fully connected layer; *i.e.* fc8. While this logic could be merged into the last fully connected layer, to observe the the principle of "Separation of Concerns (SoC)" [118], we chose to encapsulate the logic in a completely separate layer, hence making the newly added layer re-usable across different problem domains. This new layer gets as input a $n$-dimensional vector of probabilities, where $n$ represents the number of classes (275 *emotion-context* pairs). It then converts this representation to a $m$-dimensional vector of probabilities, where $m$ represents the number of decontextualized classes (8 basic emotion classes). We further show in the subsequent sections that the proposed *de-contextualization* makes our prediction model generalizable and transferable to other datasets.

### *Experimental Setup*

In this section, we provide details on our experimental setup. The experiments shown in table 4.3 are designed to help demonstrate the effect of injecting *context* to the unified training process (1) on the *precision* and *recall* balance, (2) boost in performance, and (3) the size of dataset required for training. All experiments are run using *Caffe* on 2 GPU machines; *i.e.* a GeForce GTX TITAN X (15 GB memory) and a P2 xlarge Amazon instance with 4 virtual CPUs (61 GiB memory).

### *Experiments on LUCFER*

First and foremost, to avoid the *class imbalance* problem, we sub-sample *LUCFER*. To this end, we get $t = min(S_1, ..., S_n)$, where $S_i$ represents the number of images in each category *i* and *n* represents the number of basic emotion categories; *i.e.* 8. Next we randomly select from each category, *t* images, forming a balanced dataset containing 712,664 images. This produces a uniform distribution over all categories. Next, we split our dataset into training and test sets, 80% and 20% respectively. We then configure the CNN with a batch size of 12, max iteration of 2,375,546, covering 50 epochs for our experiment on the sub-sampled *LUCFER*.

Table 4.3: Experimental setup on *LUCFER* training with different configurations along with the respective performance metrics.

| **Dataset** | **CNN Model** | **Training** | **Test** | **Batch Size** | **Max Iteration** | **Epochs** | **Training Type** | **Machine** | **Overall Precision** | **Overall Recall** | **Overall F1 Score** | **Overall Accuracy** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LUCFER | ImageNet | 570,000 | 142,000 | 12 | 2,375,546 | 50 | Context-Sensitive (275) | P2 xlarge Amazon Instance | 41.03 | 38.35 | 38.92 | 38.37 |
| LUCFER | ImageNet | 570,000 | 142,000 | 12 | 2,375,546 | 50 | Context-Free (8) | P2 xlarge Amazon Instance | 38.13 | 35.23 | 36.62 | 35.77 |
| LUCFER | Fine-tuned | 570,000 | 142,000 | 12 | 2,375,546 | 50 | Context-Sensitive (275) | P2 xlarge Amazon Instance | 73.50 | 72.73 | 72.75 | **73.12** |
| LUCFER | Fine-tuned | 570,000 | 142,000 | 12 | 2,375,546 | 50 | Context-Free (8) | P2 xlarge Amazon Instance | 65.31 | 61.10 | 63.66 | 69.87 |
| LUCFER | Fine-tuned | 380,000 | 94,000 | 12 | 1,583,333 | 50 | Context-Sensitive (275) | P2 xlarge Amazon Instance | 73.13 | 72.23 | 72.32 | 72.25 |
| LUCFER | Fine-tuned | 380,000 | 94,000 | 12 | 1,583,333 | 50 | Context-Free (8) | P2 xlarge Amazon Instance | 61.93 | 58.88 | 60.36 | 66.93 |
| LUCFER | Fine-tuned | 190,000 | 47,000 | 12 | 791,666 | 50 | Context-Sensitive (275) | GeForce GTX TITAN X | 71.20 | 70.08 | 70.63 | 70.98 |
| LUCFER | Fine-tuned | 190,000 | 47,000 | 12 | 791,666 | 50 | Context-Free (8) | GeForce GTX TITAN X | 58.82 | 56.72 | 57.75 | 59.78 |

### *Experiments on Flickr-Instagram Dataset*

To evaluate the generalizability of our method via transfer learning, we ran experiments on a sub-sampled version of the Flickr-Instagram dataset [12] using our model. Our model was trained on a *context-sensitive* dataset; *i.e. LUCFER* as opposed to the *context-free* Flickr-Instagram dataset [12]. Therefore, we designed AMT HITs to first contextualize 6,787 images extracted from Flickr-

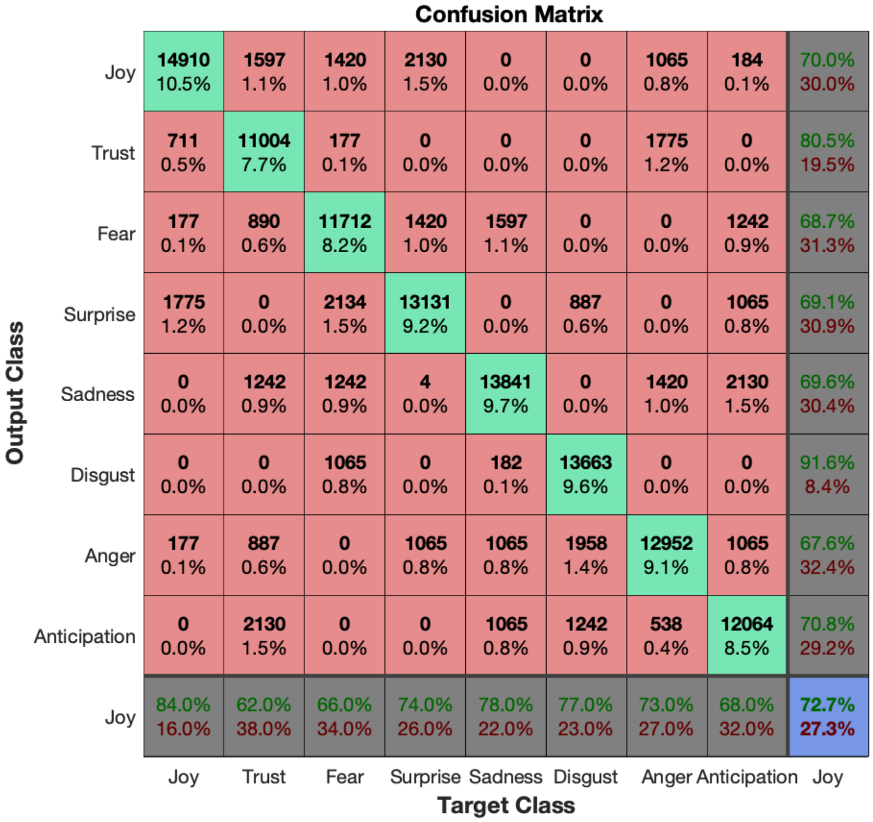
Instagram [12] covering the 4 emotion categories *Anger*, *Disgust*, *Fear* and *Sadness*. Approximately 13% of the images; *i.e*. 890 images, were labeled with contexts matching the same contexts *LUCFER* contains. We ran an experiment using our *context-sensitive* fine-tuned model (originally trained on 570,000 images) on the 890 images, yielding an accuracy of 67.5% compared to 58.3% reported in [12]. This performance boost, as high as 9.2%, indicates empirically that our model is capable of being reused on unseen datasets of similar nature. Moreover, it shows the positive effect of injecting *context* to the training process.

### *Discussion*

In this section, we discuss the result of experiments run using configurations in table 4.3. With reference to section 4, we trained an array of different configurations of the CNN, feeding the network with different inputs and loss functions. We further evaluated the models with our test sets. The experiments are performed under two main settings; i.e. *context-sensitive* and *context-free*, with the former using *LUCFER* labeled with 275 classes (*emotion-context* pairs) and the latter using *LUCFER* labeled with 8 basic emotion classes from *Plutchik*'s wheel of emotions [11].

Table 4.3 displays various performance metrics, including *accuracy*, *precision*, *recall* and *F1 score*. *F1 score* offers useful information when presented alongside *accuracy* when the class distribution is imbalanced [116]. Even though we are not challenged with this problem here as explained in section 4, F1 score is reported for the eager reader accordingly.

A key observation from table 4.3 and figure 4.7(a) is the effect of *context* and *training size* on *precision* and *recall* balance. Figure 4.7(a) helps support Eq. 4.3, showing that *precision* and *recall* becomes more balanced when *context* is used in the training phase. However, under a *context-free* training strategy, feeding the network with more training samples has a reverse effect on the *precision* and *recall* balance, moving towards a more imbalanced *precision* and *recall*.


Confusion Matrix
Output Class
Target Class

| | Joy | Trust | Fear | Surprise | Sadness | Disgust | Anger | Anticipation | Joy |
|---|---|---|---|---|---|---|---|---|---|
| Joy | 14910<br>10.5% | 1597<br>1.1% | 1420<br>1.0% | 2130<br>1.5% | 0<br>0.0% | 0<br>0.0% | 1065<br>0.8% | 184<br>0.1% | 70.0%<br>30.0% |
| Trust | 711<br>0.5% | 11004<br>7.7% | 177<br>0.1% | 0<br>0.0% | 0<br>0.0% | 0<br>0.0% | 1775<br>1.2% | 0<br>0.0% | 80.5%<br>19.5% |
| Fear | 177<br>0.1% | 890<br>0.6% | 11712<br>8.2% | 1420<br>1.0% | 1597<br>1.1% | 0<br>0.0% | 0<br>0.0% | 1242<br>0.9% | 68.7%<br>31.3% |
| Surprise | 1775<br>1.2% | 0<br>0.0% | 2134<br>1.5% | 13131<br>9.2% | 0<br>0.0% | 887<br>0.6% | 0<br>0.0% | 1065<br>0.8% | 69.1%<br>30.9% |
| Sadness | 0<br>0.0% | 1242<br>0.9% | 1242<br>0.9% | 4<br>0.0% | 13841<br>9.7% | 0<br>0.0% | 1420<br>1.0% | 2130<br>1.5% | 69.6%<br>30.4% |
| Disgust | 0<br>0.0% | 0<br>0.0% | 1065<br>0.8% | 0<br>0.0% | 182<br>0.1% | 13663<br>9.6% | 0<br>0.0% | 0<br>0.0% | 91.6%<br>8.4% |
| Anger | 177<br>0.1% | 887<br>0.6% | 0<br>0.0% | 1065<br>0.8% | 1065<br>0.8% | 1958<br>1.4% | 12952<br>9.1% | 1065<br>0.8% | 67.6%<br>32.4% |
| Anticipation | 0<br>0.0% | 2130<br>1.5% | 0<br>0.0% | 0<br>0.0% | 1065<br>0.8% | 1242<br>0.9% | 538<br>0.4% | 12064<br>8.5% | 70.8%<br>29.2% |
| Joy | 84.0%<br>16.0% | 62.0%<br>38.0% | 66.0%<br>34.0% | 74.0%<br>26.0% | 78.0%<br>22.0% | 77.0%<br>23.0% | 73.0%<br>27.0% | 68.0%<br>32.0% | 72.7%<br>27.3% |

Figure 4.6: Confusion Matrix for *LUCFER Context-Sensitive* Experiment run using 570,000 Training and 142,000 Test Images.

To better formulate this claim, the observation is:

$$|Precision_{cs} - Recall_{cs}| < |Precision_{cf} - Recall_{cf}| \tag{4.1}$$

where *cs* and *cf* stand for *context-sensitive* and *context-free* respectively.

Moreover, figure 4.7(b) depicts the effect of *context* and *training size* on the classification accuracy.

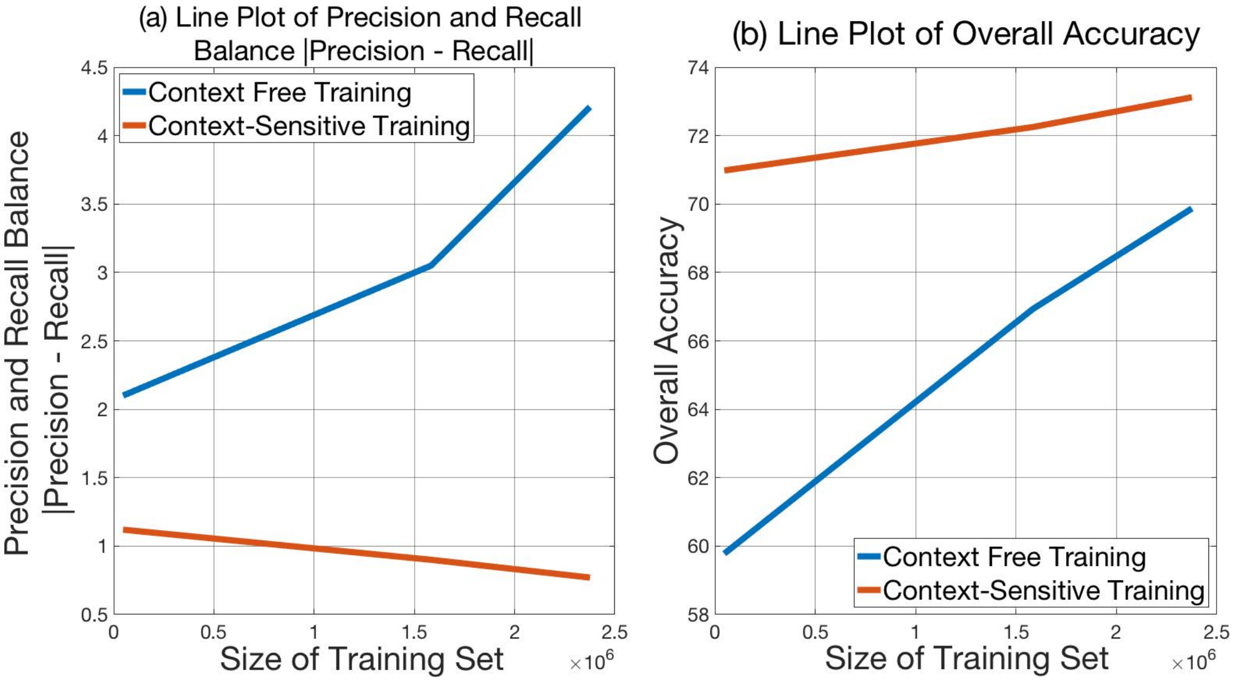


Figure 4.7: (a) Line plot showing the effect of *context* and *training size* on precision and recall balance. (b) Line plot showing the effect of *context* and *training size* on accuracy.

Under the *context-sensitive* training, less samples are required to reach a higher accuracy when compared to a *context-free* training strategy. This observation justifies our initial claim that adding *context* to the training phase alleviates the constant need to have access to large-scale datasets to some extent. This lends itself to the fact that the network tends to converge to its optimal accuracy in less iterations. Eq. 4.2 below

$$max(acc_{cf}) - min(acc_{cf}) < max(acc_{cs}) - min(acc_{cs}) \tag{4.2}$$

where acc, *cs* and *cf* stand for *accuracy*, *context-sensitive* and *context-free* respectively, yields 2.14 $<$ 10.09 when results from table 4.3 are plugged into this equation. This inequality indicates the demanding nature of *context-free* methods for more training samples.

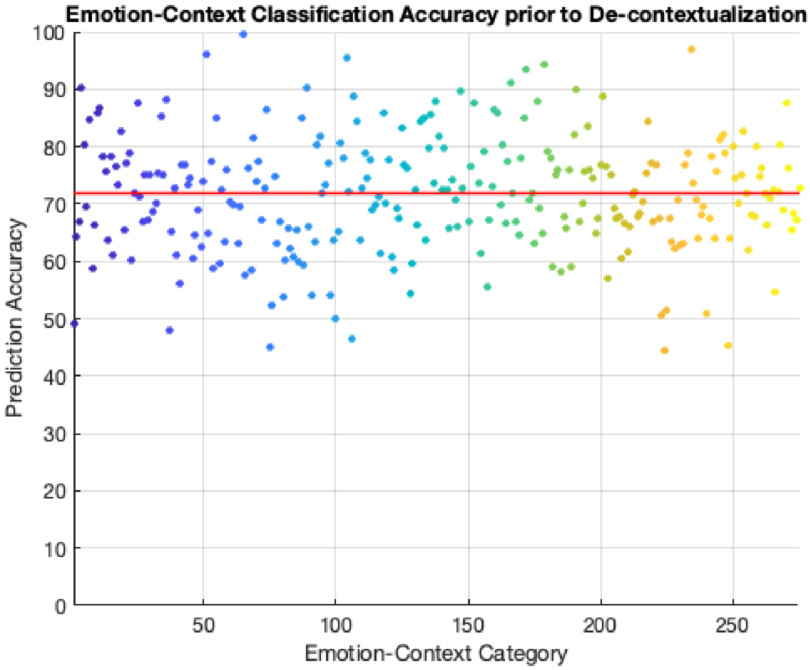


Figure 4.8: Overall accuracy scatter plot showing the overall accuracy distribution prior to de-contextualization run under the *context-sensitive* setting

To put the above observations into perspective, table 4.3 reports overall accuracies of 70.98%, 72.25% and 73.12% when fed with 791,666, 1,583,333 and 2,375,546 training images respectively under a *context-sensitive* configuration. On the contrary, when the network was trained under a *context-free* configuration, overall accuracies of 59.78%, 66.93% and 69.87% were achieved for 791,666, 1,583,333 and 2,375,546 training images respectively. Figure 4.8 portrays the overall accuracy distribution prior to applying de-contextualization run under the *context-sensitive* setting.

Table 4.3 shows our model yields superior results compared to the closest effort in *emotion recognition* in the literature by [12]. An overall accuracy of 73.12% is established using our approach as opposed to 58.3% reported by [12]. The main differences in our approach vs. that of [12]'s include: (1) our dataset size is 156 times larger, (2) our training is performed under a *context-sensitive* configuration and later de-contextualized via *dimensionality reduction*, (3) our *strongly-labeled* images under each category contain a noticeable number of visually similar images, relying on the dataset construction approach, previously explained in chapter 3. These key factors, we believe, play key roles in the boost in performance compared to state of the art. Furthermore, the confusion matrix for our *context-sensitive* experiment run on *LUCFER* using 570,000 training samples and 142,000 testing images is depicted in figure 4.6.

## Unsupervised Subspace Ranking for Continuous Emotions

Having proposed a deep learning solution to the problem of *emotion recognition* in the *categorical* domain, it is now time to approach the problem in the *unsupervised* space in the *dimensional* domain of *affective computing*. In this chapter, we first formulate the problem at hand and further provide the notations that will be used throughout the rest of this work. Next, we explain how *rank-1 cp-decomposition* can be utilized as an unsupervised ranking machine, providing proof that satisfies three important properties; *i.e. permutation invariance*, *uniqueness*, and *conformity*. We then deliver a thorough analysis on why our method works theoretically, and further apply the method to *continuous valence rank estimation* by running extensive experiments on a set of widely-used datasets used for the purpose of *valence estimation* in *affective computing*. To show the robustness of our proposed method, we perform an extensive set of ablation studies, and finally we conclude the chapter by evaluating the results of the experiments run and the ablation studies carried on.

### *Problem Formulation and Notations*

Our goal is to estimate the *valence ranking* for an unordered set of images pertaining to one *emotion category*, *e.g. joy*, *anger*, *anticipation*, etc. Let $\{\mathbf{x}_1, ..., \mathbf{x}_K\}$ denote a set of images from an *emotion* *category*, where $\mathbf{x}_k \in \mathbb{R}^{I \times J}, k = 1, ..., K$. Let also $\nu_k \in \mathcal{V}$ denote *valence*, where $\mathcal{V}$ is a continuous bounded interval in $\mathbb{R}$ describing the range of the *valence ranking*.

Assuming that $\nu_{(1)} \leq \nu_{(2)} \leq ... \leq \nu_{(K)}$, our goal is to design a ranking machine $f$ such that $f(\mathbf{x}_{(1)}) \leq f(\mathbf{x}_{(2)}) \leq ... \leq f(\mathbf{x}_{(K)})$, where the subscripts in parentheses indicate the ordered indices. Here, we do not assume availability of any training set with labeled *valence* to construct the ranking machine; *i.e*. our ranking method is unsupervised. We will prove theoretically and empirically (in the context of emotion *valence rank estimation*) that *rank-1 cp tensor decomposition* serves as an unsupervised ranking machine, if we represent the set of images as a 3-way tensor. Unlike Singular Value Decomposition (SVD) in linear algebra, there are many ways one could define tensor decomposition in multi-linear algebra. Therefore, in the next section, we start by providing a description of how *rank-r cp-decomposition* of a tensor is constructed. This will allow us to prove some important properties of *rank-1 cp-decomposition* that make it suitable for unsupervised ranking (*e.g*. for images).

We assume that *valence rank* is encoded within each set $I$, where all images in $I$ share the same emotion category with varying degrees of *valence*. We will show in section **??** why this assumption holds. In this next section, we further elaborate on the tensor formation and decomposition and how rank-1 cp decomposition is leveraged to perform *valence rank estimation*.

### *Rank-1 cp-decomposition as an Unsupervised Ranking Machine*

Let $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ be a 3-way tensor constructed by concatenating $K$ images $\{\mathbf{x}_1, ..., \mathbf{x}_K\}$ in any random order[2]. We will now prove that the *rank-1 decomposition* of $\mathcal{X}$ provides ranking information about $\mathbf{x}_k$. The rank of the tensor $\mathcal{X}$ is defined as the minimum number of rank-1 tensors that

[2] For convenience, we describe the results for 3-way tensors, although the theory holds for $n$-way tensors.

sum up to $\mathcal{X}$ [119]. A 3-way tensor is said to be rank-1 if it can be expressed as an outer product of three vectors. Although the truncation of the high-order SVD (HOSVD) of a given tensor may lead to a good low-rank approximation, it is known that this will not necessarily generate the best possible (least-squares) approximation under the given $n$-mode rank constraints [120]. Therefore, we formulate the problem of *rank-1 cp-decomposition* as:

$$\begin{aligned} \{\hat{u}, \hat{v}, \hat{w}\} &= arg \min_{\mathcal{U}} \|\mathcal{X} - \lambda u \circ v \circ w\|_F \\ \text{s.t.} \quad & \lambda = \mathcal{X}\overline{\times}_1 u \overline{\times}_2 v \overline{\times}_3 w \\ \text{and} \quad & \|u\|_F = \|v\|_F = \|w\|_F = 1, \end{aligned} \tag{4.3}$$

where $\lambda$ is a non-zero scalar, $u \in \mathbb{R}^I$, $v \in \mathbb{R}^J$, $w \in \mathbb{R}^K$, $\circ$ is the outer product, $\|\cdot\|_F$ denotes the Frobenius norm, and $\overline{\times}_i, i = 1, 2, 3$ is the multiplication between a tensor and a vector in mode-i of that tensor, whose result is also a tensor, namely,

$$\mathcal{B} = \mathcal{X}\overline{\times}_i u \Longleftrightarrow (\mathcal{B})_{jk} = \sum_{i=1}^{I} \mathcal{X}_{ijk} u_i. \tag{4.4}$$

The optimization problem in Eq. 4.3 can be solved by Generalized Rayleigh Quotient (GRQ) [121]. However, we used the alternating least squares algorithm (ALS) for optimality and rate of convergence [120, 122, 121]. The algorithm is summarized in **Algorithm** 1.

**Algorithm 1:** Rank-1 cp-decomposition

```
input  : A 3-way tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$, and an iteration termination threshold $\epsilon$
output: Three vectors $u$, $v$, and $w$ that minimize $\|\mathcal{X} - \lambda\, u \circ v \circ w\|_F$, where $u \in \mathbb{R}^I$, $v \in \mathbb{R}^J$, $w \in \mathbb{R}^K$, and $\|u\|_F = \|v\|_F = \|w\|_F = 1$
Initialize $u^{(0)}, v^{(0)}, and\ w^{(0)}$;
while $\|\mathcal{X} - \lambda^{(t)} u^{(t)} \circ v^{(t)} \circ w^{(t)}\|_F \geq \epsilon$ do
    $\widetilde{u}^{(t+1)} = \mathcal{X}\overline{\times}_2 v^{(t)} \overline{\times}_3 w^{(t)}$;
    $\widetilde{v}^{(t+1)} = \mathcal{X}\overline{\times}_1 u^{(t)} \overline{\times}_3 w^{(t)}$;
    $\widetilde{w}^{(t+1)} = \mathcal{X}\overline{\times}_1 u^{(t)} \overline{\times}_2 v^{(t)}$;
    $u^{(t+1)} = \widetilde{u}^{(t+1)} / \|\widetilde{u}^{(t+1)}\|$;
    $v^{(t+1)} = \widetilde{v}^{(t+1)} / \|\widetilde{v}^{(t+1)}\|$;
    $w^{(t+1)} = \widetilde{w}^{(t+1)} / \|\widetilde{w}^{(t+1)}\|$;
    $\lambda^{(t+1)} = \mathcal{X}\overline{\times}_1 u^{(t+1)} \overline{\times}_2 v^{(t+1)} \overline{\times}_3 w^{(t+1)}$;
end
```

Next, we prove that *rank-1 cp-decomposition* is effectively an unsupervised ranking machine. For this purpose, we prove that it satisfies three important properties, *permutation invariance*, *uniqueness*, and *conformity*[3].

**Proposition 1** *(Invariance)*

*Let* $\hat{\mathcal{X}} = \lambda\, \hat{u} \circ \hat{v} \circ \hat{w}$ *be the rank-1 decomposition of a 3-way tensor* $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ *that minimizes* $\|\mathcal{X} - \hat{\mathcal{X}}\|_F$. *We maintain that* $\hat{\mathcal{X}}_p = \lambda\, \hat{u} \circ \hat{v} \circ \hat{w}_p$ *would minimize*

$$\|\mathcal{X}_p - \hat{\mathcal{X}}_p\|_F, \tag{4.5}$$

*where* $\mathcal{X}_p = \mathcal{X} \times_1 \mathbf{P} = \mathcal{X} \times_2 \mathbf{P}$, $\mathbf{P}$ *is an arbitrary unitary transformation,* $\hat{w}_p = \mathbf{P}\hat{w}$, *and* $\times_i$, $i = 1, 2, 3$ *is a mode-*$i$ *tensor-matrix multiplication.*

What this implies is that an arbitrary permutation of the images $\mathbf{x}_1, ..., \mathbf{x}_K$ in the tensor $\mathcal{X}$ would result in the same permutation of the values in $\hat{w}$, but would not change the actual values of the components of $\hat{w}$. Also, due to the nature of the problem considered in this work, we only considered permutation along the $3^{rd}$ mode, but the proposition equally applies to all other modes.

**Proposition 2** *(Uniqueness)*

*The rank-1 cp-decomposition of a 3-way tensor* $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ *that minimizes* $\|\mathcal{X} - \hat{\mathcal{X}}\|_F$ *is unique up to a non-zero scale factor and arbitrary unitary transformation along any mode.*

Note that we eliminate scale ambiguity by explicitly enforcing $\lambda = \mathcal{X} \overline{\times}_1 u \overline{\times}_2 v \overline{\times}_3 w$

**Proposition 3** *(Conformity)*

*Let* $\hat{\mathcal{X}} = \lambda\, \hat{u} \circ \hat{v} \circ \hat{w}$ *be the rank-1 decomposition of a 3-way tensor* $\mathcal{X} = [\![\mathbf{x}_1, ..., \mathbf{x}_K]\!] \in \mathbb{R}^{I \times J \times K}$

[3] The proofs are available for review as part of the supplementary material in appendix PROOFS.

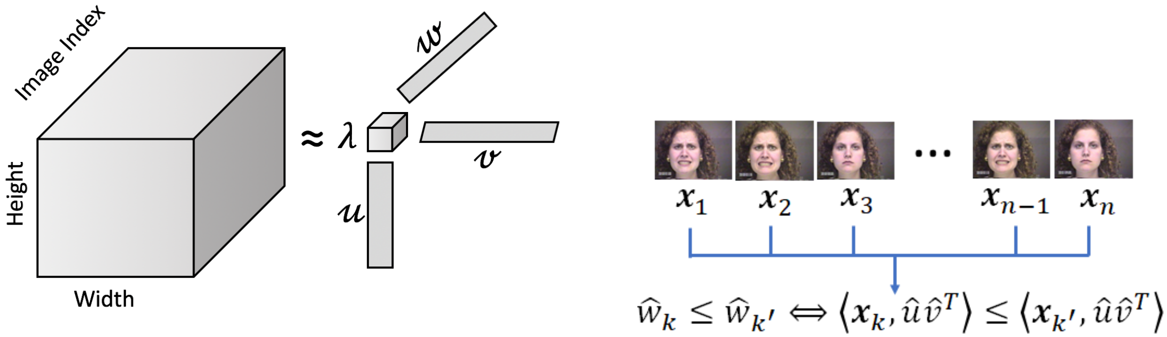


Image Index
Height
Width
$\approx \lambda$
$u$
$v$
$w$
$x_1$ $x_2$ $x_3$ … $x_{n-1}$ $x_n$
$\widehat{w}_k \leq \widehat{w}_{k'} \Leftrightarrow \langle x_k, \hat{u}\hat{v}^T \rangle \leq \langle x_{k'}, \hat{u}\hat{v}^T \rangle$

(a) Rank-1 Canonical Polyadic Decomposition. $\lambda$ is scalar, $u$, $v$ and $w$ are vectors, $w =$ *Image-indices* $\times$ 1, where 1 represents the single *emotion category* represented by the tensor; *e.g. joy*, *anger*, *surprise*, etc.

(b) Visual representation of rank-1 cp-decomposition enforcing unsupervised *valence rank estimation*.

Figure 4.9: Rank-1 Canonical Polyadic Decomposition, along with visual representation enforcing *valence rank estimation*

*that minimizes* $\|\mathcal{X} - \hat{\mathcal{X}}\|_F$*. We have* $\forall k \neq k', k, k' \in [1, ..., K]$

$$\hat{w}_k \leq \hat{w}_{k'} \quad \textit{iff} \quad \langle \mathbf{x}_k, \hat{u}\hat{v}^T \rangle \leq \langle \mathbf{x}_{k'}, \hat{u}\hat{v}^T \rangle \tag{4.6}$$

Together, these propositions prove that the cp-decomposition is a unique mapping that $\forall k \in [1, ..., K]$ measures the angle between $\mathbf{x}_k$ and the subspace of $\mathbb{R}^{I \times J \times K}$ spanned by the orthonormal basis $\{\hat{u}, \hat{v}\}$. Since the mapping is permutation invariant, it is completely unsupervised. Figure 4.9b depicts the visual representation of the concept graphically, showing intuitively why *cp-decomposition* is an unsupervised ranking machine.

## Valence Rank Estimation using Rank-1 cp-decomposition

### *Why does our method work?*

Face images are known to have subspace low-rank behavior, with primarily 4 main factors defining their low-rank representation (physiology, pose, illumination, expressions). While existing low-rank representations (eigenfaces, tensorfaces, sparse subspace representations, etc.) have their own merit, what distinguishes them is their "invariance" with respect to each of the above 4 factors. Proposition 2 states that rank-1 cp-decomposition is invariant to unitary transformations along all modes. Since most face pose changes can be modeled by rotation and mirroring, this makes rank-1 cp-decomposition resilient to pose changes. Since permutation is also unitary, it also implies that we can work with a collection of still images in any random order, or with video frames with scrambled frame order. Proposition 3 states that the elements of any of the vectors in the resulting rank-1 cp-decomposition measure the "angle" with the subspace represented by the remaining vectors. Therefore, rank-1 cp-decomposition is not measuring pairwise distances, but rather measuring the "angle" between the rank-1 representation of any slice along a given mode with a global rank-1 representation of the entire tensor in the subspace defined by remaining modes. This is an important distinction, because it implies that rank-1 cp-decomposition is robust to global illumination changes (angles do not get affected by lighting), but also the method is $O(n)$, not $O(n^2)$. This leaves rank-1 cp-decomposition primarily sensitive to the remaining two factors of facial expressions and physiological changes (facial structure/geometry, facial hair, etc.). Of course, for tests on the same subject, physiological variations are to a large extent limited or non-existent, making the method very reliable. For cross-subject tests the performance reduces, but still outperforming the state of the art by a large margin. Note that Proposition 1 states that rank-1 cp-decomposition is unique, thereby avoiding also a major ill-posedness issue.

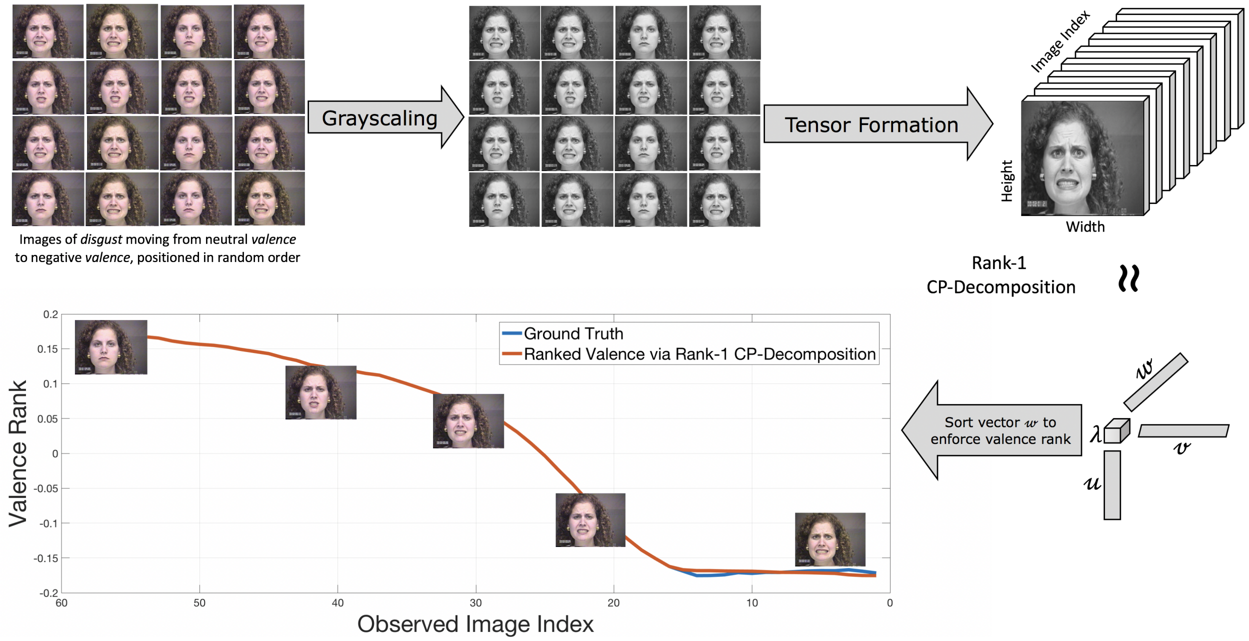


Figure 4.10: Visual representation of rank-1 cp-decomposition enforcing unsupervised *valence rank estimation.*

*Continuous Valence Rank Estimation*

Our proposed approach follows a 3-step process to perform *valence rank estimation*. The first step involves structuring images sharing the same basic *emotion category* into their own groups. The members of each group share the same emotion while representing a different but close degree of *valence* compared to the rest of the images in the group. Next, we form tensors for each group and later perform a rank-1 cp-decomposition, as explained in section 4, on each group. The last step involves estimating the *valence rank* of images in a given group, taking advantage of the ranking produced by the rank-1 cp-decomposition in section 4 above. Vector $w$ from Algorithm 1 produces a compact signature for the tensor representing the subtle changes along the *valence* dimension among the images of each group. This signature is meaningful since (1) all images within a group share the same *emotion category*, and (2) images in one group share some level of visual similarity, hence enabling each group to be used as the basis for the formation of a visual signature.

To the best of our knowledge, this is the first effort in the literature that uses a *rank-1 cp-decomposition* as an unsupervised ranking machine to estimate *valence ranking* given an unordered set of images sharing the same emotion. To validate our initial intuition proved in section 4, *i.e.* applying *rank-1 cd-decomposition* as an unsupervised ranking machine, we ran experiments on four major *emotion recognition* datasets; *i.e. CK+*, *AFEW-VA*, *SEMAINE*, and *AffectNet*, to come up with a ranking of images within each *emotion category* with respect to their levels of *valence*.

Figure 4.10 depicts the flow of actions taken starting from grayscaling of an unordered list of images pertaining to a certain *emotion category*, to tensor formation, performing the *rank-1 cp-decomposition*, followed by producing the *valence ranking*. This figure also depicts an example image sequence extracted from *CK+* dataset ranked by the *rank-1 cp-decomposition*.

We assume availability of the *emotion category* for each image group, which is a safe assumption when running experiments on any *emotion recognition* dataset. Results of experiments explored in section 4 yield promising results when compared to state of the art.

### *Datasets used in Experiments*

To evaluate our proposed method, we used four widely used *emotion recognition* datasets containing images and video frames; *i.e.* the extended Cohn-Kanade Dataset (*CK+*) [5], *AFEW-VA* [6], *SEMAINE* [7] and *AffectNet* [8]. We specifically chose this collection of datasets to cover images collected under controlled, semi-controlled and uncontrolled environments; *i.e.* (1) *CK+* and *SEMAINE*, (2) *AFEW-VA*, and (3) *AffectNet* respectively. Figure 4.11 depicts excerpts from all four datasets. The majority of methods proposed on continuous *valence estimation* are validated using

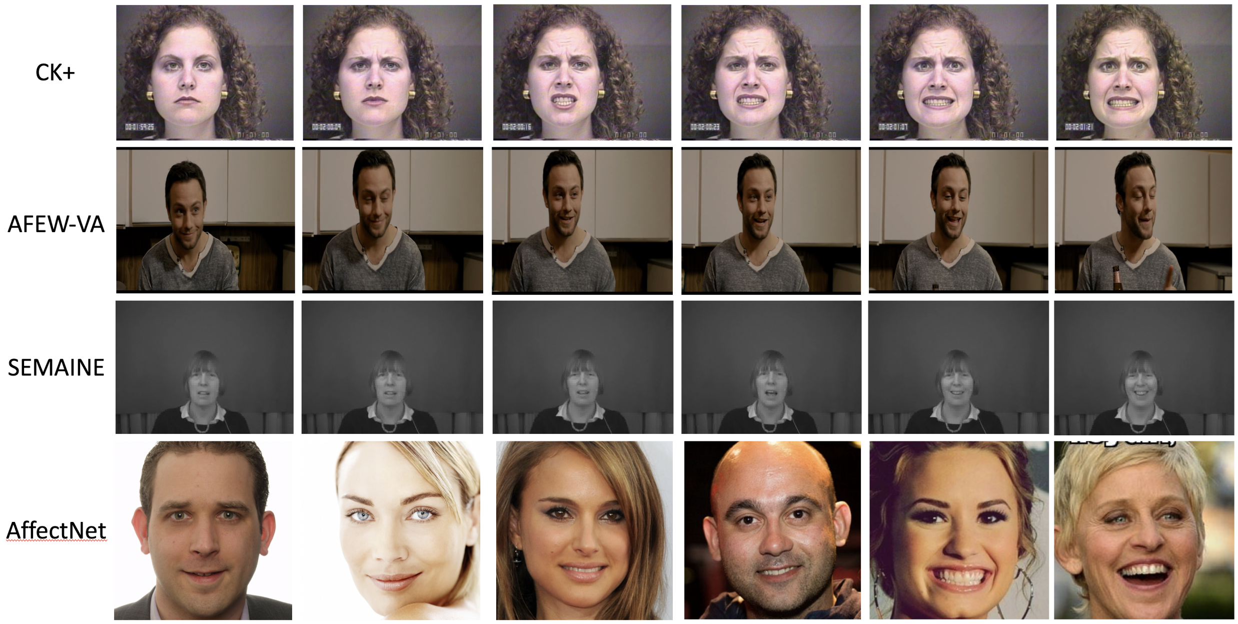


Figure 4.11: Excerpts from *CK+* [5], *AFEW-VA* [6], *SEMAINE* [7], and *AffectNet* [8] datasets showing gradual changes of *valence*. *CK+* and *SEMAINE* are collected under a controlled environment. *AFEW-VA* is extracted from feature films; *i.e.* semi-controlled environments. *AffectNet* is collected from the wild and its images are result of 15% boundary expansion of OpenCV face detector [9].

datasets captured in laboratory and under controlled settings, with a limited range of face poses and occlusions. Since state-of-the-art methods typically base their *valence estimation* on such data, it remains unclear whether these methods perform equally well on datasets collected from the wild [6]. Here we show while our method yields high accuracy when validated on datasets collected under controlled environments, it delivers high performance on datasets collected from the wild.

*CK+* includes both posed and non-posed (spontaneous) expressions and additional types of metadata. The target expression for each sequence is fully FACS coded [71]. The *CK+* distribution includes 593 sequences from 123 subjects. The image sequences vary in duration, 10 to 60 frames, and incorporate the onset (which is also the neutral frame) to peak formation of the facial expressions [5]. Each sequence is labeled with one of the seven basic emotion categories: *anger*, *contempt*, *disgust*, *fear*, *happiness*, *sadness* and *surprise*.

The *SEMAINE* database is specifically developed to address the task of achieving emotion-rich interaction with an automatic agent. It is rich in emotion and the emotions arise spontaneously in response to an activity, where activity involves a conversation with the agent. The recorded units are long enough to provide temporally extended patterns. *SEMAINE* database contains over 45 hours of material, annotated with five fully rated dimensions; *i.e*. valence, activation, power, anticipation and intensity, covering seven basic emotion categories; *i.e.* *fear*, *anger*, *happiness*, *sadness*, *disgust*, *contempt* and *amusement*. In this work, we ignore the audio information and merely focus on the video frames.

*AFEW-VA* consists of 600 videos extracted from feature films. The videos range from short (around 10 frames) to longer clips (more than 120 frames), and display various facial expressions. The clips are captured under challenging indoor/outdoor conditions such as complex cluttered backgrounds, poor illumination, large out-of-plane head rotations, variations in scale, and occlusions. In total, there are 30,000 annotated frames with per frame levels of *valence* and arousal intensities, normalized in the range of -10 to 10. Compared to *AVEC'14*, *SEMAINE* and *RECOLA*, *AFEW-VA* presents a large variation in the values of *valence* and arousal while extreme values are less frequent in most of other databases [6].

*AffectNet* is a dataset of images of facial emotions collected from the wild containing more than 1,000,000 facial images. Half of the images are strongly labeled and annotated manually for the presence of seven discrete facial expressions and the intensity of *valence* and arousal. This dataset is primarily chosen to participate in our experiments due to the fact that the images are collected from the wild under uncontrolled settings.

Figure 4.11 depicts excerpts from all four datasets.

Table 4.4: Performance comparison on AFEW-VA, SEMAINE, CK+ and AffectNet datasets

| Method | Features | AFEW-VA Dataset | | SEMAINE Dataset | | CK+ | | AffectNet | |
|---|---|---|---|---|---|---|---|---|---|
| | | Valence | | Valence | | Valence | | Valence | |
| | | Average PCC | Median PCC | Average PCC | Median PCC | Average PCC | Median PCC | Average PCC | Median PCC |
| Support Vector Machines for Regression | Norm-shape | 0.293 | - | 0.35 | - | - | - | - | - |
| Support Vector Machines for Regression | Hybrid-DCT | 0.374 | - | 0.17 | - | - | - | - | - |
| Random Forest | Norm-shape | 0.365 | - | 0.23 | - | - | - | - | - |
| Random Forest | Hybrid DCT | 0.407 | - | 0.150 | - | - | - | - | - |
| Conditional Random Field | Norm-shape | 0.244 | - | 0.275 | - | - | - | - | - |
| Conditional Random Field | Hybrid DCT | 0.137 | - | 0.173 | - | - | - | - | - |
| Deep Convolutional Neural Networks | RGB-Images | 0.17 | - | - | - | - | - | - | - |
| FT-DCNN | RGB-Images | 0.26 | - | 0.268 | - | - | - | - | - |
| Bag of Words | Hybrid-DCT | 0.124 | - | 0.166 | - | - | - | - | - |
| OR | Norm-shape | 0.25 | - | 0.18 | - | - | - | - | - |
| MKL | Shape + DCT | 0.401 | - | 0.296 | - | - | - | - | - |
| **Proposed Unsupervised Ranking** | **Feature-independent** | **0.6721** | **0.7798** | **0.7143** | **0.9245** | **0.7701** | **0.9546** | **0.6017** | **0.6671** |

## *Results and Evaluation*

To show the efficacy of our proposed method, we applied it to the task of continuous *valence rank estimation* by running experiments on the four datasets mentioned earlier in section 4. The results of these experiments are provided here in this section.

Given a ground-truth and a predicted *valence rank estimation*, here we report the performance measured using the Pearson Correlation Coefficient (PCC). PCC is a standard measure, widely used for measuring *valence* estimation accuracy. In all our experiments, we report performance in terms of PCC and compare our results against state of the art. Methods used in the literature to perform *valence estimation* mainly use *Support Vector Machine for Regression* (SVR), *Bag of Words* (BoW), *Multiple Kernel Learning* (MKL), *Conditional Random Field*, *Tree-based Random Forest* (RF), *Ordinal Regression* and *Deep learning* [6]. Our approach differs in that we tackle the problem as a fully unsupervised ranking problem performed via *rank-1 cp-decomposition* elaborated on in section 4.

It is worth mentioning that *CK+*, *SEMAINE* and *AFEW-VA* contain sequence of images with the same human subject. Conversely, *AffectNet*, a still image dataset of $1M$ images collected from the wild, includes images across different subjects per emotion category. Therefore, we first structured *AffectNet* to represent varying degrees of *valence* per emotion category so that our tensors are as

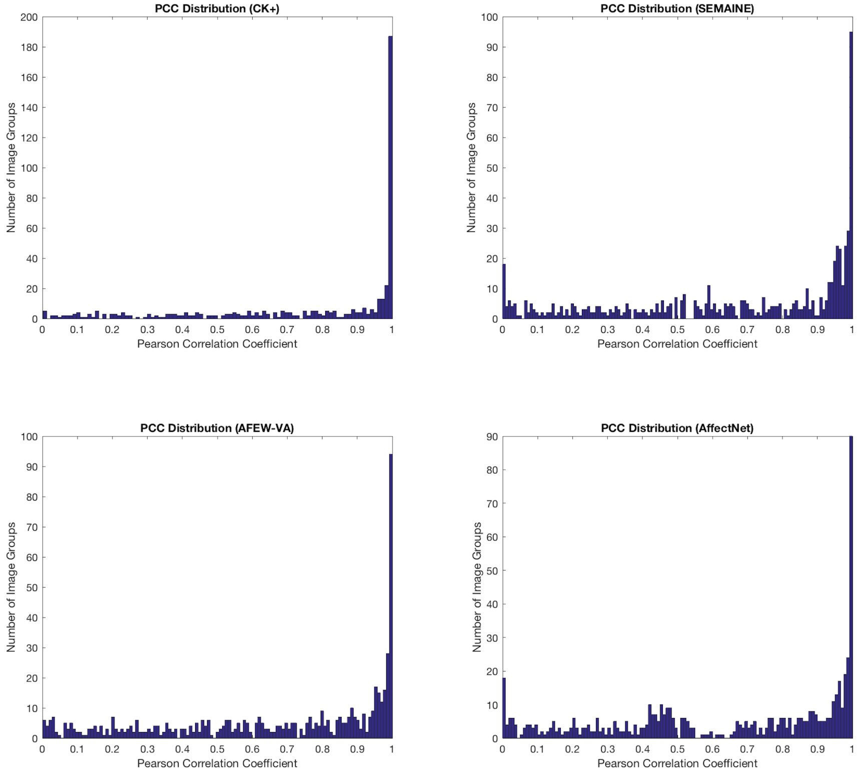


Figure 4.12: Distribution of PCC across *CK+*, *SEMAINE*, *Afew-Va*, and *AffectNet*

dense as possible, covering different degrees of *valence* in each category. The choice to run experiments on *AffectNet* stems from the fact that the majority of existing methods in the literature that report reasonable accuracy on *valence estimation* run experiments on video sequences that deal with the same subject in all frames. Here, we claim that our method performs with high performance even when run on cross-subject datasets that are not collected under laboratory settings.

Table 4.4 shows *valence estimation* results obtained by running state-of-the-art methods on *AFEW-VA* dataset. SVR performs relatively better than bag-of-words approach most probably due to the fact that the learned set of vocabulary does not offer as much information as the original shape/DCT features. Random Forest, on the other hand, perform very well, with a higher correlation than all other methods when using Hybrid-DCT. CRF does not yield good results, probably because of the short temporal dependencies in the video and due to the challenge posed by having to estimate 21 different classes. The features are not ordinal with respect to the labels and therefore the ordinal regression does not produce high performance results; *i.e.* variations in *valence* or arousal are not always directly correlated with the landmark shifts. Similarly, learning a DCNN from scratch does not provide good results. The main reason is lack of sufficient training samples needed by these networks to train a model properly. Finally, the Multiple Kernel Learning approach successfully combines shape and appearance information, producing very good results for both *valence* and arousal.

As reported in table 4.4, our unsupervised ranking method yields superior results when compared to state of the art with respect to experiments run on AFEW-VA by a large margin, improving the PCC from 0.407 reported when using Random Forest to 0.6721 using our method. Same trend is observable in experiments run on the SEMAINE dataset, with our method outperforming state of the art by a large of margin, improving the PCC from 0.35, reported when Support Vector Machines for Regression was used, to 0.7143 when our unsupervised ranking machine was applied. In other words, our method almost doubled the ranking accuracy on the SEMAINE dataset. We get the best results on *CK+* delivering a mean PCC of 0.7701 and a median of 0.9245. This is due to the fact that the images in this dataset are collected under controlled settings, hence the level of similarity between different frames in a given sequence is highlighted.

Finally, experiments run on *AffectNet* yield a PCC of 0.6017, which is relatively lower compared to our results achieved on *CK+*, *SEMAINE* and *AFEW-VA*. The reason involved is characterized by the fact that *AffectNet* emotion categories do not share the same subject, hence cp-decomposition is challenged in delivering the same performance when compared to other experiments we ran. With reference to section 4, this behavior is explained due to the fact that, for cross-subject tensors, the angle between $\mathbf{x}_k$ and the subspace of $\mathbb{R}^{I \times J \times K}$ spanned by the basis $\{\hat{u}, \hat{v}\}$ may not form an ordered set. Despite this observation, our system delivers a high performance in ranking of the *valence* under a fully unsupervised setting.

# CHAPTER 5: ABLATION STUDIES

In the area of Artificial Intelligence (AI), and specifically Machine Learning (ML), the term *ablation* refers to the removal of certain components from an AI system to monitor its performance in an attempt to deduce the level of contribution of such components in the overall performance of the system. It is especially helpful to examine a method's robustness to structural damages [123].

In this chapter, we configure and setup an array of ablation studies to measure the robustness of the unsupervised subspace ranking method proposed in chapter 4 section 4. We previously discussed the results of our experiments, elaborating on the empirical proof that our method is capable of delivering high performance under unsupervised settings. However, the experiments do not unilaterally serve as proof for our method's robustness. Hence we designed, implemented and analyzed four different ablation studies to unveil the robustness of our method accordingly.

## Ablation Study 1: Accuracy as a function of dataset size and density

In this ablation study, we investigated the accuracy of our ranking method as a function of dataset size. To do this, we split each image group into 10 different sizes and applied our method to each subset separately. All of the datasets previously used in our experiments, shared a common and crucial feature; *i.e*. the datasets are dense with respect to covering varying degrees of *valence*. "Dense" in this context refers to having a uniform distribution of varying degrees of *valence* spread across each image group, as opposed to sparse, which represents datasets having widely spaced intervals. In this study, however, we intentionally distort the density and size of each image group to further monitor the accuracy yielded by our method with respect to the dataset size and density.

We start from the full dataset and reduce the number of images in ten steps by randomly selecting

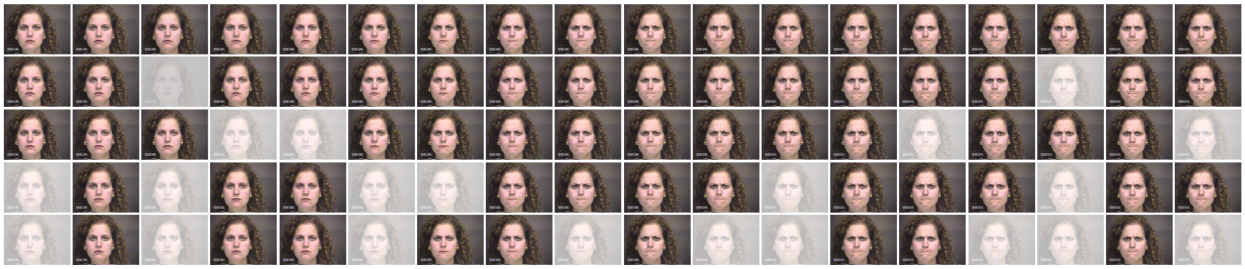


Figure 5.1: Sample set displaying different degrees of *valence* for emotion *anger* (excerpts from **CK+** dataset). Figure depicts five of the ten steps, with each step randomly removing 10% of images to (1) reduce the size, and (2) distort the density of the subset. Blurred images refer to the images randomly selected and removed in each step.

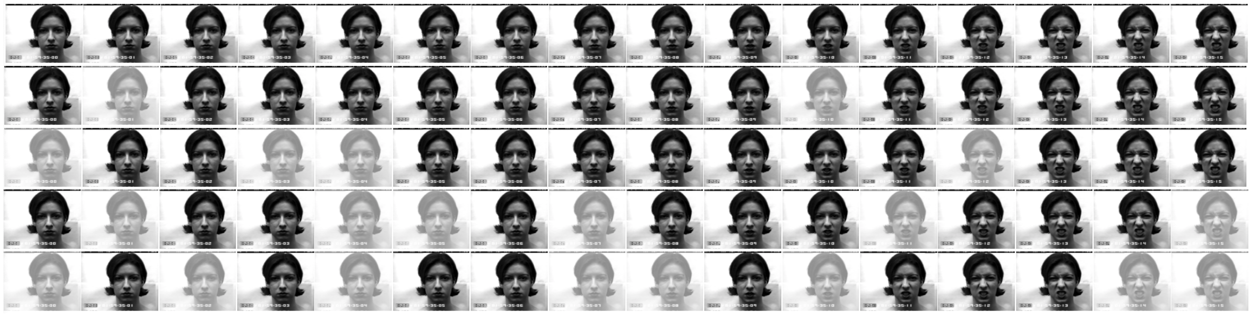
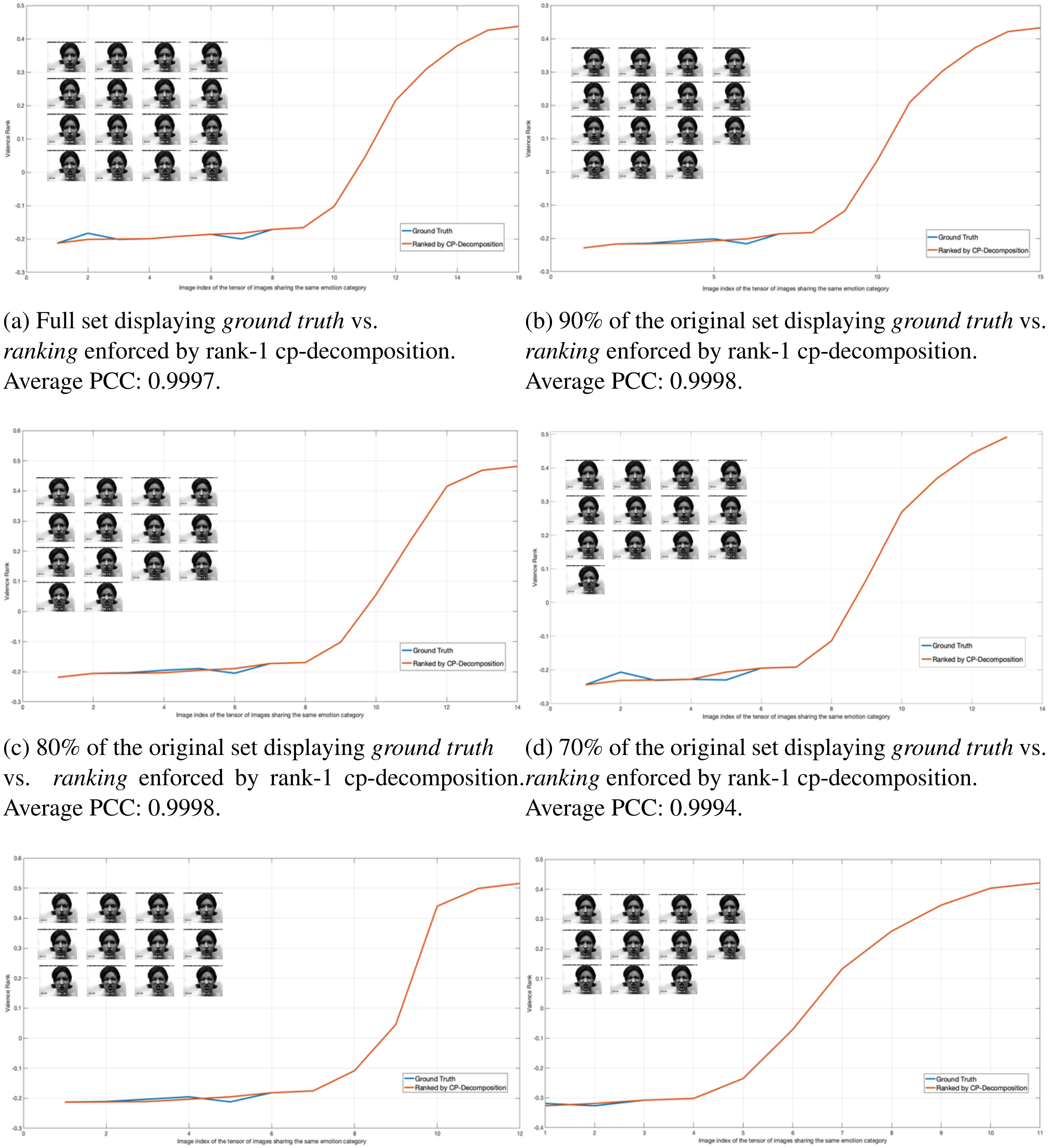


Figure 5.2: Sample set displaying different degrees of *valence* for emotion *disgust* (excerpts from **CK+** dataset). Figure depicts five of the ten steps, with each step randomly removing 10% of images to (1) reduce the size, and (2) distort the density of the subset. Blurred images refer to the images randomly selected and removed in each step.

and further removing 10% of the images in each step; *i.e*. 100%, 90%, ..., 10% of the original size accordingly. We then measure the ranking accuracy after each step of the size reduction. Considering that our sub-datasets are generated by means of a random selection process, we run this ablation study in fifty independent trials and report the average accuracy accordingly.

Figures 5.1 and 5.2 depict "dense" sample sets portraying varying degrees of *valence* for emotions *anger* and *disgust* respectively. The figures show five of the ten steps of the ablation study, with each step randomly selecting and further removing 10%, 20%, 30%, ... 90% of the images from



(a) Full set displaying *ground truth* vs. *ranking* enforced by rank-1 cp-decomposition. Average PCC: 0.9997.

(b) 90% of the original set displaying *ground truth* vs. *ranking* enforced by rank-1 cp-decomposition. Average PCC: 0.9998.

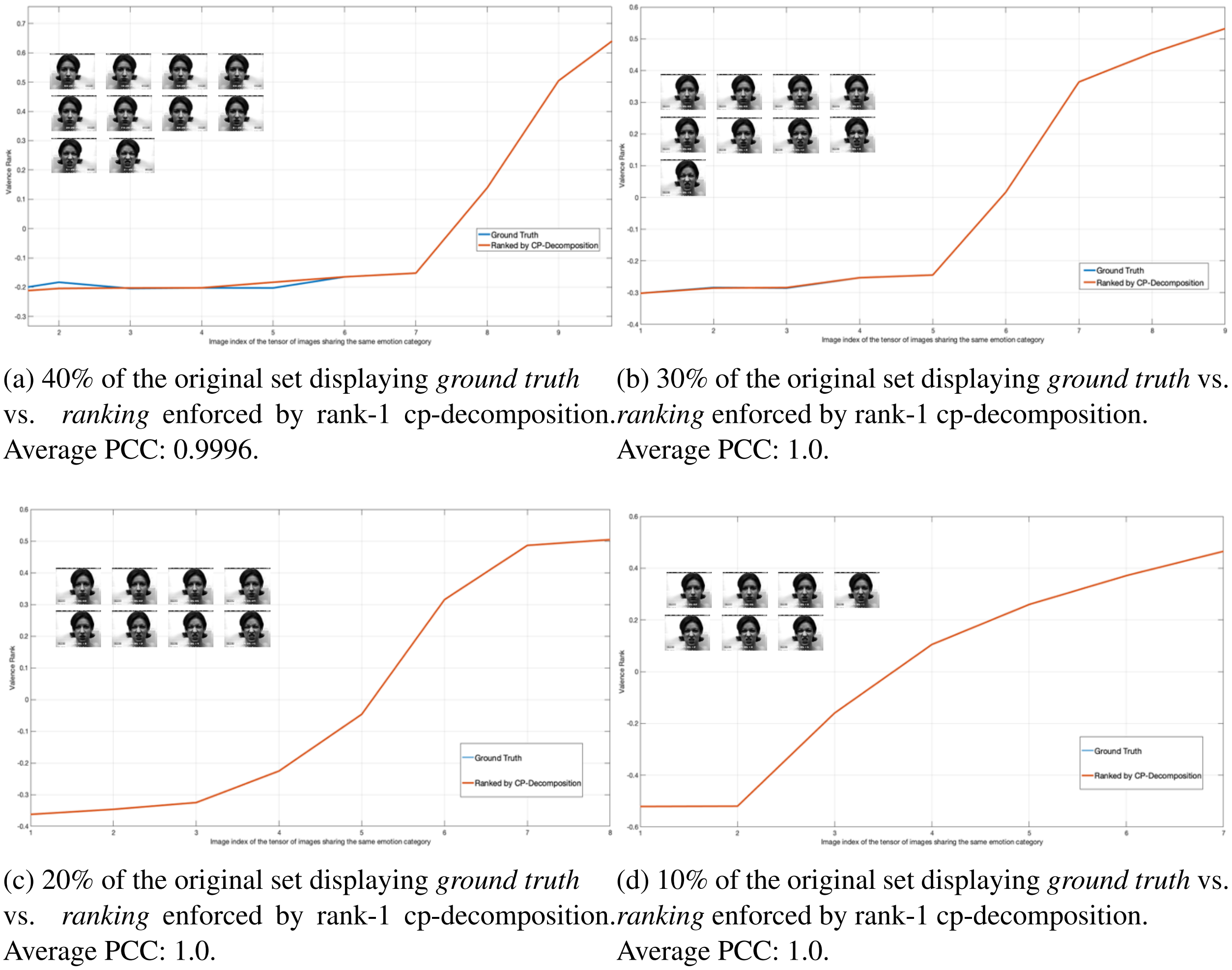
(c) 80% of the original set displaying *ground truth* vs. *ranking* enforced by rank-1 cp-decomposition. Average PCC: 0.9998.

(d) 70% of the original set displaying *ground truth* vs. *ranking* enforced by rank-1 cp-decomposition. Average PCC: 0.9994.

(e) 60% of the original set displaying *ground truth* vs. *ranking* enforced by rank-1 cp-decomposition. Average PCC: 0.9998.

(f) 50% of the original set displaying *ground truth* vs. *ranking* enforced by rank-1 cp-decomposition. Average PCC: 0.9999.

Figure 5.3: Ablation Study 1: Accuracy as a function of dataset size and density. Average PCC reported is based on fifty independent trials.



(a) 40% of the original set displaying *ground truth* vs. *ranking* enforced by rank-1 cp-decomposition. Average PCC: 0.9996.

(b) 30% of the original set displaying *ground truth* vs. *ranking* enforced by rank-1 cp-decomposition. Average PCC: 1.0.

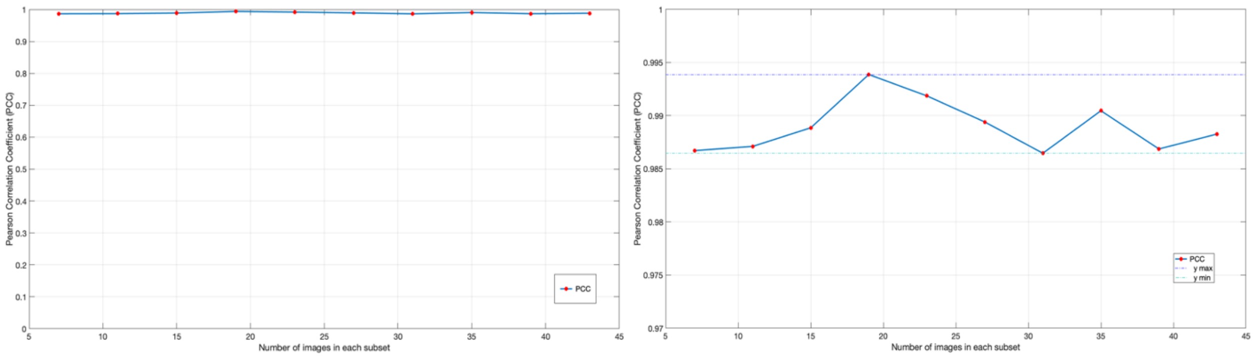
(c) 20% of the original set displaying *ground truth* vs. *ranking* enforced by rank-1 cp-decomposition. Average PCC: 1.0.

(d) 10% of the original set displaying *ground truth* vs. *ranking* enforced by rank-1 cp-decomposition. Average PCC: 1.0.

Figure 5.4: Ablation Study 1: Accuracy as a function of dataset size and density. Average PCC reported is based on fifty independent trials.

the original set. Our ranking method was further applied to each subset to measure the robustness of the method with respect to size and density of the subset. Considering that the images are removed randomly in each step, in order to prevent reporting coincidental behavior due to such randomization, we performed this ablation study in fifty independent trials and further report the average accuracy accordingly.

Figures 5.3 and 5.4 depict the ground truth vs. the *valence ranking* generated by our method for



(a) Effect of number of images on accuracy. Y axis shows the full PCC range.

(b) Effect of number of images on accuracy. Y axis zooming in for higher visibility.

Figure 5.5: Effect of the number of images on PCC. Reported PCC represents Average PCC ran in fifty independent trials.

each subset of the sample dataset displayed in figure 5.2. As observed in these figures, our method delivers highly similar rankings in different runs of the method benchmarked on different dataset sizes. This empirically proves that our ranking method is resilient to dataset size and density, where "density" refers to the distribution of varying degrees of *valence* across the test dataset. It is therefore safe to claim that our method is not challenged by small-scale dataset sizes or by datasets that do not have a full *valence* coverage pertaining to the emotion category representing the image group subject of analysis.

Figure 5.5 portrays the Pearson Correlation Coefficient (PCC) achieved in different steps of the ablation study, starting from a subset of 10% of the original set (displayed in figure 5.2) to the full dataset size. As previously claimed, this plot serves as empirical proof that our method delivers highly similar PCCs when run on datasets of varying sizes and densities. The minimum and maximum PCC is displayed as 0.9865 and 0.9939 respectively, resulting in a range amounting to 0.007, ensuring a negligible difference in PCC when run on datasets of varying size and density.

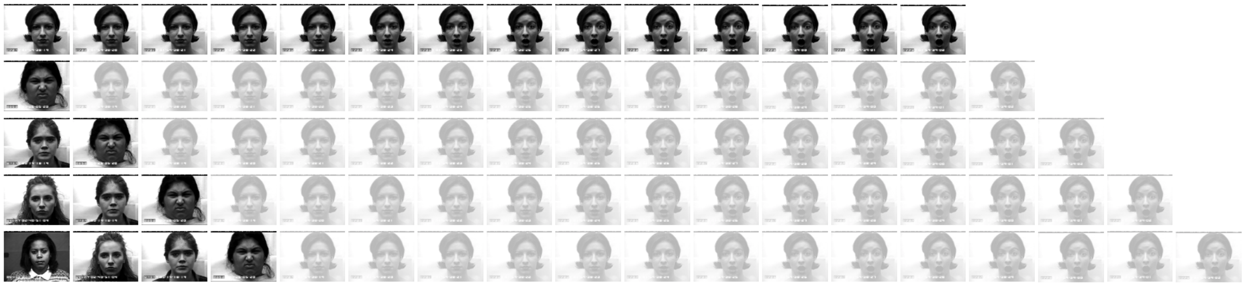


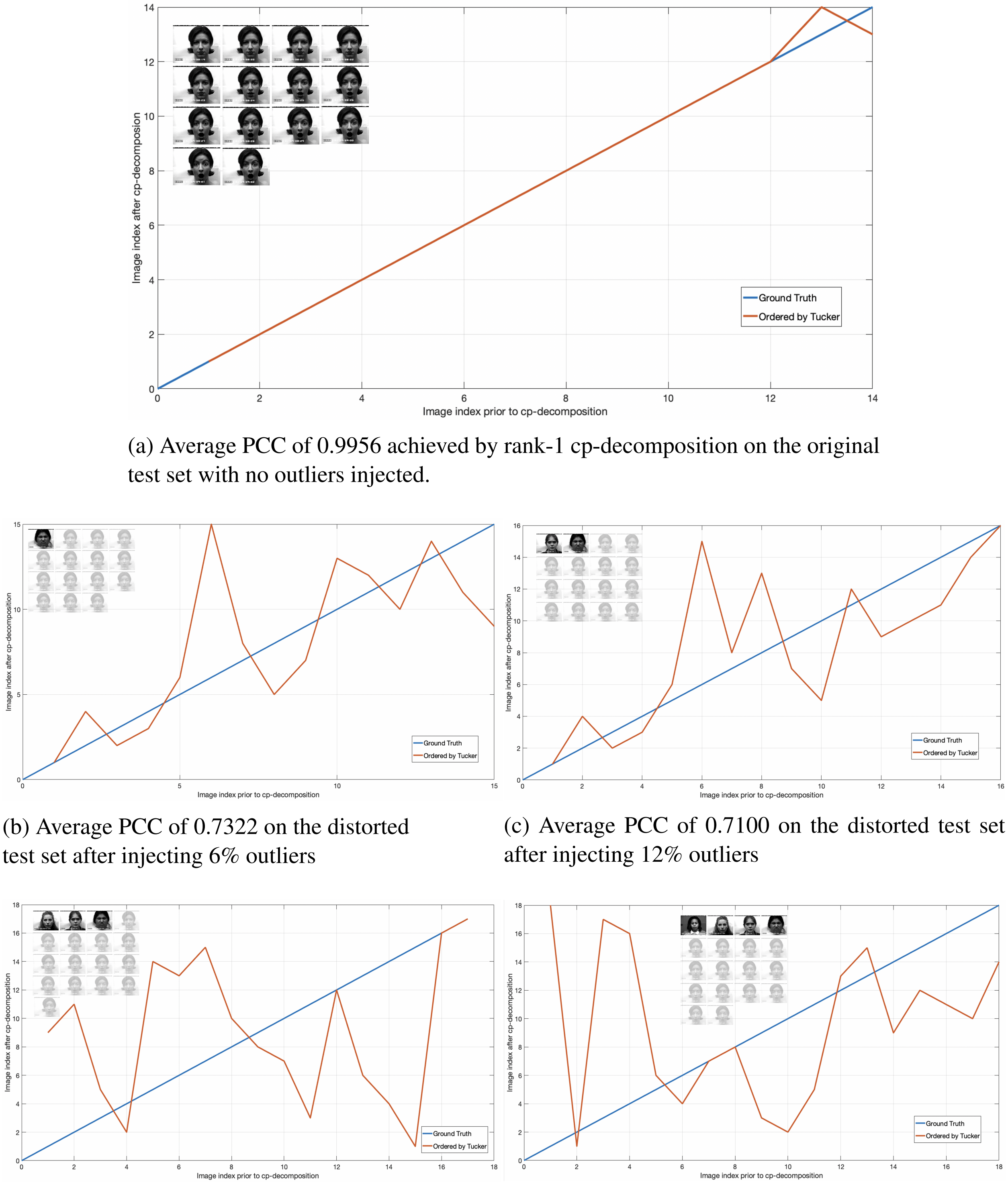
Figure 5.6: Sample set displaying different degrees of *valence* for emotion *surprise* (excerpts from **CK+** dataset). Each row displays a new outlier injected to the original set.

Ablation Study 2: Accuracy as a function of number of outliers (breakdown point)

In statistics, the “breakdown point” refers to the smallest amount of contamination, causing an estimator to become useless [124]. In machine learning paradigms and within the scope of an individual system, “fault tolerance” is achieved by encountering unexpected conditions and equipping the system such that it can cope with such unexpected events. This property of fault-tolerant systems aims for self-stabilization in an attempt for the system to converge to an error-free state [125].

In this study, we investigated the accuracy of our ranking method as a function of number of outliers injected into the test dataset. An “outlier” in this context refers to an image from a different emotion category than the one shared across all other images in the test tensor. As explained in chapter 4 section 4, we made an assumption that the images in each image group pertains to the same emotion category. This assumption is a safe one as all image datasets used in the literature include the emotion category as the main label. This ablation study, however, is designed to observe how our method would be challenged if this assumption was violated.

We start by injecting outliers to our test tensor one at a time and further applying our ranking method to derive a ranking of the emotions based on their varying degrees of *valence*. The PCC



(a) Average PCC of 0.9956 achieved by rank-1 cp-decomposition on the original test set with no outliers injected.

(b) Average PCC of 0.7322 on the distorted test set after injecting 6% outliers

(c) Average PCC of 0.7100 on the distorted test set after injecting 12% outliers

(d) PCC of 0.0715 on the distorted test set after injecting 17% outliers

(e) Average PCC of 0.0575 on the distorted test set after injecting 22% outliers

Figure 5.7: Ablation Study 2 ran on the original test set and on distorted set with up to 4 outliers injected.

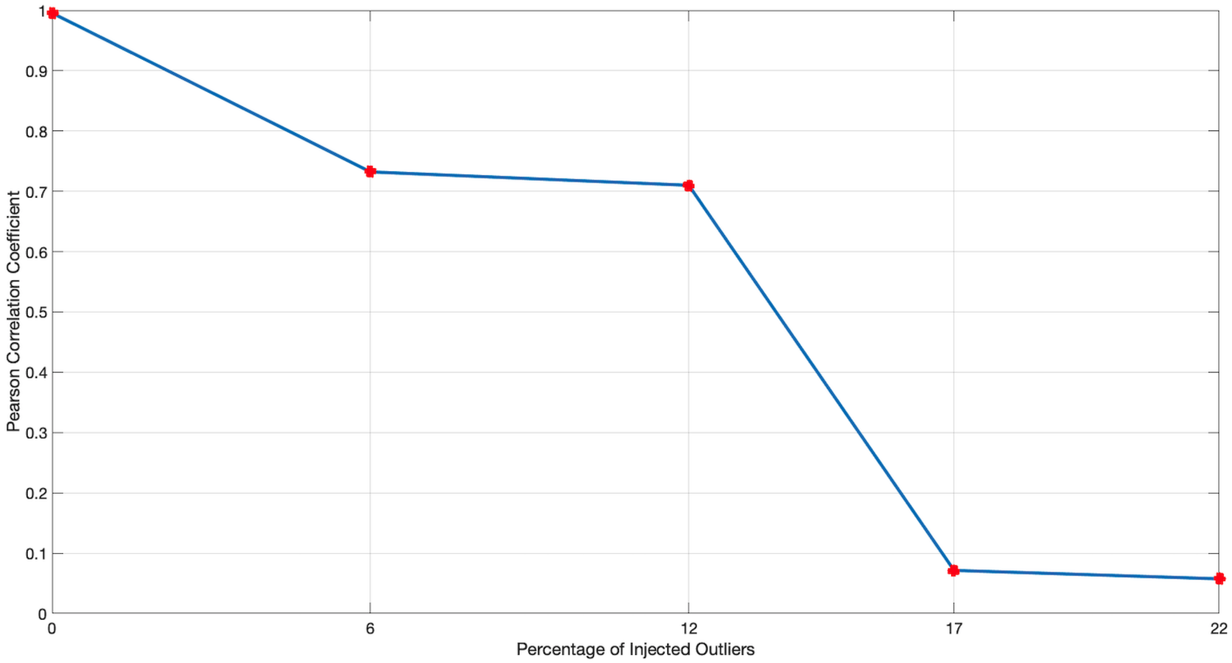


Figure 5.8: Effect of percentage of injected outliers on accuracy.

achieved on the original tensor serves as our baseline in this ablation study. We record the PCC after injecting each outlier to further analyze the fault tolerance of our method.

Figure 5.6 displays a test dataset used in our study. This figure depicts the original sample set, followed by the distorted sets, having injected outliers to the original set up to four outliers (*i.e.* 22% outliers). Figure 5.7 shows the ranking PCC achieved on the original set vs. the distorted sets including outliers. As portrayed in the figure, the system generates lower, but competitive accuracy up to 12% outliers. However, upon the injection of higher percentage of outliers, accuracy drops considerably.

Figure 5.8 depicts the plot portraying the effect of the percentage of injected outliers on accuracy. Even though the performance of our method noticeably drops after injecting more than an average of 12% outliers, our system is well equipped to accomplish the task of self-stabilization.

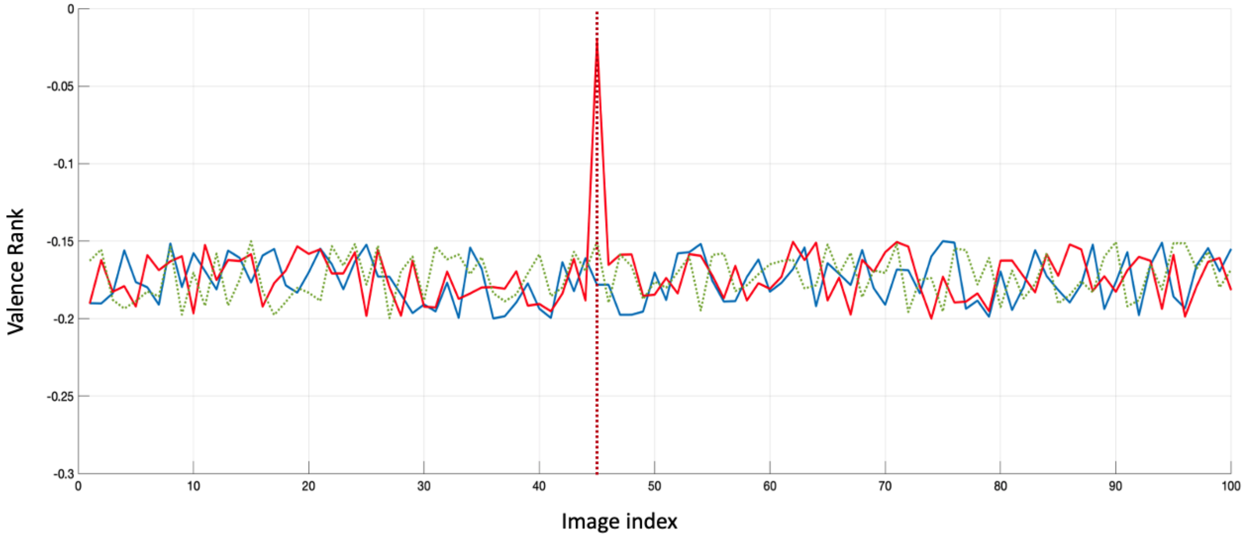


Figure 5.9: Blue line: Rank-1 cp-decomposition run on a dataset of images sharing the same emotion category
Dotted green line: Rank-1 cp-decomposition run on the same dataset after injecting a new visually similar image from the same emotion category.
Red line: Rank-1 cp-decomposition after injecting an outlier from a different emotion category.
Plot portrays visually how an outlier is detected using rank-1 cp-decomposition.

"Self-stabilization" in this context is defined as automatic detection of and auto-removal of the outliers and further re-generating the ranking accordingly once the set is free of the injected noise. With reference to our ranking method, elaborated in chapter 4 section 4, rank-1 cp-decomposition measures the distance between each image in the tensor to the ensemble representation of the remaining modes of the tensor. Figure 5.9 helps visualize the automatic outlier detection our method is equipped with. In this figure, the outlier is easily detectable where a spike is observed on the plot, representing a longer distance to the group representation of the decomposed tensor.

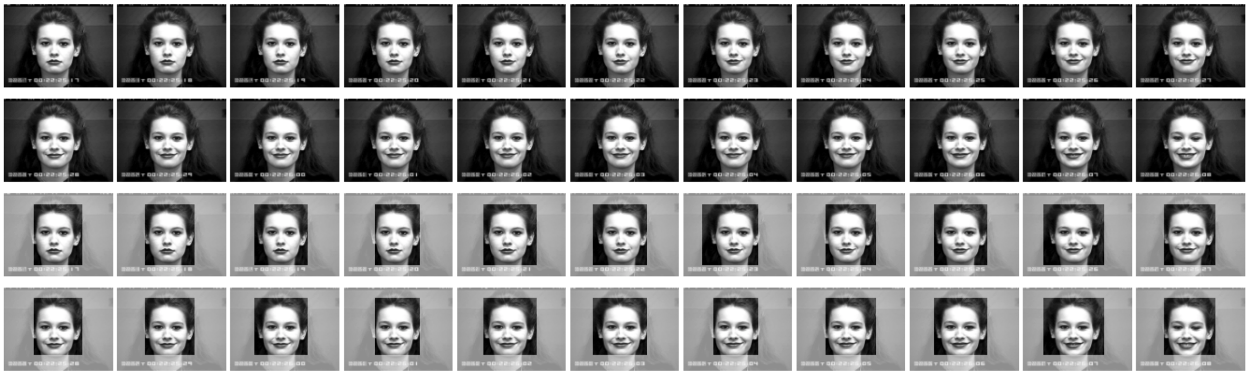


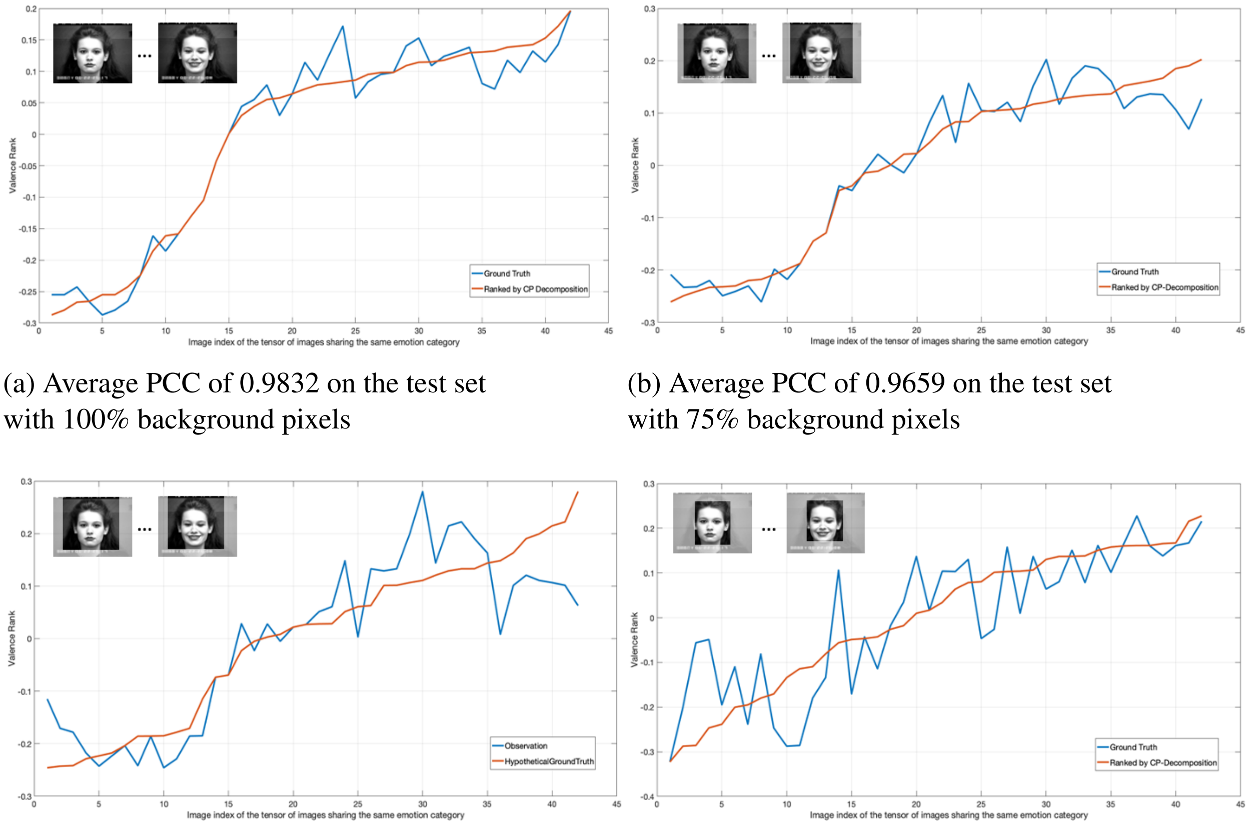
Figure 5.10: Sample set displaying images with the full background pixels included (first two rows), and the same set with background pixels removed (last two rows). Viola-Jones algorithm [10] used for face detection.

### Ablation Study 3: Accuracy as a function of percentage of foreground vs. background pixels

In this study, we further investigate the effect of the percentage of foreground vs. background pixels on accuracy. For this study, we implemented a pipeline to first detect faces, using the Viola-Jones algorithm [10]. Next, we continued by adding 25% of the background in an incremental fashion in four steps to gradually get the original image with the full background. In each step of this ablation study, we computed the PCC generated by our ranking method to monitor the accuracy as a function of the percentage of foreground vs. background pixels.

We performed this ablation study on **CK+** (Sample displayed in figure 5.10) and **AFEW** (Sample displayed in figure 5.16), with the former representing images taken under controlled environments with constant backgrounds, and the latter collected under semi-controlled environments with changing backgrounds. We specifically chose these datasets to better investigate the role of constant vs. changing background pixels on the ranking accuracy.

Figure 5.11 further portrays the Average PCC generated by our method in each step of the ablation study ran on **CK+**. Figure 5.15 portrays the plot showing the effect of the percentage of



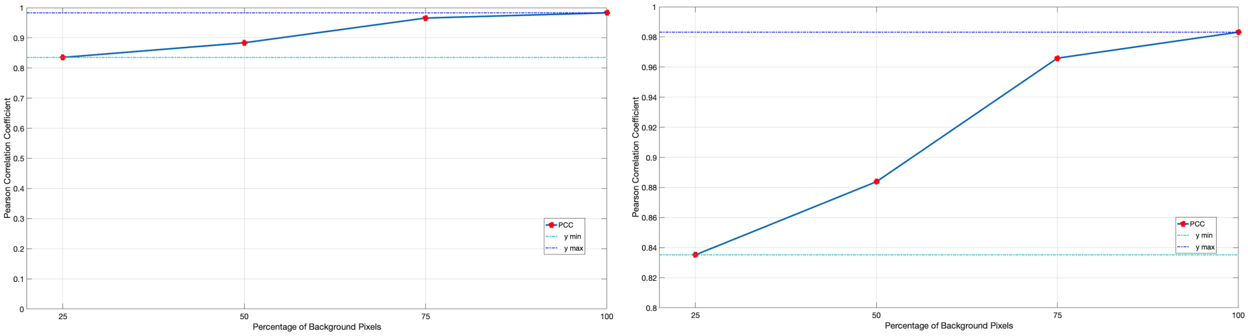
(a) Average PCC of 0.9832 on the test set with 100% background pixels

(b) Average PCC of 0.9659 on the test set with 75% background pixels

(c) Average PCC of 0.8838 on the test set with 50% background pixels

(d) Average PCC of 0.8352 on the test set with 25% background pixels

Figure 5.11: Ablation Study 3 ran on CK+ with 100%, 75%, 50% and 25% background pixels respectively. CK+ is collected under controlled environments with constant background. Average PCC is computed based on all experiments run on CK+ dataset.



(a) Accuracy plot in the scale of 0 to 1

(b) Accuracy plot rescaled for higher visibility

Figure 5.12: Ablation Study 3: Effect of percentage of foreground vs. background pixels on accuracy on **CK+** with constant background pixels.

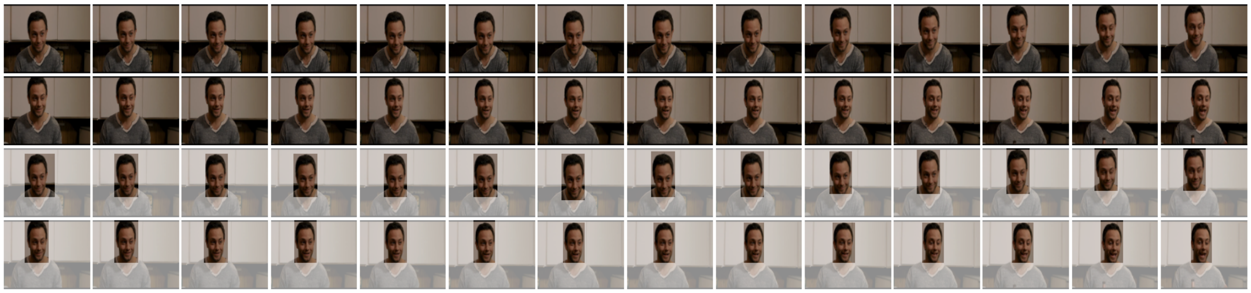
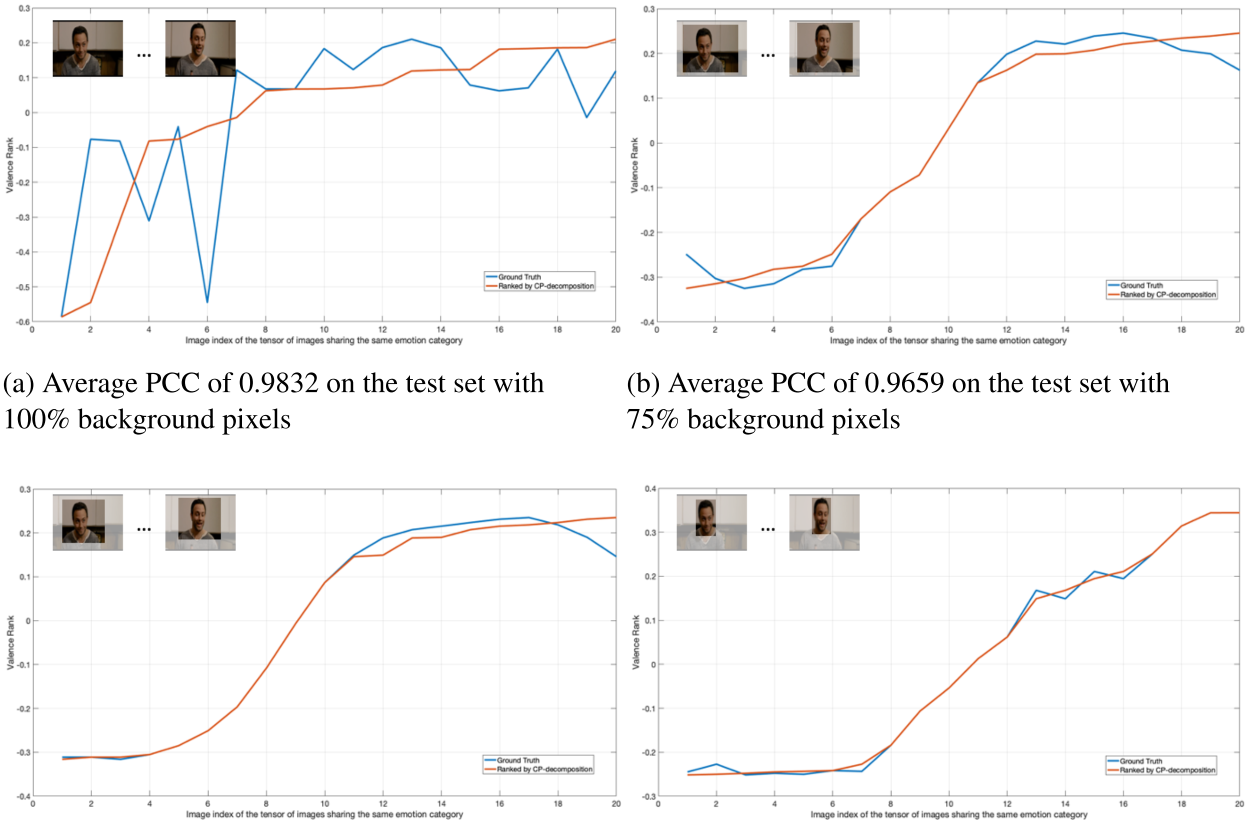


Figure 5.13: Sample set displaying images with the full background pixels included (first two rows), and the same set with background pixels removed (last two rows). Viola-Jones algorithm [10] used for face detection.

foreground vs. background pixels on the PCC our method yields. As observed in this figure, there exists a linear correlation between the percentage of background pixels and the PCC; *i.e*. the accuracy drops as the percentage of background pixels is decreased. This observation might at first be considered counter-intuitive. The reason involves the general perception that background pixels represent noise and therefore less noise in the dataset is generally expected to deliver higher accuracy. Taking a closer look at this ablation study, the backgrounds in **CK+** used in this experiment are constant with no visual changes across the images forming each test group is observed. This is different from cases where images in a given tensor have different changing backgrounds across the tensor. Accordingly, this further suggests that the ensemble representation of the decomposed tensor, formed by vectors $u$ and $v$ (from figure 4.9a) enforce a more meaningful angle (distance) between the values in vector $w$ (from figure 4.9a) and the global representation of the decomposed tensor along other modes when constant backgrounds exist across all the tensor images. To challenge this ablation study, we ran similar experiments on **AFEW**, specifically due to the fact that **AFEW** includes images per emotion category with slightly changing backgrounds.

The ablation study ran on **AFEW** shows a negative correlation between the percentage of background pixels and the ranking accuracy; *i.e*. as more background pixels are removed, higher ac-



(a) Average PCC of 0.9832 on the test set with 100% background pixels

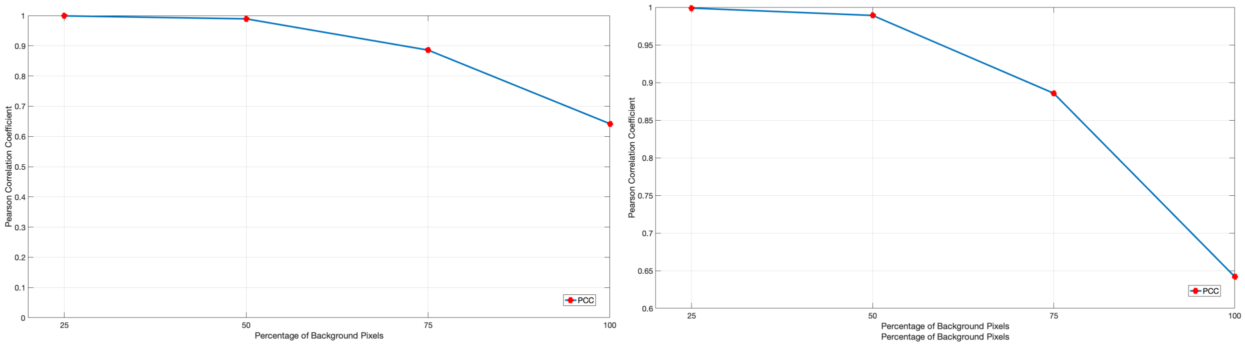
(b) Average PCC of 0.9659 on the test set with 75% background pixels

(c) Average PCC of 0.8838 on the test set with 50% background pixels

(d) Average PCC of 0.8352 on the test set with 25% background pixels

Figure 5.14: Ablation Study 3 ran on AFEW with 100%, 75%, 50% and 25% background pixels respectively. AFEW is collected under semi-controlled environments with changing background.

curacy is achieved. This is attributed to the fact that the background in **AFEW** changes across different images pertaining to the same emotion category. This property of **AFEW**'s dataset is due to the fact that it has been collected under semi-controlled environments from movie clips, hence changing of the background in different scenes is well expected. Figure 5.13 depicts a sample dataset from **AFEW** used in this ablation study. Figure 5.15 further portrays the Average PCC achieved for different percentages of background pixels. Finally, figure 5.7 depicts the plot showing the effect of the percentage of background pixels on the accuracy achieved by our ranking method. As observed in this plot, as more background pixels are removed, higher accuracy is



(a) Accuracy plot in the scale of 0 to 1

(b) Accuracy plot in the rescaled for higher visibility

Figure 5.15: Ablation Study 3: Effect of percentage of foreground vs. background pixels on accuracy on **AFEW** with changing background pixels.

achieved.

Ablation Study 4: Accuracy as a function of image resolution (image size)

In this study, we investigate the effect of image resolution (image size) on accuracy. We start by computing the PCC on the original test set to form our baseline. We, then, generate a new test set by increasing the image size by 10% and further repeat this operation in increments of 10%, eventually resizing the images by a factor of 100% (2 times the size of the original image). In each step of this ablation study, we measure the PCC generated to monitor the effect of image resolution on accuracy.

Figure 5.16 displays a test set used for this study. Figures 5.17 and 5.18 further depict the PCC achieved in each step of the ablation study for the portrayed test set. Finally, figure 5.19 displays a plot portraying the effect of image resolution (image size) on ranking accuracy. As it can be observed in this figure, our method yields almost identical accuracy in all steps of the study, proving empirically that our method is robust with respect to image resolution.

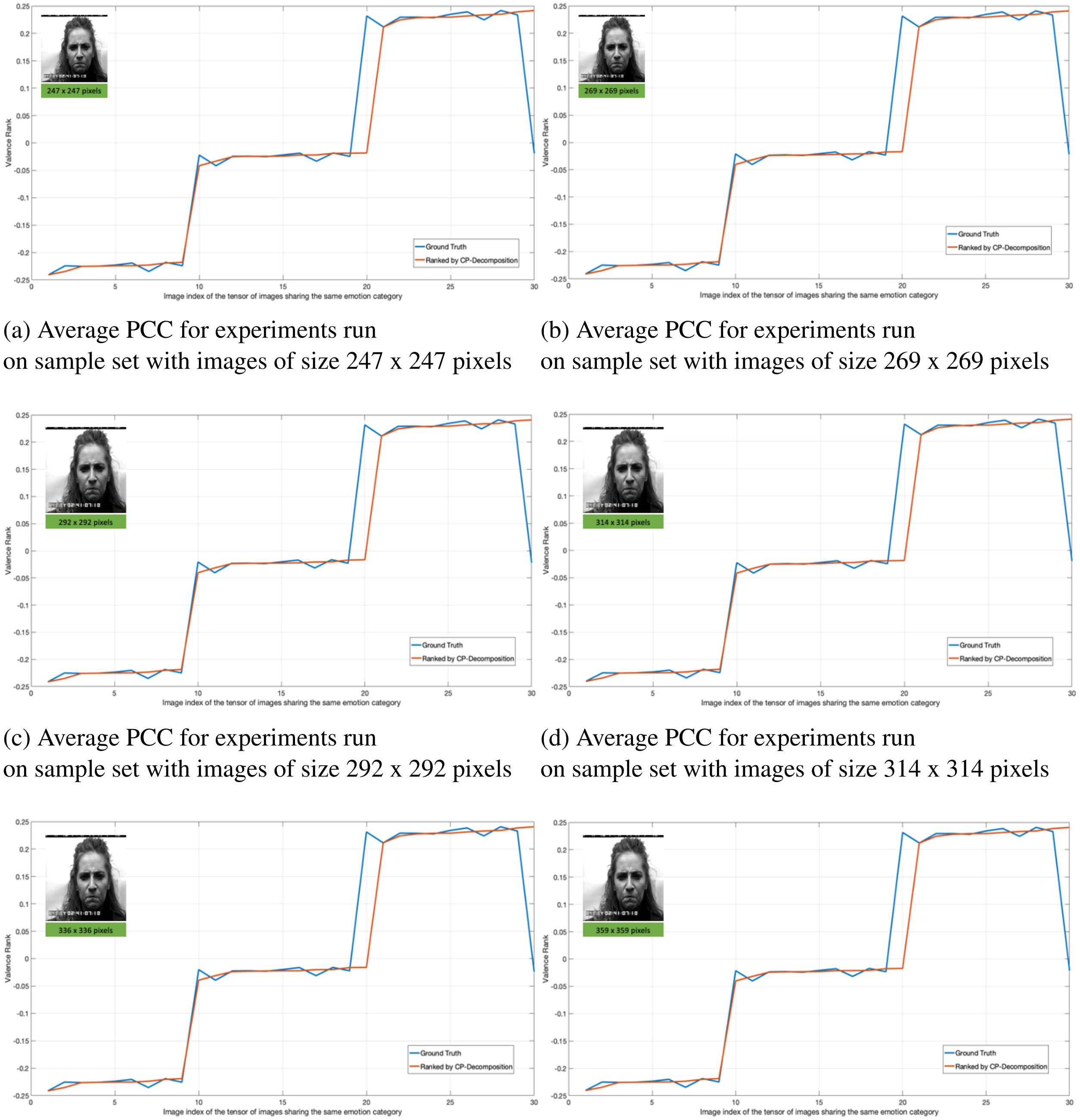
Figure 5.16: Sample set displaying varying degrees of *valence* for emotion *surprise* (excerpts from **CK+**). Different rows display the same set with varying resolutions (sizes).

Discussion

The four ablation studies discussed in this chapter further unveil numerous hidden properties of our unsupervised subspace ranking method for continuous emotions. Following, we will provide a brief discussion on each of the ablation studies and findings thereof.

**Ablation study 1 (Accuracy as a function of dataset size and density).** The results of this study provide definitive empirical proof for robustness of our method pertaining to dataset size and density. We intentionally distorted our datasets by randomly removing images from the original dataset to break the uniform distribution of *valence* in the datasets. However, our system was not challenged by the distorted subsets, still producing highly competitive accuracy. This ablation study further attests to proposition 3; *i.e. conformity*, elaborated on in chapter 4 section 4.

**Ablation study 2 (Accuracy as a function of number of outliers).** In this ablation study, we injected outliers to our test datasets to monitor the fault-tolerance of our system and further find the breakdown point. The results suggest that injecting outliers to our test sets will not break what our system claims to deliver up to a limited percentage of outliers; *i.e.* 12% of outliers. However, after the breakdown point is passed, the system's performance degrades considerably. With reference to figure 5.9, our system, however, is equipped with a self-stabilization mechanism



(a) Average PCC for experiments run on sample set with images of size 247 x 247 pixels

(b) Average PCC for experiments run on sample set with images of size 269 x 269 pixels

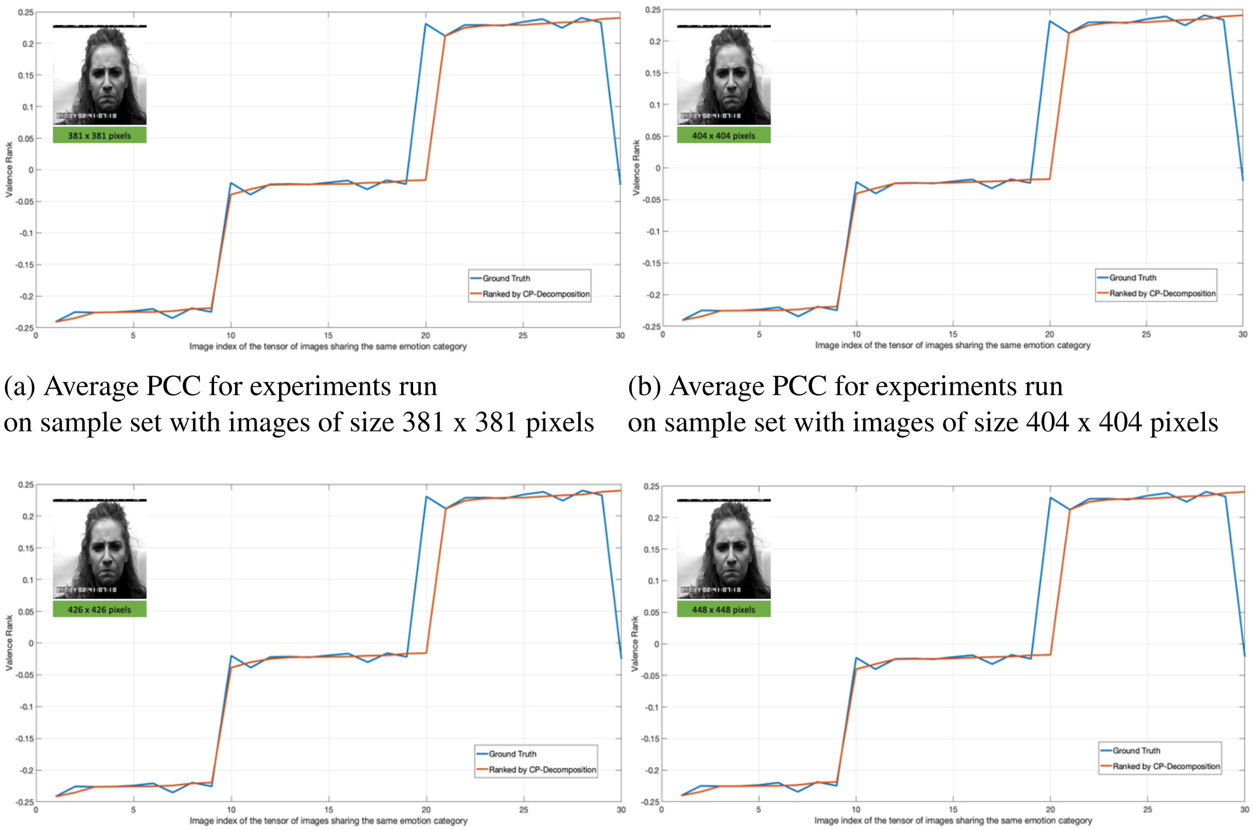
(c) Average PCC for experiments run on sample set with images of size 292 x 292 pixels

(d) Average PCC for experiments run on sample set with images of size 314 x 314 pixels

(e) Average PCC for experiments run on sample set with images of size 336 x 336 pixels

(f) Average PCC for experiments run on sample set with images of size 359 x 359 pixels

Figure 5.17: Ablation Study 4 ran on a test sample with different image resolutions



(a) Average PCC for experiments run
on sample set with images of size 381 x 381 pixels

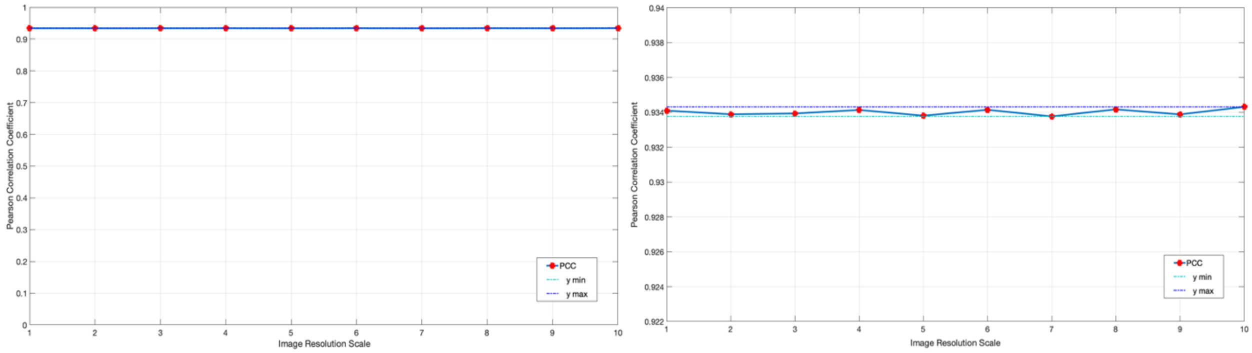
(b) Average PCC for experiments run
on sample set with images of size 404 x 404 pixels

(c) Average PCC for experiments run
on sample set with images of size 426 x 426 pixels

(d) Average PCC for experiments run
on sample set with images of size 448 x 448 pixels

Figure 5.18: Ablation Study 4 ran on a test sample with different image resolutions

to detect and further eject the outliers, hence producing a high accuracy after the ejection process. It is worth mentioning that as the percentage of outliers increase, the outliers won't confuse the ranking machine as they would in the initial phases of injecting outliers. This is due to the fact that after a certain percentage of outliers is injected into the dataset, the outliers will play the dominant role in the dataset, representing the inliers more than they represent the outliers. To avoid being challenged by this phenomena, in this ablation study, once the fault-intolerance of the system is observed, injecting more outliers was further discontinued accordingly.



(a) Plot displaying the effect of image resolution on accuracy. Y axis shows the full PCC range.

(b) Plot displaying the effect of image resolution on accuracy. Y axis zoomed in for higher visibility.

Figure 5.19: Plots displaying the effect of image resolution on accuracy.

**Ablation study 3 (Accuracy as a function of percentage of foreground vs. background pixels).** In this study, we investigated the performance of the system by gradually removing background pixels from the original images in the set, and further recording the accuracy in each step of the process. Results indicate that the percentage of the background pixels has a linear correlation with the generated accuracy when experiments are run on **CK+**. **CK+** includes images taken under controlled environments with constant backgrounds across an emotion category. The observation is that the accuracy drops as the percentage of background pixels are removed, suggesting a positive correlation between the percentage of background pixels and accuracy. This is attributed to the fact that constant backgrounds do not contribute to higher noise as opposed to images across the same emotion category with changing background pixels. To challenge this observation, a second set of experiments were run on **AFEW**. **AFEW**'s images are collected under semi-controlled environments and therefore the background of its images change across different images pertaining to the same emotion category. The observed results suggest that there exists a negative correlation between the percentage of background pixels and accuracy on such datasets; *i.e*. as background pixels are removed accuracy increases. This is due to the fact that changing background across the images forming our image tensors is considered noise, carrying no meaningful information and its removal, naturally, contributes to higher accuracy.

**Ablation Study 4 (Accuracy as a function of image resolution).** In this study, we investigated the effect of image resolution on the accuracy. We started by running experiments on the original datasets. We, then, increased the image resolution by a margin of 10% in incremental steps up to 100% (double the size of the original dataset). The results of this ablation study strongly prove that our system is robust with respect to image resolution as the generated accuracy remains almost intact in different runs of the study.

# CHAPTER 6: CONCLUSION

In this chapter, we first provide a short summary of this dissertation, followed by our main contributions to the field. We further discuss the gaps filled as a result of our contributions. Finally, we pinpoint the new opportunities our work creates, opening new doors in this area, along with the future directions the *affective computing* community is further exploring.

## Summary

In this work, we first focus on providing a comprehensive analysis on the field of *affective computing*, discussing the state of the art and further pinpointing the existing gaps in the field. Our literature review is followed by introducing our multi-angle proposed methods to the problem of *emotion recognition*. We tackle *emotion recognition* from *categorical* and *dimensional* angles, exploring supervised as well as unsupervised methods. To prove that the proposed methods deliver what they claim, extensive set of experiments were run to provide empirical proof, and theoretical proofs were delivered when applicable, as part of the supplementary material attached hereto. Result of experiments show a considerable boost in performance in the *categorical* space, while delivering high performance in the *dimensional* domain of *affective computing*, supported by extensive experiments and ablation studies run to prove the robustness of our method.

## Our Contributions

Following, we point out the main contributions delivered in this work, leveraging *deep learning*:

- we first architected a novel fully automated dataset collection pipeline, equipped with a built-in semantic sanitizer,

- we then built *UCF ER* and *LUCFER* datasets. *LUCFER* is the largest labeled *emotion recognition* dataset collected to date in the literature with more than $3.6M$ images, labeled with *emotion* and *context*, containing a rich set of metadata,
- next, we build a single-modal *context-sensitive emotion recognition* CNN models, trained on our newly constructed datasets,
- we claim and show empirically that injecting *context* to the unified training process helps achieve a more balanced *precision* and *recall*, while boosting performance, yielding an overall classification accuracy of 73.12% compared to 58.3% achieved in the closest work in the literature.

We now discuss the contributions made, leveraging *unsupervised* methodologies:

- we propose a novel unsupervised ranking method, based on *low-rank tensor decomposition*,
- we provide theoretical proof that *rank-1 cp-decomposition* can be used as a ranking machine under a fully unsupervised setting, applicable to the problem of *valence rank estimation*,
- we provide empirical proof that rank-1 cp-decomposition can be used as a ranking machine under a fully unsupervised setting, by applying the method to *valence rank estimation*, showing significant improvement in Pearson Correlation Coefficient, outperforming the state of the art by a large margin; *i.e*. 65.13% (*i.e*. difference in PCC) in one experience and 104.08% (*i.e*. difference in PCC) in another,
- we finally design and run extensive ablation studies and analyze the robustness of our ranking method with respect to (1) dataset size, (2) percentage of outliers in the dataset, (3) effect of percentage of foreground vs. background pixels on accuracy, and (4) effect of image

resolution on accuracy. Our method is robust to dataset size, density and image resolution. With respect to the role background pixels play in the system's performance, our method delivers higher accuracy as the cluttered background images are removed, and delivers lower accuracy as constant backgrounds (such as backgrounds in the **CK+** dataset) are removed. Our ranking machine has a breakdown point of 0 with respect to injecting outliers, while being equipped with self-stabilization by means of rank-1 cp-decomposition.

Doors Closed

**Largest context-sensitive emotion recognition dataset**. Existing works on *affective computing* mainly employ small-scale datasets, or in cases where relatively larger scale datasets are constructed, these datasets lack *context-sensitive* annotations, with *context-free* annotation quality not guaranteed. However, multimedia systems addressing the *affective computing* problems, are in urgent need of large-scale *context-sensitive* datasets with rich and high quality metadata. Here we address this issue by constructing the largest *emotion recognition* dataset; *i.e. LUCFER*, with $3.6M$ images labeled with *emotion* and *context*, also benefiting from a rich set of metadata, including objects, gender, caption, related searches, related images, image source, bounding boxes, hot spot coordinates among others. This dataset is available for download under the *Creative Commons Attribution 4.0 International* license.

**Empirical proof on achieving higher accuracy with *context-infused* unified training**. The *contextual* information embedded in multimedia content plays a crucial role in determining the emotion evoked when exposed to certain visual stimuli. There have been efforts, as discussed in chapter 2, in the multi-modal space, leveraging NLP techniques to tackle *context-sensitive emotion recognition* in the multi-modal domain, however, to the best of our knowledge, our work pioneers in the area of *context-sensitive still image emotion recognition* in the single-modal domain, empirically portraying the efficacy of adding *contextual information* to the unified training process. We also show that adding *context* to the training process helps achieve a more balanced *precision* and *recall*, while boosting performance, yielding an overall classification accuracy of 73.12% compared to 58.3% achieved in the closest work in the literature,

**Unsupervised *valence rank estimation* using rank-1 cp-decomposition**. The majority of proposed methods in the literature to address *emotion recognition* from a *dimensional* perspective explore solutions in the supervised space. Supervised methods, however, are challenged with the constant need for large-scale labeled datasets that are high quality and re-usable across different domains. Here in this work, we tackled the problem of *valence rank estimation* using an unsupervised approach by treating *rank-1 cp-decomposition* as a ranking machine. To the best of our knowledge, this is the first effort in the literature to perform an unsupervised ranking of emotions in the *VAD* domain using rank-1 cp-decomposition. Result of experiments run on major *emotion recognition* datasets; *i.e*. CK+, SEMAINE, AFEW-VA and AffectNet show the superiority of the proposed subspace method, showing significant improvement in the *Pearson Correlation Coefficient* (*i.e*. from 0.407 to 0.6721 in one experiment and from 0.35 to 0.7143 in another). Our extensive ablation studies prove the robustness of our method, while portraying the high level of fault-tolerance our method benefits from.

## Doors Opened

**Collecting new datasets**. We designed and built a highly reusable large-scale dataset construction pipeline, with simplicity in mind at the time of architecting the system. Lack of access to large-scale datasets is not only a challenge encountered in *affective computing*, but also one that other problem domains deal with. Our designed pipeline, available to the research community under the *Creative Commons Attribution 4.0 International* license, enables researchers in different problem domains to collect large-scale datasets. Namely, the work proposed in [126] makes use of our method to build the largest *action recognition* dataset; *i.e. UCF STAR*.

**Running multi-modal methods on *LUCFER***. Furthermore, *LUCFER*, as the largest dataset in the area of *affective computing*, and as the only dataset in the literature that is labeled not only with *emotion*, but also *context*, enriched with semantic metadata, facilitates the *affective computing* community to use this dataset by applying multi-modal methods that approach *emotion recognition* using both still images and NLP techniques. With reference to works in the literature, pointed out in chapter 2, higher accuracy is achieved using multi-modal approaches in the area of *emotion recognition*.

**Dissuading the necessity for large-scale datasets by infusing *context***. In addition, with reference to the empirical proof provided in chapter 4 suggesting that fusing *context* into the unified training process dissuades the necessity for having large-scale datasets, opens room for many empirical studies to be run on small-scale datasets by first injecting *context* as part of the unified training process.

**Reusability of our unsupervised ranking machine**. Finally, dimensional models of emotion attempt to conceptualize human emotions by defining where they lie in two or three dimensions. A widely used model of *human affect* is the *VAD* (valence, arousal, dominance) model. We showed theoretically that our unsupervised subspace ranking method is capable of delivering high performance. Considering the generalizability of our proposed ranking machine, the eager researcher is urged to apply the same method to explore other dimensions of the *VAD* model, *i.e.* arousal and dominance, or even other domains.

## Future Directions

**Viewer Profiles Clustering**. The emotion of people with different interests and backgrounds tend to be aroused differently when exposed to the same visual stimuli. However, the crowdsourcing methods adopted perform no clustering of those performing the task of labeling. Clustering viewers into their corresponding profiles based on interests and background could provide a feasible solution.

**Viewer-Image Interaction**. Current *affective computing* efforts in the literature mainly leverage direct analysis of the multimedia content and the signals conveyed such as facial expressions in order to perform *emotion recognition*. However, the joint modeling of the multimedia content and the emotion aroused in the viewers of the multimedia content being examined could better bridge the affective gap.

**Covid-19 Proof Emotion Recognition**. The majority of emotion recognition methods heavily rely on facial expressions to recognize emotions. However, wearing face masks during world pandemics, such as Covid-19, could pose serious challenges for these methods due to the partial occlusions on the face. Taking advantage of techniques in psychology that study reading emotions from the eyes, such as the one proposed by [127], could provide the *affective computing* community in constructing pandemic-proof methods.

# APPENDIX A: PROOFS

The main focus of this supplementary material is the three propositions pointed out in chapter 4 section 4, their significance, and their proofs.

## Significance of the Propositions

On theoretical ground the three propositions in our paper establish totally new results for the special case of rank-1 cp-decomposition. It is worth pointing out that these proofs are new and are not repetitions from previously published material.

Proposition 1: The closest proof in the literature is permutation invariance for minimal decomposition of a rank-R tensor (as described below). We prove a more general invariance under unitary transformation (rotation, reflection, permutation), clearly important in our application of face images.

Proposition 2: The most general proof of uniqueness is the celebrated theorem of Kruskal (as described below). However, Kruskal's theorem is not applicable here, because it proves uniqueness for minimal cp-decomposition of a rank-R tensor. We prove uniqueness for rank-1 decomposition of a rank-R tensor (see below Kruskal's theorem).

Proposition 3: Angular conformity is significant for both stability (illumination-invariance, and invariance to adding/removing data), and proof of linear complexity.

On the practical ground Section 3.3 in the paper clearly explains why the method works, directly as a result of the propositions 1-3, pointing out the fact our method is taking advantage of the known property of face images in terms of subspace behavior. This makes our work the first work recognizing that rank-1 cp-decomposition is a subspace ranking solution, and is fully unsupervised. It is a generic unsupervised subspace ranking, i.e. may be applicable to other problems of subspace nature, and in our application, it significantly outperforms state of the art. We will release the code (and all data is public) for reproducibility.

## Brief Background Theory

Before proving the three propositions that support Section 3.3 and the results of the paper, we cover some related preliminary material from multi-linear algebra, restricting to 3-way tensors for convenience (although generalizations are well-known).
We start by some basic definitions.

**Definition 1** *A 3-way tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ is said to be rank-1 if it is equal to the outer product of three nonzero vectors $u \in \mathbb{R}^I$, $v \in \mathbb{R}^J$, $w \in \mathbb{R}^K$, i.e. $\mathcal{X} = u \circ v \circ w$.*

**Definition 2** *A cp-decomposition of a 3-way tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ is given by*

$$\mathcal{X} = \sum_{r=1}^{R} u_r \circ v_r \circ w_r, \tag{A.1}$$

*where $u_r \in \mathbb{R}^I$ $v_r \in \mathbb{R}^J$, $w_r \in \mathbb{R}^K$.*

When Eq. (A.1) is the minimal sum, the constant $R$ is referred to as the rank of the tensor $\mathcal{X}$.

**Definition 3** *The matrices $\mathbf{U} = [u_1 ... u_R] \in \mathbb{R}^{I \times R}$, $\mathbf{V} = [v_1 ... v_R] \in \mathbb{R}^{J \times R}$, and $\mathbf{W} = [w_1 ... w_R] \in \mathbb{R}^{K \times R}$ are called the first, second, and third factor matrices of the tensor $\mathcal{X}$, respectively. We* *denote $\mathcal{X} = [\mathbf{U}, \mathbf{V}, \mathbf{W}]$. When $\mathcal{X}$ is rank-1, the factor matrices reduce to column vectors. Also, when the factor matrices are orthonormal, the summation in (A.1) would require a scaling factor.*

A celebrated result in multi-linear algebra is the rotational uniqueness theorem due to Kruskal [128]:

**Theorem 1** *Let $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{W}$ be the first, second, and third factor matrices of the tensor $\mathcal{X}$, respectively. Let also rank$(\mathbf{U})$ + rank$(\mathbf{V})$ + rank$(\mathbf{W}) \geq 2R + 2$. Then rank$(\mathcal{X}) = R$, and the decomposition $\mathcal{X} = [\mathbf{U}, \mathbf{V}, \mathbf{W}]$ is unique.*

Unfortunately, this theorem is not applicable in our case, because it establishes the uniqueness of the minimal cp-decomposition of a rank-R tensor, whereas our goal is to establish that the rank-1 decomposition of a rank-R tensor is unique. In fact, even when $R = 1$, Kruskal's theorem could not apply, since the Kruscal's rank condition would require $1 + 1 + 1 \geq 2 + 2$. Therefore, in the next section, we explicitly derive the uniqueness for rank-1 decomposition of a rank-R tensor.

## Proofs of propositions

We start first by the proof of Proposition 2.

**Proposition 2** *(Uniqueness)*
*The rank-1 cp-decomposition of a 3-way tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ that minimizes $\|\mathcal{X} - \hat{\mathcal{X}}\|_F$ is unique up to a non-zero scale factor and arbitrary unitary transformation along any mode.*

Let $\hat{\mathcal{X}}_1 = \lambda_1 \, \hat{u}_1 \circ \hat{v}_1 \circ \hat{w}_1$ and $\hat{\mathcal{X}}_2 = \lambda_2 \, \hat{u}_2 \circ \hat{v}_2 \circ \hat{w}_2$ be two rank-1 cp-decomposition of a 3-way tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ that minimize $\|\mathcal{X} - \hat{\mathcal{X}}_i\|_F$, $i = 1, 2$. We have:

$$\begin{aligned}
\|\mathcal{X} - \hat{\mathcal{X}}_1\|_F^2 &= \|\mathcal{X} - \lambda_1 \hat{u}_1 \circ \hat{v}_1 \circ \hat{w}_1\|_F^2 && \text{(A.2)} \\
&= \|\mathcal{X}\|_F^2 - 2\langle \mathcal{X}, \lambda_1 \hat{u}_1 \circ \hat{v}_1 \circ \hat{w}_1 \rangle + \|\lambda_1 \hat{u}_1 \circ \hat{v}_1 \circ \hat{w}_1\|_F^2 && \text{(A.3)} \\
&= \|\mathcal{X}\|_F^2 - 2\langle \mathcal{X} \overline{\times}_1 \hat{u}_1 \overline{\times}_2 \hat{v}_1 \overline{\times}_3 \hat{w}_1, \lambda_1 \rangle + \lambda_1^2 \|\hat{u}_1\|_F^2 \|\hat{v}_1\|_F^2 \|\hat{w}_1\|_F^2 && \text{(A.4)} \\
&= \|\mathcal{X}\|_F^2 - 2\lambda_1^2 + \lambda_1^2 && \text{(A.5)} \\
&= \|\mathcal{X}\|_F^2 - \lambda_1^2 && \text{(A.6)}
\end{aligned}$$

Similarly:

$$\|\mathcal{X} - \hat{\mathcal{X}}_2\|_F^2 = \|\mathcal{X}\|_F^2 - \lambda_2^2 \tag{A.7}$$

From $\|\mathcal{X} - \hat{\mathcal{X}}_1\|_F^2 = \|\mathcal{X} - \hat{\mathcal{X}}_2\|_F^2$ it therefore follows that $\lambda_1^2 = \lambda_2^2$ or $\lambda_1 = \pm\lambda_2$.

Now, let $\mathbf{T}$ be the the linear transformation that maps the orthonormal basis $\mathbf{B}_2 = [\hat{u}_2 \hat{v}_2 \hat{w}_2]$ to $\mathbf{B}_1 = [\hat{u}_1 \hat{v}_1 \hat{w}_1]$, i.e. $\mathbf{B}_1 = \mathbf{T}\mathbf{B}_2$. We have:

$$\begin{aligned}
\mathbf{B}_1 \mathbf{B}_1^T &= \mathbf{T}\mathbf{B}_2 \mathbf{B}_2^T \mathbf{T}^T && \text{(A.8)} \\
\mathbf{I} &= \mathbf{T}\mathbf{I}\mathbf{T}^T && \text{(A.9)}
\end{aligned}$$

where $\mathbf{I}$ is the identity matrix.

Therefore $\mathbf{T}\mathbf{T}^T = \mathbf{T}^T\mathbf{T} = \mathbf{I}$. It therefore follows that the rank-1 decomposition of $\mathcal{X}$ is unique up to a unitary transformation and a non-zero scale factor. $\square$

Next, we prove that the rank-1 decomposition is invariant to unitary transformations along any mode. This is important to our problem of ranking, since it makes our solution independent of the order of the input data (input images), any rotation (e.g. due to head pose), and reflection (which due to face symmetry also results in translation invariance).

**Proposition 1** *(Invariance)*

*Let $\hat{\mathcal{X}} = \lambda\, \hat{u} \circ \hat{v} \circ \hat{w}$ be the rank-1 decomposition of a 3-way tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ that minimizes $\|\mathcal{X} - \hat{\mathcal{X}}\|_F$. We maintain that $\hat{\mathcal{X}}_p = \lambda\, \hat{u} \circ \hat{v} \circ \hat{w}_p$ would minimize*

$$\|\mathcal{X}_p - \hat{\mathcal{X}}_p\|_F, \tag{A.10}$$

*where $\mathcal{X}_p = \mathcal{X} \times_1 \mathbf{P} = \mathcal{X} \times_2 \mathbf{P}$, $\mathbf{P}$ is an arbitrary $K \times K$ unitary transformation, $\hat{w}_p = \mathbf{P}\hat{w}$, and $\times_i$, $i = 1, 2, 3$ is a mode-$i$ tensor-matrix multiplication.*

Let $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{W}$ be the first, second, and third factor matrices of $\mathcal{X}$, respectively. Since $\mathbf{P}$ is a unitary transformation, we have $\|\mathbf{P}\|_F = 1$. Therefore,

$$\begin{aligned}
\|\mathcal{X} - \hat{\mathcal{X}}\|_F &= \|\mathcal{X} - \hat{\mathcal{X}}\|_F \|\mathbf{P}\|_F && \text{(A.11)}\\
&= \|\,[\mathbf{U}, \mathbf{V}, \mathbf{W}]\,\mathbf{P}^T - \lambda\,[\hat{u}, \hat{v}, \hat{w}]\,\mathbf{P}^T\|_F && \text{(A.12)}\\
&= \|\,[\mathbf{U}, \mathbf{V}, \mathbf{P}\mathbf{W}] - \lambda\,[\hat{u}, \hat{v}, \mathbf{P}\hat{w}]\,\|_F && \text{(A.13)}\\
&= \|\mathcal{X} \times_i \mathbf{P} - \lambda\, \hat{u} \circ \hat{v} \circ \mathbf{P}\hat{w}\|_F, \quad i = 1, 2 && \text{(A.14)}\\
&= \|\mathcal{X}_P - \hat{\mathcal{X}}_p\|_F && \text{(A.15)}
\end{aligned}$$

On the other hand, let $\hat{\mathcal{X}}'$ be any rank-1 tensor that minimizes $\|\mathcal{X}_P - \hat{\mathcal{X}}'\|_F$. We have:

$$\begin{aligned}
\|\mathcal{X}_p - \hat{\mathcal{X}}'\|_F &= \|\,[\mathbf{U}, \mathbf{V}, \mathbf{PW}] - \hat{\mathcal{X}}'\|_F && \text{(A.16)}\\
&= \|\,[\mathbf{U}, \mathbf{V}, \mathbf{W}]\,\mathbf{P}^T - \hat{\mathcal{X}}'\|_F && \text{(A.17)}\\
&= \|\,[\mathbf{U}, \mathbf{V}, \mathbf{W}] - \lambda'\,[\hat{u}', \hat{v}', \hat{w}']\,\mathbf{P}\|_F\|\mathbf{P}\|_F && \text{(A.18)}\\
&= \|\mathcal{X} - \lambda'\,[\hat{u}', \hat{v}', \hat{w}']\,\mathbf{P}\|_F && \text{(A.19)}
\end{aligned}$$

It follows from Proposition 2 that:

$$\begin{aligned}
\lambda'\,[\hat{u}', \hat{v}', \hat{w}']\,\mathbf{P} &= \hat{\mathcal{X}} && \text{(A.20)}\\
\hat{\mathcal{X}}' &= \lambda\,[\hat{u}, \hat{v}, \hat{w}]\,\mathbf{P}^T && \text{(A.21)}\\
\hat{\mathcal{X}}' &= \lambda\,[\hat{u}, \hat{v}, \mathbf{P}\hat{w}] && \text{(A.22)}\\
\hat{\mathcal{X}}' &= \hat{\mathcal{X}}_p, && \text{(A.23)}
\end{aligned}$$

where the choice of the mode for transformation in (A.13) and (A.22) is arbitrary. □

Finally, we prove *conformity*. This is an important property of rank-1 decomposition, because it situates our unsupervised ranking method between point-wise and list-wise methods. Essentially, each member of the input data (i.e. each image) conforms to the same group-wise ranking, but individual images are independently ranked. What this implies is that adding/removing any number of images does not affect the relative ranking of the remaining images.

**Proposition 3** *(Conformity)*

*Let* $\hat{\mathcal{X}} = \lambda\,\hat{u} \circ \hat{v} \circ \hat{w}$ *be the rank-1 decomposition of a 3-way tensor* $\mathcal{X} = [\![\mathbf{x}_1, ..., \mathbf{x}_K]\!] \in \mathbb{R}^{I \times J \times K}$

*that minimizes $\|\mathcal{X} - \hat{\mathcal{X}}\|_F$. We have $\forall k \neq k', k, k' \in [1, ..., K]$*

$$\hat{w}_k \leq \hat{w}_{k'} \quad \textit{iff} \quad \langle \mathbf{x}_k, \hat{u}\hat{v}^T \rangle \leq \langle \mathbf{x}_{k'}, \hat{u}\hat{v}^T \rangle \tag{A.24}$$

We have

$$\hat{w} = \mathcal{X} \overline{\times}_1 \hat{u} \overline{\times}_2 \hat{v} \tag{A.25}$$

$$[\hat{w}_1, ..., \hat{w}_K] = [\![\mathbf{x}_1, ..., \mathbf{x}_K]\!] \overline{\times}_1 \hat{u} \overline{\times}_2 \hat{v} \tag{A.26}$$

Therefore

$$\hat{w}_k = \mathbf{x}_k \overline{\times}_1 \hat{u} \overline{\times}_2 \hat{v} \tag{A.27}$$

$$= \left(\mathbf{x}_k^T \hat{u}\right) \hat{v}^T \tag{A.28}$$

$$= \langle \mathbf{x}_k, \hat{u}\hat{v}^T \rangle, \tag{A.29}$$

On the other hand, let $\hat{w}_{(1)}, ..., \hat{w}_{(K)}$ be the sorted elements of $\hat{w}$. Then, the indices $(1), ..., (K)$ would also sort $\mathbf{x}_1, ..., \mathbf{x}_K$ in terms of the angle between $\mathbf{x}_k$ and the subspace spanned by the orthonormal basis $\{\hat{u}, \hat{v}\}$. Furthermore from the unitary uniqueness and permutation invariance, it follows that removing any slice $\mathbf{x}_k$ from the tensor, or adding any new slice $\mathbf{x}'_k$ would not affect the ranking among other slices. Therefore:

$$\hat{w}_k \leq \hat{w}_{k'} \quad \Leftrightarrow \quad \langle \mathbf{x}_k, \hat{u}\hat{v}^T \rangle \leq \langle \mathbf{x}_{k'}, \hat{u}\hat{v}^T \rangle \tag{A.30}$$

i.e. every slice $\mathbf{x}_k$ conforms to an internal relative ranking. □

# APPENDIX B: COPYRIGHT PERMISSION FOR OWN PUBLICATION

## “Context-Sensitive Single-Modality Image Emotion Analysis: A Unified Architecture from Dataset Construction to CNN Classification” [1]

*The papers in this book comprise the proceedings of the meeting mentioned on the cover and title page. They reflect the authors' opinions and, in the interests of timely dissemination, are published as presented and without change. Their inclusion in this publication does not necessarily constitute endorsement by the editors, the IEEE Signal Processing Society, or the Institute of Electrical and Electronics Engineers, Inc.*

IEEE is committed to the principle that all persons shall have equal access to programs, facilities, services, and employment without regard to personal characteristics not related to ability, performance, or qualifications as determined by IEEE policy and/or applicable laws. To find more information about the IEEE policy visit www.ieee.org. Any person who believes that he or she has been the victim of illegal discrimination or harassment should contact IEEE Staff Director - Human Resources, at nondiscrimination@ieee.org or +1 732 465 6434. The mailing address is IEEE Human Resources, 445 Hoes Lane, Piscataway, NJ, USA.

IEEE Catalog Number: CFP18CIP-ART
ISBN: 978-1-4799-7061-2

Assembled by Conference Management Services, Inc.

# APPENDIX C: COPYRIGHT PERMISSION FOR OWN PUBLICATION “LUCFER: A large-scale context-sensitive image dataset for deep learning of visual emotions” [2]

*The papers in this book comprise the proceedings of the meeting mentioned on the cover and title page. They reflect the authors' opinions and, in the interests of timely dissemination, are published as presented and without change. Their inclusion in this publication does not necessarily constitute endorsement by the editors, the IEEE Computer Society, or the Institute of Electrical and Electronics Engineers, Inc.*

ISBN-13: 978-1-7281-1975-5
BMS Part # CFP19082-ART

*Additional copies may be ordered from*:

IEEE Computer Society
Customer Service Center
10662 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-1314
Tel: + 1 800 272 6657
Fax: + 1 714 821 4641
http://computer.org/cps
cps@computer.org

IEEE Service Center
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331
Tel: + 1 732 981 0060
Fax: + 1 732 981 9667
http://shop.ieee.org/store/
customer-service@ieee.org

IEEE Computer Society
Asia/Pacific Office
Watanabe Bldg., 1-4-2
Minami-Aoyama
Minato-ku, Tokyo 107-0062
JAPAN
Tel: + 81 3 3408 3118
Fax: + 81 3 3408 3553
tokyo.ofc@computer.org

Editorial production by Patrick Kellenberger
Cover art production by Annie Jiu

IEEE COMPUTER SOCIETY

IEEE COMPUTER SOCIETY CONFERENCE PUBLISHING SERVICES

***IEEE Computer Society***
***Conference Publishing Services* (CPS)**
http://www.computer.org/cps

# APPENDIX D: COPYRIGHT PERMISSION FOR OWN PUBLICATION

## “An Unsupervised Subspace Ranking Method for Continuous Emotions in Face Images” [3]

The **British Machine Vision Association**
and **Society for Pattern Recognition**

# The British Machine Vision Conference (BMVC)

BMVC 2020 will be held virtually as an online conference on 7th — 10th September 2020. Please see the main conference site at: http://www.bmvc2020.com/.

## Proposals to Host a Future BMVC

If you would like to host BMVC, proposals are solicited in June of each year for the BMVC two years later. Please contact the Chair for further details.

## Previous Conferences

The statistics of previous BMVCs may be of interest.

The BMVC proceedings style files are available on GitHub.

The online proceedings for past conferences are available:

- BMVC 2019, Cardiff
- BMVC 2018, Northumbria
- BMVC 2017, London
- BMVC 2016, York
- BMVC 2015, Swansea
- BMVC 2014, Nottingham
- BMVC 2013, Bristol
- BMVC 2012, Surrey
- BMVC 2011, Dundee
- BMVC 2010, Aberystwyth
- BMVC 2009, London
- BMVC 2008, Leeds
- BMVC 2007, Warwick
- BMVC 2006, Edinburgh
- BMVC 2005, Oxford
- BMVC 2004, Kingston
- BMVC 2003, Norwich
- BMVC 2002, Cardiff
- BMVC 2001, Manchester
- BMVC 2000, Bristol
- BMVC 1999, Nottingham
- BMVC 1998, Southampton
- BMVC 1997, Essex
- BMVC 1996, Edinburgh
- BMVC 1995, Birmingham
- BMVC 1994, York
- BMVC 1993, Surrey
- BMVC 1992, Leeds
- BMVC 1991, Glasgow
- BMVC 1990, Oxford

Proceedings for the BMVC's immediate predecessor, the Alvey Vision Conference, are also available:

- AVC 1989, Reading
- AVC 1988, Manchester

- AVC 1987, Cambridge

## Copyright

Please note that copyright in BMVC papers is held by the authors in every instance. The BMVA, as publisher of the proceedings, holds copyright over the collection, but the authors may make any use of papers they have authored including making an exact copy available on their own or other websites.

secretary@bmva.org

BritishMachineVisionAssociation
TheBMVA

**The British Machine Vision Association**
and Society for Pattern Recognition

# LIST OF REFERENCES

[1] P. Balouchian and H. Foroosh, “Context-sensitive single-modality image emotion analysis: A unified architecture from dataset construction to cnn classification,” in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 1932–1936.

[2] P. Balouchian, M. Safaei, and H. Foroosh, “Lucfer: A large-scale context-sensitive image dataset for deep learning of visual emotions,” in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1645–1654.

[3] P. Balouchian, M. Safaei, X. Cao, and H. Foroosh, “An unsupervised subspace ranking method for continuous emotions in face images.” in *BMVC*, 2019, p. 193.

[4] “The rise of social media,” https://ourworldindata.org/rise-of-social-media, accessed: 2020-10-18.

[5] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “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*. IEEE, 2010, pp. 94–101.

[6] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, “Afew-va database for valence and arousal estimation in-the-wild,” *Image and Vision Computing*, vol. 65, pp. 23–36, 2017.

[7] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, “The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent,” *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[8] A. Mollahosseini, B. Hasani, and M. H. Mahoor, “Affectnet: A database for facial expression, valence, and arousal computing in the wild,” *arXiv preprint arXiv:1708.03985*, 2017.

[9] "Opencv face detection," https://docs.opencv.org/3.4/d2/d99/tutorial_js_face_detection.html, accessed: 2018-11-16.

[10] P. Viola, M. Jones *et al.*, "Robust real-time object detection," *International journal of computer vision*, vol. 4, no. 34-47, p. 4, 2001.

[11] R. Plutchik, "Emotions: A general psychoevolutionary theory," *Approaches to emotion*, vol. 1984, pp. 197–219, 1984.

[12] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark." in *AAAI*, 2016, pp. 308–314.

[13] C. A. Smith and P. C. Ellsworth, "Patterns of cognitive appraisal in emotion." *Journal of personality and social psychology*, vol. 48, no. 4, p. 813, 1985.

[14] V. Shuman and K. Scherer, "Psychological structure of emotions," *International Encyclopedia of Social and Behavioral Science*, vol. 7, pp. 526–533, 2015.

[15] K. R. Scherer *et al.*, "Psychological models of emotion," *The neuropsychology of emotion*, vol. 137, no. 3, pp. 137–162, 2000.

[16] W. Wundt, "Fundamentals of physiological psychology," *Leipzig: Engelmann*, 1905.

[17] C. E. Osgood, W. H. May, M. S. Miron, and M. S. Miron, *Cross-cultural universals of affective meaning*. University of Illinois Press, 1975, vol. 1.

[18] J. Panksepp, *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press, 2004.

[19] A. Öhman, "Preattentive processes in the generation of emotions," in *Cognitive perspectives on emotion and motivation*. Springer, 1988, pp. 127–143.

[20] R. W. Picard, *Affective computing*. MIT press, 2000.

[21] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013, vol. 11.

[22] J. F. Cohn, A. J. Zlochower, J. J. Lien, and T. Kanade, “Feature-point tracking by optical flow discriminates subtle differences in facial expression,” in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 1998, pp. 396–401.

[23] G. Reevy, Y. M. Ozer, and Y. Ito, *Encyclopedia of emotion*. ABC-CLIO, 2010, vol. 1.

[24] J. A. Fulcher, “Vocal affect expression as an indicator of affective response,” *Behavior Research Methods, Instruments, & Computers*, vol. 23, no. 2, pp. 306–313, 1991.

[25] I. R. Murray and J. L. Arnott, “Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion,” *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, 1993.

[26] F. Keshtkar and D. Inkpen, “A bootstrapping method for extracting paraphrases of emotion expressions from texts,” *Computational Intelligence*, vol. 29, no. 3, pp. 417–435, 2013.

[27] K. Denecke, “Using sentiwordnet for multilingual sentiment analysis,” in *2008 IEEE 24th international conference on data engineering workshop*. IEEE, 2008, pp. 507–512.

[28] S. Poria, A. Gelbukh, E. Cambria, P. Yang, A. Hussain, and T. Durrani, “Merging senticnet and wordnet-affect emotion lists for sentiment analysis,” in *2012 IEEE 11th International Conference on Signal Processing*, vol. 2. IEEE, 2012, pp. 1251–1255.

[29] B. R. Duffy, C. Rooney, G. M. O’Hare, and R. O’Donoghue, “What is a social robot?” in *10th Irish Conference on Artificial Intelligence & Cognitive Science, University College Cork, Ireland, 1-3 September, 1999*, 1999.

[30] C. Breazeal and B. Scassellati, "How to build robots that make friends and influence people," in *Proceedings 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human and Environment Friendly Robots with High Intelligence and Emotional Quotients (Cat. No. 99CH36289)*, vol. 2. IEEE, 1999, pp. 858–863.

[31] A. Sano, J. Hernandez, J. Deprey, M. Eckhardt, M. S. Goodwin, and R. W. Picard, "Multimodal annotation tool for challenging behaviors in people with autism spectrum disorders," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 2012, pp. 737–740.

[32] S. D'mello and A. Graesser, "Mind and body: Dialogue and posture for affect detection in learning environments," *Frontiers in Artificial Intelligence and Applications*, vol. 158, p. 161, 2007.

[33] "Affective computing," https://en.wikipedia.org/wiki/Affective_computing, accessed: 2018-04-05.

[34] M. Pantic and L. J. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, 2003.

[35] G. Tiberghien, "Il context and cognition: Introduction," *Revue bimestrielle publiée avec le concours des Universités de Provenœ et Aix-Marseille II*, vol. 6, no. 2, 1986.

[36] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the international affective picture system," *Behavior research methods*, vol. 37, no. 4, pp. 626–630, 2005.

[37] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[38] "Contrasting and categorization of emotions," https://en.wikipedia.org/wiki/Contrasting_and_categorization_of_emotions, accessed: 2018-02-05.

[39] H. Schlosberg, “Three dimensions of emotion.” *Psychological review*, vol. 61, no. 2, p. 81, 1954.

[40] A. Hanjalic, “Extracting moods from pictures and sounds: Towards truly personalized tv,” *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 90–100, 2006.

[41] P. J. Lang, M. M. Bradley, B. N. Cuthbert *et al.*, “International affective picture system (iaps): Instruction manual and affective ratings,” *The center for research in psychophysiology, University of Florida*, 1999.

[42] J. Machajdik and A. Hanbury, “Affective image classification using features inspired by psychology and art theory,” in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 83–92.

[43] K.-C. Peng, T. Chen, A. Sadovnik, and A. C. Gallagher, “A mixed bag of emotions: Model, predict, and transfer emotion distributions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 860–868.

[44] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, “Emotion recognition in context,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[45] J. Yang, M. Sun, and X. Sun, “Learning visual sentiment distributions via augmented conditional probability neural network.” in *AAAI*, 2017, pp. 224–230.

[46] E. S. Dan-Glauser and K. R. Scherer, “The geneva affective picture database (gaped): a new 730-picture database focusing on valence and normative significance,” *Behavior research methods*, vol. 43, no. 2, p. 468, 2011.

[47] X. Alameda-Pineda, E. Ricci, Y. Yan, and N. Sebe, “Recognizing emotions from abstract paintings using non-linear matrix completion,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5240–5248.

[48] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 223–232.

[49] Y. Yang, J. Jia, S. Zhang, B. Wu, Q. Chen, J. Li, C. Xing, and J. Tang, "How do your friends on social media disclose your emotions?" in *AAAI*, vol. 14, 2014, pp. 1–7.

[50] S. Zhao, H. Yao, Y. Gao, R. Ji, W. Xie, X. Jiang, and T.-S. Chua, "Predicting personalized emotion perceptions of social images," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 1385–1394.

[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[52] H. Gunes and H. Hung, "Emotional and social signals: A neglected frontier in multimedia computing?" *IEEE MultiMedia*, vol. 22, no. 2, pp. 76–85, 2015.

[53] J. A. Russell, "Core affect and the psychological construction of emotion." *Psychological review*, vol. 110, no. 1, p. 145, 2003.

[54] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on affective computing*, vol. 1, no. 1, pp. 18–37, 2010.

[55] S. Zhao, G. Ding, Q. Huang, T.-S. Chua, B. W. Schuller, and K. Keutzer, "Affective image content analysis: A comprehensive survey." in *IJCAI*, 2018, pp. 5534–5541.

[56] X. Wang, J. Jia, J. Tang, B. Wu, L. Cai, and L. Xie, "Modeling emotion influence in image social networks," *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 286–297, 2015.

[57] S. Wang, J. Wang, Z. Wang, and Q. Ji, "Multiple emotion tagging for multimedia data by exploiting high-order dependencies among emotions," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2185–2197, 2015.

[58] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 47–56.

[59] J. Yang, D. She, Y.-K. Lai, P. L. Rosin, and M.-H. Yang, "Weakly supervised coupled networks for visual sentiment analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7584–7592.

[60] J. Yang, D. She, M. Sun, M.-M. Cheng, P. Rosin, and L. Wang, "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE Transactions on Multimedia*, 2018.

[61] J. Yang, D. She, and M. Sun, "Joint image emotion classification and distribution learning via deep convolutional neural network," in *Int. J. Conf. Artif. Intell*, 2017.

[62] P. Lee, Y. Teng, and T.-C. Hsiao, "Xcsf for prediction on emotion induced by image based on dimensional theory of emotion," in *Proceedings of the 14th annual conference companion on Genetic and evolutionary computation*. ACM, 2012, pp. 375–382.

[63] S. Zhao, H. Yao, and X. Jiang, "Predicting continuous probability distribution of image emotions in valence-arousal space," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 879–882.

[64] M. A. Nicolaou, H. Gunes, and M. Pantic, "A multi-layer hybrid framework for dimensional emotion classification," in *Proceedings of the 19th ACM international conference on Multimedia*. ACM, 2011, pp. 933–936.

[65] X. Lu, P. Suryanarayan, R. B. Adams Jr, J. Li, M. G. Newman, and J. Z. Wang, “On shape and the computability of emotions,” in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 229–238.

[66] B. Wu, J. Jia, Y. Yang, P. Zhao, J. Tang, and Q. Tian, “Inferring emotional tags from social images with user demographics,” *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1670–1684, 2017.

[67] B. Wu, J. Jia, Y. Yang, P. Zhao, and J. Tang, “Understanding the emotions behind social images: Inferring with user demographics,” in *Multimedia and Expo (ICME), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1–6.

[68] Q. You, J. Luo, H. Jin, and J. Yang, “Robust image sentiment analysis using progressively trained and domain transferred deep networks.” in *AAAI*, 2015, pp. 381–388.

[69] X. Zhu, L. Li, W. Zhang, T. Rao, M. Xu, Q. Huang, and D. Xu, “Dependency exploitation: A unified cnn-rnn approach for visual emotion recognition,” in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 3595–3601.

[70] J. Yang, D. She, Y. Lai, and M.-H. Yang, “Retrieving and classifying affective images via deep metric learning,” 2017.

[71] T.-Y. Liu *et al.*, “Learning to rank for information retrieval,” *Foundations and Trends® in Information Retrieval*, vol. 3, no. 3, pp. 225–331, 2009.

[72] O. Chapelle and Y. Chang, “Yahoo! learning to rank challenge overview,” in *Proceedings of the Learning to Rank Challenge*, 2011, pp. 1–24.

[73] R. Herbrich, T. Graepel, and K. Obermayer, “Large margin rank boundaries for ordinal regression,” 2000.

[74] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 89–96.

[75] C. J. Burges, "From ranknet to lambdarank to lambdamart: An overview," *Learning*, vol. 11, no. 23-581, p. 81, 2010.

[76] Z. Zheng, K. Chen, G. Sun, and H. Zha, "A regression framework for learning ranking functions using relative relevance judgments," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 287–294.

[77] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992, pp. 144–152.

[78] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[79] E. Hörster, M. Slaney, M. Ranzato, and K. Weinberger, "Unsupervised image ranking," in *Proceedings of the First ACM workshop on Large-scale multimedia retrieval and mining*. ACM, 2009, pp. 81–88.

[80] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[81] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[82] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[83] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "Avec 2011–the first international audio/visual emotion challenge," in *Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 415–424.

[84] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 3–8.

[85] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[86] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 501–508.

[87] C. Soladié, H. Salam, C. Pelachaud, N. Stoiber, and R. Séguier, "A multimodal fuzzy inference system using a continuous facial expression representation for emotion detection," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 493–500.

[88] M. A. Nicolaou, S. Zafeiriou, and M. Pantic, "Correlated-spaces regression for learning continuous emotion dimensions," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 773–776.

[89] T. Baltrušaitis, N. Banda, and P. Robinson, "Dimensional affect recognition using continuous conditional random fields," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[90] H. Chen, J. Li, F. Zhang, Y. Li, and H. Wang, "3d model-based continuous emotion recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1836–1845.

[91] S. Kaltwang, S. Todorovic, and M. Pantic, "Doubly sparse relevance vector machine for continuous facial behavior estimation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 9, pp. 1748–1761, 2016.

[92] H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, and J. P. Cosmas, "Time-delay neural network for continuous emotional dimension prediction from facial expression sequences," *IEEE transactions on cybernetics*, vol. 46, no. 4, pp. 916–929, 2016.

[93] E. Sánchez-Lozano, P. Lopez-Otero, L. Docio-Fernandez, E. Argones-Rúa, and J. L. Alba-Castro, "Audiovisual three-level fusion for continuous estimation of russell's emotion circumplex," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 31–40.

[94] M. Kächele, M. Schels, and F. Schwenker, "Inferring depression and affect from application dependent meta knowledge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 41–48.

[95] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble cca for continuous emotion prediction," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 19–26.

[96] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan, "Multimodal prediction of affective dimensions and depression in human-computer interactions," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 33–40.

[97] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Multi-scale temporal modeling for dimensional emotion recognition in video," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 11–18.

[98] Z. Huang, T. Dang, N. Cummins, B. Stasak, P. Le, V. Sethu, and J. Epps, "An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 41–48.

[99] A. Milchevski, A. Rozza, and D. Taskovski, "Multimodal affective analysis combining regularized linear regression and boosted regression trees," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 33–39.

[100] M. Kächele, P. Thiam, G. Palm, F. Schwenker, and M. Schels, "Ensemble methods for continuous affect recognition: Multi-modality, temporality, and challenges," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 9–16.

[101] S. Chen and Q. Jin, "Multi-modal dimensional emotion recognition using recurrent neural networks," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 49–56.

[102] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 65–72.

[103] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, “Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks,” in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 73–80.

[104] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang, “Multimodal audio, video and physiological sensor learning for continuous emotion prediction,” in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 97–104.

[105] M. Amirian, M. Kächele, P. Thiam, V. Kessler, and F. Schwenker, “Continuous multimodal human affect estimation using echo state networks,” in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 67–74.

[106] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, “Avec 2012: the continuous audio/visual emotion challenge,” in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 449–456.

[107] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, “Avec 2013: the continuous audio/visual emotion and depression recognition challenge,” in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 3–10.

[108] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, “Avec 2014: 3d dimensional affect and depression recognition challenge,” in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 3–10.

[109] P. Shaver, J. Schwartz, D. Kirson, and C. O'connor, "Emotion knowledge: Further exploration of a prototype approach." *Journal of personality and social psychology*, vol. 52, no. 6, p. 1061, 1987.

[110] W. G. Parrott, *Emotions in social psychology: Essential readings*. Psychology Press, 2001.

[111] J. Jia, S. Wu, X. Wang, P. Hu, L. Cai, and J. Tang, "Can we understand van gogh's mood?: learning to infer affects from images in social networks," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 857–860.

[112] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[113] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *arXiv preprint arXiv:1608.05442*, 2016.

[114] "fupes," https://linux.die.net/man/1/fdupes, accessed: 2018-04-05.

[115] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database," *International journal of lexicography*, vol. 3, no. 4, pp. 235–244, 1990.

[116] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 245–251.

[117] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Special issue on learning from imbalanced data sets," *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, 2004.

[118] "Separation of concerns," https://en.wikipedia.org/wiki/Separation_of_concerns, accessed: 2018-04-05.

[119] L. De Lathauwer, B. De Moor, and J. Vandewalle, "On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors," *SIAM journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1324–1342, 2000.

[120] P. M. Kroonenberg, *Three-mode principal component analysis: Theory and applications*. DSWO press, 1983, vol. 2.

[121] T. Zhang and G. H. Golub, "Rank-one approximation to high order tensors," *SIAM Journal on Matrix Analysis and Applications*, vol. 23, no. 2, pp. 534–550, 2001.

[122] P. M. Kroonenberg and J. De Leeuw, "Principal component analysis of three-mode data by means of alternating least squares algorithms," *Psychometrika*, vol. 45, no. 1, pp. 69–97, 1980.

[123] R. Meyes, M. Lu, C. W. de Puiseau, and T. Meisen, "Ablation studies in artificial neural networks," *arXiv preprint arXiv:1901.08644*, 2019.

[124] D. L. Donoho and P. J. Huber, "The notion of breakdown point," *A festschrift for Erich L. Lehmann*, vol. 157184, 1983.

[125] "Fault tolerance," https://en.wikipedia.org/wiki/Fault_tolerance, accessed: 2020-11-07.

[126] M. Safaei, P. Balouchian, and H. Foroosh, "Ucf-star: A large scale still image dataset for understanding human actions." in *AAAI*, 2020, pp. 2677–2684.

[127] S. Baron-Cohen, S. Wheelwright, J. Hill, Y. Raste, and I. Plumb, "The reading the mind in the eyes test revised version: a study with normal adults, and adults with asperger syndrome or high-functioning autism," *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, vol. 42, no. 2, pp. 241–251, 2001.

[128] J. B. Kruskal, "Rank, decomposition, and uniqueness for 3-way and n-way arrays," *Multiway data analysis*, pp. 7–18, 1989.