LARGE-SCALE AFFECTIVE COMPUTING FOR VISUAL MULTIMEDIA

Brendan Jou

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

ABSTRACT

LARGE-SCALE AFFECTIVE COMPUTING FOR VISUAL MULTIMEDIA

Brendan Jou

In recent years, **Affective Computing** has arisen as a prolific interdisciplinary field for engineering systems that integrate human affections. While human-computer relationships have long revolved around cognitive interactions, it is becoming increasingly important to account for human affect, or feelings or emotions, to avert user experience frustration, provide disability services, predict virality of social media content, etc. In this thesis, we specifically focus on Affective Computing as it applies to large-scale visual multimedia, and in particular, still images, animated image sequences and video streams, above and beyond the traditional approaches of face expression and gesture recognition. By taking a principled psychology-grounded approach, we seek to paint a more holistic and colorful view of computational affect in the context of visual multimedia. For example, should emotions like 'surprise' and 'fear' be assumed to be orthogonal output dimensions? Or does a 'positive' image in one culture's view elicit the same feelings of positivity in another culture? We study affect frameworks and ontologies to define, organize and develop machine learning models with such questions in mind to automatically detect affective visual concepts.

In the push for what we call "Big Affective Computing," we focus on two dimensions of scale for affect – scaling up and scaling out – which we propose are both imperative if we are to scale the Affective Computing problem successfully. Intuitively, simply increasing the number of data points corresponds to "scaling up". However, less intuitive, is when problems like Affective Computing "scale out," or diversify. We show that this latter dimension of introducing data *variety*, alongside the former of introducing data *volume*, can yield particular insights since human affections naturally depart from traditional Machine Learning and Computer Vision problems where there is an objectively truthful target. While no one

might debate a picture of a 'dog' should be tagged as a 'dog,' but not all may agree that it looks 'ugly'. We present extensive discussions on why scaling out is critical and how it can be accomplished while in the context of large-volume visual data.

At a high-level, the main contributions of this thesis include:

Multiplicity of Affect Oracles. Prior to the work in this thesis, little consideration has been paid to the affective label generating mechanism when learning functional mappings between inputs and labels. Throughout this thesis but first in Chapter 2, starting in §2.1.2, we make a case for a conceptual partitioning of the affect oracle governing the label generation process in Affective Computing problems resulting a multiplicity of oracles, whereas prior works assumed there was a single universal oracle. In Chapter 3, the differences between intended versus expressed versus induced versus perceived emotion are discussed, where we argue that perceived emotion is particularly well-suited for scaling up because it reduces the label variance due to its more objective nature compared to other affect states. And in Chapter 4 and 5, a division of the affect oracle along cultural lines with manifestations along both language and geography is explored. We accomplish all this without sacrificing the 'scale up' dimension, and tackle significantly larger volume problems than prior comparable visual affective computing research.

Content-driven Visual Affect Detection. Traditionally, in most Affective Computing work, prediction tasks use psycho-physiological signals from subjects viewing the stimuli of interest, e.g., a video advertisement, as the system inputs. In essence, this means that the machine learns to label a proxy signal rather than the stimuli itself. In this thesis, with the rise of strong Computer Vision and Multimedia techniques, we focus on the learning to label the stimuli directly without a human subject provided biometric proxy signal (except in the unique circumstances of Chapter 7). This shift toward learning from the stimuli directly is important because it allows us to scale up with much greater ease given that biometric measurement acquisition is both low-throughput and somewhat invasive while stimuli are often readily available. In addition, moving toward learning directly from the stimuli will allow researchers to precisely determine which low-level features in the stimuli are actually coupled with affect states, e.g., which set of frames caused viewer discomfort rather a broad sense that a video was discomforting. In Part I of this thesis, we illustrate an

emotion prediction task with a psychology-grounded affect representation. In particular, in Chapter 3, we develop a prediction task over semantic emotional classes, e.g., 'sad,' 'happy' and 'angry,' using animated image sequences given annotations from over 2.5 million users. Subsequently, in Part II, we develop visual sentiment and adjective-based semantics models from million-scale digital imagery mined from a social multimedia platform.

Mid-level Representations for Visual Affect. While discrete semantic emotions and sentiment are classical representations of affect with decades of psychology grounding, the interdisciplinary nature of Affective Computing, now only about two decades old, allows for new avenues of representation. Mid-level representations have been proposed in numerous Computer Vision and Multimedia problems as an intermediary, and often more computable, step toward bridging the semantic gap between low-level system inputs and high-level label semantic abstractions. In Part II, inspired by this work, we adapt it for vision-based Affective Computing and adopt a semantic construct called adjective-noun pairs. Specifically, in Chapter 4, we explore the use of such adjective-noun pairs in the context of a social multimedia platform and develop a multilingual visual sentiment concept ontology with over 15,000 affective mid-level visual concepts across 12 languages associated with over 7.3 million images and representations from over 235 countries, resulting in the largest affective digital image corpus in both depth and breadth to date. In Chapter 5, we develop computational methods to predict such adjective-noun pairs and also explore their usefulness in traditional sentiment analysis but with a previously unexplored cross-lingual perspective. And in Chapter 6, we propose a new learning setting called cross-residual learning building off recent successes in deep neural networks, and specifically, in residual learning; we show that cross-residual learning can be used effectively to jointly learn across even multiple related tasks in object detection (noun), more traditional affect modeling (adjectives), and affective mid-level representations (adjective-noun pairs), giving us a framework for better grounding the adjective-noun pair bridge in both vision and affect simultaneously.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ANP adjective-noun pair. iii, vi–viii, 5, 21, 22, 37, 43–48, 50–65, 68, 70–74, 76–90, 92, 94–98, 100, 102, 108–114, 116, 117, 142, 145

API Application Program Interface. 49, 50, 57, 67, 68, 80

AU action unit. 20

ccTLD country code top-level domain. 68, 69

CNN convolutional neural network. vii, 20, 22, 37, 77, 78, 86, 90, 92, 94, 145

CRL cross-residual learning. 5, 97, 98, 100, 103, 105–107, 116, 139, 145

EDA electrodermal activity. vii, 14, 119, 120, 122–125, 127, 128

EEG electroencephalogram. 14, 29, 120, 122

FACS Facial Action Coding System. 20

fMRI functional magnetic resonance imaging. 10, 120, 122

geodata geographical data. viii, 44, 65–67, 70–74

georef geo-reference. viii, 44, 67–70, 72, 73

GIF Graphical Interchange Format. ii, vi, viii, 26–29, 31–34, 36–40, 145, 146

GIS geographical information systems. 66

GPS global positioning system. iii, viii, 67, 68, 70, 73

GPU graphics processing unit. 81, 82, 85, 107, 111

HSV hue-saturation-value. 36, 125

IAPS International Affective Picture System. 23, 29

ICANN Internet Corporation for Assigned Names and Numbers. 69

ILSVRC ImageNet Large Scale Visual Recognition Challenge. 22, 78, 79, 84–86, 92, 94, 110, 111, 125, 134, 137

IMDb Internet Movie Database. iv, vii, 131–136

LBP local binary patterns. 20, 141

LSTM long short-term memory. iii, 104, 105, 139

MFCC Mel-Frequency Cepstral Coefficients. 126, 141

MIL multiple-instance learning. 137, 139

MSE mean squared error. 38

MTL multitask learning. 28, 32, 34–36, 39, 40, 96, 98, 100–102, 106, 107, 109, 110, 116, 117, 141

MTR multitask regression. 26, 28, 35, 38–40

MTurk Mechanical Turk. 135

MVSO multilingual visual sentiment concept ontology. ii, iii, vi–viii, 5, 43, 44, 46–48, 54, 55, 58, 59, 61, 63–68, 70–72, 74, 76–83, 85–90, 93–95, 107, 117, 132, 138, 142, 145

NER named entity recognition. 54, 68, 69, 101

nMSE normalized mean squared error. 38, 39

NN neural network. 5, 22, 40, 78, 79, 96, 98, 101, 102, 104, 139, 141

NP non-deterministic polynomial-time. 36

OLS ordinary least squares. 38

PANAS Positive and Negative Affect Schedule. 12, 132

POS part-of-speech. 47, 48, 50, 51, 74, 101, 109

RBF radial basis function. 22, 93, 94

ResNet residual network. vii, 79, 96, 97, 99, 104, 106, 110–115, 117

SGD stochastic gradient descent. 80, 84, 110

SNE Stochastic Neighbor Embedding. 65

SVM Support Vector Machine. 20–22, 37, 38, 90, 93, 94, 125, 127, 138

VA valence-arousal. 12, 13, 20, 43, 90, 131, 132, 142

VAD valence-arousal-dominance. 13, 22, 120, 142

VSO visual sentiment ontology. vi, viii, 46, 47, 49, 51, 58–60, 64, 79, 84, 107–112, 117, 132

X-ResNet cross-residual network. vii, 97, 99, 100, 102, 106, 107, 110, 112, 113, 115–117

Acknowledgments

Before any other, I am thankful to God for seeing me through this season. It often felt like the valley more than the mountain, but the Lord was ever present – leading, teaching and encouraging. He comforted me in the Psalms, "Surely God is my help; the LORD is the one who sustains me" (54:4); admonished and humbled me in the Proverbs, "The horse is made ready for the day of battle, but the victory belongs to the LORD" (21:31); and through sun and storm, He asked me to trust the anchor He laid down, the Word made flesh, Jesus the Christ. I didn't always trust (Luke 8:22-5), but He was faithful (2 Thess. 3:3). In every sorrow, He was better. In every victory, still He was better.

I am indescribably thankful for my wife Michelle, without whose love, support and prayer I would have never finished. She has been a constant source of encouragement and joy. She's sacrificed and served tirelessly – in prayer, listening to my research rants, booking my conference travel and so much more. I am humbled and grateful to be her husband. I am also indebted to my parents Brian and Angie who continue to serve us today in ways we can never repay. My sister Kaitlin and brother Darence too have given so much to Michelle and I over these past several years. Aunts and uncles, Peter, Jenny, Danny, Betty, Bernard, Jonathan, and cousins, Derrick, Angela, Debbie, Gabe, Wynne and the entire family have all been a mainstay of support in ways that I could not hope to enumerate here.

This thesis would most certainly not have been possible without the support of my doctoral advisor Professor Shih-Fu Chang, who taught me to think critically and keep pushing the boundary of the known until something new appeared. At every meeting, I was always surprised by how he navigated research details but also keep in view the 10,000-foot perspective. There is no doubt in my mind that Shih-Fu will continue pioneering research in the future. There are so many others at the Digital Video & Multimedia (DVMM) Lab, past and present, that I appreciate and learned much from – Dong Liu, Tao Chen, Subhabrata

Bhattacharya, Svebor Karaman, Rongrong Ji, Christoph Kofler, Hongzhi Li, Joseph G. Ellis, Felix X. Yu, Yan Wang, Guangnan Ye, Ming-Hen Tsai, Jie Feng, Xiaoming Wu, Zheng Shou, Xavier Giró-i-Nieto and Karin Weidner-Mubanda. I am also grateful to the Department of Defense's National Defense Science & Engineering Graduate Fellowship Program for funding three formative years of my PhD, and Edward & Janet Chen along with John Donnelly from the Wei Family Private Foundation who also helped fund some of my time at Columbia. It was also a pleasure to work with Fernando Silveira and Brian Eriksson at the Technicolor Research Center, and George Toderici, Christian Szegedy, Alexey Vorobyov and Rahul Sukthankar at Google Research during my PhD internships. I also thank John Paisley, Ching-Yung Lin, Rogerio Feris and Fred Jiang for being on my thesis committee.

So many other friends made this possible. Mat Liu and Jon Yan are still my groomsmen in life. They continue to prove themselves as men who have and still yet walk the beaten path with me, and I have deserved not their friendship; yet I am a better man because of them. Our church Trinity Grace Church Westside's pastor Derek Worthington and wife Emilie have served and blessed us more than they know. All our TGC friends, Gideon/Mel Copple, Ryan/Alison Patch, Ashley/Carly Byrd, Derek/Avery Reed, Kevin/Amy Wong, Brian/Lindsey Lee, Amaya Perea, Hilary Ribbens, Haley Szendel-Heacock, Katie Hanson-Killebrew, Simon/Jeannine McGown, Tim Hwang, Shuning Zhao and countless others deserve special mention for how they love well and are some of the most refreshing company you can find anywhere. For our small cohort of PhD students who met for prayer throughout the years, Xiaojie Zhang, Xing Xia, Nathaniel Boggs, David Noell and Ruiwen Lee, it was good to have others to struggle alongside. I am also grateful to Aron/Grace Yu who are a godly, smart and fun couple to be with; and Aron especially has been a great colleague to discuss both faith and research. Long-time family friends Mike/Joy Vega and Louis/Caroline/Erin Miu also have been a faithful support. Finally, but not lastly, I am thankful for Brendan Kiu, Valerie Chang, Danica Chiu-Kwok, Jason/Andrea Huang, Andrew Lee, Benson Tsai, Nathan Wan, Andrew Yeager, Jack Mao and many others, who have each served me in one way or another throughout my doctorate.

Dedicated to Michelle, Brian, Angie, Kaitlin and Darence.

Chapter 1

Introduction

Affections were never part of the original design of the computer. Since at least the Middle Ages, the process of calculating, or accounting, has been understood as a deeply rational and cognitive task. And "calculating" is precisely what the computer and indeed, the Turing machine [Turing, 1950], were originally designed for. Ever since, the rational horsepower of the computational machine has been a source of societal wonder and academic intrigue. Along the way, numerous perceptual tasks were annexed to the computer, including solving partial differential equations, simulating physical systems, network fault detection, and recognizing real-world objects in digital imagery. Yet the range of computational tasks has most always been limited to the numerable, calculable and quantifiable – and very rarely the sometimes irrational, often hard-to-quantify "soft" elements of our lives and interactions.

Shortly after the rise and popularity of sentient machines in science fiction literature and television, serious research discussions around the mid-1990s began on the possibility of engineering "emotionally intelligent" machines. In her seminal 1997 book *Affective Computing*, Picard paved the way for a new line of research focused on responsibly and ethically incorporating human "feelings" into machine intelligence [Picard, 1997]. The field of Affective Computing has since found itself in a wide range of applications including content engagement [Wang and Cheong, 2006; Schaefer *et al.*, 2010; Fleureau *et al.*, 2012; McDuff *et al.*, 2013; Silveira *et al.*, 2013], market research [Ahn and Picard, 2014], social network marketing [Chen *et al.*, 2014c; Wang *et al.*, 2015a], and human-computer interaction [Simon, 1967; Oviatt and Cohen, 2000; Hoque *et al.*, 2013; Marsella and Gratch, 2014].

Figure 1.1: Simple comparative example illustrating the ability of visuals to evoke emotions distinctly compared to linear text. Adapted from [Hipps, 2009].

In this thesis, we focus on how we can engineer machines to better understand human affections in and with respect to visual multimedia at scale.

1.1 Motivations

Sight is among the most fundamental of human senses. And photography, especially in its modern digital form, has been one of the most influential agents reaffirming this law of nature. Being able to actually *see* an object, a scene, a person or an idea has way of capturing our attention and affections unlike any other medium before it has. One striking example of how visuals differ vastly as a form of affective communication is in a simple side-by-side comparison. John Dyer writes,

> Photography, which became more common beginning around 1850, implies a [different] kind of thinking [compared to the printed book]. . . . one of the key differences that media scholars often point to is that while printed text is particularly good at conveying linear, abstract data, images are much better at drawing us into a concrete story and evoking an emotional reaction. Shane Hipps, in his book *Flickering Pixels*, points out that the old adage, 'A picture is worth a thousand words,' isn't really accurate. 'It would seem a picture is actually worth a thousand emotions' [Hipps, 2009]. He gives the example of the difference between our reactions to a printed sentence like 'The boy is sad' and a picture of a starving child, crying in the middle of the scored African plain. The printed sentence presents us with abstract concepts, but the picture immediately pulls on our hearts. When we see words, they cause us to think; but when we see a picture, we react first and then think about our reaction afterward. [Dyer, 2011]

As seen pervasively in film, advertisements and journalism, visuals have become a pow-

erful tool in digital storytelling and evidentiary arguments. And yet, as to what exact emotional mechanisms are at work in our visual faculties is still a hot ground for research debate and study (q.v. §1.1.2 and §2.1.1.2). Even so, just as machine vision research is developing in parallel with research in biological human vision, i.e., and not *after* the latter is fully understood, so also significant research strides can still and are being made in affect modeling and sensing both in machines and human concurrently [Picard, 2003]. Likewise, there are many correlations and empirical heuristics that only become apparent as problems are studied at scale.

1.1.1 Large-Scale and Ubiquitous Visual Data

Fueled by innovations like the Web and cheap storage, every generation since has continued to produce higher throughput and volume of digital data than the last. Visual multimedia in particular has experienced an explosive growth in volume, spurred on in part by the social media revolution. At the time of this writing, approximately 400 hours of video are uploaded to YouTube every minute[1], Instagram boasts over 80 million photos uploaded per day[2], and some 728 million images are uploaded to Flickr yearly[3]. While these numbers have attracted computer vision and multimedia research for years, of equal importance is the "value" of these data. An often unspoken assumption is that these research problems are important *not only* because photos and video are being uploaded in bulk but also because they are shared, seen, (dis)liked, tagged, modified and repurposed massively.

The "value" of visual data can be as much a financial or reputational entity, e.g., as measured by the number views a video generates or royalties claimed from licensing an image, as much it can be a deeply personal entity, e.g., a wedding video or photo capturing a memorable moment. For example, several high view count videos on YouTube now boast over one billion views[4] and Instagram claims 40 billion total shares as well as 3.5 billion likes daily. If we wanted to value a photo, how would we go about it? Explicit signals for measuring

[1] `https://www.youtube.com/yt/press/statistics.html`

[2] `https://www.instagram.com/press`

[3] `https://www.flickr.com/photos/franckmichel/6855169886`

[4] `https://en.wikipedia.org/wiki/List_of_most_viewed_YouTube_videos`

this "value" such as view, like and share counts are undoubtedly valuable metrics; however, they fall short in many scenarios, such as in metric inflation spam [Cha *et al.*, 2007; Benevenuto *et al.*, 2009], the "cold start problem" in personalization and recommender systems [Schein *et al.*, 2002], and social science or psychology research on human-computer interactions [Lang *et al.*, 1997; Mikels *et al.*, 2005; Vessel *et al.*, 2014]. In such cases, content-based methods have risen an effective solution and increasingly also as a complementary signal, serving as an *implicit signal* for capturing the "value" of visual data. And it can be argued that no other content-based paradigm cuts to the center of the valuation problem for visual multimedia than Affective Computing.

1.1.2 The Affective Gap

It may come as a surprise to some that a clear, succinct and agreed upon definition of affect (or emotion) remains elusive today among philosophers, psychologists, sociologists and neuroscientists. One broad and loose definition of human affect is our experience of feelings and emotions, usually in response to some stimulus. However, there are several models of emotion as well as multiple frameworks for understanding their onsets, transitions and transactions (q.v. §2.1.1 and §2.2). In this thesis, we tend most to agree with and adopt the position taken by sociologist Arlie R. Hochschild of "emotion as social," who wrote,

> What is emotion? Emotion, I suggest, is a biologically given sense, and our most important one. Like other senses – hearing, touch, and smell – it is a means by which we know about our relation to the world... Emotion is unique among our senses, however, because it is related not only to an orientation toward *action* but also to an orientation toward *cognition*. ... [F]eeling signals perception and expectation to us, and turning this around, different patterns of perception and expectation correspond to different feeling names. Since culture directs our seeing and expecting, it directs our feeling and our naming of feeling. [Hochschild, 1983]

Though Affective Computing had yet to come into formation when Hochschild wrote this in 1983, this idea of "naming of feeling" would become foundational to a core goal in teaching machines to understand human affect.[5]

[5]Indeed, Hochschild alone should not be credited for this idea of "naming" as other works such as [Katz, 1980; Plutchik, 1980] also developed similar lines of thinking around the same time.

One of long-standing engineering goals of Affective Computing is to bridge what is known as the "affective gap." In essence, the *affective gap* is the conceptual divide between low-level features from a signal of interest and high-level human affective states [Calvo and D'Mello, 2010]. The signal of interest can be the stimulus itself or it can be psycho-physiological measurements of a human subject. And on the other side of the gap, the high-level affect states are most often understood to be categorical semantic entities by which we "name" feelings or emotions. Formally, we can understand this as a traditional machine learning problem where the goal is to formulate and learn a functional mapping $\mathcal{H}(\mathbf{X}, \mathbf{y})$, where $\mathbf{X} \in \mathbb{R}^{n \times d}$ corresponds to the low-level features and $\mathbf{y} \in \mathbb{R}^n$ are the high-level affect states. However, as we shall see in this thesis, though the problem can be cast in this well-understood form, the acquisition, condition, qualities and most especially interpretations of both the inputs $\mathbf{X}$ and outputs $\mathbf{y}$ differ and vary greatly from the traditional setting when in the Affective Computing context.

1.1.3 Computer Vision and Affective Science

In Computer Vision and Multimedia, a concept pre-dating the "affective gap" is the "semantic gap." In the *semantic gap* problem, we seek to bridge the conceptual divide between low-level content features and high-level semantics describing the content. For example, given an input image depicting a chair, the goal of a machine might be to output the semantic label 'chair' or 'furniture'. Similarly, in Affective Computing, we might desire a machine to output a label like 'sad' for image of the sad boy in Figure 1.1. In some contexts, the "affective gap" problem can be understood to be equivalent to the "semantic gap" problem; however, the target output in the "affective gap" problem need not always be semantic categories, e.g., they can be psychological affect states.

Much of current Affective Computing focuses on detecting and recognizing the affect states of individuals, either within or in response to a stimulus. As a result, a major line of vision-based research in Affective Computing is devoted to tasks like face expression and gesture recognition, e.g., [Tao and Tan, 2005; Zeng *et al.*, 2009; Calvo and D'Mello, 2010; McDuff *et al.*, 2013; Kim *et al.*, 2013; Mediratta *et al.*, 2013; Marsella and Gratch, 2014]. As important as these problems are, we often want to understand affect in much broader

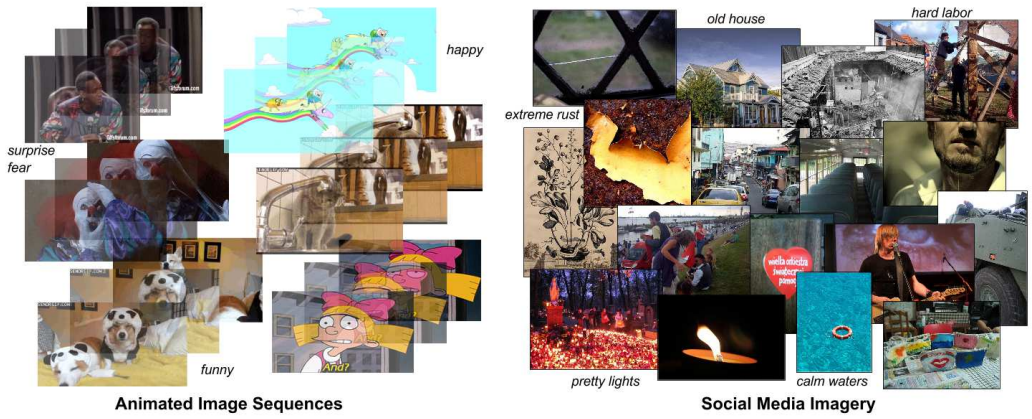


Figure 1.2: Examples of the diverse visual multimedia content we are interested in. Note that while there are indeed some faces and gestures that exist in visual multimedia, there are also far more diverse content that is of affective interest. Face and gesture recognition can thus only be one part of the Visual Affective Computing equation moving forward.

contexts than simply knowing there is a smile or frown in a photograph. For example, even if we could detect that the boy in Figure 1.1 was 'sad,' we do not know why or what caused that expression. Likewise, if there was an image without a human entity present, could machines still detect that the depicted scene was 'dark' or 'eerie' or simply 'negative'. In this thesis, we cast the net much wider to expand the range of affective targets that vision-based multimedia systems are able to understand, thus while we do include elements like facial expression recognition we do not limit our techniques to it (e.g., see Figure 1.2).

1.2 Overview of the Thesis

In Chapter 2, an overview of the Affective Computing and its particular relation to Computer Vision and Multimedia is presented. A framework for understanding these related work through multiple affect oracles and their inter-relationships is proposed in §2.2. In Part I of the thesis, we illustrate content-driven affect detection in visual multimedia using a psychology-grounded affect representation. In particular, in Chapter 3, we present emotion detection experiments in animated image sequences along with advocating perceived emotion detection. Part II of the thesis focuses on mid-level representations for Visual Affective

Computing, where a framework for understanding multilingual visual affect is proposed in Chapter 4 and accompanying visual detection methods are discussed in Chapter 5, and further, a novel cross-task learning method applicable to affective visual concept detection is proposed in Chapter 6. Finally, open challenges in Visual Affective Computing are discussed in Part III where some inroads along the lines of leveraging psycho-physiological signals (Chapter 7) and highly curated professional films (Chapter 8) are presented, and a concluding discussion with a summary of contributions and concluding remarks on applications and ethics is given in Part IV.

Chapter 2

Overview of Affective Computing and Visual Concept Detection

Arguably the first work on Affective Computing was by Hebert A. Simon in [Simon, 1967], although the field was not formed until Rosalind W. Picard popularized it in [Picard, 1997]. Simon proposed that one critical function of intelligent systems, and organisms in general, was that they could switch between deliberate processing of information as well as reactionary processing in order to achieve some end goal. Principally, this meant that if a human interacting with a machine became frustrated, say because of usability issues, an emotionally intelligent machine should be able to detect and factor the frustration into its behavior, hopefully taking some action to make the goal the user easier to achieve. Picard aptly wrote on this, "Affect, like weather, is hard to measure; and like weather, it probably can't be predicted or controlled with perfect reliability. But, if we can do significantly better than random, then people will at least be less likely to get caught in a thunder storm without an umbrella" [Picard, 2003].

In this chapter, we provide an abbreviated summary of Affective Computing with emphasis on vision-based multimedia systems. We briefly discuss affective representations from a social science and neuroscience perspective and their subsequent relation to engineering problems for computationally modeling emotion. We also propose a curator-content-user relation paradigm for intuitively understanding and organizing prior art and that also helps

define directions for future research in affect understanding. We also briefly review methods in visual concept detection with focus on works targeting affective applications.

2.1 Affective Science

Affective Computing is deeply rooted at the intersection of the computational, biological and social sciences. And so, to develop a bearing for what is actually to be computed, it is instructive to briefly first discuss the historical and modern understanding of affect. Affective science research largely comes out of three streams of study, each with slightly different emphases: psychology, social science and neuroscience (listed in no particular order). Much of the work in this thesis hinges on perspectives in psychology and social science, but the indirect contributions of neuroscience are also acknowledged.

2.1.1 Affective Mechanisms and Models

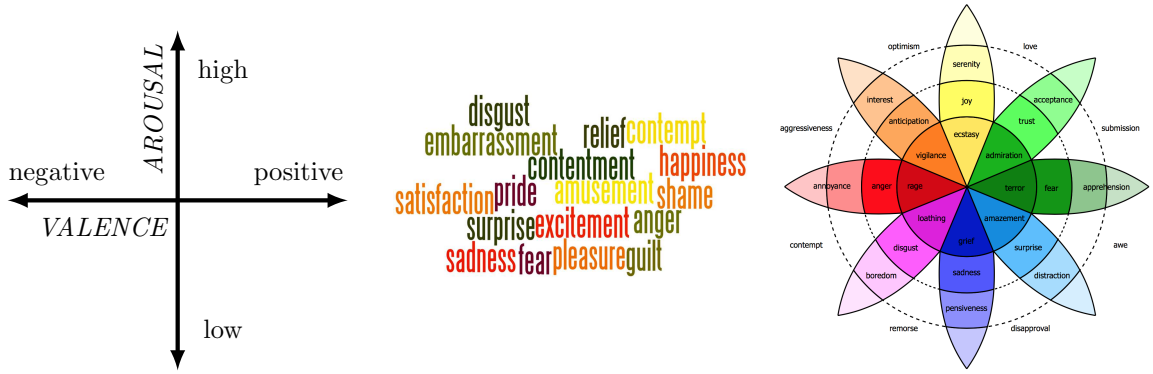
Affections (or emotion) operate significantly differently from the more cognitive processes in our psyche, i.e., memory, attention, language, problem solving, and planning. For example, psychologist Paul Ekman writes, "Emotions can have a very fast onset, beginning so quickly that they can happen before one is aware that they have begun. ... Emotions are unbidden, not chosen..." [Ekman, 1999]. Likewise, some research shows an individual may not even be conscious of a stimulus when the affective region of their brain activates [Öhman, 2002]. One of the more popular illustrations of this dichotomy between cognition and emotion is found in a psychological phenomenon known as the *mere-exposure effect*. The *mere-exposure effect* occurs when individuals develop a preference simply on a basis of familiarity or repeated exposure. In such cases, it is clear that no cognitive decision is made to bias toward preference and yet the "affective reaction [is] likely to become increasingly positive" [Kunst-Wilson and Zajonc, 1980]. And so, as far as we currently understand, the processes governing everything from the elicitation to sustaining of emotion does depart significantly from those of cognition (though they are not mutually exclusive processes).

2.1.1.1 Affective Neuroscience and Visual Perception

The primary biological processes that we are concerned with in this thesis originate principally in the ventral (base), anterior (front) and posterior (back) regions of the brain. Specifically, the prefrontal cortex and limbic system are associated with processing and regulating mood/personality/emotions; and the visual cortex, located in the occipital lobe, is responsible for processing visual information. Briefly, there have been a number of "locationist approaches" to methodically isolate regions or networks in the brain associated with emotions [Lieberman, 2007]. For example, in [Chikazoe *et al.*, 2014], fine-grained patterns of neural activity in functional magnetic resonance imaging (fMRI) results were discovered in the orbitofrontal cortex which hint at a standard or canonical "code" by which our brains derive emotions. On the other hand, works like [Barrett and Satpute, 2013] argue against localizing psychological faculties in favor of a "constructionist approach" where high-dimensional brain states and their interactions are the focus of study. Respectively, for vision, it is generally well understood now that much of our visual processing comes from the primary visual cortex (a.k.a. V1), which is especially tuned for pattern recognition tasks, and the secondary visual cortex (a.k.a. V2) in the visual association area, where some visual attentional and memory processes occur [Gazzaniga, 2009].

2.1.1.2 Models of Emotion

There are two basic models of emotion – one leans toward the psychologists and neuroscientists, and the other toward the sociologists [Hochschild, 1983]. The "organismic model" arises from the work of those like Charles Darwin, William James and Sigmund Freud, who essentially argued that emotions are principally biological and psychological processes [Darwin, 1872; James, 1883]. In this line of thinking, research is fixated far more on drivers like instinct and libido. The "interactional model" comes to us from work by those like John Dewey and Erving Goffman, where the focus is the *meaning* of the psychological processes and the role of social exchanges [Goffman, 1967]. Thus, it is sufficient for the interactionalist to say that affect always has some innate biological component, but they are concerned more with social structures, e.g., "framing" and "emotional labor" in the social sciences. While there are some other works that try to strike a middle ground between these two



(a) Valence-Arousal Space. (b) Ekman Emotions. (c) Plutchik Wheel of Emotions.

Figure 2.1: Example Affect Representation Visualizations. A dimensional representation is shown in **(a)**, basic emotion semantics in **(b)**, and a dimensional-semantic hybrid in **(c)**.

models, these emotion models offer two different "levels" of affect granularity that can serve as oracles during computational learning.

2.1.2 Affective Representations

Affect can be represented in many different ways. And from an engineering perspective, we might imagine some Euclidean space of unknown dimensions onto which we might assign regions to affective categories. However, in order to remain faithful to psychological models of affect, we need to consider that the actual dimensionality of the space considered will depend on the chosen affect representation, the affect semantics of corresponding dimensions may not necessarily be common emotion names, and/or the affect axes may not necessarily even be orthogonal. Fortunately, there are several already established representations of affect drawn from existing psychology and sociology research. Here, we discuss four popular methods for representing affect in Affective Computing:

Sentiment. The simplest way to represent affect is along a single axis representing the positivity or negativity. In this *sentiment* context, we might talk about affect in a binary sense of "feeling positive" or a "negative feeling," or a trinary if we also consider "neutral" as a state. Beyond n-ary extensions, the sentiment axis also allows us to compare stimuli in a relative, pairwise fashion, e.g., "that movie scene gave me a

more positive feeling than the ones before it." Due to the simple nature of this representation, sentiment is where the bulk of Affective Computing modeling and applications focus, e.g., [Turney, 2002; Bautin *et al.*, 2008; Wan, 2009; Baccianella *et al.*, 2010; Siersdorfer *et al.*, 2010; Thelwall *et al.*, 2010; Liu, 2012; Yuan *et al.*, 2013; You *et al.*, 2014; Xu *et al.*, 2014]. Another advantage of the sentiment representation is that the small number of output dimensions also can reduce the variance in performance in automatic systems. It is also worth noting that some more broadly consider sentiment as referring to personal dispositions and as being related to opinions [Soleymani *et al.*, 2016].

Valence, Arousal and Dominance (and its Variants). While convenient, a fundamental problem with treating affect as sentiment alone is when, for example, "[w]e lose the distinction between a fearful dislike [and] an angry dislike..." [Hochschild, 1983]. Extended two- and three-dimensional models, along with several variants, were proposed in lieu of this and are currently the dominant representation in psychological studies. One popular two-dimensional representation assigns *valence* to one axis, denoting *sentiment* or positiveness/negativeness[1], and *arousal* to the other, referring to the degree of stimulation[2]. This valence-arousal (VA) space allows us to refer to further qualify *sentiment* in a much richer fashion by essentially introducing magnitudes (arousal). In [Bradley *et al.*, 1992; Dietz and Lang, 1999], a 'boomerang'-shaped feasible region in the VA space was proposed where affects like a highly negative feeling (valence) with a low arousal were hypothesized to be unrealizable. Several other works also tried to map regions the VA space to namable emotions, e.g., [Russell, 1990] proposed a circumplex model where emotions exist on a circle approximately equidistant from the origin, or in the Positive and Negative Affect Schedule (PANAS) System of [Watson *et al.*, 1988], where positive and negative affect are treated as separate "dimensions" where the VA axes lie at a 45° angle to the positive-negative axes. Many psychologists also extend into a third dimension called *dominance*, denoting the volition an individual can maintain when affecting[3], e.g., consider the difference between "anger" and "rage" [Mehrabian, 1980; Russell, 1991] (and some like [Fontaine *et al.*, 2007],

[1]In some literature, *valence* is also referred to as *pleasure* or "pleasantness/unpleasantness".

[2]In some literature, *arousal* is also referred to as "arousing/subduing" or "level of activation".

[3]In some literature, *dominance* is also referred to as *control* or "attention/rejection".

even extend further, dividing *dominance* into *potency* and *unpredictability*). Several Affective Computing works have since adopted these valence-arousal (VA) and valence-arousal-dominance (VAD) spaces, e.g., [Hanjalic and Xu, 2005; Yang *et al.*, 2006; Yang *et al.*, 2008; Gunes and Pantic, 2010; Schaefer *et al.*, 2010; Koelstra *et al.*, 2011; Xu *et al.*, 2012].

Ekman Emotions. While sentiment, VA and VAD spaces garner support from a large body of works, they are often difficult to apply when concrete semantics are necessary, like in user facing applications. Ekman developed a set of six emotions that he believed to be "basic" and universal: anger, disgust, fear, happiness, sadness and surprise (sometimes, contempt was also included in early writings); and later expanded the set to 17 emotions, also including amusement, contentment, embarrassment, excitement, guilt, pleasure, pride, relief, satisfaction and shame [Ekman, 1992; Ekman, 1999]. Ekman argued that other affect were simply derivatives or composites of these core emotions. These *Ekman emotions* rose to popularity especially in affective multimedia systems because they provided a semantic grounding for multimodal approaches, e.g., [Shin and Kim, 2010; Teixeira *et al.*, 2011; Wang *et al.*, 2013a; Peng *et al.*, 2015; Wang *et al.*, 2015a; Xu *et al.*, 2015].

Plutchik Emotions. Contrary to Ekman's suggestion, psychologist Robert Plutchik posited that there are instead eight primary emotions: anger, fear, sadness, disgust, surprise, anticipation, trust and joy [Plutchik, 1980]. He suggested that these emotions were bipolar and could be organized into contrasting pairs, for example, surprise versus anticipation or trust versus disgust. In addition, from these eight primary emotions, Plutchik proposed that each emotion could have three levels of intensity, where a primary emotion like joy can have an emphasized form, ecstasy, and a mellower form, serenity, resulting in 24 total, non-composite emotions. To organize these bipolar emotions, their intensities, as well as some composite emotions, Plutchik proposed a conic illustration called the "Wheel of Emotions" (unfolded cone view in Figure 2.1c). Of all the affective representations discussed, *Plutchik emotions* have had the least amount of traction in psychology, partially due to criticisms about omitting pride/shame, but they are still widely applied in Affective Computing, e.g., [Borth *et al.*, 2013b; Jiang *et al.*, 2014; Xu *et al.*, 2015]. Notably, there are also some other similar representations like [Kiesler, 1983] that extend to interpersonal

affect structures.

2.2 Aspects of Affective Gaps

As discussed in §1.1.2, the task of bridging of the *affective gap* remains at the cornerstone of Affective Computing. Many Multimedia and Computer Vision, and even modern Affective Computing, works naïvely reduce this problem to approximating a ground truth label distribution from some input features without much thought to where these labels come from. However, because affections are deeply subjective and faceted entities, this reductionist assumption that ground truth labels are generated by some *universal* oracle actually gets broken. One analogy to consider is that in a classical object detection problem almost no one might disagree that an image contains a 'cup,' however, not everyone will agree that the shot of the 'cup' was 'beautiful'. As a result, *one of the main positions and proposals of this thesis that we first clearly delineate the social oracles involved in the generation of the affect targets that we seek to model and infer.* To illustrate this, we propose divisions along curation roles as well as along cultural lines beginning in §2.2.2.

2.2.1 Affective Computing Paradigms

The reductionist Affective Computing view is illustrated in Figure 2.2a, where there is some input feature set $\mathbf{X}$ we want to map to some affect label $\mathbf{a}$. The reductionist is satisfied say that the input feature set $\mathbf{X}$ can generally represent anything from pixels, audio, biological signals or some derivative, and is equally satisfied to simply say there is an affect label $\mathbf{a}$ without clarifying where it came from.[4] However, historically, Affective Computing has considered the input feature set $\mathbf{X}$ to specifically refer to the psycho-physiological features measured from an individual $\mathbf{X}_{\text{physio.}}$ in response to a stimulus $\mathbf{X}_{\text{content}}$ (see Figure 2.2b). In this setting, the input $\mathbf{X}_{\text{physio.}}$ could be anything from a web camera recording facial expression images to electroencephalogram (EEG) or electrodermal activity (EDA) signals

[4]In fact, there have been many peer-reviewed works in Computer Vision and Multimedia have unfortunately presented methods for recognizing emotion categories, e.g., "sad", "happy" and "angry", with little mention of where the affective taxonomy even came from or why they were chosen.

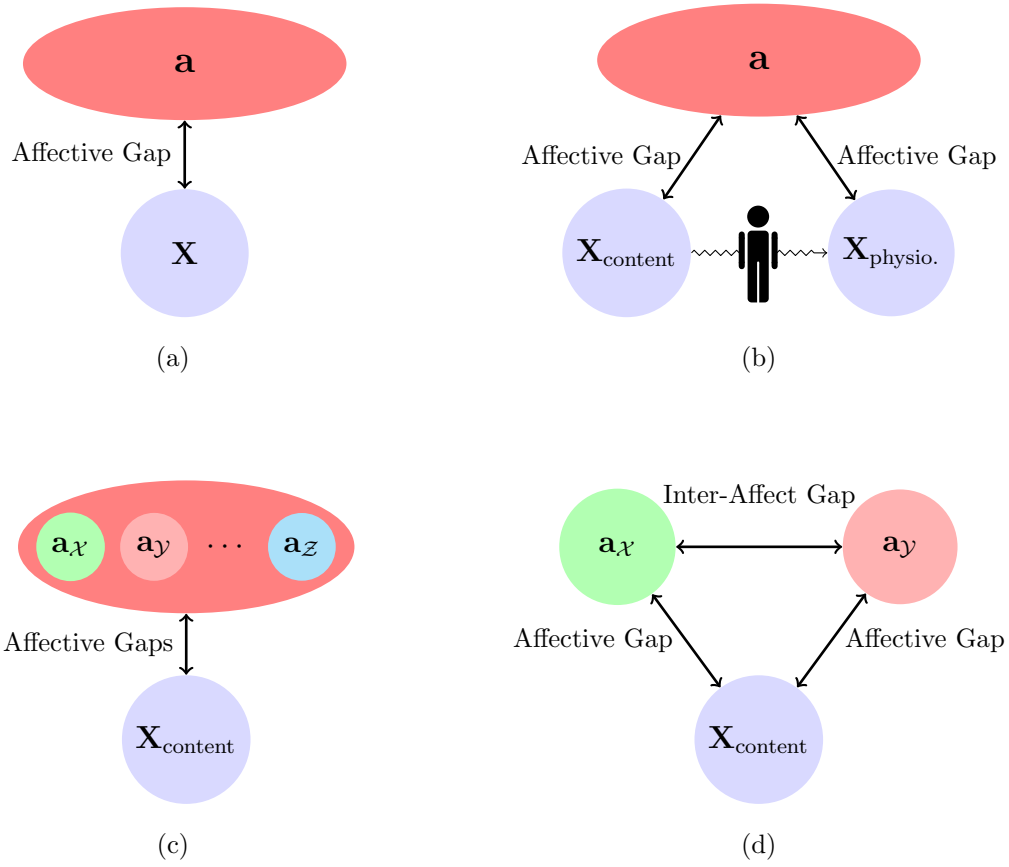


Figure 2.2: Affective Computing Paradigms: **(a)** Reductionist view of the *affective gap*, where *ambiguous* input features $\mathbf{X}$ are mapped to an *ambiguous* affect oracle $\mathbf{a}$. **(b)** In actuality, some Affective Computing works try to bridge the *affective gap* between an affect oracle $\mathbf{a}$ and stimulus $\mathbf{X}_{\text{content}}$ and/or physiological signals $\mathbf{X}_{\text{physio.}}$, where $\mathbf{X}_{\text{content}}$ is the stimulus that induces $\mathbf{X}_{\text{physio.}}$ response in a human. **(c)** However, there can be multiple affect oracles $\mathbf{a}_{\mathcal{X}}, \mathbf{a}_{\mathcal{Y}}, \ldots, \mathbf{a}_{\mathcal{Z}}$, especially with respect to the stimulus $\mathbf{X}_{\text{content}}$, whereas (a) assumed $\mathbf{a}_{\mathcal{X}} = \mathbf{a}_{\mathcal{Y}} = \cdots = \mathbf{a}_{\mathcal{Z}}$. **(d)** And subsequently, there can also be gaps between affect oracles as well, e.g., the *inter-affect gap* between $\mathbf{a}_{\mathcal{X}}$-$\mathbf{a}_{\mathcal{Y}}$.

measured from the subject. In addition, here, the affective label **a** specifically refers to the affect of the single, affecting, measured human subject.

With the recent rising success of content-based methods, a growing body of Affective Computing works have honed in on bridging the *affective gap* from the stimulus $\mathbf{X}_{\text{content}}$ directly (including this thesis). As a result, one of the increasingly apparent challenges is the need for a large number of instance-label pairs for learning. However, unlike tasks in object recognition where labels from multiple individuals, like crowdsourcing workers, can simply be pooled together, labeling tasks in Affective Computing settings cannot be so simply be aggregated because each individual can be a legitimate affect oracle in their own right. This can make it difficult to distinguish between crowdsourcing spammers and a legitimate, but minority affect. And so, as shown in Figure 2.2c, with respect to a stimulus $\mathbf{X}_{\text{content}}$, there is generally more than one affect oracle. Naturally, this observation that there are multiple affect oracles then leads to *inter-affect gaps* between oracles (Figure 2.2d).

2.2.2 Affect Oracles and Targets

While some of the differentiations between affect oracles in §2.2.1 may initially seem like little more than a philosophical juggling act, they have direct consequences to how we build affective computational systems. Two examples of how dividing or splitting up the affect oracle aid the design of visual affect prediction are presented in Chapter 3 and 4. Briefly though, consider an experimental setup where an *experimenter* or *content curator* has produced a piece of *content*, or *stimulus*, that will be presented to a *user* or *subject*. In this setup, there are four affective targets that we might consider predicting:

- **Intended Affect.** A film director (*curator*) might intend for his audience to feel a certain emotion at a particular point in a scene. A conceivable learning task then would be to predict a director's *intended emotion* given particular scene, but such curator labels are typically very difficult to procure, especially at scale since it requires the originator's direct input.

- **Expressed Affect.** Given the affect state that the film director intends his audience to feel, he might choose to express it by curating a scene to contain the appropriate

color schemes, auditory cues and story narrative. Alternatively, the director might be choosing to express his emotional interpretation of a scene without actually even intending for an audience to feel anything. Much like intended affect, *expressed affect* labels are typically difficult to procure and are oftentimes even indistinguishable from intended affect. As a result, though it is worth noting they can be different, we tend to treat this the same as intended affect in this thesis.

- **Induced/Evoked Affect.** In response to the film (*stimulus*) created by the director, some affect may be induced in an individual (*subject*) in the audience. This *induced affect* may not necessarily be the same as the one the director (*curator*) intended, but it is nonetheless an affect. If the individual were say a movie pre-screener, they might fill out a post-showing survey and describe what they felt and computational models could be trained to predict how this same individual might react to a new scene. And in fact, this setting of *induced affect* prediction is where the overwhelming majority of Affective Computing lies (though not often acknowledged).

- **Perceived Affect.** Although film scenes may have evoked a particular affect in the individual (*subject*), they may have the ability to perceive what affect the director (*curator*) intended. This *perceived affect* might, for example, be an understanding that the director intended the audience to be surprised, but the individual felt apathetic. Generally, this particular type of affect is both easy to acquire labels for and considerably less subjective than *induced affect*. And to the best of our knowledge, this affect had not been explored computationally before [Jou *et al.*, 2014] (q.v. Chapter 3).

Each of the above affective targets can be represented in any number of the ways from §2.1.2. And indeed this particular four-way division of the affect oracle is actively studied in psychology and social science, e.g., [Katz, 1980; Hochschild, 1983]. Further, this is *not* the only way to partition the affect oracle as we might consider culture (q.v. Chapter 4), gender, age, occupation and more as other ways to partition the affect oracle, preserving the subjective nature of affect while reducing label variance.

2.3 Visual Affect Detection

Early Affective Computing works that specifically focused on affect in visual content/stimuli naturally began with color-based features. Some proposed color-based emotion detection for image retrieval [Wang *et al.*, 2006; Solli and Lenz, 2008], while others suggested it for aesthetics prediction [Datta *et al.*, 2006b], webpage design [Shin and Kim, 2010] and image database organization [Solli and Lenz, 2010]. Likewise, for video, [Hanjalic and Xu, 2005] and [Wang and Cheong, 2006] explored affect understanding in film clips along with multimodal approaches combining visual and auditory cues. For the most part though, the broader visual affect detection problem had little traction and was eclipsed by face expression and gesture recognition until about the early 2010s.

In Table 2.1, we organize some selected Affective Computing (or related) works and show their relation to one conceptual partitioning of the affect oracle along intended, expressed, induced/evoked and perceived affect dimensions (q.v. §2.2.2). This catalog is organized to the best of our understanding of each work since none of these prior works considered such an affect partitioning nor ever used such terminology in their works. Early face expression and gesture recognition works tend to fall under the 'expressed affect' category since they often used acted expressions while later in-the-wild works along with most all other Affective Computing work fall under the 'induced affect' category. Some recent studies do fall into the 'intended affect' category based on how data was gathered, but we leave this for expanded discussions later, e.g., in §2.3.2 and Part II. Some of our work in 'perceived affect' detection is also discussed later in Chapter 3.

2.3.1 Face Expression and Gesture Recognition

Though not the primary focus of this thesis, it is worth highlighting several seminal works in face expression and gesture recognition since this is where much of Visual Affective Computing has focused the majority of its effort in the past two decades. In biometrics, facial expression recognition has been a cornerstone problem for years, before "Affective Computing" was even coined [Samal and Iyengar, 1992; Yacoob and Davis, 1996; Essa and Pentland, 1997; Yin *et al.*, 2000; Sim *et al.*, 2002; Chang *et al.*, 2006; Zeng *et al.*, 2009; Gross *et al.*, 2010;

Citation	Intended	Expressed	Induced	Perceived	Citation	Intended	Expressed	Induced	Perceived
[Lang *et al.*, 1997]			✓		[Jia *et al.*, 2012]			✓	
[Lyons *et al.*, 1998]		✓			[Soleymani *et al.*, 2012]			✓	
[Kanade *et al.*, 2000]		✓			[Xu *et al.*, 2012]			✓	
[Sim *et al.*, 2002]		✓			[Borth *et al.*, 2013b]	✓		✓	
[Fasel and Luettin, 2003]		✓			[Canini *et al.*, 2013]			✓	
[Lisetti and Nasoz, 2004]		✓			[Kim *et al.*, 2013]		✓		
[Chanel *et al.*, 2005]			✓		[McDuff *et al.*, 2013]			✓	
[Hanjalic and Xu, 2005]			✓		[Mediratta *et al.*, 2013]			✓	
[Tao and Tan, 2005]			✓		[Silveira *et al.*, 2013]			✓	
[Wang *et al.*, 2006]			✓		[Yuan *et al.*, 2013]			✓	
[Wang and Cheong, 2006]			✓		[Chen *et al.*, 2014c]			✓	
[Yang *et al.*, 2006]			✓		[Ellis *et al.*, 2014a] †				✓
[Castellano *et al.*, 2007]			✓		[Jou *et al.*, 2014] †				✓
[Solli and Lenz, 2008]			✓		[Martínez *et al.*, 2014]			✓	
[Yang *et al.*, 2008]			✓		[Yang *et al.*, 2014]	✓			
[Yanulevskaya *et al.*, 2008]			✓		[You *et al.*, 2014]			✓	
[Kipp and Martin, 2009]			✓		[Xu *et al.*, 2014]			✓	
[Zeng *et al.*, 2009]			✓		[Zhao *et al.*, 2014]			✓	
[Calvo and D'Mello, 2010]			✓		[Campos *et al.*, 2015] †			✓	
[Machajdik and Hanbury, 2010]			✓		[Jou *et al.*, 2015] †	✓			✓
[Schaefer *et al.*, 2010]			✓		[Wang *et al.*, 2015a]			✓	
[Shin and Kim, 2010]			✓		[Baveye *et al.*, 2015b]			✓	
[Siersdorfer *et al.*, 2010]			✓		[Vandal *et al.*, 2015]			✓	
[Solli and Lenz, 2010]			✓		[Campos *et al.*, 2016] †			✓	
[Dan-Glauser and Scherer, 2011]			✓		[Jou *et al.*, 2016] †	✓			
[Koelstra *et al.*, 2011]			✓		[Jou and Chang, 2016a] †	✓			
[Morency *et al.*, 2011]		✓			[Liu *et al.*, 2016] †	✓			✓
[Teixeira *et al.*, 2011]			✓		[Pappas *et al.*, 2016] †	✓			✓

Table 2.1: Breakdown of selected Affective Computing works along intended, expressed, induced and perceived affect. Listed alphabetical by year. The † indicates selected work that we have contributed to the field.

Lucey *et al.*, 2010]. These methods for facial expression recognition span a broad spectrum from Gabor wavelets [Lyons *et al.*, 1998] to local binary patterns (LBP) [Shan *et al.*, 2009], and more recently, to convolutional neural networks (CNNs) [Tang, 2013]. Other works adopted a localization approach originally developed in [Ekman *et al.*, 1980] called the Facial Action Coding System (FACS) which details a lookup table of craniofacial regions called action units (AUs) that compositionally form a facial expression; so, for example, AU15 denotes a "Lip Corner Depressor". Here, the efforts are typically focused on the proxy problem of recognizing AUs rather than an ultimate facial expression since the former is a more fine-grained challenge and the expressions can be derived by psychology-grounded rules [Kanade *et al.*, 2000; Tian *et al.*, 2001; Fasel and Luettin, 2003; Bartlett *et al.*, 2005; Lucey *et al.*, 2010; Mediratta *et al.*, 2013; McDuff *et al.*, 2013].

Gesture recognition [Wu and Huang, 1999; Mitra and Acharya, 2007] (not be confused with hand sign recognition), like facial expression recognition, is also an area of active research. Recently, some commercial systems have begun to incorporate gesture recognition for video games, controlling augmented reality environments and general-purpose computer interfacing. Pure vision-based gesture recognition though has been approached with methods like oriented histograms of texture [Freeman and Roth, 1995], skin color tracking via particle filtering [Bretzner *et al.*, 2002] and optical flow [Zhang and Kender, 2013]. And specifically in connection to emotion, works like [Castellano *et al.*, 2007] used body movement in addition to gesture expressions to recognize semantic emotions. Also, work in [Kipp and Martin, 2009], gesture forms like hand shape, palm orientation and motion direction were used to detect emotion in the VA space.

2.3.2 Visual Affective Concept Detection

By treating emotions as categorical entities, many vision and multimedia methods can also be applied to the general problem of visual affect detection. In [Yanulevskaya *et al.*, 2008], "holistic" image features with Support Vector Machines (SVMs) [Cortes and Vapnik, 1995] were proposed for sentiment prediction on an image dataset used largely in psychology [Lang *et al.*, 1997]. The "holistic" features were essentially a codebook over local color histogram and Gabor features. Likewise, in [Siersdorfer *et al.*, 2010], global and local color

histograms as well as SIFT [Lowe, 2004] features were used with SVMs on a larger dataset of social images also for sentiment. Several other works like [Jia *et al.*, 2012] followed the trend and applied these generic vision features to other image domains, e.g., paintings.

While these works were mostly a carve-and-copy of existing methods to the new application of visual affect recognition, the pendulum swung in the other direction when in [Machajdik and Hanbury, 2010], psychology and art theory inspired features were proposed. The work proposed specialized color features, composition indicators like the rule-of-thirds, facial features, and more to aid visual emotion detection. And this idea was extended several times, including in [Shin and Kim, 2010] using color composition, in [Wang *et al.*, 2013b] with color aesthetics, and in [Zhao *et al.*, 2014] where elements like artistic balance, emphasis and harmony were leveraged. Some recent works have also studied the multimodal integration of these features for visual sentiment analysis [Soleymani *et al.*, 2016].

As attribute learning and mid-level feature representations began to rise in popularity, e.g., ObjectBank [Li *et al.*, 2010] and Classemes [Torresani *et al.*, 2010], several mid-level features began to be proposed for visual affect detection. Early work in [Wang *et al.*, 2006] that went largely noticed by the community proposed 12 mid-level adjective-adjective word pairs like *warm-cool*, *brilliant-gloomy* and *vibrant-desolute*. In [Yuan *et al.*, 2013], 102 mid-level features called Sentribute were proposed for sentiment which were simply derived from scene attributes in the SUN Attribute dataset [Patterson and Hays, 2012], i.e., semantic attributes like *still water*, *ice* and *hiking* were included. In [Borth *et al.*, 2013a; Borth *et al.*, 2013b], a bank of visual classifiers forming mid-level representations were proposed called SentiBank. The representation consisted of a set of 1,200 linear SVM outputs where SVMs were trained using a taxonomy of adjective-noun pairs (ANPs). The ANPs combine a "noun" for visual detectability and an "adjective" for affective modulation of the noun, resulting in pairs like *cute dog*, *beautiful sunset*, *disgusting food* and *terrible accident*. Some image localization for a subset of these ANPs was also later proposed [Chen *et al.*, 2014b]. Though the authors were not aware they did so, this was one of the first works to depart from classical 'induced affect' line of works and investigate 'intended affect,' albeit in a weak label setting, and we show later, in Part II, how such mid-level features can be scaled even further.

Moving into the modern day, with the success of convolutional neural networks (CNNs) [LeCun *et al.*, 1998], spurred on by the performance of AlexNet [Krizhevsky *et al.*, 2012] on image classification in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [Russakovsky *et al.*, 2015], there has been a rush to apply CNNs, and more broadly neural networks (NNs), in a wide range of other applications. For visual affect detection, [Xu *et al.*, 2014] used AlexNet simply as a feature extractor [Razavian *et al.*, 2014] with SVMs and logistic regression classifiers, while in [You *et al.*, 2014], AlexNet was fine-tuned, both works for sentiment detection on social media images. And in [Chen *et al.*, 2014a], a network called DeepSentiBank was proposed that also fine-tuned from AlexNet but was trained to recognize the ANP visual concepts from [Borth *et al.*, 2013b]. Also, though not using deep networks, some work in [Wang *et al.*, 2015c] explored unsupervised emotion detection in social images.

It is interesting to note that many of the above works focused specifically on the sentiment prediction task, and mostly, from constrained and small image datasets. For video multimedia, [Teixeira *et al.*, 2011] experimented with early- and late-fusion of audio and visual features with Bayesian networks to detect viewer affective states in a VAD space representation using professional movie clips. In [Canini *et al.*, 2013], they defined "film grammar" features for movie recommendation that corresponded to scene distinctives capturing a director's storytelling style. In [Ellis *et al.*, 2014a], a person-specific pilot sentiment analysis study was performed on broadcast video news streams they recorded and mined from cable and over-the-air television [Jou *et al.*, 2013; Li *et al.*, 2013]. In [Jiang *et al.*, 2014], audio, visual and mid-level features were combined using kernel-level fusion with χ^2 radial basis function (RBF) SVM for recognizing semantic emotions in unconstrained user-generated videos. And in [Pang and Ngo, 2015], extending from [Srivastava and Salakhutdinov, 2012], a multimodal deep Boltzmann machine was proposed that trained on the same audio, visual and mid-level features and dataset from [Jiang *et al.*, 2014].

2.4 Big Affective Computing

The majority of seminal visual affect detection worked within the bounds of small, controlled datasets and were sometimes repurposed psychology studies not intended for computational learning. Among the earliest of these was the International Affective Picture System (IAPS) which consisted of a psychology-curated set of about 1,000 still images [Lang *et al.*, 1997]. Originally, these images were intended as stimuli to induce emotions in subject for measuring physiological signals. The Geneva Affective PicturE Database (GAPED) [Dan-Glauser and Scherer, 2011] consisted of 730 pictures meant to supplement IAPS and tried to narrow visual themes across images. This small-scale phenomenon is not limited to early works; for example, in the DeepSent [You *et al.*, 2014] dataset, only about 1,270 images from a social media platforms are available for sentiment prediction.

One reason for this scaling limitation was discussed earlier in §2.2.1, namely that the subjective nature of affect can make it difficult to conceptually tease apart noise from ground truth. However, another reason for this historical ceiling on scaling up visual affect datasets is that affect can break the assumption of *visual consistency*. If the former limitation of affect subjectivity can be thought of as increasing label variance or noise, then the visual inconsistency property of affect increases the data variance or noise. Consider that we have a corpus of images labeled 'sad,' even if we considered the generating process for these labels to be trustworthy, the images could range drastically from pictures of a 'graveyard' to a 'wilting flower' to a seemingly arbitrary 'building' that evokes personal memories. This raises significant difficulties compared to traditional vision tasks of recognizing a 'laptop,' particular dog breed, or even more complex categories like a 'birthday party' because there is no guarantee that visual elements are shared even within the same categorical label.

In this thesis, we advocate a movement toward what we call "Big Affective Computing". In large-scale computing problems, increasingly popularized under the category of "Big Data" problems, it is becoming common to differentiate between different kinds of "scale". One partitioning of different "scales" is known as the "Four V's": *volume*, *variety*, *velocity* and *veracity*.[5] Many of the scaling limitations mentioned with visual affect so far can

[5]This is only one partitioning and many remark that other categories are missing like *cost* and *variability*.

actually be subsumed in this paradigm as well, e.g., *visual inconsistency* with *veracity*. And some initial works like [Siersdorfer *et al.*, 2010; Soleymani *et al.*, 2012; Borth *et al.*, 2013b; Baveye *et al.*, 2015b; Vandal *et al.*, 2015] have already begun to scale up visual affect detection in *volume*, yet as visual data and computation becomes more accessible, it will become necessary to also consider other types of "scale" in Visual Affective Computing.

However, one other "V" often discussed is *value* which we have already mentioned in §1.1.1.

Part I

Content-driven Visual Affect Detection

Chapter 3

Perceived Emotion Prediction in Animated Image Sequences

Animated GIFs are everywhere on the Web. In this chapter, we present an exercise in the computational prediction of emotions perceived by viewers after they are shown animated Graphical Interchange Format (GIF) images [Jou *et al.*, 2014]. We evaluate our results on a dataset of over 3,800 animated GIF images gathered from a social Web platform called GIFGIF, each with scores for 17 discrete semantic emotions aggregated from over 2.5M user annotations – the first computational evaluation of its kind for content-based prediction on animated GIFs to our knowledge. One of our objectives was to systematically compare different types of content features for emotion prediction, including low-level, aesthetics, mid-level semantic and face expression features. As a part of this, we propose and advocate a conceptual paradigm in affect prediction that shows that delineating distinct types of affect is important and is useful to be concrete about the affective target (q.v. §2.2 and Fig. 2.2). We also formulated a multitask regression problem to evaluate whether viewer perceived emotion prediction can benefit from jointly learning across emotion classes compared to disjoint, independent learning.

3.1 Introduction

Animated Graphical Interchange Format (GIF) images are a largely unexplored media in Computer Vision and Multimedia research. In the pre-2000s, animated GIFs garnered a niche popularity, largely on the platforms of Internet forums and blogs, where common elements to animate include burning fires or waving flags. After falling out of popularity for a period, in the recent decade, the format became massively popular again with even more wide-spread use in the Web 2.0 era. Animated GIFs, usually characterized by relatively short sequences of image frames, have quickly become a channel for visually expressing emotion in our modern society.[1] Their role in popular culture has contributed to the rise of widely, rapidly spread cultural references called *memes* as well as new art forms, and some businesses even employ GIFs as a part of their Web marketing.

The use of animated GIFs particularly for conveying emotions is prevalent on the Web today, and GIFs are now massively found on forums, message boards, social media and websites of every genre. In moving to large-scale, many learning problems for other media suffer the *sampling problem* given the enormity of the Web, that is, the range of Web content available is so vast that it is often difficult to determine feasible sampling schemes. To counteract this, most works simply settle for a single domain, which, in turn, simply results in a "domain transfer" problem where a learned mapping $\mathcal{H}$ trained in one vertical context is not useful in another. Fundamentally, this problem branches from the wide range of uses of specific media, e.g., videos are as commonly used for education and documentaries as they are for car chases and explosions in a movie, likewise for still images and audio streams, etc. While these still are worthwhile media to explore (some of which we will, in fact, explore in later chapters), it is also important for the field not to ignore media with narrower popular applications that are no less massively used. Animated GIFs are one such media, but their almost exclusive use as emotional expression tools on the Web are precisely what gives us confidence that nearly any GIF we sample from the Web today will have some affective bias, making them in powerful media context to study and perform Affective Computing. In this chapter, we focus on the computational prediction, or recognition, of emotions perceived

[1] E.g., see `http://www.pbs.org/video/2207348428`

by viewers in animated GIF images, the first study of its kind to the best of our knowledge [Jou *et al.*, 2014].

The key contributions of this chapter include: (1) the first modern work to formulate a Computer Vision task using animated GIFs to the best of our knowledge, (2) the introduction of the paradigm of different types of affect oracles for Affective Computing, while explicitly advocating for perceived affect, the first Affective Computing work explicitly for this emotion type to the best of our knowledge, and (3) the prediction of 17 discrete multi-label semantic emotions perceived by viewers using multitask learning on an animated GIFs dataset created using annotations from over 2.5 million users.

3.2 Related Work

In [Borth *et al.*, 2013b], an ontology of visual concepts are proposed which model a semantic structure for mid-level visual affective detection (highlighted previously also in §2.3.2). These concepts were mined from the social multimedia platform, Flickr, where users uploaded images along with tag metadata to describe their content. Although not acknowledged by the authors in [Borth *et al.*, 2013b], because of the way that SentiBank [Borth *et al.*, 2013a] was trained, the representation actually *describes* an uploader's or originator's *intended* emotion, not emotion in general; this we argue in fact is a type of *perceived* emotion, which we expound on later. Recent work in [Chen *et al.*, 2014c] studied the metadata of images and used a text-based model to extract "publisher affect concepts" and "viewer affect concepts," but tries to divide emotion along the axis of human roles, e.g., publishers and viewers, but does not acknowledge that emotion can also be divided along how they arise in those persons, e.g., intended, perceived and induced. We set a more ambitious goal of predicting viewer perceived emotion using content-based methods where such metadata and comments are not available, as is often the case. In addition, while nearly all the works highlighted in §2.3.2 use independently trained classifiers for detecting emotion. We use multitask regression to learn multiple regressors jointly across emotions, resulting in a fast, low-cost linear projection at test time.

There are several existing public datasets available to the community for recognizing

emotion in visual media, but nearly all are focused on induced emotion to best of our knowledge. The IAPS dataset [Lang *et al.*, 1997] discussed in §2.4 falls into this category of induced emotions. The DEAP dataset [Koelstra *et al.*, 2011] consists of 40 one-minute excerpts from music videos and spontaneous physiological signals from 32 subjects. The MAHNOB-HCI Tagging dataset [Soleymani *et al.*, 2012] has 20 short film clips with ∼27 subjects with physiological signals like EEG and audio recordings of the subjects. FilmStim [Schaefer *et al.*, 2010] and LIRIS-ACCEDE [Baveye *et al.*, 2015b] are two other datasets consisting also of short film clips, 64 and 9,800 clips, respectively, but use scores and rankings by participant ratings instead of strict classifications. And in [Jiang *et al.*, 2014], they study discrete emotions from 1,101 user-generated videos, but labels data using 10 annotators following an unspecified "detailed definition of each emotion". Meanwhile, we present baselines on a dataset of animated GIFs annotated by over 2.5M users where each example has soft labels on 17 discrete semantic perceived emotions, i.e., instead of one-hot labels, we have multi-labeled examples.

3.3 The Case for Perceived Affect Detection

Here, we make an explicit distinction among different types of affect and justify the importance of perceived emotion in affective study, expanding on the discussion in §2.2.2. Affective Computing today is largely rooted in the task of recognizing "induced" human emotion [James, 1883; Dietz and Lang, 1999]. Much of research in this direction follows a canonical process of presenting stimuli to human subjects, measuring their physiological signals (including facial, voice and body expression), and then manually or computationally analyzing the outcome, e.g., [Haag *et al.*, 2004; Lisetti and Nasoz, 2004; Chanel *et al.*, 2005; Zeng *et al.*, 2009; Calvo and D'Mello, 2010; Gunes and Pantic, 2010; Fleureau *et al.*, 2012; Marsella and Gratch, 2014; Ellis *et al.*, 2014b; Vandal *et al.*, 2015]. Subjects are often asked to complete a survey to qualitatively or quantitatively describe their emotions to the stimuli post hoc [Schaefer *et al.*, 2010; Koelstra *et al.*, 2011; Soleymani *et al.*, 2012]. A key issue with this approach is it assumes there is only one type of emotion to study when, in fact, *not all affect is induced*. While traditional Affective Computing has assumed a single emotion

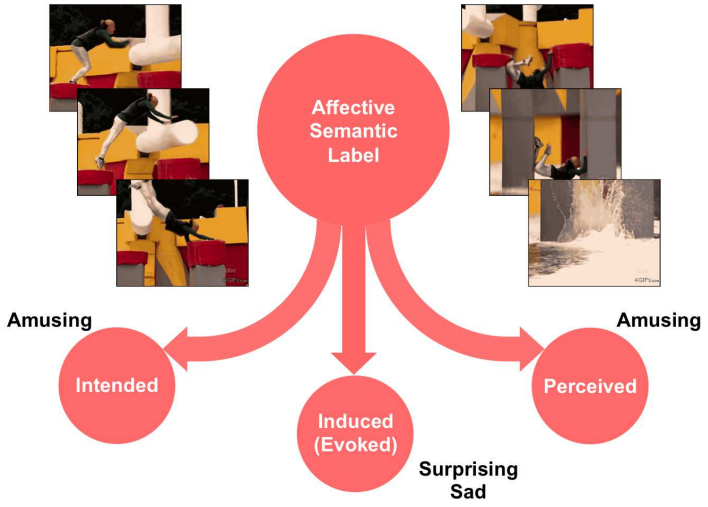


Figure 3.1: Illustration of Different Affect Oracles. The image sequence shows a game show contestant jumping an obstacle only to fall after an attempt. Such a sequence may have originally been cut and shared for the purpose of evoking amusement (*intent*), but actually resulted in recipients thinking the originator had a cruel sense of humor since they actually pitied the contestant (*induced*). Even so, they understood that the originator intended it for humor (*perceived*).

oracle as the target, we propose that there are in fact, many different types of emotion oracles, or equivalently, that the single affect oracle can be partitioned.

Perceived affections are an important phenomena to study because they are more concrete and objective than the more commonly studied induced affections, where labels are less reliable due to their subjectivity [Hochschild, 1983; Barrett, 2006]. In addition, computationally recognizing what an artist or director or author intended for the audience to affect is often challenging since such labels likely do not exist or are difficult to acquire. For example, it is unlikely to have access to what a YouTube[2] uploader meant for their viewers to feel about a video they uploaded, and even less likely to have that information at the a scene, shot or frame level. One illustration of this is shown in Figure 3.1. And in fact, in psychology, perceived emotions have indeed been experimentally been shown to be different [Gabrielsson, 2002; Kallinen and Ravaja, 2006], and yet, for computing, there are no works

[2] `https://www.youtube.com`

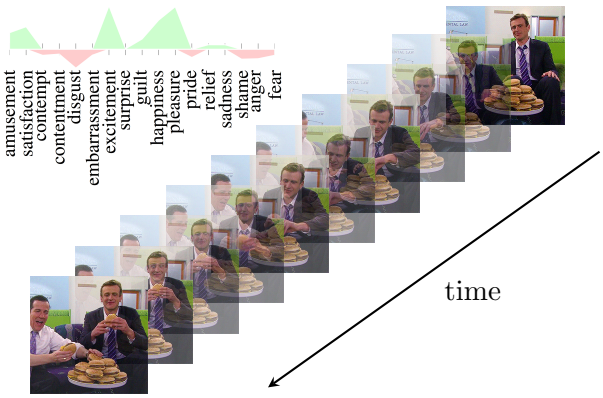


Figure 3.2: Example Animated GIF Image with Emotion Scores. Ten frames of an animated GIF sequence are shown with its respective emotion scores from user annotation on the upper-left. Scores that reflect positive presence of emotions are shown in green and negative are shown in red. The diversity of emotions illustrates the various types of emotions involved in multimedia interactions, e.g., intended, perceived and induced.

to our knowledge that have yet adopted this explicit distinction.

In summary, perceived affect or emotions are *cognitive assessments about non-cognitive entities*, i.e., affections. And as a result, since many of our modern large-scale machine learning problems today rely on datasets with easy-to-acquire *cognitive* assessments, or labels, focusing on 'perceived affect' invariably enables large-scale datasets for Affective Computing while simultaneously reducing subjectivity and label noise in the learning problem. Certainly though, it is worth noting that this is not to claim that 'subjectivity' is not an important area of study, even computationally, but when seeking to scale Affective Computing, we believe it is critical to understand subjectivity as a entity that we *can* trade-off on rather than an on-off switch that defines our research field.

3.4 Perceived Emotions in Animated GIF Images

In Figure 3.2, we show another example of how there may be different categories of emotion, or affect. The example animated GIF shows a scene in the American sitcom "How I Met Your Mother" where two characters, one in white, one in black, are seated in front of a

large pile of hamburgers. As they dig into this pile of burgers, the character dressed in black makes a wide-eyed expression, while the character in white makes a wide-mouthed expression (e.g., especially in the second frame shown). Perhaps the author of this image sequence designed it for a setting where they *intended* to make their audience feel 'disgusted' at gluttony of these two characters. What we show at the top-left of Figure 3.2 though is that viewers can *perceive* that the GIF expresses 'pleasure,' 'happiness,' 'excitement,' and 'satisfaction' – that is, they cognitively understand that the image sequence portrays those emotions. Ultimately though, despite what they perceive, they may actually feel (or, be *induced* with) something different – 'guilt' or 'shame' for having done something similar before or, for others, a sense of 'amusement' because they find the GIF to be funny. Additionally noteworthy from both Figure 3.1 and 3.2 is that emotions like 'amusement,' 'excitement' and 'surprise' are not necessarily mutually exclusive and are actually affectively related, indicating that such correlations should be accounted for.

In this chapter [Jou *et al.*, 2014], we aim to evaluate features and computational models for predicting perceived emotions of GIFs. Specifically, we compare features of different types, including low-level features like color histograms, aesthetics, mid-level semantic features inspired by emotion modeling and face features. Also, in view of the relatedness of multiple emotions, we apply and evaluate multitask learning to model multiple emotion tasks jointly. In Figure 3.3, we provide an overview of our experimental setup for performing studying perceived emotions in animated GIFs. Given over 3,800 animated GIFs and emotional scores along 17 perceived semantic emotions, we seek to computationally predict these scores in a regression setting.

3.5 GIFGIF Dataset

We gathered data from a website created by human-computer interaction researchers at the MIT Media Lab called GIFGIF[3]. The researchers originally designed the website using GIFs to understand human social interactions through the popular media form of GIFs gathered

[3] `http://gifgif.media.mit.edu` or `http://www.gif.gf`

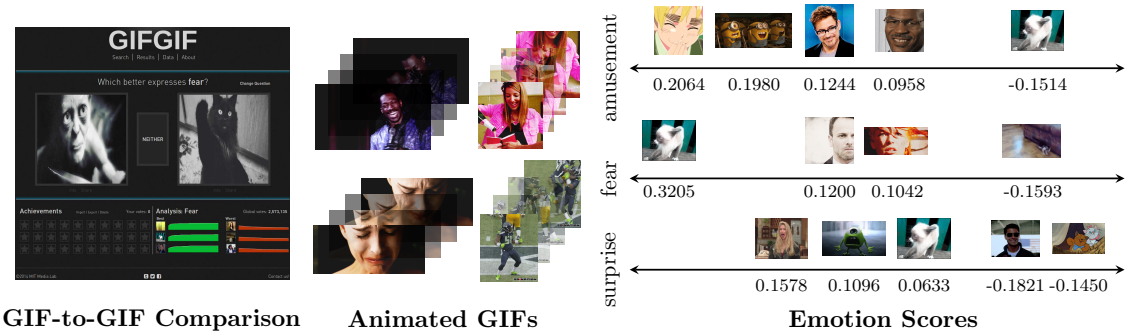


Figure 3.3: GIFGIF Dataset. **Left:** Screenshot of the GIFGIF interface designed at the MIT Media Lab (circa April 2014). Users are asked to vote which GIF best expresses an emotion. **Center:** Several example animated GIFs from the dataset collected from GIFGIF. **Right:** Last frame of example GIFs for three emotions along with their emotion scores, shown sorted by their score on a [-1,1] scale where 0 is neutral.

from Giphy[4], and thus were not focused on the computational emotion prediction task. We collected 3,858 GIFs on April 29, 2014 from GIFGIF along with their crowdsourced annotations. The GIFs, which have at most 303 frames each, are sourced from a large variety of domains from films, television shows, cartoons/animations, sports, video games, advertisements, user generated content, and user-edited content. The GIFs span a wide range of camera angles, illumination, special effects, humans and non-humans, black/white and color, zooming, resolution, and some original content from as early as the 1930s. Also, since animations and cartoons are generally considered as a separate media, we note briefly that there are 860 total of such GIFs in GIFGIF (∼22.3%).

As shown in Figure 3.3 (left), the GIFGIF website presents users with a pair of GIFs and asks "Which better expresses X?" where X is one of 17 emotions inspired by [Ekman, 1999]: amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, happiness, pleasure, pride, relief, sadness, satisfaction, shame and surprise. Users answer with the GIF they perceive expresses the emotion best or select neither. This particular question of "expression" is an effort to capture emotions users *perceive* in the content rather than what users "feel" after seeing the GIF, which would reflect *induced* emotion.

[4] `https://giphy.com`

This simply modification in the annotation process indicates how drastically different affect oracles can sometimes be targeted without even knowing. Likewise, it points to how simply we can gather 'perceived affect' labels by more carefully designing annotation questions.

At the time of our data collection, GIFGIF aggregated over 2.5M user annotations to produce a 17-dimensional vector for each GIF containing a score between 0 and 50 for each emotion where 25 is neutral. GIFGIF calculates these scores using the TrueSkill algorithm [Herbrich *et al.*, 2006], a method originally designed for ranking video game players given the outcome of a game. Here, GIFs replace players and GIFGIF users provide outcomes of GIF pair match-ups. The range [0, 50] comes from heuristic parameters in the TrueSkill algorithm that define a mean performance $\mu = 25$ and uncertainty $\sigma = 25/3$ per player or GIF. In all our experiments and analysis, we normalize this range to [-1, 1] for convenience. We note that we are only 'allowed' to aggregate these annotations because we are specifically targeting perceived emotions since they are cognitive assessments of affects. The GIFGIF labels are unique in that each GIF has soft labels for each emotion whereas other datasets [Baveye *et al.*, 2015b; Jiang *et al.*, 2014; Koelstra *et al.*, 2011; Schaefer *et al.*, 2010; Soleymani *et al.*, 2012] only have one categorical emotion assignment per image or video, i.e., single label versus multi-label.

3.6 Multitask Emotion Regression

Classical independent, or single-task, learning ignores the potential for classes to be related – for example, in emotion detection, the sensation of 'surprise' is not necessarily orthogonal to the feeling of 'fear'. In multitask learning (MTL) [Caruana, 1997; Zhou *et al.*, 2012], as illustrated in Figure 3.4, we may have multiple inputs per task where each may differ in number of samples such that we have inputs $\mathbf{X} \in \mathbb{R}^{n_i \times d}$ where n_i is the number of samples for the i^{th} task and we have d features, the same follows likewise for the label vectors $\mathbf{Y} \in \mathbb{R}^{n_i}$. The goal then is to learn a weight matrix $\mathbf{W} \in \mathbb{R}^{d \times t}$ consisting of t tasks via the optimization $\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) + \Omega(\mathbf{W})$, where $\mathcal{L}(\mathbf{W})$ is the empirical training loss and $\Omega(\mathbf{W})$ is a regularization encoding task-relatedness and the prediction output is $y = \mathbf{x}^T \cdot \mathbf{W}_i$.

In terms of the label space, the simplest case of MTL to consider is when each task

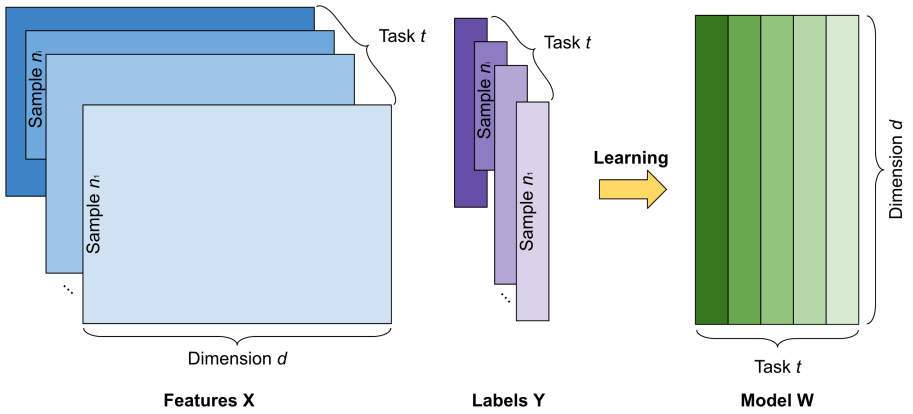


Figure 3.4: Multitask Learning. Adapted from [Zhou *et al.*, 2012].

is binary, i.e., we have t binary prediction tasks. Unlike in single-task learning where each learner would be trained per task to produce a *single* binary prediction model, or a traditional multiclass approach of predicting multiple classes with a *single* model but without cross-class relationship modeling, multitask learning jointly learns across all tasks and produces *multiple* models specialized to each task but with cross-task information integrated. In addition, unlike multiclass learning, MTL is naturally well suited to scale when the prediction tasks are not binary or even when tasks have different ultimate usages, e.g., one task may be reconstruction while another may be classification and still another for ranking.

In our context, as shown in Figure 3.5, we focus on the application of MTL, and specifically, multitask regression (MTR), to emotion prediction where we learn to predict over 17 binary tasks where "tasks" correspond to discrete semantic emotion classes from [Ekman, 1999]. The best of our knowledge, we are the first to apply MTR as an application toward emotion prediction to explicitly treat affect classes as correlated entities that can contribute to each other. To model the relationship between emotions in $\mathbf{W}$, one approach is to constrain the regressors for different emotions to share a low-dimensional subspace. Formally, this means we seek a low-rank weight matrix $\mathbf{W}$ and need to solve the rank minimization

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) + \lambda \cdot \text{rank}(\mathbf{W}), \tag{3.1}$$

i.e., so we regularize over the rank of $\mathbf{W}$ so that $\Omega(\mathbf{W}) = \text{rank}(\mathbf{W})$. This problem is

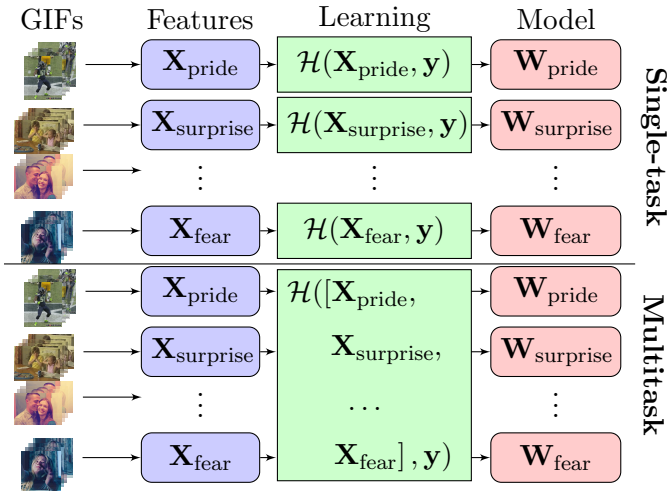


Figure 3.5: Single-task and Multitask Learning for GIF Emotion Detection. **Top:** Single-task learning pipeline for semantic emotion recognition where learner $\mathcal{H}$ is learned independently for each emotion with training labels $\mathbf{y}$. **Bottom:** Multitask learning jointly learns over all emotion classes yielding multiple models.

NP-hard in general [Vandenberghe and Boyd, 1996], and a popular solution is to minimize the trace (or nuclear) norm $\|\cdot\|_*$ instead of the rank function. Trace norm regularization [Ji and Ye, 2009] in multitask learning for a data matrix $\mathbf{X}$ takes the form

$$\min_{\mathbf{W}} \sum_{i=1}^{t} \left\|\mathbf{W}_i^T \mathbf{X}_i - \mathbf{y}_i\right\|_F^2 + \rho_1 \left\|\mathbf{W}\right\|_*, \tag{3.2}$$

where a least squares loss is used for $\mathcal{L}(\mathbf{W})$.

3.6.1 Feature Representations of Emotional Animated GIFs

To better understand what features aid the prediction of viewer perceived emotion, we used four image feature representations, each chosen for their previous use or connection to emotion recognition in visual content in prior art:

- **Color Histograms:** We compute frame-level color histograms in HSV color space for its classical use in vision and affective computing.
- **Face Expression:** Facial expressions in GIFs can impact how viewers perceive conveyed emotions. Note that these are facial expressions of characters in the con-

tent, not subjects responding to the GIF. We use a convolutional neural network (CNN) with top-level one-vs-all SVMs [Cortes and Vapnik, 1995] with squared hinge losses which achieved the best validation performance of ∼69.4% in a public evaluation [Tang, 2013]. We use an implementation by [Tang, 2013] which trained on 28,708 48×48 "in the wild" black/white face images over the seven Ekman emotions [Ekman, 1999]: angry, disgust, fear, happy, sad, surprise and neutral. We perform face detection using a Haar-like cascade [Lienhart and Maydt, 2002; Viola and Jones, 2001] as implemented in OpenCV[5], resize as needed and apply the facial expression recognition on the largest face for a six-dimensional vector of SVM decision outputs as a feature.

- **Image-based Aesthetics:** A number of works have shown that emotion has some intrinsic correlation with visual aesthetics [Datta *et al.*, 2006a; Datta *et al.*, 2006b; Machajdik and Hanbury, 2010; Jia *et al.*, 2012; Bhattacharya *et al.*, 2013]. We use a subset of image-based aesthetic features described in [Bhattacharya *et al.*, 2013]. We divide the GIF frames into 3×3 cells from which cell-level statistics are computed including the dark channel, luminosity, sharpness, symmetry, white balance, colorfulness, color harmony and eye sensitivity. The normalized area of the dominant object and distances from the dominant object's centroid to grid-line intersections are also computed in a per-frame basis. Together these form a 149-dimensional feature vector for each frame.

- **SentiBank:** We also use the mid-level visual representation composed of visual sentiment detectors, SentiBank [Borth *et al.*, 2013b]. Again, the representation consists of a 1,200-dimensional vector of linear SVM score outputs over the presence of adjective-noun pair concepts. SentiBank has been shown to work well in various affective applications like in [Borth *et al.*, 2013a; Bhattacharya *et al.*, 2013; Jiang *et al.*, 2014; Chen *et al.*, 2014b; Chen *et al.*, 2015; Gelli *et al.*, 2015; Szekely *et al.*, 2015]. Intuitively, we might expect that adjective-noun pair detections like *scared cat* or *scary horror* indicate semantic emotions like 'surprise' or 'fear'.

[5] `http://opencv.org`

3.6.2 Animated GIF Emotion Regression Experiments

We sample ten equally spaced frames for each GIF, or less for shorter GIFs, and apply the emotion labels from GIFGIF's TrueSkill algorithm as weakly supervised labels $\mathbf{y}$ to each of the frames. We ran our experiments over five random repetitions of a 20/80% train/test ratio, find regularization parameters using cross-validation, and report results on the test set where regression outputs are computed by averaging frame-level scores for GIFs in each emotion category. We adopt the normalized mean squared error (nMSE) used in previous studies [Argyriou *et al.*, 2008; Chen *et al.*, 2011] for our experiments. The nMSE is defined as the mean squared error (MSE) divided by the variance of the target vector and assures that the error is not biased toward models that over or under predict.

We compare our approach to classical linear regression with ordinary least squares (OLS), i.e., no weighting, and logistic regression. We note that methods like support vector regression [Cortes and Vapnik, 1995] could have been used in place of multiple linear regression, but they would still be learning emotion models independently. Also, despite the pairwise comparisons in the GIFGIF annotation, methods incorporating pairwise constraints like RankSVM [Joachims, 2003] or relative attributes [Parikh and Grauman, 2011] could not be used, at least directly, because only TrueSkill score outputs were exposed by GIFGIF, not pairwise comparison outcomes.

	Ordinary Least Squares	**Logistic Regression**	**Multitask Regression**
Color Histogram	1.7398 ± 0.1868	1.6618 ± 0.2991	1.4641 ± 0.1935
Face Expression	0.8925 ± 0.0036	0.9130 ± 0.0030	0.8955 ± 0.0024
Aesthetics	1.0440 ± 0.0133	1.0571 ± 0.0116	1.0361 ± 0.0093
SentiBank	1.5694 ± 0.0614	1.4944 ± 0.0593	2.2901 ± 0.1981

Table 3.1: Perceived Emotion Prediction on GIFGIF. The normalized mean squared error (nMSE) is reported across 17 emotions over five random repetitions of a 20/80% train/test split for color histograms, face expression, aesthetics and SentiBank features with OLS linear regression, logistic regression, and low-rank multitask regression. Lower nMSE indicates better performance.

From Table 3.1, we see that color histograms expectedly performs poorly since emo-

tion is more complex than can be captured by color. The aesthetics feature, though not designed specifically for emotion recognition, still does encode some information related to the perceptual emotion and achieved the second best nMSE. Surprisingly, the SentiBank feature which is tailored as a mid-level semantic feature for aspects of emotion does not perform well. This may be explained by the fact that there is a cross-domain issue from the training image set that SentiBank used to the GIFGIF dataset, and that SentiBank has a conservative F-score of 0.6 for its detectors [Borth *et al.*, 2013b]. It is possible that features like SentiBank would benefit from some kind of (group) sparsity for feature selection in the multitask regression. Of all features, the face expression feature performs the best, where we note that the training label average was predicted when no faces were detected. The possible simple intuition is that humans express and perceive much of their emotions through their faces, partially confirmed by some other work in popularity trends in social media [Bakhshi *et al.*, 2014].

Overall, we consistently observed that the best performing emotion was 'happiness' followed by 'amusement' for all regressors using face expression features, and the worst performing emotion was also consistently 'embarrassment'. Inspecting the GIFs in the 'embarrassment' category, we found that this emotion was heavily dominated by sequences where a person or cartoon would hide their faces with their hands or another object, or would look down hiding their faces via their pose. In these cases, as one would expect, face detection fails because of occlusions. We believe that the 'embarrassment' emotion would benefit most from gesture recognition as many of the occlusions are due to hand movement. The good performance on the 'happiness' and 'amusement' emotions are unsurprising as both emotions are visually expressed with smiles and laughter. However, due the subtle differences between these two emotions, we also believe that this is also one of the reasons why multitask learning sometimes performs marginally worse due to ambiguities like this in the model. We believe it maybe important for future multitask emotion models to also regularize on the similarity and not just the dissimilarity of emotion tasks.

3.7 Conclusions

In this chapter, we showed that there are different delineations of emotions and presented a computational approach to predicting viewer perceived emotions on a dataset of animated GIF images. Additionally, we showed that emotions need not be decoupled from each other by presenting a multitask regression approach for jointly learning over 17 discrete semantic emotions. Facial expression features proved to be particularly useful and applicable to the setting of perceived emotion detection in GIFs in comparison to some other traditional vision features. It is important to note that this may only be true in the context of animated GIF images and there were some categories of emotion like 'embarrassment' that facial expression recognition performed poorly on.

In the future, following recent work by [Martínez *et al.*, 2014], instead of regressing or classifying animated GIFs over emotion categories, it would be interesting explore learning to rank GIFs. This would allow us to make relative affect statements that, for example, GIF `A` is funnier than GIF `B` but not more than `C`. In addition, given the looping nature of animated GIFs, we would like to explore using recurrent neural networks (RNNs) [Hochreiter and Schmidhuber, 1997; Gers *et al.*, 2002], perhaps in conjunction with multitask learning and attention, to improve emotion detection and even develop affective trends over time. Also, since our work [Jou *et al.*, 2014], work on even larger datasets like [Li *et al.*, 2016b] have been proposed, so it would be interesting to benchmark affect detection tasks on such datasets as well.

Part II

Mid-level Representations for Visual Affect

Chapter 4

Multicultural Visual Affective Computing

Every culture is unique. In this chapter, we continue unpacking the Big Affective Computing tenant that deals with the multi-faceted or multimodal nature of visual affect data (i.e., "variety") while also scaling up the "volume" and "veracity" of the Visual Affective Computing problem (q.v. §2.4). We seek to provide a computational view into the uniqueness of culture along the lines of *language* and *geography* in relation to human sentiment and emotion, and how they manifest in visual social multimedia [Jou *et al.*, 2015; Jou *et al.*, 2016; Liu *et al.*, 2016; Pappas *et al.*, 2016].

As we have seen several times in this thesis already, one major challenge with existing visual affect computing approaches is that the *context* within which affect is elicited, felt and perceived is often not properly considered (q.v. §2.2 and Fig. 2.2). For example, it is often assumed that visual affect is a universal entity, and factors like culture are not often taken into account. Here, we explicitly tackle computational visual affect understanding from a multi-cultural perspective. In particular, we believe language and regional localization are "lens" through which we can observe these cultural differences. We developed a large-scale "in the wild" image corpus on a popular, multilingual social multimedia platform to study these phenomena.

Following the trend in many other fields, the advent of high-volume and weakly-supervised

data are driving increased interest in *large-scale* sentiment studies in Affective Computing. However, directly studying visual affect in a dimensional representation (e.g., valence-arousal) or discrete semantics (e.g., [Ekman, 1999; Plutchik, 1980]) tend to suffer from the problem of data sparsity since such specialized psychology terminology are unlikely to be found in large volume from the Web. As a result, in this chapter, we use a mid-level semantic representation that serves a surrogate target that allows for visual affect to be explored at scale.

4.1 Introduction

If you scoured the world and took several people at random from major countries and asked them to fill in the blank "_____ love" in their native tongue, how many unique adjectives would you expect to find? Would people from some cultures tend to fill it with *twisted*, while others *pure* or *unconditional* or *false*? All over the world, we daily express our thoughts and feelings in culturally isolated contexts; and when we travel abroad, we know that to cross a physical border also means to cross into the unique behaviors and interactions of that people group – its cultural border. How similar or different are our sentiments and feelings from this other culture? Or the thoughts and objects we tend to talk about most? Motivated by questions like this, our work explores the computational understanding of human affect along cultural lines, with focus on visual content. In particular, we seek to answer the following important questions: (1) how are images in various cultures used to express affective visual concepts, e.g., *beautiful place* or *delicious food*? And (2) how are such affective visual concepts used to convey different emotions and sentiment across cultures?

We develop sets of sentiment- and emotion-polarized visual concepts using semantic structures called adjective-noun pairs (ANPs), originally introduced in [Borth *et al.*, 2013b], but extended into a multilingual context. We propose a new language-dependent method for automatic discovery of these adjective-noun constructs. We show how this pipeline can be applied on a social multimedia platform for the creation of a large-scale multilingual visual sentiment concept ontology (MVSO). Unlike the flat concept structure found in [Borth *et al.*, 2013b], our unified ontology is organized hierarchically by multilingual clus-

ters of visually detectable nouns and subclusters of emotionally biased versions of these nouns [Jou *et al.*, 2015; Pappas *et al.*, 2016]. In addition, we mine a large number of images from the Web using these visual concepts and associate them to geographical data [Jou *et al.*, 2016]. A new, publicly available dataset of over 15.6K sentiment-biased visual concepts across 12 languages with language-specific detector banks, over 7.36M images and their metadata with 1.7M geo-references is released as a part of this effort to understand visual affect multiculturally.

We make the following contributions in this chapter: (1) a principled, context-aware pipeline for designing a multilingual visual sentiment concept ontology (MVSO), (2) an incarnation of a multilingual visual sentiment concept ontology mined from social multimedia data end-to-end across 12 languages, (3) a MVSO of 15.6K visual concepts organized hierarchically into noun-based clusters and sentiment-biased adjective-noun pair subclusters, (4) a regional localization of associated images yielding over 1.7M geo-references covering 237 countries, (5) two interactive and fluid browsing visualization systems called **Complura** and **SentiCart** for ontology and geodata navigation, respectively, and (6) the release of a dataset containing MVSO and large-scale collection of 7.3M images with dense metadata.[1]

4.2 Related Work

In psychology, there are two major schools-of-thought on the connection between cultural context and human affect. Some believe emotion to be culture-specific [McCarthy, 1994; Jack *et al.*, 2012], that is, emotion is dependent on one's cultural context, while others believe emotion to be universal [Ekman, 1999; Haselton and Ketelaar, 2006], that is, emotion and culture are independent mechanisms. For example, while this thesis is written in English, there are emotion words/phrases in other languages for which there is no exact translation in English [Lomas, 2016], e.g., *Schadenfreude* in German refers to pleasure at someone else's expense. Do English-speakers not feel those same emotions or do they simply refer to them in a different way? Or even if the semantic reference is the same, perhaps the underlying emotion is different?

[1]The MVSO project webpage and data are available at `http://mvso.cs.columbia.edu`.

The study of emotions across culture has long been a topic of research in psychology. A main contention [Ekman *et al.*, 1987] concerns whether emotions are culture-specific [McCarthy, 1994; Jack *et al.*, 2012], i.e., their perception and elicitation varies with the context, or universal [Ekman, 1999; Haselton and Ketelaar, 2006]. In [Russell, 1991], a survey of cross-cultural work on semantics surrounding emotion elicitation and perception was presented, showing that there are still competing views as to whether emotion is pan-cultural, culture-specific, or some hybrid of both. Inspired by research in this domain, we are the first to investigate the relationship between visual affect and culture from a multimedia-driven and computational perspective, as far as we know [Jou *et al.*, 2015; Jou *et al.*, 2016; Liu *et al.*, 2016; Pappas *et al.*, 2016].

Other work in cross-lingual research comes from text sentiment analysis and music information retrieval. In [Bautin *et al.*, 2008] and [Mihalcea *et al.*, 2007], they developed multilingual methods for international text sentiment analysis in online blogs and news articles, respectively; and in [Brooke *et al.*, 2009] and [Wan, 2009], they studied online product reviews. In [Lee *et al.*, 2005] and [Hu and Yang, 2014], they presented approaches to indexing digital music libraries with music from multiple languages. Specific to emotion, [Hu and Yang, 2014] tried to highlight differences between languages by building models for predicting the musical mood and then cross-predicting in other languages. Unlike these works, we propose a multimedia-driven approach for cross-cultural visual sentiment analysis in the context of online image collections.

One close relative to the work in this chapter is [Vandal *et al.*, 2015] where a large-scale collection of over 1.5M facial expression videos sourcing from over 94 countries was presented. Subjects were shown one of about 8K online videos as stimuli and their reactions were captured via a web camera. Although there is a wide variety in originating locations of subjects, this study focuses more on region-based differences rather than cultural differences. In addition, [Vandal *et al.*, 2015] also relies on *evoked* sentiment in facial expression videos while we use weakly supervised semantic cues for *intended* sentiment.

Given that we build our work on the adjective-noun pair (ANP) construct, it is also worth mentioning that work in pure vision like [Divvala *et al.*, 2014] for learning attributes across different object categories has some similarities to our own work, although they do not focus

Figure 4.1: Example images from "around the world" organized by affective visual concepts. The top set shows images of the *old market* concept from three different cultures/languages; and the bottom, images of *good food*. Even though the conceptual reference is the same, each culture's sentimental expression of these concepts may be adversely different.

on affectively biased attributes. As a result, attributes studied in [Divvala *et al.*, 2014] can include verbs and adjectives as well as form word pairs that are named entities, whereas we take definitive steps for filtering such cases (q.v. §4.3.2) in order to form affectively biased word pairs that have a one-word object grounding via the 'noun' and sentimental component through the 'adjective'. Also, attribute ambiguity and so-called "schools of thought" have been investigated in works like [Kovashka and Grauman, 2015] for subjectivity attachment to products like shoes which have a flavor of affective detection for specific noun verticals.

4.3 Multilingual Visual Sentiment Ontology (MVSO)

Recall that in Affective Computing, we often refer to the *affective gap* as the conceptual divide between the low-level visual stimuli, like images and features, and the high-level, abstracted semantics of human affect, e.g., *happy* or *sad* (q.v. §1.1.2). In one attempt to bridge sentiment and visual media, [Borth *et al.*, 2013b] developed a visual sentiment ontology (VSO), a set of 1,200 mid-level concepts using structured semantics called adjective-noun pairs (ANPs). The noun portion of the ANP allows for computer vision detectability and

the adjective serves to polarize the noun toward a positive or negative sentiment, or emotion, e.g., so instead of having visual concepts like *sky* or *dog*, we have *beautiful sky* or *scary dog*. Most prior Visual Affective Computing works built algorithms, models and datasets on single languages with the assumption that emotions are universal. However, while such works provide great research contributions in their target language, their applicability and generalization to other languages and cultures remains largely unexplored. Meanwhile, we present a large-scale multilingual visual sentiment concept ontology (MVSO) and dataset including adjective-noun pairs from 12 languages of diverse origins.

It is important to distinguish our work from that of visual sentiment ontology (VSO)[2] [Borth *et al.*, 2013b] and its associated detector bank, SentiBank [Borth *et al.*, 2013a]. The proposed adjective-noun pair mid-level representation approach has proven effective in a wide range of applications in emotion prediction [Jou *et al.*, 2014], social media commenting [Chen *et al.*, 2014c], etc. However, in addition to lack of multilingual support, there are several technical challenges with VSO [Borth *et al.*, 2013a; Borth *et al.*, 2013b] that we seek to improve on via (1) detection of adjectives and nouns with language-specific part-of-speech taggers, as opposed to a fixed list of adjectives and nouns, (2) automatic discovery of adjective-noun pairs correlated with emotions, as opposed to "constructed" pairs from top frequent adjectives and nouns, and (3) stronger selection criterion based on image tag frequency, linguistic and semantic filters and crowdsource validation. Our proposed MVSO [Jou *et al.*, 2015] discovery method can be easily extended to any language, while achieving greater coverage and diversity than VSO [Borth *et al.*, 2013b].

4.3.1 Adjective-Noun Pair Discovery

An overview of the proposed method for multilingual visual sentiment concept ontology construction is shown in Figure 4.2. In the first stage, we obtain a set of images and their tags using seed emotion keyword queries, selected according to emotion ontologies from psychology such as [Plutchik, 1980] or [Ekman, 1999]. Next, each image tag is labeled automatically by a language-specific part-of-speech tagger and adjective-noun combinations are discovered from words in the tags. Then, the combinations are filtered based on lan-

[2] `https://visual-sentiment-ontology.appspot.com`

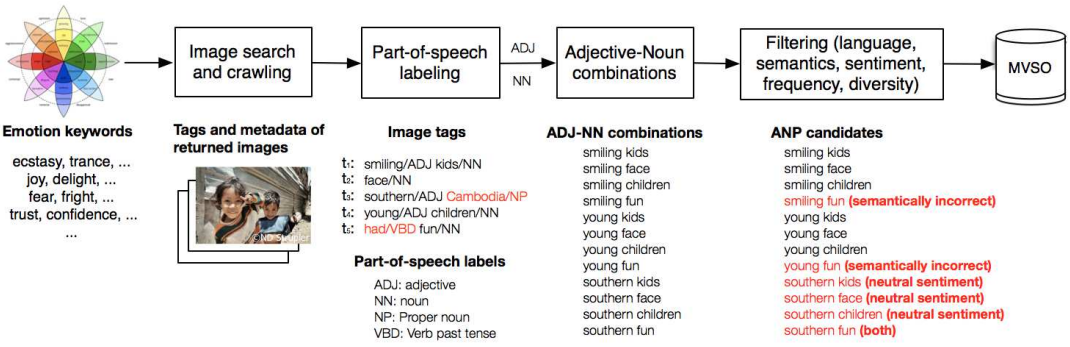


Figure 4.2: The construction process of our multilingual visual sentiment concept ontology (MVSO) begins with crawling images and metadata based on emotion keywords. Image tags $(t_1, \ldots, t_5)$ are labeled with part-of-speech tags, and adjectives and nouns are used to form candidate adjective-noun pair (ANP) combinations, while others are ignored (in red). Finally, these candidate ANPs are filtered based on various criteria which help remove incorrect pairs (in red), forming a final MVSO with diversity and coverage.

guage, semantics, sentiment, frequency and diversity filters to ensure that the final set of ANPs have the following properties: (a) are written in the target language, (b) they do not refer to named entities or technical terms, (c) reflect a non-neutral sentiment, (d) are frequently used, and (e) are used by a non-trivial number of users of the target language.

The discovery of affective visual concepts for these languages using adjective-noun pairs poses several challenges in lexical, structural and semantic ambiguities, which are well-known problems in natural language processing. Lexical ambiguity is when a word has multiple meanings which depend on the context, e.g., *sport jaguar* or *forest jaguar*. Structural ambiguity is when a word might have different grammatical interpretation depending on the position in the context, e.g., *ambient light* or *light room*. Semantic ambiguity is when a combination of words with the same syntactic structure have different semantic interpretation, e.g., *big apple*. We selected languages in our MVSO according to the availability of public natural language processing tools and sentiment ontologies per language so that automatic processing was feasible. In addition, we sought to cover a wide range of geographic regions from the Americas to Europe and to Asia (q.v. §4.6). We settled on 12 diverse languages: Arabic, Chinese, Dutch, English, French, German, Italian, Persian,

Polish, Russian, Spanish and Turkish.

We applied our proposed data collection pipeline to a popular social multimedia sharing platform, Yahoo! Flickr[3], and collected public data from November 2014 to February 2015 using the Flickr API[4]. We selected Flickr because there is an existing body of multimedia research using it in the past, and in particular, [Jin *et al.*, 2010] describes how Flickr satisfies two conditions for making use of the "wisdom of the social multimedia": popularity and availability. We do not repeat the argument in [Jin *et al.*, 2010], but note that in addition to those benefits, Flickr has multilingual support and the use of Flickr facilitates a natural comparison to the seminal VSO [Borth *et al.*, 2013b] work.

For our seed keywords, we selected the *Plutchik's Wheel of Emotions* [Plutchik, 1980]. This psychology ontology was selected because it consists of graded intensities for multiple basic emotions providing a richer set of emotional valences compared to alternatives like [Ekman, 1999] and allows for better comparison to VSO [Borth *et al.*, 2013b]. It is worth noting that the ultimate ontology is generally applicable to other emotional semantics like Ekman's emotions though since our seed queries get expanded in the process. As illustrated also in Figure 2.1c, the Plutchik emotions are organized by eight basic emotions, each with three valences: ecstasy > joy > serenity; admiration > trust > acceptance; terror > fear > apprehension; amazement > surprise > distraction; grief > sadness > pensiveness; loathing > disgust > boredom; rage > anger > annoyance; and, vigilance > anticipation > interest.

4.3.1.1 Multilingual Query Construction

To obtain seeds for each language, we recruited 12 native and proficient language speakers to provide a set of translated or synonymous keywords to those of the 24 Plutchik emotions [Plutchik, 1980]. Speakers were allowed to use any number of keywords per emotion since the possible synonyms per emotion and language can vary, but they were asked to rank their chosen keywords along each emotion seed. They were also allowed to use tools like Google Translate[5] or other resources to enrich their emotion keywords. Table 4.2 lists top

[3] `https://www.flickr.com`

[4] `https://www.flickr.com/services/api`

[5] `https://translate.google.com`

	#Images	#Tags	#Candidates	#ANPs (final)
Arabic	116,125	958,435	15,532	29
Chinese	895,398	3,919,161	50,459	504
Dutch	260,093	4,929,581	1,045,290	348
English	1,082,760	26,266,484	2,073,839	4,421
French	866,166	22,713,978	1,515,607	2,349
German	528,454	10,525,403	854,100	804
Italian	548,134	10,425,139	1,324,076	3,349
Persian	128,546	1,304,613	103,609	15
Polish	294,821	5,261,940	141,889	70
Russian	60,108	1,518,882	30,593	129
Spanish	827,396	15,241,679	925,975	3,381
Turkish	332,609	4,717,389	73,797	231
Totals	5,940,610	107,782,684	8,154,766	15,630

Table 4.1: Ontology refinement statistics over 12 languages. Beginning with many images from seed emotion keywords denoted by (#Images), we extracted tags from these images (#Tags), and performed adjective-noun pair (ANP) discovery for candidate combinations (#Candidates). Through a series of filters – frequency, language, semantics, sentiment and diversity – and after crowdsourcing, we get our final visual sentiment concepts #ANPs.

ranked keywords according to speakers for 7 out of 12 languages in each emotion.

Given the set of keywords $E^{(l)} = \{e_{ij}^{(l)} \mid i = 1 \dots 24, j = 1 \dots n_i\}$ describing each emotion i per language l, where n_i is the number of keywords per emotion i, we performed tag-based queries on tags with the Flickr API to retrieve images and their related tags. Like [Borth *et al.*, 2013b], for each emotion, we chose to sample only the top 50K images ranked by Flickr relevance to simply limit the size of our results, but if an emotion had less than 50K images, we extended the search to additional metadata, i.e., title and description.

4.3.1.2 Part-of-speech Labeling & ANP Discovery

To identify the type of each word in a Flickr tag, we performed automatic part-of-speech (POS) labeling using pre-trained language-specific taggers which achieve high accuracy (over

English	joy	trust	fear	surprise	sadness	disgust	anger	anticipation
Spanish	alegría	confianza	miedo	sorpresa	tristeza	asco	ira	previsión
Italian	gioia	fiducia	paura	sorpresa	tristezza	disgusto	rabbia	anticipazione
French	bonheur	confiance	peur	surprise	tristesse	dégoût	colère	prévision
German	Freude	Vertrauen	Angst	Überraschung	Traurigkeit	Empörung	Ärger	Vorfreude
Chinese	歡樂	信任	害怕	震驚	悲	討厭	憤怒	預期
Dutch	vreugde	vertrouwen	angst	verrassing	verdriet	walging	woede	anticipatie

Table 4.2: Most representative keywords according to native/proficient speakers for eight basic emotions and for 7 of our 12 languages, chosen and shown top-to-bottom in decreasing number of discovered visual affect concepts, or adjective-noun pairs.

95% for most languages), in particular, we used TreeTagger [Schmid, 1994], Stanford tagger [Toutanova *et al.*, 2003], HunPos tagger [Halácsy *et al.*, 2007] and morphological analyzer for Turkish [Güngördü and Oflazer, 1994]. Though not all the tags contained multiple words, the average number of words was always greater than the average number of tags for all languages, so word context was almost always taken into account. From the full set of part-of-speech labels, we retained identified nouns, adjectives and other part-of-speech types which can be used as adjectives, such as simple or past participle (e.g., *smiling face*) in English.

We then based our discovery strategy for ANPs on co-occurrence in image tags, that is, if an adjective-noun pair is relevant to the specific emotion it should appear at least once as that exact pair phrase in the crawled images for that emotion. To validate the completeness of our strategy we compared with VSO and found that ∼86% of ANPs discovered by VSO [Borth *et al.*, 2013b] overlap with the English ANPs discovered by our method.

4.3.2 Filtering Candidate Adjective-Noun Pairs

From these discovered ANPs, we applied several filters to ensure they satisfied the following criteria: (a) written in the target language, (b) do not refer to named entities, (c) reflect a non-neutral sentiment, (d) frequently used and (e) used by multiple speakers of the language.

4.3.2.1 Language and Semantics

We used a combination of language dictionaries[6] instead of language classifiers to verify the correctness of the ANP as the performance of using the latter was low for short-length text, especially for Romance languages which share characters. All of the English ANPs were classified as indeed English by the dictionary, while for other languages, ANPs were removed if they passed the English dictionary filter but not the target language dictionary. The intuition for this was that most candidate ANPs in other languages were mixed mostly with English. We removed candidate pairs which referred to named entities or technical terms, where named entities were detected using several public knowledge bases such as Wikipedia and dictionaries for names[7], cities, regions and countries[8], and technical terms were removed via a manually created list of words specific to our source domain, Flickr, containing photography-related (e.g., *macro*, *exposure*) and camera-related words (e.g., *DSLR*, *Canon*, *lens*).

4.3.2.2 Non-neutral Sentiment

To filter out neutral candidate adjective-noun pairs, each ANP was scored in sentiment using two publicly available sentiment ontologies: SentiStrength [Thelwall *et al.*, 2010] and SentiWordNet [Baccianella *et al.*, 2010]. The SentiStrength ontology supported all the languages we considered, but since SentiWordNet could only be used directly for English, we passed in automatic translations in English from all other languages to it, following previous research on multilingual sentiment analysis in machine translation [Banea *et al.*, 2008; Balahur and Turchi, 2012].[9] We computed the ANP sentiment score $S(anp) \in [-2, +2]$ as:

$$S(anp) = \begin{cases} S(a) & : sgn\{S(a)\} \neq sgn\{S(n)\} \\ S(a) + S(n) & : \text{otherwise} \end{cases} \tag{4.1}$$

[6] `http://www.winedt.org`

[7] `https://www.wikipedia.org` and `https://www.ssa.gov`, respectively.

[8] `http://geobytes.com`

[9] For four non-English languages with the highest ANP counts, we verified a small percentage of non-neutral ANPs (less than 2%) reverse sentiment polarity after translation, confirming similar observations in the previous work.

where $S(a) \in [-1, +1]$ and $S(n) \in [-1, +1]$ are the sentiment scores of the individual adjective and noun words, respectively, each of which are given by the arithmetic mean of SentiStrength and SentiWordNet scores on the word, and *sgn* is the sign of the scores. The piecewise condition essentially says that if the signs of the sentiment scores of the adjective and noun differ, then we ignore the noun. This highlights our belief that adjectives are the dominant sentiment modifiers in an adjective-noun pair, so for example, even if a noun is positive, like *wedding*, an adjective such as *horrible* would completely change the sentiment of the combined pair. And so, for these sign mismatch cases, we chose the adjective's sentiment alone. In the other case, when the sign of the adjective and noun were the same, whether both positive (e.g., *happy wedding*) or both negative (e.g., *scary spider*), we simply allowed the ANP sentiment score to be the unweighted sum of its parts. ANP candidates with zero sentiment score were filtered out.

4.3.2.3 Frequency

Desirable adjective-noun pairs are those which are actually used in colloquial interactions together. Here, we loosely define an ANP's "frequency" of usage as its number of occurrences as an image tag on Flickr. When computing counts for each pair, we accounted for language-specific syntax like the ordering of adjectives and nouns. Following anthropology research [Dryer and Haspelmath, 2013], we followed two dominant orderings (91.5% of the languages worldwide): adj-noun and noun-adj. We also "merged" simplified and traditional forms in Chinese by considering them to be from the same language pool but distinct characters sets. In addition, we considered the possible intermediate Chinese character 的 during our frequency counting. For all non-English languages, we retained all ANPs that occurred at least once as an image tag; but for English, since Flickr's most dominant number of users are English-speaking, we set a higher frequency threshold of 40.

4.3.2.4 Diversity

The frequency of an adjective-noun pair's occurrence alone was not sufficient to ensure a pair's pervasive use in a language. We also checked if the ANP was used by a non-trivial number of distinct Flickr users for a given language. We identified the number of users

contributing to uploads of images for each ANP and found a power law distribution in every language. To avoid this uploader bias, we removed all ANPs with less than three uploaders. Many removed candidate pairs came from companies and merchants for advertising and branding, who dominated certain image tags. Some power law trends for uploader bias are shown in Figure 4.3 per language.

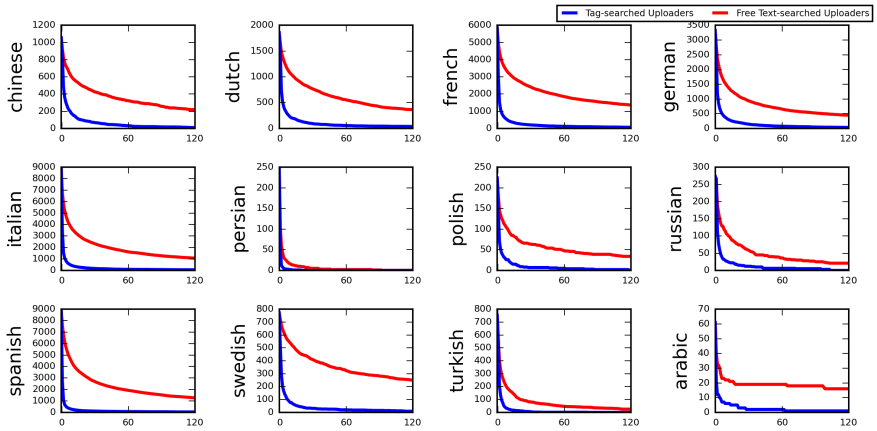


Figure 4.3: Uploader bias by language with #ANPs versus #uploaders, i.e., such that #ANPs with $\geq k$ uploader(s) is shown for $\forall\, k \in [1, 120]$.

To further ensure diversity in our MVSO, we subsampled nouns in every language by limiting to the 100 most frequent ANPs per adjective so that we do not have, for example, the adjective *surprising* modifying every possible noun in our corpus. In addition, we performed stem unification by checking and including only the inflected form (e.g., singular/plural) of an ANP that was most popular in usage as a tag on Flickr. This unification also filtered some candidate ANPs as some "duplicates" were present but simply in different inflected forms.

4.3.3 Crowdsourcing Validation

A further inspection of the corpus after the automatic filtering process showed that some issues could not completely be solved in an automatic fashion. Common errors included many fundamental natural language processing challenges like confusions in named entity recognition (e.g., *big apple*), language mixing (e.g., adjective in English + noun in

Turkish), grammar inconsistency (e.g., adj-adj, or verb-noun) and semantic incongruity (e.g., *happy happiness*). So to refine our multilingual visual sentiment concept ontology, we crowdsourced a validation task. For each language, we asked native speaking workers to evaluate the correctness of ANPs post automatic filtering. We collected judgements using CrowdFlower[10], a crowdsourcing platform that distributes small tasks to a large number of workers, where we limited workers by their language expertise.

We required that each ANP was evaluated by at least three independent workers. To ensure high quality results, we also required workers to be (1) native speakers of the language, for which CrowdFlower had its own language competency and expertise test for workers, and (2) have a good reputation according to the crowdsourcing platform, measured by workers' performance on other annotation jobs. For whatever reason, for three languages (Persian, Polish and Dutch), the CrowdFlower platform did not evaluate workers based on their language expertise, so we filtered them by provenience, selecting the countries according to the official language spoken (e.g., Netherlands, Belgium, Aruba and Suriname for Dutch).

4.3.3.1 Crowdsourcing Task Interface

The verification task for workers consisted of simply evaluating the correctness of adjective-noun pairs. At the top of each page, we gave a short summary of the job and tasked workers: "*Verify that a word pair in* `<Language>` *is a valid adjective-noun pair.*" Workers were provided with a detailed definition of what an adjective-noun pair is and a summary of the criteria for evaluating ANPs, i.e., it (1) is grammatically correct (adjective + noun), (2) shows language consistency, (3) shows generality, that is, commonly used and does not refer to a named entity, and (4) is semantically logical. To guide workers, examples of correct and incorrect ANPs were provided for each criteria, where these ground truths were carefully judged and selected by four independent expert annotators. In the annotation interface, aside from instructions, workers were shown five ANPs and simply chose between "yes" or "no" to validate ANPs.

[10] `http://www.crowdflower.com`

	#Candidates	#Users	#Countries	%Correct	%Agree
Arabic	81	10	7	0.57	0.90
Chinese	1055	56	24	0.63	0.83
Dutch	1874	45	2	0.23	0.92
English	5369	223	52	0.78	0.84
French	5840	152	37	0.43	0.86
German	3360	119	27	0.32	0.90
Italian	4996	216	42	0.57	0.88
Persian	65	6	6	0.37	0.86
Polish	159	6	1	0.52	0.93
Russian	294	13	3	0.70	0.89
Spanish	4992	190	30	0.70	0.89
Turkish	701	61	22	0.66	0.84

Table 4.3: Crowdsourcing results via number of input candidate ANPs (#Candidates), users (#Users), countries (#Countries), and percentage of ANPs accepted (%Correct) and annotator agreement (%Agree).

4.3.3.2 Crowdsourcing Quality Control

Like some other crowdsourcing platforms, CrowdFlower provided a quality control mechanism called *test questions* to evaluate and track the performance of workers. These test questions come from pre-annotated ground truth, which in our case, correspond to ANPs with binary validation decisions for correctness. To access our task at all, workers were first required to correctly answer at least seven out of ten such test questions. In addition though, worker performance was tracked throughout the course of the task where these test questions were randomly inserted at certain points, disguised as normal units. For each language, we asked language experts to select ten correct and ten incorrect adjective-noun pairs from each language corpus to serve as the test questions.

4.3.3.3 Crowdsourcing Results

To measure the quality of our crowdsourcing, we looked at the annotator agreement along each validation task. For all languages, the agreement was very strong with an average

annotator agreement of 87%, where workers agreed on either the correctness or incorrectness of ANPs. We found that workers tended to agree more that ANPs were correct than that they were incorrect. This was likely due to the wide range of possible criteria for rejecting an ANP where some criteria are easy to evaluate (e.g., language consistency), while others, such as general usage versus named entity, may cause disagreement among users due to the cultural background of the worker. For example, not all workers may agree that an ANP like *big eyes* or *big apple* refers to a named entity. However, for languages where the agreement on the incorrect ANPs was high, namely Arabic, German, and Polish, the average annotator agreement as a percentage of all ANPs for that language were greater than 90%.

On average, our crowdsourcing validated that a vast number of the input candidate ANPs from our automatic ANP discovery and filtering process were indeed correct ANPs. English, Spanish and Russian were the top three for which the automatic pipeline performed the best, where every three in five ANPs were approved by the crowd judgements. However, for certain languages, including German, Dutch, Persian and French, the number of ANPs rejected by the crowd was actually greater than accepted ANPs due to a higher occurrence of mixed language pairs, e.g., *witzig humor*. In Table 4.3, we summarize statistics from our crowdsourcing experiments according to the number of ANPs, percentage of correct/incorrect ANPs by worker majority vote, and average agreement.

4.3.4 Ontology-structured Image Mining

Having acquired a final set of adjective-noun pairs for each of the 12 languages, we downloaded images by querying the Flickr API with ANPs using a mix of tag and metadata search. To limit the size of our dataset, we downloaded no more than 1,000 images per ANP query and also enforced a limit of no more than 20 images from any given uploader on Flickr for increased visual diversity. The selected 1,000 images were selected from the pool of retrieved image tag search results, but in the event that this pool is less than 1,000, we also enlarged the pool to include searches on the image title and description, or metadata. Selections from the pool of results were always randomized and a small number of images which Flickr or uploaders removed or changed privacy settings midway were removed. In

total, we downloaded 7,368,364 images across 15,630 ANPs for the 12 languages, where English (4,049,507), Spanish (1,417,781) and Italian (845,664) contributed the most images. In Chapter 5, we will show how this large-scale image corpus is used to train multilingual visual concept detector models.

4.4 Ontology Analysis and Statistics

Here, we briefly provide a discussion on the difference between our MVSO [Jou *et al.*, 2015] and VSO [Borth *et al.*, 2013b] as well as discuss sentiment and emotion distributions in the multilingual ontology.

4.4.1 Comparison with Other Visual Sentiment Ontologies

To verify the efficacy of our MVSO, we provide a comparison of our extracted English visual sentiment ontology with that of VSO [Borth *et al.*, 2013b] along dimensions of size (number of ANPs) and diversity of nouns and adjectives (Figure 4.4). In Figure 4.4a, the overlap of English MVSO with VSO is compared with VSO alone after applying all filtering criteria except subsampling which might exclude ANPs belonging to VSO. As mentioned previously, about 86% overlaps between them (2,304 out of 2,681 ANPs when $t > 0$). As we vary a frequency threshold t (as described in §4.3.2) over image tag counts, the overlap converges to 100%. This confirms that the popular ANPs covered by VSO are also covered by MVSO, an interesting finding given the difference in the crawling time periods and approaches. In Figure 4.4b, we show that there are far greater number of ANPs in our English MVSO compared to VSO ANPs throughout all the possible values of frequency threshold, after applying all filtering criteria. Similarly, as shown in Figure 4.4c, given there are more adjectives and nouns in our English MVSO, we also achieve greater diversity than VSO.

In Figure 4.4d, we compare the number of ANPs for the remaining languages in MVSO with VSO after applying all filtering criteria. The curves show that VSO has more ANPs than all the languages for most of the languages over all values of t, except from Spanish, Italian and French in the low values of t. Our intuition is that this is due to the popularity of English on Flickr compared to other languages. In Figures 4.4e and 4.4d, we observe that

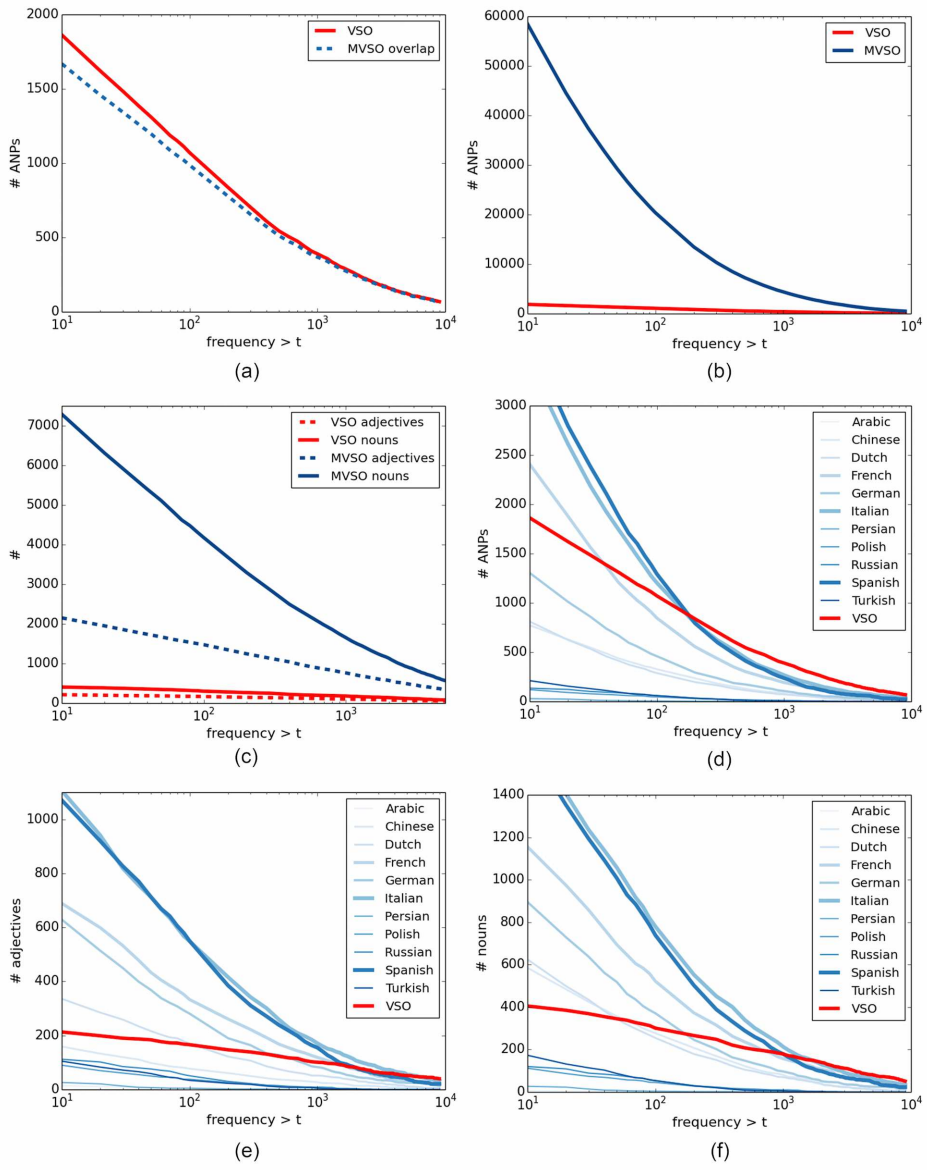


Figure 4.4: Comparison of our English MVSO and VSO in **(a)**, **(b)** and **(c)**, in terms of ANP overlap, number of ANPs, adjectives and nouns; and with all other languages in **(d)**, **(e)** and **(f)**, in terms of the number of ANPs, adjectives and nouns when varying the frequency threshold t from 0 to 10,000 (on log-scale), respectively.

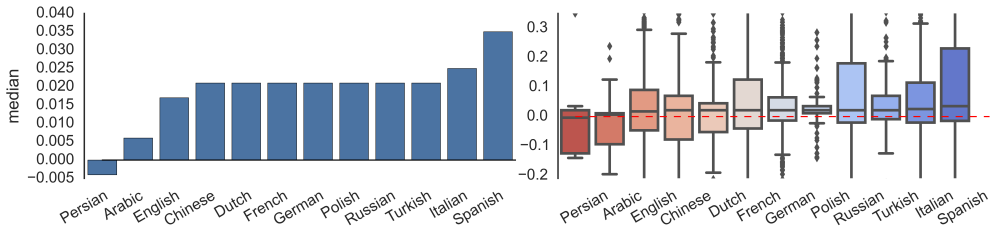


Figure 4.5: Median sentiment computed over all ANPs per language is shown on left, and the sentiment distribution using box plots on the right (zoomed at 90% of the distributions). On right, languages are sorted by median sentiment in ascending order (from the left).

these three languages have greater diversity of adjectives and nouns than VSO for $t \leq 10^3$, German and Dutch have greater diversity than VSO for smaller values of threshold t, while the rest of the languages have smaller diversity over most values of t.

4.4.2 Sentiment Distributions

Returning to our original research motivation from §4.1, an interesting question to ask is which languages tend to be more positive or negative in their visual content. To answer this, we computed the median sentiment value across all ANPs of each language and ranked languages as in Figure 4.5. Here, to take into account the popularity difference among ANPs, we replicated each ANP k times, with k equal to the number of images tagged with the ANP, up to an upper limit $L = \alpha \times \text{Avg}_i$, where Avg_i is the average image count per ANP in the ith language. Varying α value will result in different medians and distributions, but the trend in differentiating positive languages from negative ones was quite stable. We show the case when $\alpha = 3$ in Figure 4.5, indicating that there is an overall tendency toward positive sentiment across all languages, where Spanish demonstrates the highest positive sentiment, followed by Italian. This surprising observation is in fact compatible with some previous research showing that there is a universal positivity bias over languages with Spanish being the most relatively positive language [Dodds *et al.*, 2015]. The languages with the lowest sentiment were Persian and Arabic, followed by English.

The sentiment distributions (Figure 4.5: right) also showed interesting phenomena: Spanish being the most positive language also has the highest variation in sentiment, while

German has the most concentrated sentiment distribution. Even for languages that have the lowest median sentiment values, the range of sentiment was concentrated in a small range near zero (between 0 and -0.1).

4.4.3 Emotion Distributions

Another interesting question arises when considering co-occurrence of ANPs with the emotions in different languages. While our adjective-noun pair concepts were selected to be sentiment-biased, emotions still represent the root of our framework since we built MVSO out from seed emotion terms. So aside from sentiment, which focuses on only positivity/negativity, what are probable mappings of ANPs to emotions for each language? What emotions are most frequently occurring across languages? Given the set of keywords $E^{(l)} = \{e_{ij}^{(l)} \mid i = 1 \dots 24, j = 1 \dots n_i\}$ describing each emotion i per language l, where n_i is the number of keywords per emotion i, the set of ANPs belonging to language l, noted as $x \in X^{(l)}$, and the number of images tagged with both ANP x and emotion keyword e_{ij}, $C^{(x)} = \{c_{ij}^{(x)} \mid i = 1 \dots 24, j = 1 \dots n_i\}$, we define the probabilities of emotion for each ANP x in language l as:

$$\text{emo}^i(x) = \frac{\frac{1}{n_i}\sum_{j=1}^{n_i} c_{ij}^{(x)}}{\sum_{i=1}^{24}\frac{1}{n_i}\sum_{j=1}^{n_i} c_{ij}^{(x)}} \in [0,1] \tag{4.2}$$

Note the model in (4.2) does not take into account correlation among emotions, where for example, by an image tagged with "ecstasy," users may also imply "joy" even though the latter is not explicitly tagged. These correlations can be easily accounted for by smoothing co-occurrence counts c_{ij} over correlated emotions, e.g., the co-occurrence counts of an ANP tagged with "ecstasy" can be included partially in the co-occurrence count of "joy". Regardless, still based on (4.2), we compute a normalized emotion score per language l and emotion i as:

$$\text{score}^i(l) = \frac{\sum_{x=1}^{|X^{(l)}|}\text{emo}^i(x)\cdot\text{count}(x)}{\sum_{i=1}^{24}\sum_{x=1}^{|X^{(l)}|}\text{emo}^i(x)\cdot\text{count}(x)} \in [0,1] \tag{4.3}$$

In Figure 4.6, we show these scores per language and Plutchik emotion [Plutchik, 1980] on a heatmap diagram. Scores in each row sum to 1 (over 24 emotions). The emotions

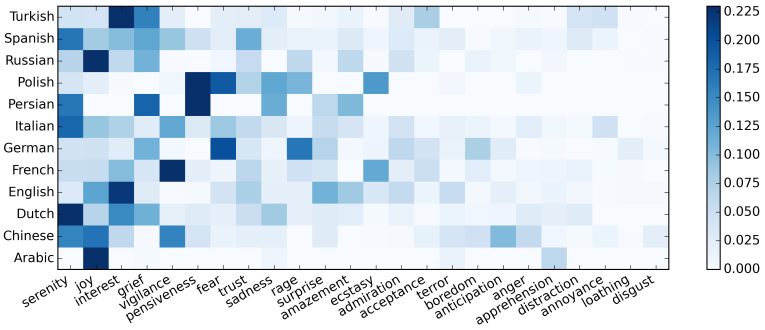


Figure 4.6: Probabilities of emotions per language with respect to their visual sentiment content. Emotions are ordered by the sum of their probabilities across languages (left to right) and clipped for better visualization. Each row sums to 1.

are ordered by the sum of their scores across languages. The top-5 emotions across all languages are *joy*, *serenity*, *interest*, *grief* and *fear*. And the highest ranked emotion is *joy* in Russian, Chinese and Arabic. Two other emotions in the top-5 were also positive: *serenity*, being high ranked emotion for Dutch, Italian, Chinese and Persian, and *interest* for English, Turkish and Dutch. The remaining two emotions in the top-5 were negative: *grief* for Persian and Turkish, and *fear*, which was high ranked in German and Polish. We also observed that *pensiveness* was top ranked for Persian and Polish, *vigilance* for French, *rage* for German, while *apprehension* and *distraction* for Spanish. We note that these results are more concrete for languages with many ANPs (>1000) and less conclusive for those with few ANPs like Arabic and Persian.

4.5 Cross-lingual Matching

To get a gauge of topics commonly mentioned across different cultures and languages, we analyzed alignments of translations for each ANP to English as a basis. Two approaches were taken to study this: exact and approximate alignment. An example cross-lingual connectivity mapping is shown in Figure 4.7[11], the generation process for which we now describe in detail. We ensured that translations of ANPs also passed all our validation

[11] Figure from co-author work in [Pappas *et al.*, 2016].

filters described in §4.3.2 for this analysis.

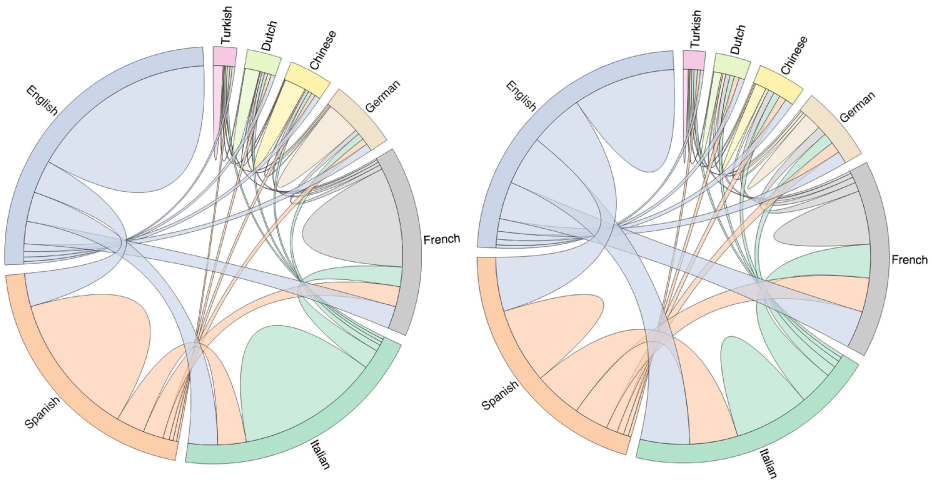


Figure 4.7: ANP clustering connectivity across the top-8 most popular languages in MVSO measured by the number of ANPs in the same cluster of a given language with other languages represented in a chord diagram. On the left **(a)**, the clusters based on exact matching are mostly dominated by a single language, while on the right **(b)**, based on approximate matching, the ANP clusters are enriched with multiple languages.

4.5.1 Exact Alignment

Here, we grouped ANPs from each language that have the exact same translation[12]. For example, *old books* was the translation for one or more ANPs from seven languages, including 老書 (Chinese), *livres anciens* (French), *vecchi libri* (Italian), Старые книги (Russian), *libros antiguos* (Spanish), *eski kitaplar* (Turkish). The translation covered by the greatest number of languages was *beautiful girl* with ANPs from ten languages. Figure 4.8 (left) shows a correlation matrix of the times ANPs from pairs of languages appeared together in a set with the exact same translation, e.g., out of all the translations that German ANPs were translated to (782), more were translated to the same phrase with the ANPs used by Dutch speakers (39) than with the ANPs used by Chinese speakers (23). This was striking given that there were less (340) translation phrases from Dutch than from Chinese (473).

[12] `https://cloud.google.com/translate`

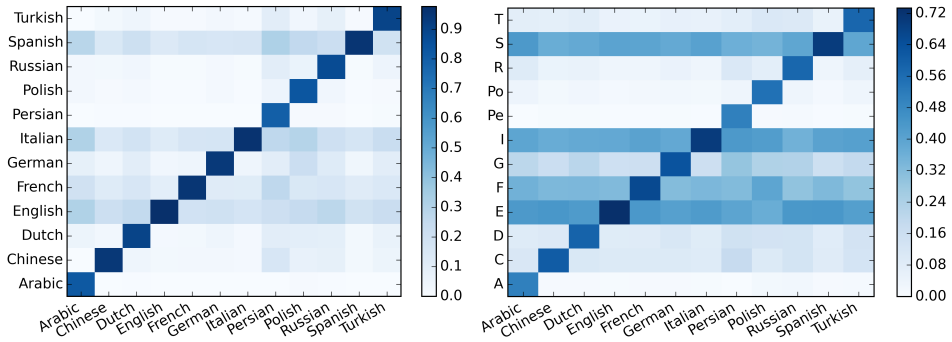


Figure 4.8: Percentage of times adjective-noun pairs from one language (columns) were translated to the same phrase **[Left]**, or to a phrase in the same cluster as in another language (rows) **[Right]**.

4.5.2 Approximate Alignment

Translations can be inaccurate, especially when capturing underlying semantics where context is not provided. And so, we relaxed the strict condition for exact matches by approximately matching using a hierarchical two-stage clustering approach instead. First, we extracted nouns using TreeTagger [Schmid, 1994] from the list of translated phrases and discovered 3,099 total nouns. We then extracted word2vec features [Mikolov *et al.*, 2013], a word representation trained on a Google News corpus, for these translated nouns (188 nouns were out-of-vocabulary), and performed k-means clustering ($k = 200$) to get groups of nouns with similar meaning [Pappas *et al.*, 2016]. The number of clusters was picked based on the coherence of clusters; and we picked the number where the inertia value of the clustering started saturating while gradually increasing k. In the second stage of our hierarchical clustering, we split phrases from the translations into different groups based on the clusters their nouns belonged to. We extracted word2vec [Mikolov *et al.*, 2013] features from the full translated phrase in each cluster and ran another round of k-means clustering (adjusting k based on the number of phrases in each cluster, where phrases in each noun-cluster ranged from 3 to 253). This two-stage clustering enables us to create a hierarchical organization of our ANPs across languages and form a multilingual ontology over visual sentiment concepts (MVSO) [Jou *et al.*, 2015], unlike the flat structure in VSO [Borth *et al.*, 2013b]. We discovered 3,329 sub-clusters of ANP concepts, e.g., resulting in

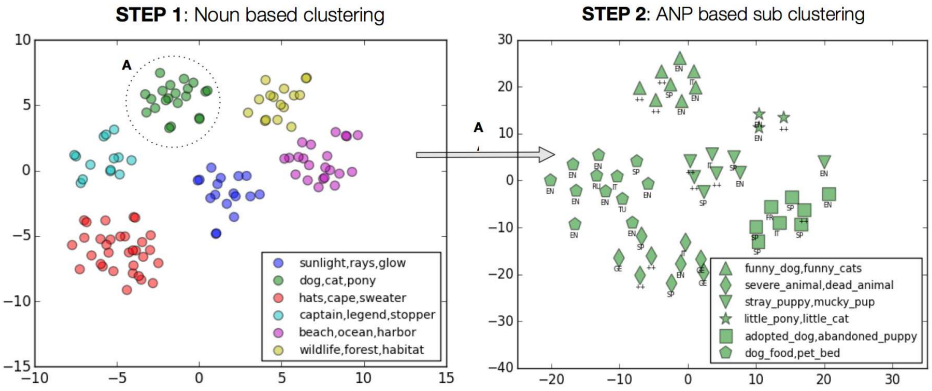


Figure 4.9: Examples of noun clusters **[Step 1]** and ANP sub-clusters **[Step 2]** from our two-stage clustering for cross-lingual matching. For visualization, word2vec vectors were projected to 2-D using t-SNE [van der Maaten and Hinton, 2008].

clusters containing *little pony* and *little horse* as in Figure 4.9. This approach also yielded a larger intersection between languages, where German and Dutch share 118 clusters, and German and Chinese intersect over 101 ANP clusters.

The correlation matrix from this approximate matching is shown in Figure 4.8, along with one subtree from our ontology by hierarchical clustering in Figure 4.9. We projected data to $\mathbb{R}^2$ using t-SNE dimensionality reduction [van der Maaten and Hinton, 2008] in the visualization. On the left, six clusters composed of different sets of nouns are shown with clusters of *sunlight-rays-glow* and *dog-cat-pony*. On the right, we show the sub-clustering of ANPs for the *dog-cat-pony* cluster in **A**, giving us noun groupings modified by sentiment-biasing adjectives to get ANPs like *funny dog-funny cats* and *adopted dog-abondoned puppy*.

4.6 Geographical Variety

The goal so far with the MVSO [Jou *et al.*, 2015] was to expand affective variety by introducing cultural diversity, accomplishing this via language diversity. However, language is only one of many signals to capture multicultural visual sentiment, and geographical data (geodata) is another signal not previously explored for visual sentiment understanding, to the best of our knowledge. Since any one language may be spoken across a multitude of countries and with varying density, to truly understand visual sentiment around the world,

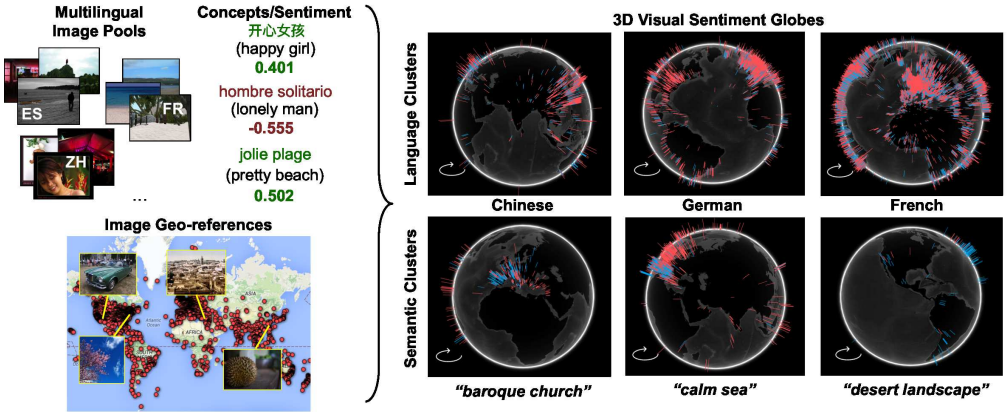


Figure 4.10: Example globe visualizations of geographical data using **SentiCart** from sentiment-biased images in MVSO from around the world by language and semantics.

geodata is a necessary complement to the language dimension. How geographically diverse are our sentiments in social multimedia? And specifically, how diverse (or localized) are our sentiments in the images and concepts we use everyday along linguistic and geographical lines? Motivated by such questions, we also augment MVSO by incorporating geographical data signals for multicultural affect. To assist the study of visual sentiment variety in geodata, we also developed a visualization system called **SentiCart** [Jou *et al.*, 2016] discussed later in §4.7.2 to chart out multilingual visual sentiment around the world.

Contextualizing social multimedia with explicit location or implicit geographical data has been an area of research for over a decade [Backstrom *et al.*, 2010; Luo *et al.*, 2011; Ji *et al.*, 2015]. Among the first for digital photography, [Toyama *et al.*, 2003] presented several frameworks for location tag acquisition, geo-referencing and image media browsing interfaces. In Computer Vision, [Zheng *et al.*, 2009] used geo-tagged photos to assist visual landmark detection over about 20M images. In [Hao *et al.*, 2011] and [Moore *et al.*, 2014], Twitter posts are used to study geographic sentiment and music preference differences, respectively, while in [Singh *et al.*, 2010], Twitter and Flickr posts are analyzed spatiotemporally to map out social visual interests. These geographical information systems (GIS) each lack a multilingual, visual grounding, or sentiment biasing component we seek here.

In order to collect geo-localization data for social photos, we use a combination of two multi-source methods – one with high reliability, but low coverage and another with

GPS	Image Title	Image Description	Geodata From?
✓	✓or ✗	✓or ✗	GPS
✗	✓	✓	Title
✗	✓	✗	Title
✗	✗	✓	Description
✗	✗	✗	<Omit>

Table 4.4: Geo-reference Data Sourcing. When GPS data are available, we always use it as our geo-localization for a given image regardless of user-provided metadata. Otherwise, we prefer locations in the image title over those in descriptions.

lower reliability, but higher coverage [Jou *et al.*, 2016]. We root our geodata collection and analysis on the same MVSO implementation discussed in §4.3 [Jou *et al.*, 2015].

4.6.1 GPS Coordinate Data

Global positioning system (GPS) geo-localization provides highly reliable latitude-longitude coordinates (usually within several meters worst-case depending on satellite-receiver precision) and may be encoded in image headers of some digital photos. In November 2015, we queried the Flickr API over the entire MVSO corpus [Jou *et al.*, 2015] of 7,368,364 images and acquired GPS coordinate data for 1,410,892 images. The remaining images were either not GPS-tagged or privacy permissions were not granted for public querying. The top three languages with GPS-tagged images in order are: Persian (32.24%), French (25.35%) and Italian (24.40%). Although the largest language by image count is English (4,049,507), it only had a 17.48% coverage in GPS-tagged images.

4.6.2 Metadata-inferred Location Data

In most social media platforms, including our setting on Flickr, users can and do often provide image title and descriptions to add additional context for their media posts. Locations are commonly found in these user metadata because they provided a concrete grounding for *where* an event or memory took place. For geo-localization though, this data can often be very uninformative, e.g., user input text like "our new house," as well as arbitrary, e.g., "Lit-

Language	#ANPs	#Georefs	Language	#ANPs	#Georefs
Arabic	22	99	Italian	3,184	206,315
Chinese	395	11,553	Persian	10	92
Dutch	315	16,292	Polish	67	3,873
English	4,407	707,846	Russian	95	2,014
French	2,241	163,193	Spanish	3,241	259,138
German	717	38,544	Turkish	137	1,933

Table 4.5: Number of GPS-based geo-references (georefs) collected from MVSO images by language. Since not every visual concept (ANP) had images with GPS data, we also show the number of remaining ANPs with at least one geo-referenced image.

tle Rock" could refer to a literal small rock in the image content or a town in Arkansas, USA. These two streams of user text from title and description offer greater coverage of images than depending on the presence of GPS data, but introduces noise is less reliable in localizing. As result, when GPS data are available, we always use it to geo-localize a given image regardless of user-provided metadata. Otherwise, we prefer automatically extracted locations in the image title over those in descriptions.

Using image metadata streams, we extracted location fields by performing named entity recognition (NER) [Finkel *et al.*, 2005] on user-provided text. We translated all text into English and used only English NER models. We note that translation allowed us to get a higher recall of tagged locations compared to native-language NER models due to the frequency at which users posted descriptions with mixed languages, e.g., an image title "Small Alley in 香港". We extracted metadata-inferred locations for all languages with under 200K GPS-tagged images to bolster their geo-reference count.

We applied NER-tagged locations as queries to Google Maps' Geocoding API[13] to retrieve latitude-longitude coordinates as well as to filter incorrect NER detections or ambiguous locations. Since geocoding queries can be region-biased, we performed multiple searches over all relevant country code top-level domains (ccTLDs) per language according to official

[13] `https://developers.google.com/maps/documentation/geocoding`

Language	ccTLDs	Corresponding Countries
Arabic	.ae, .dz, .eg, .jo, .ma, .om, .qa, .sa, .sd, .sy, .tn, .ye	United Arab Emirates, Algeria, Egypt, Jordan, Morocco, Oman, Qatar, Saudi Arabia, Sudan, Syria, Tunisia, Yemen
Chinese	.cn, .hk, .mo, .sg, .tw	China, Hong Kong, Macau, Singapore, Taiwan
Dutch	.nl, .nu	Netherlands, Niue
French	.fr	France
German	.de	Germany
Persian	.ir	Iran
Polish	.pl	Poland
Russian	.by, .ru	Belarus, Russian
Turkish	.tr	Turkey

Table 4.6: Country code top-level domains (ccTLDs) queried per language during geocoding along with their corresponding countres, listed respectively.

Internet Corporation for Assigned Names and Numbers (ICANN) listings[14], e.g., .pl for Polish was included, .ir for Persian, .ru for Russian, etc. This increased the likelihood that matches were found in regions that speak the target language and also gave queries as many chances to succeed overall as there were ccTLDs per language, e.g., "Main Street"; for example, we queried across over 10 different ccTLDs for Arabic given the diversity of regions it is spoken in. When multiple named locations were detected by NER, we queried for geocodes in descending string length order taking the first query that succeeded. And when multiple geocoding results matched, e.g., the query "Gulf Coast" can match to "Gulf Coast Airport, Tivoli, TX", "Gulf Coast, Missouri City, TX" and "Gulf Coast, Naples, FL", we selected one at random only if all matches came from the same country, and otherwise, omitted the query result altogether for its ambiguity. We also ruled out geo-references to

[14]A crowdsource summarized table of Internet Corporation for Assigned Names and Numbers (ICANN) listings can be found at https://en.wikipedia.org/wiki/List_of_Internet_top-level_domains. Last accessed on June 21, 2016.

Language	**#Locations**(T)	**#Georefs**(T)	**#Locations**(D)	**#Georefs**(D)
Arabic	318	102	1,317	137
Chinese	27,631	9,110	179,407	28,579
Dutch	8,537	3,344	73,315	13,848
French	107,481	41,386	794,183	116,139
German	23,867	13,947	207,630	52,159
Persian	124	49	545	140
Polish	4,048	854	23,222	2,721
Russian	2,731	1,168	20,586	3,539
Turkish	4,235	734	28,890	2,781

Table 4.7: Number of named locations recognized and geo-references (georefs) coded from user-provided title (T) and description (D) metadata in MVSO images by language. Only *new* geo-references counts are given, i.e., #Georefs columns are mutually exclusive with each other and also mutually exclusive to GPS-tagged images in Table 4.5.

continents, e.g., "South America," and large bodies of water, e.g., "Mediterranean Sea," for being too broad for our setting. If all queries failed to geocode an image, we determined that the image lacked sufficient information needed to localize.

On the subset of nine languages that we extracted metadata-inferred geo-references for, a total of 290,737 new geodata points were extracted compared to 237,593 GPS coordinates, i.e., in relative, 22.37% total greater image coverage and 79.05% greater image coverage for the eight smallest languages (excluding French). The combination of GPS and metadata-inferred geodata accounted for 23.09% of the total 7,368,364 images in the MVSO [Jou *et al.*, 2015] image dataset.

4.7 Navigation Interfaces

Given the large-scale nature of the MVSO – in volume, variety and veracity – we developed two visualization user interfaces to facilitate navigation through the data. In a system called **Complura** [Liu *et al.*, 2016], we implemented a web-based ontology browser that allowed for intuitive navigation through multilingual ANP clusters and image corpus. And

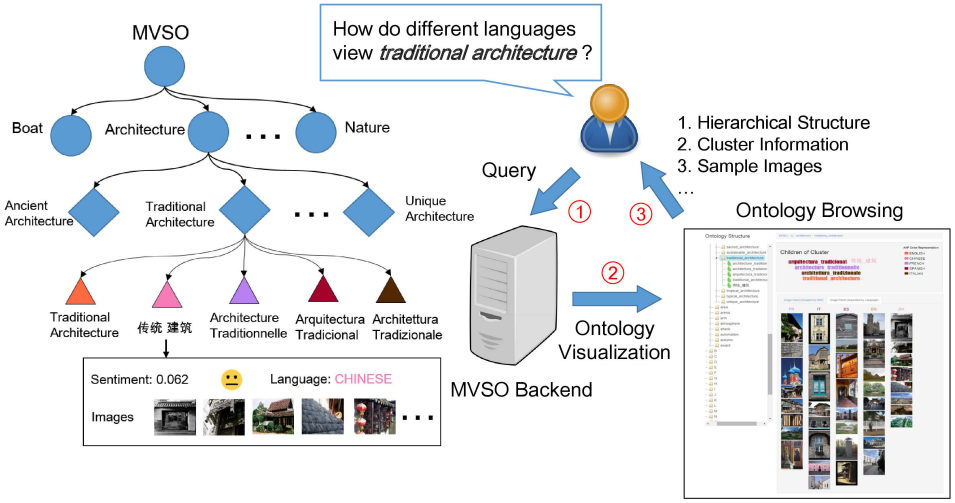


Figure 4.11: **Complura** Ontology Browser Web Interface Process Illustration. A user with a question or hypothesis about how visual affective concepts are understood across languages or sentimentally portrayed can explore the MVSO corpus by querying **Complura** and exploring the ontology. **Complura** visualizes the hierarchical structure, cluster information and shows related sample images in mined from Flickr.

in a system called **SentiCart** [Jou *et al.*, 2016], we developed two geodata visualizations for exploration of multilingual visual sentiment along geographic lines.

4.7.1 Complura Ontology Browser

The **Complura** ontology browser [Liu *et al.*, 2016] enables users to choose controlled ANP clusters and compare visual differences between images of different languages. Given a scenario like the one shown in Figure 4.11[15], if a user wants to research the cultural difference of the visual concept *traditional architecture*, they can use the ontology browser to visualize MVSO in a succinct, intuitive and interactive browser interface. In our web interface, shown in the bottom-right of Figure 4.11, the ontology browser consists of three main panels: the ontology-structure panel (left), word cloud panel (top-right) and image comparison panel

[15]Figure from equal contribution co-author work in [Liu *et al.*, 2016].

(bottom-right)[16].

The hierarchical structure of MVSO is visualized in the ontology-structure panel using an interactive top-down tree view, providing folder-like expansions for the user to navigate through the ontology via a library called jsTree[17]. For non-leaf nodes in the ontology, the word cloud panel allows user to quickly survey the content of certain node. On a mouseover of a word cloud item, a hoverbox is shown and displays additional contextual information for that item like English translations and number of child nodes. In **Complura**, we use the exact match alignment strategy (q.v. §4.5.1) where ANPs are grouped by their direct English translations and also nested by common noun semantics; however, in addition, we also merge noun groups that are synonyms or related, e.g., 'clothes' and 'dress' are merged, to enable additional compactness during visualization. As shown in the Figure 4.11, in the image comparison panel, each column shows images related to a specific ANP, where a language may occupy multiple columns (since there may be multiple ANPs from the same language in a single ANP cluster). This column-based panel view enable users the ability to simultaneously compare inter- and intra-language correlations within a certain ANP cluster.

4.7.2 SentiCart Geodata Visualizations

We developed two interactive visualizations in **SentiCart** [Jou *et al.*, 2016].[18] One visualization is a flat, point-wise, low-interaction view of geodata points. The lower interactivity in this visualization allowed for batch processing and thus precise localization to geodata points at our large input scale. The other visualization is a three-dimensional, fluid, high-interaction globe of geo-references. This visualization allowed us to quickly and easily compare across data modalities and also gives us an additional dimension along which to interact with the geodata. The two modes of visualization provide a trade-off in exploring geographical distributions of visual sentiment. In Figure 4.12, we show examples from our

[16]The **Complura** web demonstration can be accessed at `http://mvso.cs.columbia.edu/complura`.

[17]`https://www.jstree.com`

[18]The geographical data we collected and **SentiCart** web demonstration can be accessed at `http://www.ee.columbia.edu/ln/dvmm/senticart` and a video showcasing the functionality of **SentiCart** can be accessed at `https://youtu.be/cI-2llSErSo`.

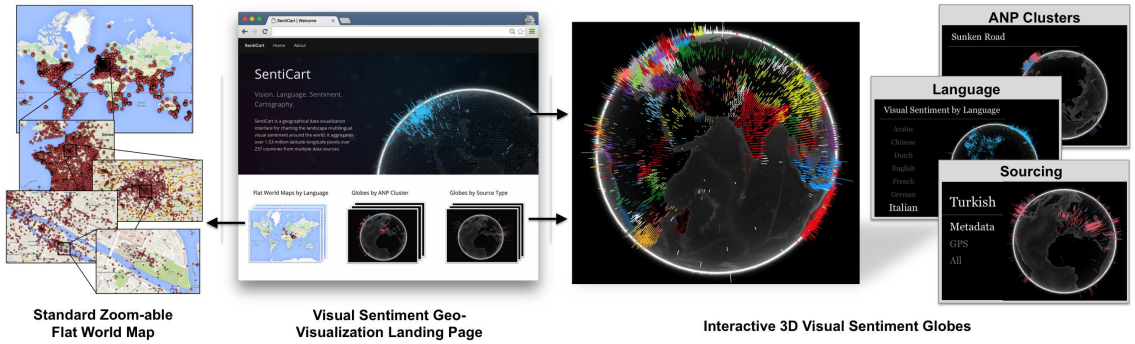


Figure 4.12: Example visualizations in the **SentiCart** system of multilingual visual sentiment around the world, showing the main landing page along with a classic flat world map view, here on the French GPS sub-corpus, and 3D globe visualizations of visual sentiment. The colorful globe shows country-colored sentiment for the English sub-corpus and adjacent cards show additional display modes in **SentiCart** along source, semantics and language.

visualizations. For the flat world map view, our interface matches many other canonical map interfaces but scales to hundreds of thousands of geodata points by linking with Google Fusion Tables [Gonzalez *et al.*, 2010]. The interface allows for a zoomable, albeit flat view of the data and can render elements like political borders without burdening usability.

For the 3D globe view, we enable fast and fluid browser-based rendering using WebGL and a base globe library[19]. To maintain low-latency interaction in this rendering, we reduced the resolution of the geo-references by performing geohashing, where latitude-longitude coordinates are hashed into geodesic spatial bins and where hash lengths, or precisions, correspond directly to geographical distances. In our visualization and scale of 1,701,629 geo-references, we found that quantizing coordinates to within 19.55 km (or 12.14 mi), which corresponds to 10 hash bits for both latitude and longitude, provided the best latency-to-resolution visualization trade-off on most modern browsers and machines. The added dimension of the 3D globe visualization compared to the flat world map, also allowed us to visualize the sentiment magnitudes at given geo-localization points. Since we perform geohashing, to aggregate sentiment in a given spatial bin, we take the weighted average of the image sentiment values from adjective-noun pairs.

[19] `https://www.chromeexperiments.com/globe`

In the globe view, we provide multiple slice views of the ontology and geodata we collected. We can visualize sentiment by language around the world and easily compare regional differences. On the surface of the globe, bar heights correspond to sentiment strengths scaled $[0, h \in \mathbb{R}]$ where colors represent positive or negative (usually red/"hot" and blue/"cool"). In addition, since we had multiple sources for our geodata, we can visualize each source's geographical origins to compare localization consistency. We also enable visualization along ANP clusters using the ontology structure in MVSO [Jou *et al.*, 2015] where ANPs were gathered into semantically coherent groups, e.g., we can visualize geodata and sentiment of images in the multilingual cluster "old town square".

4.8 Conclusions

In this chapter, toward computationally diversifying visual affect along multicultural lines, we extended a mid-level semantic representation called adjective-noun pair in both linguistic and geographic dimensions. A new multilingual discovery method was presented for visual sentiment concepts and we showed its efficacy on a social multimedia platform for 12 languages [Jou *et al.*, 2015]. We based our approach on some psychology theory that emotions are culture-specific and carry inherent linguistic context, and so we showed how to use language-specific part-of-speech labeling along with progressive filtering to achieve coverage and diversity of visual affect concepts in multiple languages. We presented a two-stage hierarchical clustering approach to unify our ontology across languages [Pappas *et al.*, 2016] and mined geographic localization cues for associated imagery using both explicit and implicit signals [Jou *et al.*, 2016]. We also made our multilingual visual sentiment concept ontology (MVSO) and associated image, geodata and metadata corpus available to the public. In addition, we also presented two web-based navigation interface systems for ontology browsing and geographical data exploration with relation to multilingual visual sentiment [Liu *et al.*, 2016; Jou *et al.*, 2016].

In the future, we plan to explore differences along other human factors which can be collected from self-reported user metadata like age group, gender, profession, etc. We also plan to adapt our approach to other language-specific social multimedia platforms to counter

the insufficient data for some languages like Arabic, Persian and Chinese.

Chapter 5

Multilingual Visual Sentiment Prediction

Mid-level visual affective concepts are conceptually an attractive compromise between abstract affective states and visual object groundings. But how well can we actually detect the presence of such concepts in images? And can they be used to computationally model abstract affective states like sentiment? In Chapter 4, we developed a multilingual ontology with 15.6K affective visual concepts across 12 languages and spanning over 235 countries, along with a corpus of 7.3M images and associated metadata [Jou *et al.*, 2015; Jou *et al.*, 2016; Pappas *et al.*, 2016]. Given these adjective-noun pair (ANP) concepts, we develop an image-based prediction task to evaluate well modern visual classifiers can detect the presense ANPs as well as a sentiment prediction task to determine how generalizable language-specific models are in a multilingual sentiment context.

Beyond image-based prediction tasks for ANP and sentiment detection, we also seek to develop real-world useful systems using MVSO and our detector banks [Jou *et al.*, 2015; Jou and Chang, 2016b]. We propose two applications which are part of the **Complura** system, complementary to the ontology browser discussed in §4.7.1, which enable multilingual sentiment assessment as well as an image-based query expansion engine with multilingual semantic coherence and sentiment sensitivity [Liu *et al.*, 2016].

5.1 Introduction

While constructing computational affective models to infer states like Ekman emotions [Ekman, 1999] or sentiment, e.g., Chap. 3, require far more care given the lack of model interpretability, mid-level affective representations benefit from semantics that allow for interpreting decision outputs and concrete language for understanding what drove a model toward a particular output. In this chapter, we develop visual concept detector banks [Li *et al.*, 2010; Torresani *et al.*, 2010] in the context of the multilingual visual sentiment concept ontology (MVSO) [Jou *et al.*, 2015] using modern CNNs to recognize the presence of adjective-noun pairs (ANPs) in social images from Flickr. Our visual concept detectors are trained to detect 9,918 sentiment-biased concepts from six major languages: English, Spanish, Italian, French, German and Chinese.

In addition, since the original image pool in the MVSO corpus [Jou *et al.*, 2015] was mined from a mix of tag-only and free text queries on Flickr (q.v. §4.3.4), giving rise to label noise for images from user's social interactions and content description behaviors, we partitioned out a sub-corpus of MVSO images based on tag-restricted queries for higher fidelity labels. We show that as a result of these higher fidelity labels, higher performing ANP detectors can be trained using the tag-restricted image subset as compared to models from the hybrid corpus, trading off a smaller coverage of ANPs [Jou and Chang, 2016b].

To test the effectiveness of a vision-based approach for mid-level visual affect understanding when crossing languages, we also implemented language-specific sentiment predictors on two social multimedia platforms [Jou *et al.*, 2015; Campos *et al.*, 2016]. Inspired by work in [Hu and Yang, 2014], we studied the extent to which the visual sentiments of a given language can be predicted by sentiment models of other languages. We chose to focus on a sentiment prediction task, i.e., predicting whether an image is of positive or negative sentiment, because the large body of work on sentiment, e.g., [Yanulevskaya *et al.*, 2008; Borth *et al.*, 2013b; You *et al.*, 2014], and for its simplicity. By focusing on sentiment, we also reduce the number of output variables to be analyzed for the primary goal of uncovering cross-lingual similarities and differences. Additionally, while there have been cross-lingual sentiment studies in natural language processing, e.g., [Mihalcea *et al.*, 2007; Bautin *et al.*, 2008; Brooke *et al.*, 2009; Wan, 2009; Boyd-Graber and Resnik, 2010], there

have not been any for image-based sentiment to the best of our knowledge.

The contributions of this work include (1) multilingual, sentiment-driven visual concept detector banks over six major languages on ∼10K concepts, (2) a sub-corpus of the multilingual visual sentiment concept ontology (MVSO) image dataset based on tag-restricted queries for higher fidelity labels, (3) adjective-noun pair (ANP) detector banks based on this tag-restricted subset, and (4) two applications using the multilingual ANP detectors for sentiment analysis and image query expansion, and (5) a cross-lingual sentiment detection benchmark on MVSO using our multilingual ANP detector banks.

5.2 Related Work

Convolutional neural networks (CNNs) [LeCun *et al.*, 1998] are class of deep neural networks (NNs) that have risen to extradinary popularity over the past decade due to its success in a variety of visual recognition tasks, most notably on CIFAR [Krizhevsky, 2009] and ILSVRC [Russakovsky *et al.*, 2015]. Many empirical tricks and improvements have been proposed since including on activation functions [Nair and Hinton, 2010], generalization techniques [Srivastava *et al.*, 2014], initialization schemes [He *et al.*, 2015], normalization [Ioffe and Szegedy, 2015], etc.

The wide application of CNNs in vision and multimedia was in part propelled by the success of [Krizhevsky *et al.*, 2012] on ILSVRC, now known as AlexNet. AlexNet consisted of five convolutional layers, some followed by normalization and max pooling, and three fully-connected layers with a final 1000-D softmax. In [Jia *et al.*, 2014], a variant called CaffeNet was implemented where pooling is done before normalization; and in [Zeiler and Fergus, 2014], another variant, independently dubbed ZF Net, was proposed where some middle layers in AlexNet were expanded to free up representational bottlenecks. In [Simonyan and Zisserman, 2015], a much deeper network called VggNet of 16 convolutional/fully-connected layers was proposed that was the runner-up in ILSVRC2014, bested by GoogLeNet [Szegedy *et al.*, 2015] which used a unique network-in-network approach [Lin *et al.*, 2014] where mini-networks were stacked on top of each other where each contained shortcut connections [Raiko *et al.*, 2012] with projections, global pooling, as well

as traditional convolutional layers. And most recently, in [He *et al.*, 2016a], a residual network called ResNet using many shortcut connections was used to develop very extremely deep neural networks winning ILSVRC2015.

Using these pre-trained networks on ILSVRC turns out to be effective in numerous transfer learning tasks [Donahue *et al.*, 2014; Razavian *et al.*, 2014], i.e., where the pre-trained networks are used for fine-tuning or simply as feature extractors. In this chapter, we take the approach of the former to perform fine-tuning from pre-trained models to evaluate their performance on language-dependent ANP detection. We then use the latter approach where our fine-tuned models are used for a visual sentiment classification task.

5.3 Visual Sentiment Concept Detectors

We construct a bank of visual concept detectors like in [Borth *et al.*, 2013a] for our final MVSO adjective-noun pairs in §4.3. For sufficient data for model training, we focused on the six languages with the most ANPs and associated images in our dataset: in decreasing order, English, Spanish, Italian, French, German and Chinese. Combined these six languages account for 94.7% of the ANPs in MVSO and 98.4% of the images in our dataset. However, to further ensure that there were enough training images for each ANP, only the ANPs with no less than 125 images were selected for model training and prediction. This reduced the combined ANP coverage to 63.5% but still ensured 92.0% coverage for images. For each ANP, the images were split randomly 80/20% train/test, respectively.

5.3.1 Hybrid-pool ANP Detectors

To construct our banks of ANP detectors, we begin with a CaffeNet [Jia *et al.*, 2014] structure, a modified AlexNet-styled architecture [Krizhevsky *et al.*, 2012]. To train our detector banks, we fined-tuned six models, one for each language, where network weights were initialized with DeepSentiBank [Chen *et al.*, 2014a], an AlexNet model trained on the VSO [Borth *et al.*, 2013b] dataset which was fine-tuned from a model trained on ILSVRC2012 [Russakovsky *et al.*, 2015]. This fine-tuning approach attempts to "affectively bias" each network's training process by initializing with weights that solve a similar problem from

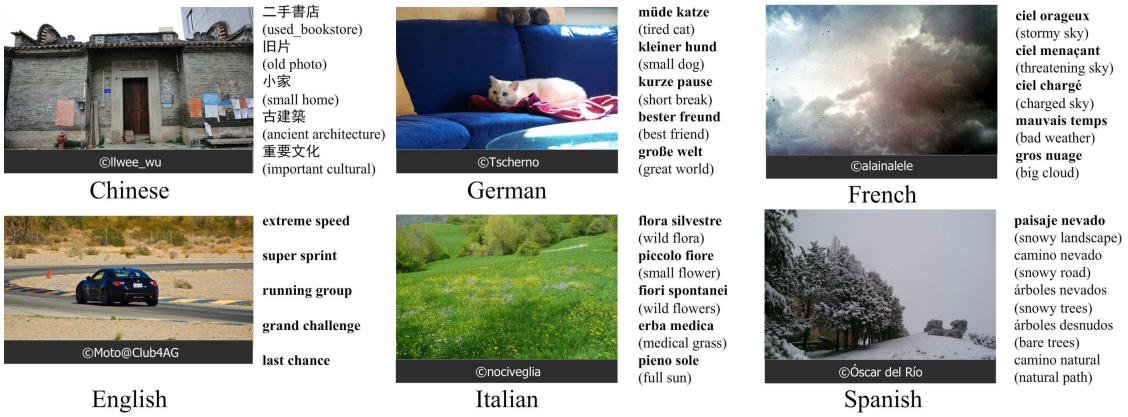


Figure 5.1: Example top-5 hybrid-pool classification results from our multilingual visual sentiment concept detector banks. Original attributions are included below each image. Translations to English provided for convenience.

the same domain. Networks were learned using stochastic gradient descent (SGD) where initial learning rates were set to 0.001 and decreased by a factor of 10 when performance plateaued. The number of output neurons in the last fully connected layer were set to the number of training ANPs of each language. Step sizes for reducing the learning rate were set proportional to the number of training images per language.

In the original MVSO image mining detailed in §4.3.4, when we downloaded data from Flickr, we first queried the Flickr API using our ANPs. Generally, there are two types of ways to search for photos on Flickr: asking the Flickr API to search for the query in the (1) tags of a photo, or (2) anywhere in the photo text data, including photo title, description and tags. We call these "tag search" and "free-text search," respectively. Tag search yields less but more precise results, and free-text search will give more but noisier data, i.e., leading to weaker supervision. The intuition is that if a user uses an ANP as a tag, then it is more likely to describe what is actually visually present in the image, while an ANP that occurs in, say the description, may be relevant, but not visually present. In our original data collection, we used an upper bound of 1,000 images per ANP where we first queried via tag search and if the upper bound had not been reached, we then pulled results from free-text search on Flickr. As a result, the original MVSO image dataset constitutes a "hybrid-pool" of Flickr images with respect to how the querying was performed. Here,

we first discuss the classification results from using this hybrid-pool of images constructing our MVSO detector banks.

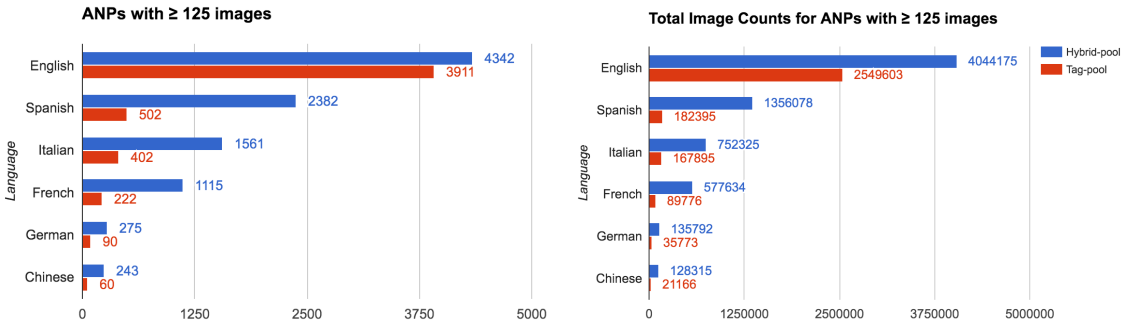
In Figure 5.1, we show several detection results from our fine-tuned networks for multilingual ANP detection. For each language, fine-tuning took between 12 and 95 hours for convergence on a single NVIDIA GTX 980 GPU implemented with Caffe [Jia *et al.*, 2014]. From Table 5.1, as expected we achieve higher top-1 and top-5 accuracies than DeepSentiBank [Chen *et al.*, 2014a], even when the number of output units are higher than DeepSentiBank as in English and Spanish. We note that we experimented with larger choices of dropout [Srivastava *et al.*, 2014] in our networks especially for the smaller languages like German and Chinese, but found that the standard ratio of 0.5 consistently performed the best in this architecture. Top-k accuracy refers to the percentage of classifications for which the true class is in the top k predicted ranks.

Language	**#ANPs**	**#Params**	**#Train**	**#Test**	**Time**	**Top-1**	**Top-5**
English	4,342	74.66	3,236,728	807,447	95	10.13%	21.06%
Spanish	2,382	66.63	1,085,678	270,400	45	12.35%	25.36%
Italian	1,561	63.26	602,424	149,901	30	17.01%	30.93%
French	1,115	61.44	462,522	115,112	26	17.66%	35.46%
German	275	57.99	108,744	27,048	12	30.11%	52.78%
Chinese	243	57.86	102,740	25,575	15	27.07%	45.06%
DeepSentiBank	2,089	65.43	826,806	41,113	-	8.2%	19.1%

Table 5.1: ANP classification performance on the hybrid-pool images for six major languages in MVSO [Jou *et al.*, 2015] with comparisons to DeepSentiBank [Chen *et al.*, 2014a]. Number of visual sentiment concepts #ANPs, network parameters #Params (in millions), #Train and #Test images are shown with training walltimes (in hours), Top-1 and Top-5 accuracies (%).

5.3.2 Tag-pool ANP Detectors

Since the image corpus in the hybrid-pool setting, where images come from both tag and free-text search on Flickr, can be noisy, we perform a separate set of experiments where we



(a) (b)

Figure 5.2: Hybrid `vs.` Tag-pool MVSO Coverage. In **(a)**, we show the coverage of ANPs and in **(b)**, images, across the six major languages in MVSO. English suffers the least from the tag restriction while all other language experience a large drop in ANP and image coverage.

train our ANP detector banks just from the subset of images that come from tag-restricted queries on Flickr. We call this the "tag-pool" of MVSO images. Though it is a subset, each ANP still has a maximum of 1,000 images to learn from and we split the dataset as before 80/20% per ANP. However, since we still enforce the requirement that there are at least 125 images per ANP, some languages experienced significant decrease in ANP coverage. In Figure 5.2, we show the change in ANP and image coverage when going to the tag-restricted subset in MVSO. We observe that English experiences a small 9.92% loss in ANPs and a 63.04% loss in image count when restricting to tag-only queries while all other languages experience about 80% coverage loss for both ANPs and images. Though the ANPs in MVSO are useful given their pervasive occurrence in social media, this indicates that there are a fair number of images in hybrid-pool have a considerable amount of weak supervision. At the same time, it is worth noting that it is wrong to assume all images from the free-text search are not useful for visual recognition of adjective-noun pairs.

To train our tag-pool ANP detectors, we follow the same scheme as before in §5.3.1. We fine-tuned CaffeNet networks, pre-training with DeepSentiBank [Chen *et al.*, 2014a], using Caffe [Jia *et al.*, 2014] with the same optimization strategies, with the only slight difference being that we use a NVIDIA GeForce GTX Titan X GPU here. In Table 5.2, we present

Language	#ANPs	#Params	#Train	#Test	Time	Top-1	Top-5
English	3,911	72.89	2,294,411	255,192	85	19.00%	33.81%
Spanish	502	58.92	164,119	18,276	12	29.07%	52.86%
Italian	402	58.52	151,088	16,807	9	33.69%	55.76%
French	222	57.78	80,790	8,986	6	34.70%	63.32%
German	90	57.24	32,195	3,578	2.5	47.04%	74.62%
Chinese	60	57.11	19,044	2,122	1.5	45.05%	71.35%

Table 5.2: ANP classification accuracies (%) on the tag-pool images for the six major languages in MVSO [Jou *et al.*, 2015]. The number of parameters (millions), #ANPs, #Train and #Test images, and training process walltime (hours) are also shown.

results for our ANP detectors using the tag-restricted subset of ANPs and images. Again, note that the number of ANP classes each detector learns over has reduced compared to those seen earlier in Table 5.1 due to the tag-based pool restriction.

At first glance, the top-1 and top-5 accuracy rates seem much higher than those we observed with a similar architecture in Table 5.1. However, it is still difficult to determine if restricting to the tag-pool actually contributed to more reliable labels and thus better performing ANP detectors, especially since the number of output classes changed. To investigate, we evaluated the hybrid-pool ANP detectors from §5.3.1 on the test set of the tag-restricted pool of images as well as the the intersection of the test sets. Since random shuffling was done during train/test splitting for both hybrid- and tag-pool image datasets, the latter ensures that no images in the test set (#Test) were used in either model's training. The results from this evaluation are shown in Table 5.3. We observe that in both cases, the classification performances are both lower than those in Table 5.2 by about 10% absolute on each language, indicating that the tag-pool ANP detectors do indeed achieve a comparably higher accuracy on the same set of ANPs. Given these observations, we believe it is a user-level trade-off decision of whether to prefer a smaller coverage of ANPs with a higher automatic detection precision (tag-pool) or a larger coverage set of ANPs with relatively weaker detection performance (hybrid-pool).

Language	#ANPs	#Test	Top-1	Top-5	#Test	Top-1	Top-5
English	3,911	756,243	10.47%	21.67%	51,062	12.04%	24.58%
Spanish	502	91,002	14.86%	30.48%	3,740	17.91%	34.39%
Italian	402	64,668	22.51%	39.44%	3,386	24.25%	42.85%
French	222	38,107	21.78%	43.10%	1,835	25.29%	46.54%
German	90	13,249	36.18%	58.99%	769	39.14%	62.68%
Chinese	60	9,093	29.47%	48.78%	412	34.95%	51.94%
			(a)			(b)	

Table 5.3: ANP classification performance (%) at Top-1 and Top-5 of the hybrid-pool ANP detectors from Table 5.1 [Jou *et al.*, 2015], but now evaluated on the ANPs and test images **(a)** of the tag-pool in Table 5.2, and **(b)** intersection of both the hybrid-pool and tag-pool test sets.

5.3.3 Going Deeper with Convolutions for Visual Affect

To improve the detection performance of ANPs, we also experimented with a much deeper and more complex architecture called GoogLeNet, or Inception [Szegedy *et al.*, 2015]. The original GoogLeNet design consists of mini-networks, also called Inception modules, that are composed of $1\times1C$, $1\times1C$-$3\times3C$, $1\times1C$-$5\times5C$, and $3\times3MP$-$1\times1C$ towers where C corresponds to convolutions and MP denotes max pooling, i.e., so $1\times1C$-$3\times3C$ denotes one path in the modules that has a 1×1 convolution followed by a 3×3 convolution. Note that 1×1 convolutions are actually $1\times1\times d$ where d is the number of filters, or channels. These fan-in-fan-out modules are stacked in an architecture referred to as network-in-network [Lin *et al.*, 2014]. GoogLeNet additionally uses two auxiliary branches in the network to prevent gradients from vanishing during backpropagation [Szegedy *et al.*, 2015].

For training, we preserve the same training setting as in §5.3.1 using the same train/test split of the hybrid-pool of images. Unlike before we now pre-train from a ILSVRC2012 model since the architecture is significantly different from DeepSentiBank [Chen *et al.*, 2014a] and no comparable Inception-like network was previously trained on VSO. In addition, since the output space for languages like English and Spanish is large, in the auxiliary heads of Inception, we widen the second-to-last fully-connected layer which is originally 1,024 since it would otherwise become a representational bottleneck. We use SGD with a sigmoid decay

learning rate decay with a base learning rate of 0.001, a batch size of 64 and a decay factor of 0.1 and unlike before use the true training data mean image during mean subtraction, instead of the ILSVRC mean image. We implemented this in Caffe [Jia *et al.*, 2014] on a NVIDIA GeForce GTX Titan X GPU. While a different GPU and different training settings were used here compared to the CaffeNet setting, we argue that much like ILSVRC where advances over the years have also had different hyperparameters, we can still claim that the performances are approximately comparable given same test sets, targets and more-similar-than-different optimization settings.

Language	**#ANPs**	**#Params**	**#Train**	**#Test**	**Time**	**Top-1**	**Top-5**
English	4,342	30.50	3,236,728	807,447	370	13.64%	26.63%
Spanish	2,382	20.84	1,085,678	270,400	106	13.86%	27.98%
Italian	1,561	7.57	602,424	149,901	57	17.46%	32.25%
French	1,115	7.12	462,522	115,112	36	16.76%	34.27%
German	275	6.26	108,744	27,048	11	31.08%	54.10%
Chinese	243	6.22	102,740	25,575	10	25.96%	47.11%

Table 5.4: ANP classification accuracies (%) using an Inception-based architecture [Szegedy *et al.*, 2015] on the MVSO hybrid-pool [Jou *et al.*, 2015], i.e., same dataset as in §5.3.1. The classes, training and testing sets match Table 5.1. The number of parameters (millions), #ANPs, #Train and #Test images, and training process walltime (hours) are also shown.

In Table 5.4, we show the results of fine-tuning with Inception networks for ANP detection across six languages in MVSO. Compared to previous experiments in §5.3.1 and §5.3.2 which we fine-tuned from an affectively biased set of network weights [Chen *et al.*, 2014a], training takes noticeably longer in most cases since we now fine-tune from a model trained on ILSVRC. Nonetheless, we observe that we are able to get improved ANP classification rates on most languages with Inception [Szegedy *et al.*, 2015], particularly when there is more data for the networks to take advantage of; for example, with English ANP detection, we get a 35.05% relative improvement at top-1. Overall, except for French, we see a consistent improvement at top-5, indicating that more semantically relevant ANPs are

being surfaced into the top ranks by Inception compared to the CaffeNet model. While we do not see this same consistency at top-1, given that these detectors currently treat all ANPs as if they came from a flat taxonomy, many of the ANPs may in fact be semantically close to each other as indicated by the hierarchical grouping studied in the original ontology construction process [Jou *et al.*, 2015] (q.v. Chap. 4). As a result, from a purely empirical standpoint, a top-1 accuracy metric may treat a 'pretty flower' prediction for a 'beautiful flower'-labeled image to be a misclassification. A top-5 metric in these ontology-structure-agnostic settings may then be a more reasonable indicator of detector bank performance than a top-1 performance metric. We plan to investigate semantic hierarchy-aware detector bank training in the future.

Briefly, we also trained a massive multi-class CNN on 5,675,460 training images using the VggNet [Simonyan and Zisserman, 2015] architecture, fine-tuning from a ILSVRC2014 [Russakovsky *et al.*, 2015] model after widening some fully-connected layers, and classifying over 10,159 ANPs across all 12 languages in MVSO (all ANPs in MVSO with $\geq$125 images). The model took over 31 days to train and achieved a surprising top-1 accuracy of 15.34% (top-5 of 27.67%) over 1,414,530 test images. For training, we used the same strategy as we had with Inception [Szegedy *et al.*, 2015], e.g., sigmoid learning rate policy, etc. In Figure 5.3, we show a slice of the training process of this massive multi-class ANP classification network over time. We note that this model may be biased toward English ANPs though simply because they dominate most of MVSO, and this is an area we seek to explore in the future, either by stratifying over classes or a first-stage visual language detector. Also, to test the same setting as before, since a softmax loss was used, to compensate for the class imbalance, experimenting with another loss like an information gain loss will be important.

5.4 Applications of Multilingual ANP Detectors

We show the application of the unique ontological structure of MVSO and our language-specific ANP detectors in two novel, pivot applications: language-specific image sentiment analysis and culturally-coherent image query expansion. These applications are components of the **Complura** system discussed in §4.7.1 [Liu *et al.*, 2016].

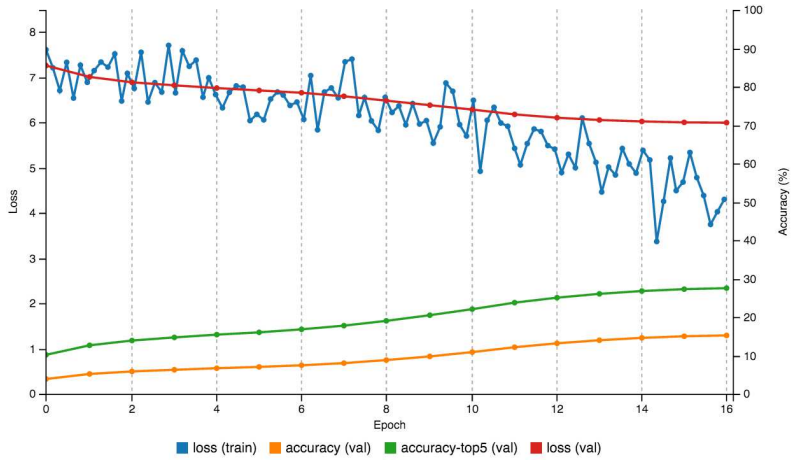


Figure 5.3: Example network training and validation trends for fine-tuning VggNet [Simonyan and Zisserman, 2015] over time (in epochs) on a massive multi-class learning problem on MVSO over 10,159 ANP classes and 5.68M training images.

5.4.1 Multilingual Sentiment Analysis

In MVSO, each ANP is associated with a sentiment score, which is calculated using SentiStrength [Thelwall *et al.*, 2010] and SentiWordNet [Baccianella *et al.*, 2010], and verified by crowdsourcing validation [Pappas *et al.*, 2016] (q.v. §4.3). Using these sentiment scores, we develop a multilingual sentiment analysis tool where, given an image, we use the mid-level affect representations via our ANP detector banks as a sentiment proxy. Essentially, after detecting ANPs in an image, we use the sentiment scores we have pre-assigned to each ANP as the output of our system for multilingual sentiment assessment by the user.

As in Figure 5.4[1], consider a scenario where a Chinese businessperson is designing a website to advertise 'good food' or 'modern art' but wants to expand to markets of other cultures. It is critical that culturally-sensitive images are chosen for use in these other markets. One application of MVSO and our ANP detectors is to enable the image-based cultural sentiment comparisons. A user uploads an image to find if the image is suitable for foreign markets, and our ANP detector banks return the top detected ANP in multiple languages, and the user makes their own judgment on its cultural sensitivity, e.g., based

[1]Figure from equal contribution co-author work in [Liu *et al.*, 2016].

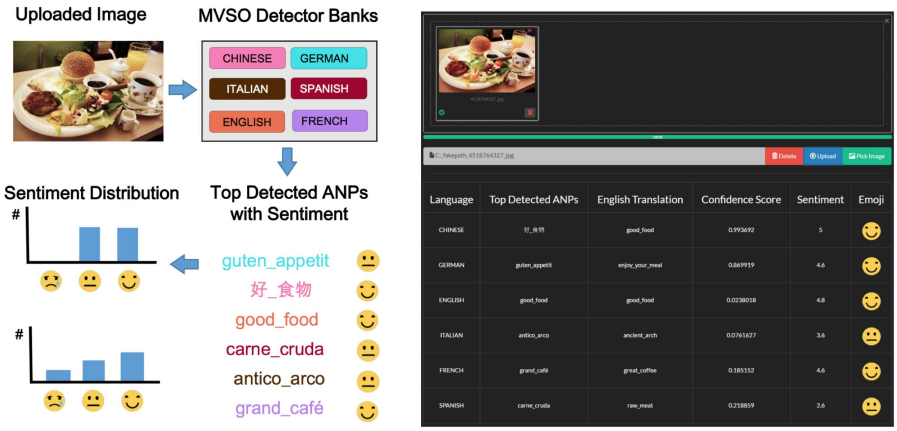


Figure 5.4: Multilingual image sentiment analysis in **Complura**. Given an input image, top detected ANPs of different languages with corresponding sentiment scores are returned.

on the sentiment of the top detected ANP. This application, part of the **Complura** system, also enables the user to find additional images using the detected ANPs in foreign languages. The resulting distribution of sentiment values of the retrieved ANPs may be vary based on the input image where some images produce more uniform sentiment distributions across languages, while others are more polarized. In our example scenario of the Chinese businessperson, they would likely prefer a more polarized distribution, specifically polarized toward a sentiment they intend, so that the do not upset the sentiments of other cultures when moving into those markets. By investigating the sentiment distribution of detected multilingual ANPs for a query image, users can target markets of foreign culture with greater with sentimental sensitivity.

5.4.2 MVSO Image Query Expansion

Another application and a feature in the **Complura** system is an image query expansion engine with cultural and sentiment coherence using the semantic structure in MVSO and detectors. As illustrated in the Figure 5.5[2], in this application, users begin by choosing a language and upload a seed image. The system then detects ANPs in specified language, and ANPs with culturally-coherent semantics are retrieved through traversing a related

[2]Figure from equal contribution co-author work in [Liu *et al.*, 2016].

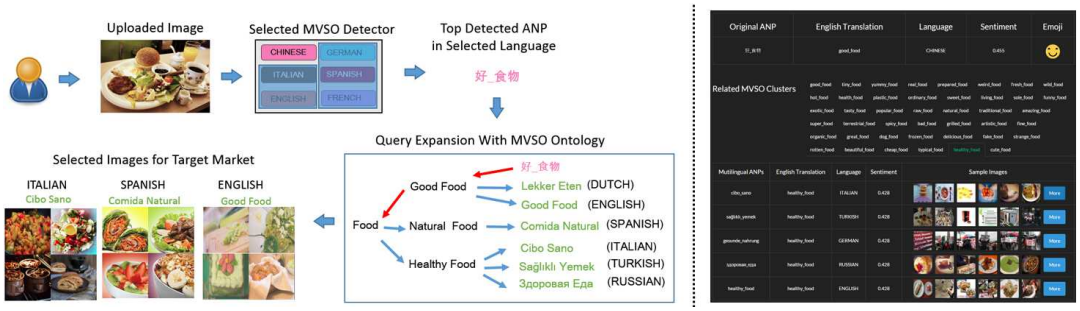


Figure 5.5: Culturally-coherent, sentiment-sensitive image query expansion in **Complura**. Given an image and selected language, the top detected ANP using detectors of that language is applied toward query expansion by discovering related ANPs in the hierarchical structure of the ontology. Sample images and related ANPs are returned to user for further comparison and selection.

sub-tree in the ontology. Some results are then pruned by ANP detector confidence scores as well as sentiment scores of retrieved ANPs. The system is then able to return ANPs of the specified language, together with other ANPs that are semantically and sentimentally coherent in other languages. Image query expansion is achieved then by returning images associated with the retrieved multilingual ANPs.

Following the same scenario in Section 5.4.1, the businessperson can also perform image query expansion using **Complura**. After uploading their target image and selecting the origin language, e.g., Chinese, **Complura** will then run the Chinese ANP detectors on the input image. The system maps detected Chinese ANPs to top MVSO clusters that elicit similar semantics but are both multilingual and have uniform sentiment. The multilingual nature of our clusters allows for a query to have diverse, but semantically consistent expression in multiple languages beyond the original language of the query, and enforcing a uniform sentiment constraint ensures that we maintain the sentiment of the original query even when expanding to other language semantics. From each of these mapped ANP clusters, the user can then view the translated meanings and the sentiment of the retrieved multilingual ANPs. We also show sample images related to retrieved ANPs from the other languages. Through this process, the image search experience can be culturally and sentimentally diversified.

5.5 Sentiment Prediction in Social Multimedia

In Part I, we discussed several approaches for recognizing visual affect directly using affective representations common in psychology (q.v. §2.1.2). Although in this Part II, we expressly focus on mid-level affective representations called adjective-noun pair, it is interesting to consider whether these mid-level concepts can be used to detect the more traditional affective representations like valence-arousal or sentiment. In addition, given the development of a multilingual visual sentiment concept ontology, we naturally would also like to determine if there are interesting cross-lingual phenomena that we can see through the application of computational models. Motivated by these, we present visual sentiment prediction tasks on two social multimedia platform, Flickr and Twitter[3], with particular focus on using our MVSO ANP detector banks and on cross-lingual effects.

5.5.1 Sentiment in Flickr

Here, we use our CaffeNet-based visual concept models (i.e., those from Table 5.1) trained for each language to extract image features and use the sentiment scores of ANPs in MVSO as supervised labels to learn sentiment prediction models. We compare different layers of the CNN models as image features. To simplify the analysis, we binarized the ANP sentiment scores computed via Eq. (4.1), i.e., into positive and negative classes, and learn a binary classifier using linear SVMs, one for each language. The images used are those associated with ANPs of strong sentiment scores taken from the test set of the MVSO hybrid-pool (absolute values higher than 0.05), i.e., so we threshold on 807,477 images for English, 270,400 for Spanish, 149,901 for Italian, etc. Splits of training and test sets are stratified across all languages so that the amount of training and testing for positive and negative sentiment classes was the same for fair cross-lingual experiment comparison.

We found that the softmax output features from the penultimate layer outputs of each language's CNN model performed the best for all languages, and we show resulting sentiment prediction results in Figure 5.6. Each language expectedly did better in predicting test samples from its own language, but in addition, Chinese generally was the most difficult to

[3] `https://twitter.com`

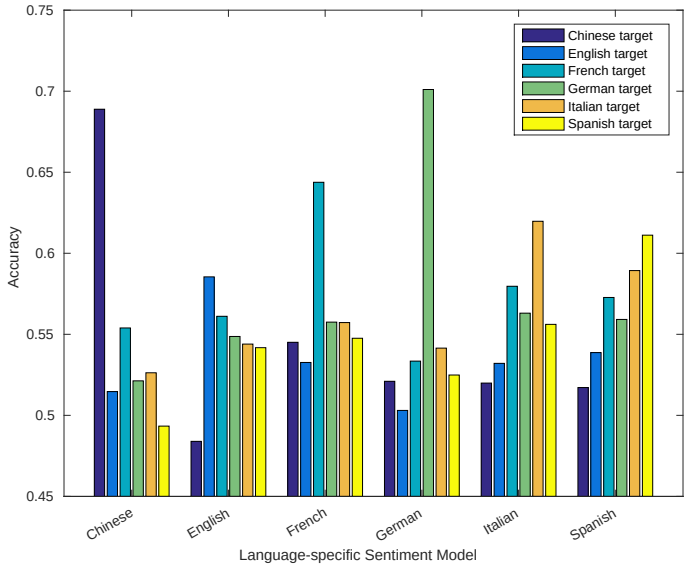


Figure 5.6: Image-based, cross-lingual domain transfer sentiment prediction results with language-specific models applied on cross-lingual examples.

predict by models trained from other languages; and using a sentiment model trained over Chinese images to predict the sentiment in other languages was also the worst in average. We speculate that this is due to the difference in the visual sentiment portrayal from Eastern and Western cultures. Interestingly, the classification of French and Italian sentiments was the most consistent using models from all languages. We also observed good performance in cross-lingual prediction for Latin languages, i.e., Spanish, Italian and French, where Italian was the best cross-lingual classifier for Spanish and French sentiment, and Spanish was best for Italian sentiment, followed by French. Despite not performing as well as others in average, the English-specific sentiment model had the least variance in its accuracy across all languages, likely from the pervasiveness of English worldwide and across cultures.

In Figure 5.7, we show three classification example results from our cross-lingual sentiment prediction. On the left, an image from the Italian test set representing the *costumi tradizionali* concept was labeled as positive via sentiment scoring, but was predicted by the German model to be negative; this may be due to differences in cultural perceptions of traditional clothing. In the center, the Chinese model wrongly predicted an image from the English test set of *foggy morning* as positive, possibly for its resemblance to a Chinese

Original Language	Italian	English	French
Adj-Noun Pair	costumi tradizionali (traditional costume)	foggy morning	beau village (beautiful village)
Prediction Model	German	Chinese	Spanish
Truth/Predicted	positive/negative	negative/positive	positive/positive

Figure 5.7: Classification examples from cross-lingual sentiment prediction. The model from a source language is used to predict the sentiment of a target language image where the true label comes from the sentiment of the associated ANP.

painting. And on the right, an image of a *beau village* from the French test set was successfully classified as positive with the Spanish sentiment predictor. These examples and preliminary experiments highlight some similarities and differences in how visual sentiment is expressed and perceived by various cultures.

5.5.2 Sentiment in Twitter

Like in §5.5.1, here, we also use the CaffeNet ANP detector banks (Table 5.1) for sentiment prediction, but now on a set of photographs collected from Twitter. In [You *et al.*, 2014], the DeepSent dataset was introduced for visual social sentiment prediction benchmarking; it is a small dataset consisting of 1,269 Twitter images with crowdsourced sentiment annotations where labels are majority voted with variable agreement in two classes: positive and negative. [You *et al.*, 2014] proposed a CNN where the second-to-last fully-connected layer consisted of 24 output units, inspired by Plutchik's emotions [Plutchik, 1980]. However, not much intuition as to why Plutchik's emotions should necessarily translate into number of fully-connected layer output units was given in their design, and so in [Campos *et al.*, 2015], we performed an extensive analysis of various strategies for fine-tuning networks and found that even simply fine-tuning a CaffeNet trained on ILSVRC2012 [Russakovsky *et al.*, 2015] could achieve a 4.34% relative improvement over the results in [You *et al.*, 2014].

In Table 5.5, we show a summary of results reported on the DeepSent dataset. The "agreement" columns correspond to how many annotators during crowdsourcing agreed

		Five Agree		≥ Four Agree		≥ Three Agree	
		Acc	F1	Acc	F1	Acc	F1
SentiBank [Borth *et al.*, 2013b]		0.709	0.776	0.675	0.734	0.662	0.721
DeepSentiBank [Chen *et al.*, 2014a]		0.804	–	–	–	–	–
Sentribute [Yuan *et al.*, 2013]		0.738	0.805	0.709	0.771	0.757	0.783
DeepSent [You *et al.*, 2014]		0.783	0.846	0.755	0.811	0.715	0.779
[Campos *et al.*, 2015]	Standard	0.817	–	–	–	–	–
	Oversampled	0.830	–	–	–	–	–
MVSO [Jou *et al.*, 2015] Chinese Detector Bank	Fine-tuning	0.797	–	–	–	–	–
	fc7+RBF-SVM	0.718	0.812	0.665	0.760	0.648	0.742
	fc8+RBF-SVM	0.684	0.789	0.663	0.760	0.617	0.728
	prob+RBF-SVM	0.659	0.794	0.617	0.763	0.606	0.754
MVSO [Jou *et al.*, 2015] English Detector Bank	Fine-tuning	**0.839**	–	–	–	–	–
	fc7+RBF-SVM	0.718	0.806	0.687	0.768	0.664	0.753
	fc8+RBF-SVM	0.696	0.792	0.684	0.764	0.653	0.747
	prob+RBF-SVM	0.655	0.787	0.609	0.751	0.604	0.713
MVSO [Jou *et al.*, 2015] French Detector Bank	Fine-tuning	0.825	–	–	–	–	–
	fc7+RBF-SVM	0.712	0.806	0.673	0.754	0.640	0.737
	fc8+RBF-SVM	0.709	0.800	0.659	0.752	0.629	0.727
	prob+RBF-SVM	0.657	0.790	0.615	0.761	0.607	0.755
MVSO [Jou *et al.*, 2015] German Detector Bank	Fine-tuning	0.837	–	–	–	–	–
	fc7+RBF-SVM	0.710	0.802	0.665	0.757	0.623	0.726
	fc8+RBF-SVM	0.701	0.800	0.653	0.751	0.637	0.739
	prob+RBF-SVM	0.659	0.794	0.617	0.763	0.606	0.755
MVSO [Jou *et al.*, 2015] Italian Detector Bank	Fine-tuning	0.838	–	–	–	–	–
	fc7+RBF-SVM	0.700	0.794	0.665	0.759	0.661	0.756
	fc8+RBF-SVM	0.704	0.796	0.668	0.755	0.648	0.741
	prob+RBF-SVM	0.659	0.794	0.626	0.759	0.606	0.755
MVSO [Jou *et al.*, 2015] Spanish Detector Bank	Fine-tuning	0.833	–	–	–	–	–
	fc7+RBF-SVM	0.715	0.798	0.700	0.773	0.668	0.750
	fc8+RBF-SVM	0.727	0.810	0.686	0.762	0.667	0.747
	prob+RBF-SVM	0.659	0.793	0.634	0.750	0.630	0.753

Table 5.5: DeepSent Twitter Sentiment Prediction. Accuracy (Acc) and F1 scores are reported across various methods.

with each other on the sentiment label of an image. The higher the agreement, the less noisy the sentiment labels are, but less examples become available for training/testing. At full agreement across all five annotators, there are 880 images (580 positive and 301 negative) which are divided into five folds for cross-validation. The average cross-validation accuracy and F1 scores are reported.

With respect to our MVSO ANP detector banks, we observe that, as a straightforward feature extractor, all of the languages detectors perform about the same when used in conjunction with a RBF SVM [Fan *et al.*, 2008], with the *fc7* feature map generally performing the best. The feature maps *fc7*, *fc8* and *prob* correspond to the second-to-last, last and softmax output scores of the network, respectively [Jia *et al.*, 2014]. Recently, in co-authored work in [Campos *et al.*, 2016], we also performed a set of fine-tuning experiments using the ANP models from six languages in MVSO [Jou *et al.*, 2015] to fine-tune, i.e., instead of fine-tuning from ILSVRC, and we found that this further improves the sentiment prediction, yielding a 7.15% relative improvement over [You *et al.*, 2014]. Notably, the English model performs the best for fine-tuning, outperforming all other languages as well as DeepSentiBank [Chen *et al.*, 2014a], likely due to the fact that annotators were English and also that Twitter is dominated by English-speaking users.

5.6 Conclusions

In this chapter, we presented multilingual ANP detector banks across six major languages in MVSO using modern CNN structures, including CaffeNet [Jia *et al.*, 2014], VggNet [Simonyan and Zisserman, 2015], and GoogLeNet [Szegedy *et al.*, 2015]. We showed how network classification performance could be further improved by restricting the sourcing of images to tag-based queries from Flickr compared to a hybrid-pool of tag and free-text search results. In addition, two applications using these multilingual ANP detector banks was presented for multilingual image sentiment analysis and image-based query expansion with semantic and sentiment coherence. Lastly, we presented two sentiment prediction tasks to evaluate whether these mid-level affective concepts, ANPs, can also be used for traditional affective representation modeling.

Our cross-lingual analyses of our large-scale MVSO and image dataset using semantic matching (q.v. §4.5) and visual sentiment prediction (q.v. §5.5) hint again that human affect is not necessarily best modeled computationally as culturally universal. These preliminary results show that there are indeed commonalities, but also distinct separations, in how visual affect is expressed and perceived, where other works thus far assumed only commonalities. We believe these point to the colorful diversity of our world, rather than our cultural inability to understand one another.

In the future, we seek to explore more complicated network structures for ANP detection at scale as well as explore training sentiment and emotion detectors jointly with our mid-level ANP concepts. In addition, in this chapter, we have treated the ANPs in MVSO as a flat taxonomy rather than a nested set of concepts as defined in §4.3. Network structures that take advantage of these hierarchical semantics are likely to lead to interesting future insights into multicultural visual affect.

Chapter 6

Cross-task Affective Visual Concept Detection

As we scale out the diversity (or "variety") in Visual Affective Computing, computationally modeling and exploiting relationships between visual affective targets becomes important. In Chapter 3, we explored how such relationships might be leveraged through the use of multitask learning on discrete emotion classes (q.v. §3.6). In the context of mid-level representations, multitask learning can likewise be used to great benefit to enable compact and generalized feature representations as well as improved accuracy as a result of joint learning. In this chapter, we extend a recent class of deep neural networks to perform multitask visual affective recognition of mid-level representations, specifically, adjective-noun pairs (ANPs).

Residual learning [He *et al.*, 2016a] has recently surfaced as an effective means of constructing very deep neural networks for object recognition. However, current incarnations of residual networks do not allow for the modeling and integration of complex relations between closely coupled recognition tasks or across domains. Such problems are often encountered in multimedia and vision applications involving large-scale content recognition. We propose a novel extension of residual learning for deep networks that enables intuitive learning across multiple related tasks using cross-connections called cross-residuals [Jou and Chang, 2016a]. These cross-residuals connections can be viewed as a form of in-network regularization and

enables greater network generalization. We show how cross-residual learning (CRL) can be integrated in multitask networks to jointly train and detect visual concepts across several tasks. We present a single multitask cross-residual network with >40% less parameters that is able to achieve competitive, or even better, detection performance on a visual sentiment concept detection problem normally requiring multiple specialized single-task networks. The resulting multitask cross-residual network also achieves better detection performance by about 10.4% over a standard multitask residual network without cross-residuals with even a small amount of cross-task weighting.

6.1 Introduction

In concept detection, leveraging the complex relationships between learning tasks remains an open challenge in the construction of many multimedia and vision systems. While some recent approaches have begun to model these relationships in deep architectures [Deng *et al.*, 2014; Wu *et al.*, 2014], still many solutions tend to have multiple parts that specialize rather than a more versatile, general solution that leverages cross-task dependencies. As an illustration, visual sentiment prediction is a rising topic of interest in multimedia and vision. In [Borth *et al.*, 2013b], a semantic construct called adjective-noun pairs (ANPs) was proposed such that there are visual concept pairs like 'happy girl,' 'misty woods' and 'good food'. These semantic concepts serve as a bridge between vision-based tasks that are focused on object (or "noun") recognition and affective computing tasks that are focused on qualifying the affective capacity or strength of multimedia, e.g., through the "adjective" in the ANP. However, even though the tasks of object recognition, affect prediction and ANP detection all have some relation to each other, the construction of classifiers for each is treated independently. In this chapter, we propose a novel method for jointly learning and generalizing across tasks which can be easily and very efficiently integrated into a deep residual network and show how it can be used for visual sentiment concept detection.

To understand how "relatedness" is both important and applicable to visual concept detection, consider several example images and concepts from [Borth *et al.*, 2013b] in Figure 6.1. In the example, we observe that the ANP 'shiny cars' can be superclassed by both the

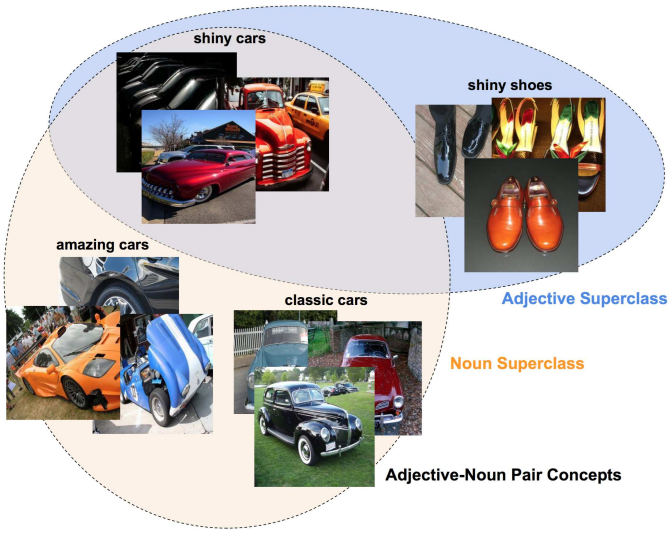


Figure 6.1: Example of related visual concept detection tasks that directly benefit from our proposed cross-residual learning (CRL). Adjective-noun pairs can be superclassed by their noun or adjective components. Exploitable visual and semantic similarities exist within (intra-relatedness) as well as between superclasses (inter-relatedness).

'shiny' adjective category and 'cars' noun category. Within the 'shiny' adjective category, there are other concepts like 'shiny shoes' that bear both semantic and visual similarities to the 'shiny cars' ANP. This *intra-relatedness* also exists within the noun superclass which includes ANPs like 'amazing cars' and 'classic cars'. In addition to relatedness within the same (super)class, we observe that there are visual similarities also present between classes of different superclasses, e.g., 'classic cars' and 'shiny shoes'. This *inter-relatedness* between (super)classes illustrates how in settings like concept detection, classifiers can benefit from exploiting representational similarities across related tasks. Both of these senses of *relatedness* show that visual representations across related tasks can be shared to a degree. We develop a multitask learning problem for visual affective concept detection to illustrate how our proposed method can be applied. We design a deep neural network with a stack of shared low-level representations and then higher level representations that both specialize and mix information across related tasks during learning. We then show how such a multitask network architecture with cross-task exchanges can be used to simultaneously learn classifiers to detect adjective, noun and adjective-noun pair visual concepts.

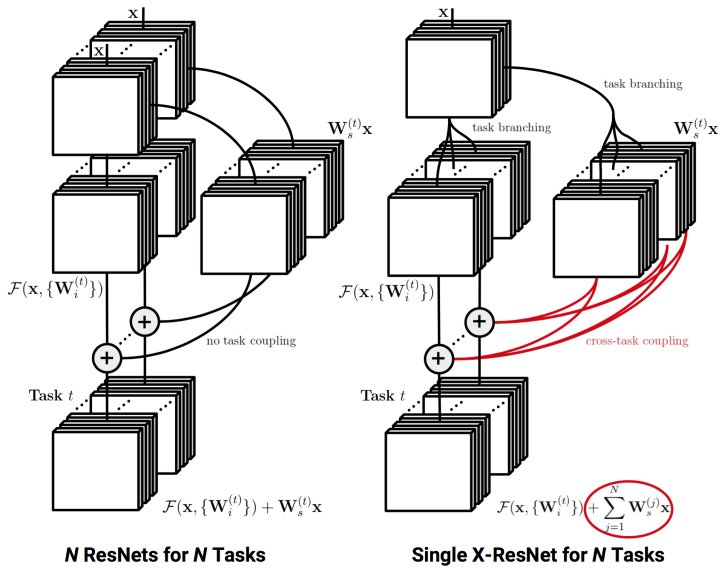


Figure 6.2: Feature Map Illustration of Residual Network (ResNet) and Cross-residual Network (X-ResNet) Layers. X-ResNet extends ResNet to enable structures like multitask networks where a single network can jointly perform multiple related tasks instead of requiring one network per task. Our network uses cross-task connections indicated in red to simultaneously enable specialization per task and overall generalization.

In [He *et al.*, 2016a], residual learning is proposed as an approach to enable much deeper networks while addressing the *degradation* problem where very deep networks have a tendency to *underfit* compared to shallower counterpart networks [Srivastava *et al.*, 2015]. In residual learning, an identity mapping via the use of shortcut connections [Raiko *et al.*, 2012] is proposed where an underlying mapping $\mathcal{H}(\mathbf{x}) = \mathcal{F}(\mathbf{x}) + \mathbf{x}$ is learned given that $\mathcal{F}(\mathbf{x}) = \mathcal{H}(\mathbf{x}) - \mathbf{x}$ represents another mapping fit by several stacked layers. One interpretation is that $\mathcal{F}(\mathbf{x})$ represents a noise term and the model is fitting the input plus some additive nonlinear noise. Thus, if we were performing reconstruction, a trivial solution to the residual learning problem is that an identity mapping is optimal, i.e., $\mathcal{F}(\mathbf{x}) = 0$. However, in [He *et al.*, 2016a], it is argued that optimization software may actually have difficulty with approximating identity mappings with a stack of nonlinear layers, and also that for prediction problems, it is unlikely that the strict identity is optimal. They also argue that fitting residual mappings can enable deeper networks given the information boost achieved via the

shortcut connection and thus reduces the likelihood of model degradation. Recently, there is even work suggesting a relationship between residual networks, recurrences and the primate visual cortex [Liao and Poggio, 2016]. Our work extends residual learning [He *et al.*, 2016a] to also integrate information from other related tasks enabling cross-task representations. Specifically, we hypothesize and experimentally show that reference components from correlated tasks can be synergistically fused in a residual deep learning network for *cross-residual learning*.

Our contributions include (1) the proposal of a novel extension of residual learning [He *et al.*, 2016a] using cross-connections for coupling multiple related tasks in a setting called cross-residual learning (CRL), (2) the development of a multitask network with a fan-out architecture using cross-residual layers, and (3) an evaluation of cross-residual networks on a multitask visual sentiment concept detection problem yielding a single network with very competitive or even better accuracy compared to individual networks on three classification tasks (noun, adjective, and adjective-noun pair detection) but uses >40% less model memory, while also outperforming the predictive performance of a standard multitask configuration without cross-residuals by about 10.4%.

6.2 Related Work

This work broadly intersects three major lines of research areas: transfer learning, deep neural architectures for vision, and affective computing. In traditional data mining and machine learning tasks, we often seek to statistically model a collection of labeled or unlabeled data and apply them to other collections. In general, the distributions of these sets of data collections are assumed to be the *same*. In transfer learning [Pan and Yang, 2010], the domain, tasks and distributions are allowed to be *different* in both training/source and testing/target. In this work, we specifically focus on a subset of transfer learning problems that assume some *relatedness* between these collections. Specifically, in multimedia and vision contexts, *relatedness* may refer to settings where groups of tasks have semantic correlation, e.g., classifying dog breeds and bird species, or visual similarity, e.g., jointly classifying and reconstructing objects, and is often referred to as multitask learning (MTL) [Caruana, 1997;

Zhou *et al.*, 2012] (q.v. §3.6). Likewise, relatedness may also refer to the same source task but applied in different domains, e.g., classifying clothing style across cultures, and is sometimes called cross-domain learning [Jiang *et al.*, 2008] or domain transfer/adaptation [Glorot *et al.*, 2011; Jiang *et al.*, 2009]. Nonetheless, the hypothesis of explicitly learning from related tasks is that we can learn more generalized representations with minimal performance cost or in some cases, leading to gains from learning jointly.

Multitask networks are recently becoming a popular approach to multitask learning, riding on successes of deep neural networks, and have several recent applications in vision and multimedia [Huang *et al.*, 2015; Sudowe *et al.*, 2015; Rudd *et al.*, 2016; Wang *et al.*, 2016]. One early work in [Collobert and Weston, 2008] showed how a single network could be trained to solve multiple natural language processing tasks simultaneously like part-of-speech tagging, named entity recognition, etc. Multitask networks have since proven effective for automated drug discovery [Dahl *et al.*, 2014; Ramsundar *et al.*, 2015], query classification and retrieval [Liu *et al.*, 2015], and semantic segmentation [Dai *et al.*, 2016]. Recently, [Ghifary *et al.*, 2015] proposed multitask auto-encoders for generalizing object detectors across domains; and in [Luong *et al.*, 2016], multitask sequence-to-sequence learning is proposed for text translation. Also, architectures like [Rasmus *et al.*, 2015; Yim *et al.*, 2015] can be categorized as multitask networks since they reconstruct and classify simultaneously. Unlike other multitask networks but similar to ladder networks [Rasmus *et al.*, 2015], instead of a single branching point in our network that creates forked paths to only specialize to individual tasks, we continue mixing information even after branching via our cross-skip connections.

Whereas multitask learning can generally be understood as a fan-out approach where a (usually, single) shared representation is learned to solve multiple tasks, an analogous complement is a fan-in approach where multiple either features or decision scores are fused together to solve a single-task. For example, graph diffusion can be used smooth decision scores for leveraging intra-relatedness between categories [Jiang *et al.*, 2009]. In [Deng *et al.*, 2014], instead of an undirected graph, explicit directed edges were used to model class relationships like exclusion and subsumption. And with some semblance to our work, in [Wu *et al.*, 2014], a multimodal neural network structure is developed where inter-

class (but still intra-task) relationships are integrated as an explicit regularizer. Although inspired from multitask learning, the network design in [Wu *et al.*, 2014] still operates in a single-task context as there is only a single output network head. Additionally, because the network integrates multiple input feature towers, the overall memory and training burden of the image-to-decision pipeline is much greater than a fan-out network alternative.

In [Narihira *et al.*, 2015], "factorized" neural network are proposed, which are essentially just a multitask network with two networks heads for predicting noun and adjective targets topped off with a simple matrix-multiply operation to get an adjective-noun matrix that has as many rows as adjectives and columns as nouns. A fundamental problem with this "factorization" approach is that not all the entries of the resulting matrix are not valid adjective-noun pairs in the sense that not all adjectives can be paired with every noun and vice-versa. As a result, there is a sparse subset of elements in the output matrix that are actually semantically valid, sentimentally-biased and colloquially popular. It is worth noting that in [Narihira *et al.*, 2015], they also propose the use a "Fork-Net" which omits the matrix-multiply operation, resulting in a standard two-head multitask network with adjective and noun prediction. However, for all their baselines as well as their "Fork-Net" and factorized network, the target task is ANP detection and all the evaluations presented reflect this. Meanwhile, we develop a deep multitask cross-residual network able to simultaneously predict noun, adjective and adjective-noun visual concepts.

6.3 Cross Residual Learning

Given an input $\mathbf{x}$ and output $\mathbf{y}$ vector to a residual learning layer and the mapping function $\mathcal{F}(\mathbf{x}, \{\mathbf{W}_i\})$ to fit, where for vision problems this might represent, for example, a stack of convolutional operations with batch normalization [Ioffe and Szegedy, 2015] and ReLU activation [Nair and Hinton, 2010], we have the following formulation in residual learning [He *et al.*, 2016a]:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{\mathbf{W}_i\}) + \mathbf{W}_s\mathbf{x}, \tag{6.1}$$

where $\mathbf{W}_s$ is an optional linear projection, but required when matching dimensions, on the shortcut connection. For identity shortcut connections, $\mathbf{W}_s = \mathbf{I}$.

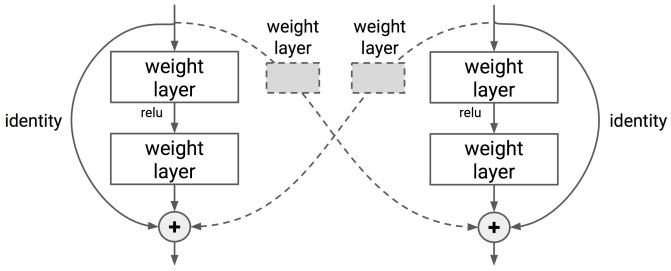


Figure 6.3: Cross-residual Building Block (with two tasks). Cross-residual weight layers and cross-skip connections are dashed and allow for network-level flexibility over task specialization.

Here, we propose a simple and efficient extension of [He *et al.*, 2016a] when fitting across multiple related learning tasks which we refer to as *cross-residual learning* (CRL). Given a task t and $N-1$ other related tasks, we define the task output of the cross-residual module as:

$$\mathbf{y}^{(t)} = \mathcal{F}(\mathbf{x}^{(t)}, \{\mathbf{W}_i^{(t)}\}) + \sum_{j=1}^{N} \mathbf{W}_s^{(j)} \mathbf{x}^{(j)}, \tag{6.2}$$

where the superscript $(\cdot)$ indexes the target task and a normalization factor is omitted for simplicity and can be lumped with the shortcut weights $\mathbf{W}_s^{(j)}$. As also illustrated in Figure 6.3, the other target tasks additively contribute to the current target task t by $\sum_{j \neq t} \mathbf{W}_s^{(j)} \mathbf{x}^{(j)}$. The cross-residual contributions can also more generally be stacks of operations $\mathcal{C}(\mathbf{x}^{(j)}, \{\mathbf{W}_{s,m}^{(j)}\})$, but here, we only illustrate the simple weighted once case $\mathbf{W}_s^{(j)} \mathbf{x}^{(j)}$. In addition, we note that in (6.2), we have task-specific inputs $\mathbf{x}^{(j)}$ which can also be the equivalent, e.g., in the first layer after branching in a multitask network.

6.3.1 "Early" Regularization Interpretation

In optimization, when minimizing a loss $\mathcal{L}(f(\mathbf{x}), \mathbf{y})$, we often add a regularization term $\mathcal{R}(f(\mathbf{x}))$ to constrain the "badness" of the solution, factor in assumptions of our system, and reduce overfitting. For example, in solving deep networks, the squared 2-norm is a common choice to penalize large parameter values and smooth network mappings. Cross-residual units can be viewed as a way of regularizing the solution of a specific task by other related tasks, i.e., we do not want the learned mapping $\mathcal{F}(\mathbf{x}^{(t)}, \{\mathbf{W}_i^{(t)}\})$ to be too far from a weighted

combination of task-specialized transformations of the input $\sum_j \mathbf{W}_s^{(j)}\mathbf{x}^{(j)}$. For example, when learning to visually recognize species of birds, we may want to introduce regularization to ensure the mapping fit is not too far from the separate, but related task of recognizing types of mammals. While such a regularization usually takes place in the loss layer of a neural network, using cross-residual layers we can introduce this task conditioning "earlier" in the network and also stack them for additional information mixing. Cross-residual layers thus serve as a type of in-network regularization much like dropout [Srivastava *et al.*, 2014], though with less stochasticity.

6.3.2 Connection to Highway Networks and LSTMs

As also discussed in [He *et al.*, 2016a], residual networks can be seen highway networks [Srivastava *et al.*, 2015] that do not have transform or carry gate. In highway networks, an output highway layer is defined as

$$\mathbf{y} = \mathcal{H}(\mathbf{x}, \mathbf{W}_{\mathcal{H}})\mathcal{T}(\mathbf{x}, \mathbf{W}_{\mathcal{T}}) + \mathbf{x} \cdot \mathcal{C}(\mathbf{x}, \mathbf{W}_{\mathcal{C}}), \tag{6.3}$$

where $\mathcal{T}$ and $\mathcal{C}$ are the transform and carry gates, respectively. Clearly, when both gates are on, this is precisely the same as a residual layer. By extension, a cross-residual layer can be thought of as an ungated highway layer with multiple "highways" merging onto the same information path. Cross-residual weighting layers then are carry gates which govern the amount of cross-task pollination.

Similarly, it has been argued that residual layers can also be viewed as a feed-forward long short-term memory (LSTM) [Hochreiter and Schmidhuber, 1997] units without gates. Specifically, consider the LSTM version from [Gers *et al.*, 2002]:

$$\left.\begin{aligned}
\mathbf{i}_k &= \sigma(\mathbf{W}_{\mathbf{xi}}\mathbf{x}_k + \mathbf{W}_{\mathbf{hi}}\mathbf{h}_{k-1} + \mathbf{b}_{\mathbf{i}}) \\
\mathbf{f}_k &= \sigma(\mathbf{W}_{\mathbf{xf}}\mathbf{x}_k + \mathbf{W}_{\mathbf{hf}}\mathbf{h}_{k-1} + \mathbf{b}_{\mathbf{f}}) \\
\mathbf{c}_k &= \mathbf{f}_k\mathbf{c}_{k-1} + \mathbf{i}_k \tanh(\mathbf{W}_{\mathbf{xc}}\mathbf{x}_k + \mathbf{W}_{\mathbf{hc}}\mathbf{h}_{k-1} + \mathbf{b}_{\mathbf{c}}) \\
\mathbf{o}_k &= \sigma(\mathbf{W}_{\mathbf{xo}}\mathbf{x}_k + \mathbf{W}_{\mathbf{ho}}\mathbf{h}_{k-1} + \mathbf{b}_{\mathbf{o}}) \\
\mathbf{h}_k &= \mathbf{o}_k \tanh(\mathbf{c}_k)
\end{aligned}\right\}, \tag{6.4}$$

where k indexes the timestep, $\mathbf{i}$, $\mathbf{f}$ and $\mathbf{o}$ are the input, forget and output gates, $\mathbf{c}$ and $\mathbf{h}$ are the cell and output states, all respectively, and peephole connections and some bias terms

are omitted for simplicity. By ignoring recurrent connections $k-1$ for the feed-forward case and making the LSTM completely ungated, i.e., $\mathbf{i} = \mathbf{f} = \mathbf{o} = \mathbf{I}$, and initializing the cell state to the input $\mathbf{c}_{k-1} = \mathbf{x}$, we are left with a residual layer. Again by extension then, cross-residual layers can be thought of as feed-forward, ungated LSTMs whose cell states are additively coupled. LSTM forget gates then are analogous to cross-residual weight layers. And indeed, this is much like highway networks' carry gate, since highway layers can be viewed as feed-forward LSTMs with only forget gates [Srivastava *et al.*, 2015]. A major difference to note though is that cross-residual layers couple the transformed input $\mathcal{H}$ with *multiple* and usually *different* prior cell states $\mathbf{c}_{k-1}^{(t)}$ or information highways $\mathbf{x}^{(t)}$.

6.3.3 Similarities to Ladder Networks

Structurally, the building blocks of cross-residual learning bears some resemblance to the layout in ladder networks [Rasmus *et al.*, 2015]. In ladder networks, two encoders and one decoder joined via lateral connections are used to jointly optimize a weighted sum over a cross-entropy and reconstruction loss and have thus proven successful in semi-supervised learning. As part of the reconstruction process, a Gaussian noise term is injected in one of the encoders and the decoder receives a combination of this noisy signal via a lateral connection and a vertical "feedback" connection to reconstruct the original input into the noisy encoder. Since the mapping term $\mathcal{F}(\mathbf{x})$ in residual learning can be viewed as noise term, albeit learned unlike in ladder networks, both models essentially are trying to fit the input subject to some additive nonlinear noise. For cross-residual learning, although we use shortcut connections instead of lateral connections as in ladder networks, both designs operate on the principle that combining channels of information at the same structural level in the network can ultimately result in a model with higher learning capacity under less constraints, e.g., for ladder networks, less labeled data requirements since it is semi-supervised.

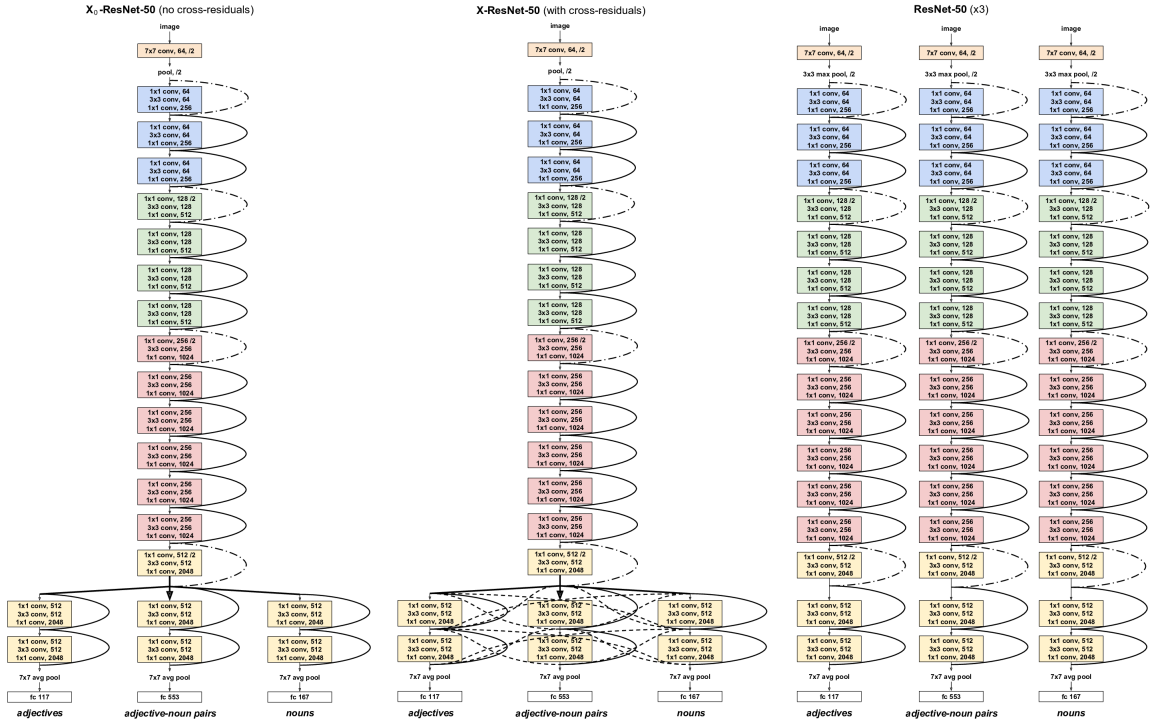


Figure 6.4: Example network architecture layouts for a standard multitask residual network, a multitask cross-residual network, and single-task residual networks, respectively, with 50 layers. Solid shortcuts (—) indicate identity, dash-dotted (-·-·-) shortcuts indicate 1×1 projections, and dashed (- - -) shortcuts indicate cross-residual weighted connections. Residual weight blocks show three convolutions grouped.

6.4 Multitask Cross Residual Networks

While there may be a number of settings that would benefit from cross-residual learning, we focus on one natural setting here in multitask learning (MTL) [Caruana, 1997]. To implement a multitask network, a common approach is to introduce a branching point in the architecture that leads to one network head per task [Collobert and Weston, 2008; Ghifary *et al.*, 2015; Liu *et al.*, 2015; Ramsundar *et al.*, 2015], e.g., see Figure 6.2. In Table 6.1 and Figure 6.4, we show 50-layer multitask residual networks with a branching point at the last input size reduction. The earlier in the network this branching point is introduced the larger the input feature map size is to the individual network heads, often resulting in multitask networks with a large memory footprint. On the other hand, if the branching

point begins deeper in the network, the representational specialization available for each task is limited to a small space of high-level abstract features. In our design of a multitask cross-residual network (X-ResNet), we address this latter problem by allowing additional cross-task mixing via cross-residual weights which cheaply increases late-layer representational power without requiring large input feature spaces. While it is possible to completely forego a branching point in the network design and simply couple multiple network towers using cross-residual skip connections, this results in a composite network that is very memory intensive and only feasible in a multi-GPU environment (though this could be somewhat alleviated by freezing weights, e.g., in combination with greedy layerwise training).

In addition, to introduce some task specialization, at the branching point in our multitask network design and before the cross-residual layers, we move the last ReLU activation and batch normalization canonically present inside the residual building block outside, placing it after the elementwise addition such that there is one per task. This helps to produce a slightly different normalization for each task branch and in practice, slightly improves performance. As in most multitask networks with a branching point, the total network loss is taken to be a combination of each of the individual network head losses. While some tune the loss weight for each of these network heads, we simply use the unweighted sum over all the network head losses.

6.5 Multitask Visual Sentiment

To illustrate the utility and effectiveness of cross-residual layers when used in multitask networks, we pose the mid-level visual sentiment concept detection in a multitask context. In particular, we use the visual sentiment ontology (VSO) [Borth *et al.*, 2013b] and cast affective mid-level concept detection as a multitask learning problem. We chose the VSO dataset for our experiments over multilingual visual sentiment concept ontology (MVSO), i.e., Chap. 4, simply because VSO is more dated and there are a larger body of works to benchmark against for VSO compared to MVSO. We note that more general vision problems could also have been use to illustrate cross-residual learning, e.g., CIFAR-100 [Krizhevsky, 2009] where we might choose to predict classes and superclasses simultaneously,

Output Size	Adjective	Adj-Noun Pair	Noun
112×112	7×7, 60 /2		
56×56	3×3 max pool /2		
56×56		$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	
28×28		$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	
14×14		$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	
7×7	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
1×1	avg pool	avg pool	avg pool
	117-d fc	553-d fc	167-d fc
	softmax	softmax	softmax

Table 6.1: Multitask Residual Network with 50 layers (without cross-residuals). Bracketed blocks are stacked residual building blocks. Downsampling is performed by stride 2 after stacked residual blocks.

but here we are expressly interested in affective mid-level representations. The ANP detection problem in VSO can be recast to naturally fit the multitask setting with a sufficiently large accompanying image corpus over three tasks, i.e., adjective, noun and adjective-noun pair, while other general vision image datasets are often smaller and/or only consist of two learning tasks which yield a small number of task interactions.

Given the diversity of adjective-noun pairs, including concepts like 'cute dress,' 'gentle smile,' 'scary skull,' 'wild rose' and 'yummy cake,' there is both a considerable amount of semantic variance in VSO as well as inter-class visual variance due to the image data being gathered from social media streams. As a result, to cope with this diversity and

variance, we believe that exploiting cross-task correlations as part of the network design is important, especially when the tasks are tightly related as they are with noun, adjective, and adjective-noun pair concept detection.

We additionally note that even though VSO [Borth *et al.*, 2013b] argues that the noun component of the ANP serves to visually ground the mid-level concept, no experiments were actually ever run to determine the performance of detecting adjective (or even, noun) concepts separately.[1] Our evaluation thus also serves as the first evaluation on the VSO dataset to benchmark noun-only and adjective-only detection performance.

6.5.1 Multitask-structured Visual Sentiment Ontology

Briefly, recall that the data in VSO [Borth *et al.*, 2013b] was originally collected from the social multimedia platform, Flickr, using psychology-grounded seed queries from *Plutchik's Wheel of Emotions* [Plutchik, 1980] which consists of 24 basic emotions, such as *joy*, *terror*, and *anticipation*. The query results yielded images with user-entered image tags which were annotated using a part-of-speech tagger for identifying adjective and noun components and parsed for sentiment strength. The identified adjective and noun components were combined, checked for semantic consistency and filtered based on sentiment strength then used to feed back as queries to Flickr to filter based on frequency of usage. A subsampling of adjective-noun pair combinations is then done to prevent many adjective variations on any one noun, resulting in the final visual sentiment ontology. The adjective-noun pairs were then used to query and pull down an image corpus from Flickr, limiting to at most 1,000 images per concept.

The image dataset in VSO [Borth *et al.*, 2013b] has a long tail distribution where some adjective-noun pair concepts are singletons and do not share any adjectives or nouns with other concept pairs. As a result, we use a subset of VSO and use it to perform adjective, noun, and ANP concept detection in social images, specifically, as a multitask learning problem. The original VSO dataset [Borth *et al.*, 2013b] consists of a refined set of 1,200 ANP concepts. Since there are far less adjectives that serve to compose these adjective-noun pairs, and also some nouns that are massively over-represented in the ontology, we

[1] From independent communication with the authors.

filtered to keep concepts that matched the following criteria: (1) adjectives with ≥3 paired nouns, (2) nouns that are not overwhelmingly biasing, v.s. *face* or *flowers*, and non-abstract, unlike *loss*, *adventure* or *history*, and (3) ANPs with ≥500 images. It is helpful to think of ANPs as a bipartite graph with nouns and adjectives on either side and valid ANPs as edges. From these conditions, we obtained a visual sentiment sub-ontology, suitable for multitask learning, that normalized the number of adjective and noun nodes while ensuring maximal ANP edge coverage. The final multitask-flavored VSO contains 167 nouns and 117 adjectives which form 553 adjective-noun pairs over 384,258 social images from Flickr.

6.5.2 Experiments & Discussion

In our experiments, we use an 80/20 partition of the multitask VSO data stratified by adjective-noun pairs resulting in 307,185 images for training and 77,073 for test at 224×224. All our residual layers use "B option" shortcut connections as detailed in [He *et al.*, 2016a] where projections are only used when matching dimensions (stride 2) and other shortcuts are identity. Except for cross-residual weight layers, projections are performed with a 1×1 convolution with ReLU activation and batch normalization as in [He *et al.*, 2016a]. For our cross-residual weight layers $\mathbf{W}_s^{(j)}$, we use the identity on self-shortcut connections $\mathbf{W}_s^{(t)} = \mathbf{I}$ and a cheap channelwise scaling layer for cross-task connections $\mathbf{a} \odot \mathbf{x}$, $\forall\ j \neq t$ which adds no more than 2,048 parameters each, i.e., so in our case, after branching we have $\mathbf{x} \in \mathbb{R}^{7\times7\times2048}$ and so $\mathbf{a} \in \mathbb{R}^{1\times1\times2048}$ for scaling.

For training multitask networks, we initialized most layers using weights from a residual network (ResNet) model trained on ILSVRC-2015 [Russakovsky *et al.*, 2015], but done such that for layers *after* the branching point in our network we initialize them to the *same* corresponding layer weights in the original ResNet model. For cross-residual weight layers, we follow [He *et al.*, 2016a] and initialize them as in [He *et al.*, 2015], i.e., zero mean random Gaussian with a $\sqrt{2/n_l}$ standard deviation where we set n_l to be the average of input and output units layerwise. No dropout [Srivastava *et al.*, 2014] was used in residual or cross-residual networks. We use random flips of the input at training. We trained our cross-residual networks with stochastic gradient descent (SGD) using a batch size of 24, momentum of 0.9 and weight decay of 0.0001. We used a starting fixed learning rate of

0.001 and decreased it by a factor of ten whenever the loss plateaued until convergence. All networks and experiments were run using a single NVIDIA GeForce GTX Titan X GPU and implemented with Caffe [Jia *et al.*, 2014].

We baseline against four single-task architectures: a variant of the AlexNet architecture [Krizhevsky *et al.*, 2012] swapping pooling and normalization layers called CaffeNet [Jia *et al.*, 2014], the first iteration of the GoogLeNet architecture [Szegedy *et al.*, 2015] denoted as Inception-v1 which uses a bottlenecked 5×5 convolution in the sub-modules, the 16-layer version of VggNet [Simonyan and Zisserman, 2015] (VggNet-16), and the ResNet architecture [He *et al.*, 2016a] with 50-layers (ResNet-50). Each of these single-task architectures were fine-tuned from an ImageNet-trained model and represent competitive baselines that achieved top ranks in ILSVRC tasks in the past. In addition, we also evaluated against DeepSentiBank [Chen *et al.*, 2014a], also an AlexNet-styled model trained on the full, unrestricted VSO data [Borth *et al.*, 2013b] to detect 2,089 ANPs. We did not retrain [Chen *et al.*, 2014a] but rather re-evaluated their model on the subset of 553 ANP concepts we focus on here; however, since we do not know the train and test image splits that they used, the result provided for DeepSentiBank [Chen *et al.*, 2014a] could still be an overestimate. In Figure 6.4 (rightmost), we show the learning and inference paradigm represented by these single-task architectures with residual networks (ResNets) used as an example. Each of these baselines treat the adjective, noun and adjective-noun recognition tasks as independent targets.

We summarize network parameter costs and top-k accuracy on the multitask VSO tasks in Table 6.2. For network parameter costs, note that for Inception-v1 [Szegedy *et al.*, 2015] we did not count the parameters from auxiliary heads although they are used during training. Top-k accuracy denotes the percentage of correct predictions within the top k ranked decision outputs.

6.5.2.1 Adjective vs. Noun vs. ANP Detection

In general, as originally posited in [Borth *et al.*, 2013b], in terms of problem difficulty ordering, noun prediction is indeed "easier" as visual recognition task than adjective prediction. However, though not in stark contrast to [Borth *et al.*, 2013b], and although there are indeed

	Task	#Parameters	Top-1	Top-5
Chance	Noun	–	0.60	2.96
	Adj	–	0.86	4.20
	ANP	–	0.18	0.90
DeepSentiBank [Chen *et al.*, 2014a]	ANP	65.43	7.86	11.96
CaffeNet [Jia *et al.*, 2014]	Noun	57.55	36.11	63.48
	Adj	57.35	23.84	51.20
	ANP	59.13	18.84	41.57
Inception-v1 [Szegedy *et al.*, 2015]	Noun	10.82	39.93	67.98
	Adj	10.66	26.32	55.57
	ANP	12.00	20.48	45.01
VggNet-16 [Simonyan and Zisserman, 2015]	Noun	134.94	41.64	69.51
	Adj	134.74	28.45	57.77
	ANP	136.53	22.68	47.70
ResNet-50 [He *et al.*, 2016a]	Noun	23.90	41.64	69.81
	Adj	23.80	28.41	57.87
	ANP	24.69	22.79	47.82
X_0-ResNet-50	Noun	43.16	40.06	68.06
	Adj		26.81	56.09
	ANP		20.74	45.46
X_I-ResNet-50	Noun	43.16	28.61	56.52
	Adj		17.98	43.10
	ANP		12.56	31.49
X_s-ResNet-50	Noun	43.18	**42.18**	**70.04**
	Adj		**28.88**	**58.50**
	ANP		**22.89**	**48.54**

Table 6.2: Number of Parameters (millions) and Top-k Accuracy (%) on Multitask VSO. Note that X-ResNet-50 are multitask networks so classifiers are trained jointly in a single network while other methods train one specialized network per classification task.

more ANP classes than nouns and adjectives, we still did expect to observe higher accuracy rates for ANP concept detection than we did, expecting that the rates would be much closer to that of noun detection and not lower than adjective detection since [Borth *et al.*, 2013b] argues that adjectives lack visual grounding. We suspect that this difference by almost a half at top-1 between noun and ANP detection may point to the difficulty of the ANP detection problem in a slightly different sense than difficulty for the adjective detection problem. For adjective detection, visual recognition difficulty is likely to arise from visual variance, e.g., there may be a wide range of visual features required to describe the concept 'pretty'. However, for ANP detection, we believe that visual recognition difficulty is more likely due to visual nuances than overall visual variance. Much like fine-grained classification, this may imply that in ANP concept detection, concepts like 'sad dog' and 'happy dog' may share many visual characteristics but differ on few but highly distinguishing traits. The hope then is that by using a scaling layer, which acts as a soft gating mechanism in cross-residual connections, these few but distinguishable characteristics are accentuated.

6.5.2.2 Effects of Cross-residual Weighting

In Table 6.2, we also show results for multitask cross-residual networks with different types of weighting: no cross-residual weighting ($\mathrm{X}_{\mathbf{0}}$-ResNet-50), with all identity cross-residual weights ($\mathrm{X}_{\mathbf{I}}$-ResNet-50) and with identity on the self-task connections and channelwise scaling on just cross-residuals as described earlier ($\mathrm{X}_{\mathbf{s}}$-ResNet-50). The multitask cross-residual networks without and with cross-residuals are illustrated in Figure 6.4 (leftmost and center, respectively), and all of these multitask networks use a residual network with 50-layers (ResNet-50) as the basis and branch as described in Section 6.4. As we might expect, when *all* cross-residual weights are identity ($\mathrm{X}_{\mathbf{I}}$-ResNet-50), the accuracy of the multitask network across all tasks drastically reduces since the "amount" of cross-task mixing is forced to be equally weighted. Even as related as tasks might be forcing cross-residual weights equal across all tasks makes it difficult during learning for any single task to specialize and determine discriminative patterns useful for that specific task. It may be tempting to then assume that the other extreme of making the cross-residual weights zero where $\mathbf{W}_s^{(j)} = \mathbf{0}, \forall\, j \neq t$, i.e., equivalent to a multitask network without cross-residuals ($\mathrm{X}_{\mathbf{0}}$-ResNet-50), allows more

specialization and would naturally achieve the best discriminative performance. However, we found that this actually consistently achieves lower accuracy across all tasks compared to its single-task equivalents (ResNet-50), e.g., ∼9% worse relative on ANP detection. We hypothesize that without cross-residuals the performance case becomes upper-bounded by the shared representation learned before the branching the multitask network.

Once we allow for even some simple learned weighting on the cross-residuals, like a channelwise scaling (X_s-ResNet-50), the predictive performance of the multitask network improves, outperforming both the case when no cross-residuals are used as well as equally weighted cross-residuals. In general, we observed that multitask networks achieved comparable performance to the three specialized single-task networks with just a single network while requiring less than 60% of the combined parameters of the three single-task networks (∼43.2M vs. ∼72.4M). This confirms our original hypothesis that the low-level representations can be shared across these related tasks and can be generalized to perform well across all tasks. However, in order to ensure that we do not take a hit in accuracy by generalizing, weighted cross-residuals layers can be used which, at a very marginal parameter cost, enable the multitask network to match the performance of specialized single-task networks. Notably, as we had hoped, the highest gain from using cross-residuals was on the most difficult of the three tasks: ANP detection. We observe that adding scaling cross-residual weights improves the concept detection performance by as much as ∼10.37% relative on the ANP detection task compared to without any weighting.

Though we do not claim that our cross-residual multitask network (X_s-ResNet-50) definitively achieves a significantly higher accuracy over the single-task networks, we do note that we observed marginally better concept detection rates with our network across all tasks. Since we only used two cross-residual layers in our multitask network (c.f. Figure 6.4), it is possible that increasing the number of stacked cross-residual layers or beginning the branching in the network earlier could improve the overall cross-task performance; however, doing so would naturally come at increased parameter cost. Nonetheless, we believe that all of these observations show that jointly learning across related tasks with cross-task information mixing even at the late layers of a network can simultaneously improve the network's capacity to discriminate and generalize.

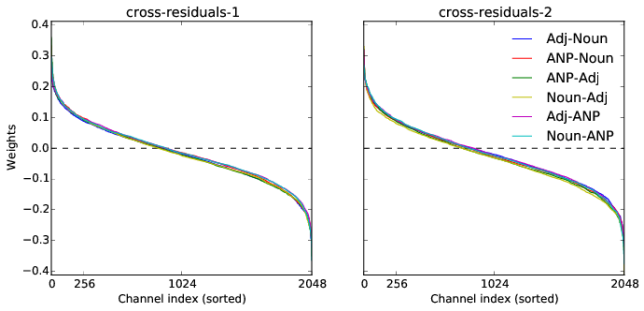


Figure 6.5: Example learned unnormalized cross-residual weights (sorted). Legend notation refer to cross-residual connections as `SourceTask-TargetTask`. Left and right plots show cross-residual weights of the first and second (feed-forward direction) cross-residual layers as in Figure 6.4 (center), respectively.

To further reinforce the fact that the optimal weighting for cross-residual connections are unlikely to be zero or identity, in Figure 6.5, we show the unnormalized weight magnitudes of a learned multitask cross-residual network sorted by channel index for two cross-residual layers in a network structured as in Figure 6.4 (center). If an all zero or identity cross-residual connection were to be optimal, we would expect to see a plateau with many weights near zero or one. Instead, we observe that mostly non-negative cross-task weights were learned across all shortcut connections such that the overall network objective was optimized. Additionally, we note that though the weight magnitudes are indeed small, this also follows from intuition in the original residual network work [He *et al.*, 2016a] that these small, but non-zero weights are precisely what enable residual networks to be made very deep.

6.5.2.3 Example Multitask Detection Results

In Figure 6.6, we show example classification results from our multitask cross-residual network. Note the presence of both intra- and inter-relatedness between tasks in the top detected concepts. In many cases, the cross-residual network is able to surface concepts not visually present but intuitively related; for example, in the first image, 'spider' is a detected noun which may be a result of either the branches in the image or the visual co-occurrence of the 'spider' concept in the training set with other top ranked concepts like 'tiny' (adjective)

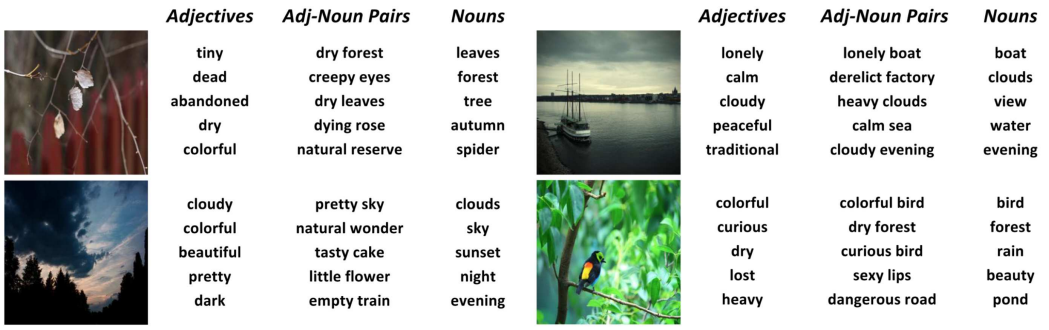


Figure 6.6: Example top-5 classification results of adjective, noun, and adjective-noun pair concepts using our multitask cross-residual network.

and 'leaves' (noun). As a potential failure case, in the last image, the ANP 'sexy lips' was ranked highly possibly due to relatedness learned with the 'colorful' adjective concept. In these cases, just as with over regularization in other learning settings, the network may have indeed have learned a more general representation, but as a result is unable to decouple certain learned relationships. Such cases may be easily addressed in cross-residual networks by giving cross-task weighting layers more computational budget, e.g., convolutional projections, to model more complicated task relationships. Overall, we observe here that the multitask cross-residual network is able to successfully co-detect concepts across multiple related visual recognition tasks.

6.6 Conclusions

In this chapter, we presented an extension of residual learning enabling information mixing between related tasks called *cross-residual learning* (CRL) achieved by coupling the residual to other related tasks to ensure the learned mapping is not too far from other task representations. This enables more generalized representations to be learned in a deep network that are useful for multiple related tasks while preserving their discriminative power. We also showed how cross-residuals can be used for multitask learning by integrating cross-residual layers in a fan-out multitask network. We showed how such a multitask cross-residual network can achieve competitive, or even better, predictive performance on a visual sentiment concept detection problem as compared to specialized single-task networks but with >40%

less parameters, while also outperforming a standard multitask residual network with no cross-residuals by about 10.4% relative on adjective-noun pair detection, the hardest of the three related target tasks. Without cross-residual connections, we observed a ∼9% drop in accuracy on ANP detection, indicating the importance of using cross-residuals. In addition, we showed the importance of cross-residual weighting over simply forcing identity cross-residual connections since equally weighting cross-task connections bottlenecks the information flow in the network.

Here, we only presented experiments on a subset of the VSO image corpus suitable for multitask learning, but in the future, we would also like to extend this to the multilingual visual sentiment concept ontology (MVSO) image corpus. In the multilingual context, using multitask networks would allow for exploration of cross-lingual relationships as well as correlations between ANP clusters. We also believe cross-residual networks are also applicable to other learning settings and domains, and can be extended in several ways. Cross-residual networks can be applied to other multitask learning settings where we are not only interested in classification but also other tasks like reconstruction, object segmentation, etc. Likewise, cross-residual networks are likely to be useful in domain transfer and adaptation problems where, for example, network tower weights are frozen but cross-residual weights are learned. Architecturally, while we only explored the canonical shortcut connections of [He *et al.*, 2016a] and used a channelwise scaling layer for the cross-residual, there is recent work exploring different types of transforms and gating on shortcuts [He *et al.*, 2016b] that can also be extended to the self- and cross-connections in cross-residual networks. We plan to explore these learning settings and network architectures in the future.

Part III

Open Challenges and Future Work

Chapter 7

Implicit Affect Detection in Full-length Films

In this ongoing work, we discuss a novel affective gap which leads to what we call an *implicit affective computing* problem. We propose that physiological signals are themselves a measurement of our human affect, however less interpretable and semantic. By reducing these physiological signals to a set of detectable traits though, we can re-utilize familiar content-based prediction tools to learn functional mappings from stimuli to traits like biometric markers. We discuss preliminary experiments of a pilot study using audiovisual stimuli from data collected in an unconstrained environment with long-running stimuli of >40 minutes with multiple subjects using electrodermal activity, a skin conductance measurement known be correlated with affect arousal.

7.1 Introduction

The early 20$^{\text{th}}$ century American poet, E. E. Cummings, famously wrote in his 1973 poem titled "Since feeling is first":

> *Since feeling is first*
> *who pays any attention*
> *to the syntax of things*
> *will never wholly kiss you;*

While Cummings was no psychologist or affective computing researcher, he hinted that in the order of events in our human experience is that we feel something, and then use words to convey what was felt. In psychology and social science, this poetic example and observation has been referenced and debated repeatedly over the years [Zajonc, 1980]. Today, this notion that "feeling" comes before "thinking" is widely accepted in the scientific community [Ekman, 1999; Davis and Lang, 2003; Lieberman, 2007]. In computational research, this order prominence can partially be attributed to the development of research and engineering efforts in Affective Computing where "feelings" are inputs, like those measured by physiological signals, and outputs are "thought"-related concepts, such as emotion semantics. While this focus on "the syntax of things," e.g., affect semantics like *sadness*, *liking*, *high arousal* and *low valence*, remains a critical area of study, it has also become so exclusively the area of focus in recent years that a wider perspective can be lost. Though these traditional affective representations are important for research to communicate the state of affect (q.v. §2.1.2), they are still one degree of abstraction removed from the affect itself. In this ongoing work, we propose a novel intermediate, and even uniquely separate, computational problem that we argue is worth investigating and that takes a conceptual step back from semantically grounded visual affect concept prediction. Essentially, we relax the semantic or dimensional conditioning on the representation of affect, and instead explore the relationship between the stimulus and our feelings directly as measured by *markers in* or *traits of* our physiological signals.

Recalling the discussion from §2.2, to bridge the "affective gap" there are two established approaches: (1) we present a subject with a stimulus, e.g., audiovisual or textual media, hoping to evoke a bodily response that we can measure through biometrics like electrodermal activity (EDA), electroencephalogram (EEG), functional magnetic resonance imaging (fMRI), face video, etc, or, (2) we use the original stimulus directly using content-based methods from computer vision, audio processing, natural language processing, etc. As we had summarized in Figure 2.2b and re-illustrated in Figure 7.1a, the affective gap can refer to the divide between either physiological signals and affect state, or between stimulus and affect state. Generally, such affect states are described in terms of discrete emotions as in [Ekman, 1999; Plutchik, 1980] or in the dimensional representations of valence-arousal-

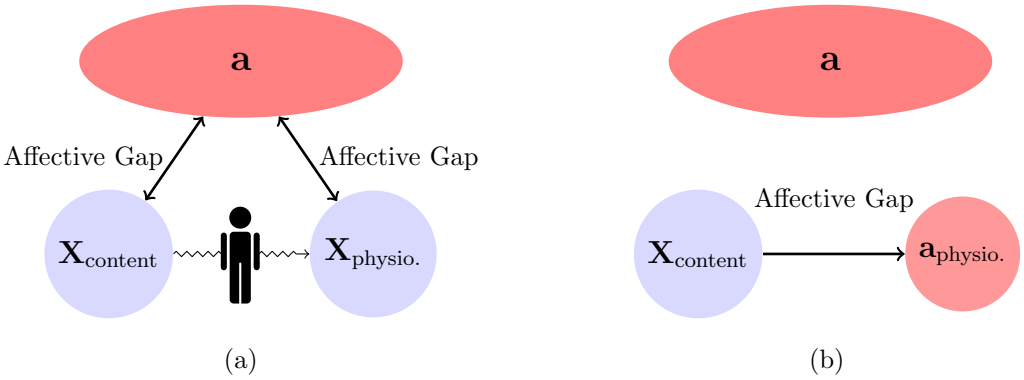


Figure 7.1: Implicit Affective Computing: **(a)** We can try to bridge the *affective gap* by using content-based methods from the stimuli $\mathbf{X}_{\text{content}}$ or use features extracted from physiological signals $\mathbf{X}_{\text{physio.}}$ to predict the affect $\mathbf{a}$; or **(b)** as we propose, we can also treat traits of physiological signals as a type of affect state $\mathbf{a}_{\text{physio.}}$ and bridge from the stimuli.

dominance [Gunes and Pantic, 2010].

However, whether taking the route of learning a mapping from physiological signals to affect states or from stimulus to affect states directly (or even, a fusion of the two), the actual relationship between stimulus and physiological signals had remained largely unexplored in computation. As shown in Figure 7.1b, we propose that there is another type of affective gap that may be considered between stimulus and physiological signals. The psycho-physiological signals then "becomes" an implicit representation of affect, as compared to the explicit representations discussed in §2.1.2. However, since the mapping between a potentially high-dimensional and continuous stimulus with a simultaneously continuous biometric is likely difficult, we hypothesize that reducing the continuous physiological signal to discrete traits like binary spiking activity can serve as learnable "labels" or "markers" using content-based methods. By this, we do not make any claims of a new psycho-physiological framework, but rather, explore a new computational framework not previously explored, as far as we know.

The main goals of this work include: (1) the conceptualization of an "affective gap" previously ignored in Affective Computing between stimuli and physiological signals, (2) a proposal to bridge this gap by distilling physiological signals down to detectable, learnable traits for use with familiar machine learning tools, and (3) benchmarking content-based

experiments with audiovisual stimuli for learning and predicting such physiological traits.

7.2 Related Work

As discussed in §2.3.1 some early attempts for affect recognition relied on only face and body expressions of subjects, but others focused on physiological signals [Chanel *et al.*, 2005] like electromyography [Haag *et al.*, 2004] and fMRI [Cunningham *et al.*, 2004]. Some works embraced more non-invasive measurement approaches [Lisetti and Nasoz, 2004], enabling user-facing applications like video entertainment [Fleureau *et al.*, 2012] and audience engagement monitoring [Silveira *et al.*, 2013]. Additionally, several datasets have become publicly available from this research effort, including DEAP [Koelstra *et al.*, 2011] and MAHNOB HCI-Tagging [Soleymani *et al.*, 2012], which both focus on multiple biometric responses to audiovisual stimuli (mentioned briefly also in §3.2). Also, as already discussed in other areas of this thesis, there have been a number of prior successes in recognizing affect states from stimuli using content-based methods (e.g., see §2.3.2). Here, we seek to use non-invasive EDA measurements of subjects for the task of predicting traits extracted from the measurements via the stimulus itself.

For the paradigm shown in Figure 7.1b, work in [Koelstra *et al.*, 2011] is likely the most similar. In their experiments, they fused content-based features consisting of low-level audio features, like energy, pitch and zero crossing rate, as well as visual features, like color, shadow proportion and shot length variance along with feature from physiological signals, like average skin resistance, band-limited energy, eye blink rate and median peak-to-peak time. For their DEAP dataset based on music video stimuli, they found that decision, or late, fusion of content-based and EEG modalities performed best for arousal classification, and fusion of content-based and peripheral physiological signal modalities performed the best for valence classification. Like the previous paradigm, while this bares close relationship to our problem setting, our target variable is inherently different since we do not seek to predict arousal, valence, or any such affect state that can necessarily directly tied to semantics. Instead, we seek a mapping between stimuli and physiological signals by first reducing the physiological signals to detectable traits and then using content-based

methods to learn a mapping from stimuli to these traits.

7.3 Spike Detection from Electrodermal Activity

In our preliminary experiments, we focus on audiovisual stimuli using a full-length movie and a television episode in an uncontrolled setting. We chose to focus on audiovisual stimuli for their rich expressiveness as well as the temporal dimension that they have that add to the experience. We used two full-length motion pictures: one movie title, *Flight* (2012)[1], and a television episode from the American police drama series *NCIS: Naval Criminal Investigative Service*, specifically Episode 9 of Season 11 titled "Gut Check" (2013)[2]. These two motion picture films were chosen for their drama/triller/mystery rooted plot lines. Both films were rendered at 23.98 frames/sec.

We showed both films in their unedited, original form to subjects in an unconstrained theater setting and measured subjects' electrodermal activity (EDA) responses with the Affectiva[3] Q Sensor. In the case of Flight, we also measured subjects' heart rate and accelerometer data, but for comparison with the NCIS data, we ignored their contributions. We calibrated all devices and synchronized the start of the stimulus with our measurements.

Film	**Dur** (min)	**Subjects** (M/F)	**Age Range**
Flight (2012)	138	22 (9/13)	20-49
NCIS S11E09	44	27 (14/13)	23-58

Table 7.1: Film and subject statistics from full-length feature film showings for electrodermal activity (EDA) response recording toward implicit affect detection.

In this initial study, we specifically focus on the use of EDA as our "target affect", e.g., as illustrated in Figure 7.2. Electrodermal activity has been studied for affect and engagement to motion picture widely over the years [Fleureau *et al.*, 2012; Kaiser and Roessler, 1970]. For our context, because we seek to predict traits of the EDA, we need to reduce the

[1] `http://www.imdb.com/title/tt1907668`

[2] `http://www.imdb.com/title/tt3268322`

[3] `http://www.affectiva.com` (Note: the Q Sensor product has since been retired)

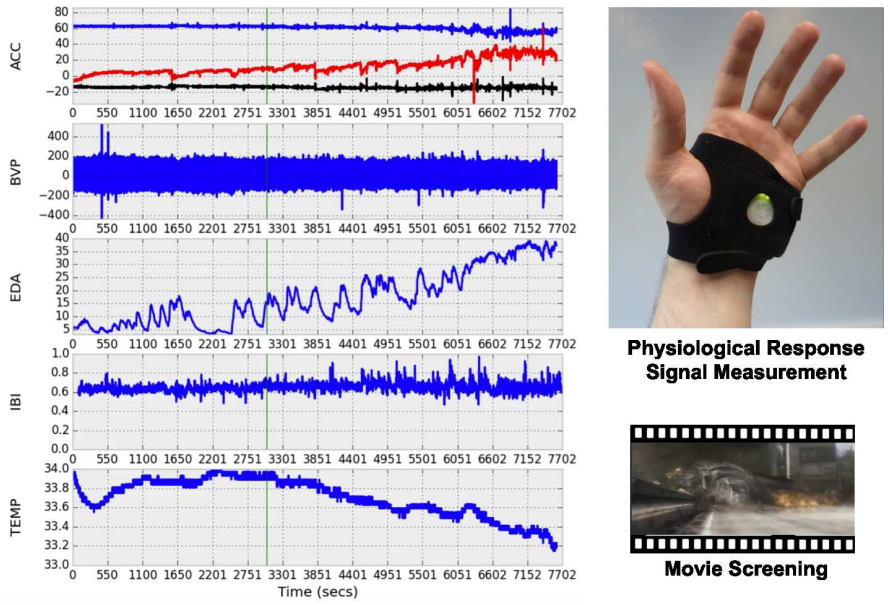


Figure 7.2: Example of the author's own physiological signal capture data in response to a full-length feature film. A synchronized frame from the film is shown which corresponds in time to the green vertical marker to sensor measurements like accelerometer (ACC), blood volume pulse (BVP), electrodermal activity (EDA), inter-beat intervals (IBI) and skin temperature (TEMP) data.

dimensionality of the signal. To do this, we apply the spike detection method proposed in [Silveira *et al.*, 2013] which uses an adaptive decomposition method to greedily decompose the input EDA signal into a set of dictionary components which contains a temporal offset parameter. The dictionary covers a wide range of feasible exponential skin conductance response shapes which can be parameterized as:

$$\mathbf{d}_{\lambda_1,\lambda_2,t_0}(t) = \begin{cases} \lambda_1^{-\lambda_2(t-t_0} & t \geq 0 \\ 0 & t < t_0 \end{cases} \tag{7.1}$$

where λ_1 corresponds to the impulse decay of the skin conductance, λ_2 is a log-linear decay, and t_0 is the response start time. In [Silveira *et al.*, 2013], the decomposition is performed by beginning with a high-pass filtered EDA signal and initializing the dictionary with a single component, determining the best fit using orthogonal matching pursuit, updating the dictionary, computing a residual signal, and repeating for some fixed number of iterations.

7.4 Inferring Physiological Spiking Activity from Stimuli

Motivated by a recent study advocating ranking ratings of affect over classification and regression [Martínez *et al.*, 2014], we formulate our problem as a learning to rank problem over slices of the audiovisual stimuli, which we call micro-videos. Our micro-film audiovisual slices are obtained by windowing over the temporal audiovisual sequence and treating each window as an instance in our ranking framework. Inspired by work in [Redi *et al.*, 2014], where they specifically studied creative versus non-creative videos on the popular social media platform Vine[4], we similarly restrict our focus to window sizes of six (6) seconds. In the context of motion picture films and affective understanding, this length choice is motivated by two other intuitions. First, the delay between electrodermal activity and stimulus onset is known to be between 1-3 seconds and since we are using temporal stimuli we conservatively double the longest delay for a window size of six seconds. In addition, through a commercial user study in social multimedia Vine determined that six seconds is an artistically and expressively good choice[5].

7.4.1 Multimedia Content Analysis Features

For the full-length film, Flight, at 23.98 frames/sec, we had 199,367 frames over an input frame size of 1280×536 and 48kHz audio at 95kb/s, while for the television episode from NCIS, we had 63,280 frames at an input size of 720×404 and 48kHz audio at 116kb/s. At the most low-level description, we used standard color histograms in HSV color space using 18×3×3 bins, respectively. We also experimented with Gist [Oliva and Torralba, 2001], a global image feature accepted widely for its ability to estimate the "shape of a scene". We used a set of aesthetic features [Bhattacharya *et al.*, 2013] shown to be effective for in application to videos, albeit applied on a frame-by-frame basis. In addition, we use two flavors of the mid-level representation SentiBank [Borth *et al.*, 2013a]: the vanilla SVM-based detectors as well as DeepSentiBank [Chen *et al.*, 2014a]. We also used feature representations extracted using AlexNet [Krizhevsky *et al.*, 2012] which was trained on ILSVRC12

[4] `https://vine.co`

[5] `http://www.npr.org/player/embed/213846816/213902185`

Descriptors	**Modality**	**Dimensionality**
Color Histogram	Visual	162
Gist [Oliva and Torralba, 2001]	Visual	512
Aesthetic [Bhattacharya *et al.*, 2013]	Visual	139
SentiBank [Borth *et al.*, 2013a]	Visual	1200
DeepSentiBank [Chen *et al.*, 2014a]		
∗ softmax prob	Visual	2089
∗ fc8		2089
∗ fc7		4096
CaffeNet [Jia *et al.*, 2014]		
∗ softmax prob	Visual	1000
∗ fc8		1000
∗ fc7		4096
openSMILE [Eyben *et al.*, 2013]		
∗ emobase	Audio	988
∗ emobase2010		1582

Table 7.2: Summary of input multimedia content analysis features for implicit visual affect detection in full-length feature films.

[Russakovsky *et al.*, 2015], specifically the Caffe variant [Jia *et al.*, 2014]. For both representations based on deep network architectures, CaffeNet [Jia *et al.*, 2014] and DeepSentiBank [Chen *et al.*, 2014a], we further experimented with features from the output softmax layer which correspond to probabilistic semantic decisions as well as outputs from last two fully connected layers (*fc8* and *fc7*). All the above features focus on the representations from the visual modality of the input stimulus, for the audio stream, we extracted a set of features using openSMILE [Eyben *et al.*, 2013], which includes low-level descriptors like Mel-Frequency Cepstral Coefficients (MFCC), auditory model based loudness, F_0 envelope, and zero crossing rate with various functionals applied such as min/max, standard deviation, skewness, etc. We extracted two emotion-based feature sets using openSMILE: *emobase*, a seminal reference set used in many other prior speech-based emotion prediction tasks, and *emobase2010* [Schuller *et al.*, 2011].

7.4.2 Ranking Micro-videos from Electrodermal Activity

For learning to rank, we used a ranking SVM [Joachims, 2003], specifically an efficient kernel implementation [Kuo *et al.*, 2014], to rank windows of six seconds with preference labels given by binarized events detected from electrodermal activity [Silveira *et al.*, 2013]. In our preliminary experiments, the EDA event labels were aggregated across the entire audience. Specifically, given training label/query/instance tuple sets $(y_i \in \mathbb{R}, q_i \in S \subset \mathbb{Z}, \mathbf{x}_i \in \mathbb{R}^n)$, $i = 1, \ldots, n$ and a set of preference pairs $P = \{(i,j) \,|\, q_i = q_j, y_i > y_j\}$, RankSVM [Joachims, 2003] solves

$$\begin{aligned} \min_{\omega,\xi} \quad & \frac{1}{2}\omega^T\omega + C \sum_{(i,j)\in P} \xi_i \\ s.t. \quad & \omega^T(\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)) \geq 1 - \xi_{i,j}, \\ & \xi_{i,j} \geq 0, \ \forall\, (i,j) \in P \end{aligned} \tag{7.2}$$

where $C > 0$ is a regularization parameter, ϕ is a kernel function that maps data into a higher dimensional space, and $\xi_{i,j}$ is called the ℓ_1 loss.

For the color histogram feature, we use the chi-squared kernel $\phi(\mathbf{x}, \tilde{\mathbf{x}}) = 1 - \sum_{i=1}^{n} \frac{(\mathbf{x}_i - \tilde{\mathbf{x}}_i)^2}{(\mathbf{x}_i + \tilde{\mathbf{x}}_i)/2}$ while for all other features, we use the radial basis function kernel $\phi(\mathbf{x}, \tilde{\mathbf{x}}) = \exp(-\gamma|\mathbf{x} - \tilde{\mathbf{x}}|^2)$. Parameters are learned by cross-validation over a grid search. In addition, due to the combinatorial number of constraints that are introduced to the optimization and the class imbalance of far more negatives than positive instances, we discard negatives samples with 50% probability during training.

When windowing, we allow for a 50% overlap and ensured that we did not randomly shuffle our data for train and test partitioning to avoid temporal overlap bias. Trials over the audiovisual stimulus came from adjacent data partitions where we remove boundary test data instances overlapping training splits. For frame-based visual features, we experiment with several pooling schemes including taking the maximum response in each feature dimension, taking the average, and selecting the feature vector from the center of the window. Audio-based feature vectors are extracted strictly within the designated six-second windows, i.e., they are given no additional context. We measure our performance by computing the average precision within each partition.

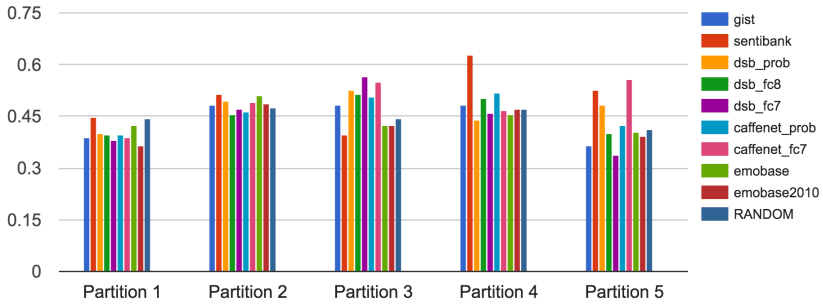


Figure 7.3: Average precision ranking performance from EDA spike events for an episode from the television show, NCIS, with a window size of six seconds and average pooling for visual features across several adjacent partitions of the film.

In the preliminary experiments shown in Figure 7.3, it is difficult to tell if one feature performances significantly better than another even in the context of a single partition. In addition, we observe that in general our learned rankers are only able to predict marginally better than chance on each partition. Interestingly, we do observe that the SentiBank features [Borth *et al.*, 2013b], which are the most affectively biased features evaluated, consistently perform well on a majority of the film's partitions. We believe that even though the problem of predicting traits of human bio-signals from content-based visual multimedia is an ambitious task, the fact that there is some marginal improvement over chance at this early stage of investigation shows that we can still push the boundary further.

7.5 Open Issues

One main challenge with the current setup is that it is difficult to decouple sources of errors in the implicit affect detection process. For example, although it is known that EDA has a correlation with affect arousal, it does not necessarily mean that a single spiking event is correlated to an individual's arousal, e.g., it could be spiking frequency instead. In addition, the spike detection method [Silveira *et al.*, 2013] is not without noise and since the method operates greedily, it tends to produce false detections. Even assuming such issues could be resolved, there is also no guarantee that a windowed portion of a film that is labeled as a 'spiking event' will bare semblance to another window. While the reasons for

this are many, in a full-length film, narrative context could have caused an individual to be aroused much like visual elements which are much more challenging to extract automatically. Alternatively, this could be solved by reducing the stimulus length; however, doing so would introduce subject calibration noise from film showing to showing that may be impossible to normalize out due to the unconstrained recording environment. Despite these challenges, we believe this ambitious new proposed problem setting of *implicit affective computing* poses great potential, paving the way for computationally modeling biological affect states jointly with content-based multimedia analysis.

In the future, since we focused on audience-level ranking here, we plan to experiment with and study the effects of subject-specific and demographic-specific ranking, e.g., aggregating by age or gender. In addition, we believe there are opportunities for further improving the performance of our rankers by multimodal fusion, e.g., by simply performing kernel-level fusion and setting kernel weights for each representation by cross-validation or multiple kernel learning. It may be interesting to also determine whether the detection of these physiological signal traits can serve as a mid-level representation for ultimately detecting semantic/dimensional affect as well.

Chapter 8

Mid-level Movie Concepts for Visual Affect Detection

Thus far in this thesis, outside of ongoing work in Chapter 7, we have given little treatment to affective computing in the context of video streams – everywhere else, we have focused on images since the visual affect detection problem is already a largely untreated and difficult problem in that context. In this ongoing work, we seek to explore hierarchically-organizable *mid-level affective concepts* (q.v. Part II and §4.3) with specific connections to *perceived* emotion (q.v. §2.2 and §3.3) in the vertical of movies/films (much like in Chap. 7), all at scale, using deep networks for visual affect (q.v. Chap. 5 and 6). The hope is to apply many of the ideas proposed already in this thesis together in a single framework for affect recognition in videos.

8.1 Introduction

Cinematography has always sought to move alongside and guide the affections of an audience throughout the course of a film. From a carefully composed score on the backdraft of rolling green hills to a dramatic fight scene laced with explosions, the role of the film director is in some sense crafting the affective experience of their viewers frame by frame. The hope then is that we can develop computational machinery that can not only detect the visual affect of a movie scene, but also output the specific semantic characteristics that led to

that decision in the inference process. Consider the example of the fight scene, perhaps the presence of 'explosions,' 'gunshots' and a 'dark alley' are all semantic indicators that should cue an affect detector to aspects of 'fear' and/or 'high arousal' are in this visual clip.

Recall that in bridging the "affective gap" (q.v. §1.1.2), one approach is to use mid-level representations that allow for grounding machine detectable concepts as well as an affective bias (q.v. Part II). In this ongoing work, we seek to develop useful mid-level representations in the form of audiovisual concepts commonly occurring in movies and film to allow for more interpretable and improved classification performance affect predictors. Such a system would allow for fine-grained affective analysis of events, entities and affective states throughout the temporal progression of an audiovisual film stream. In particular, we build upon the popular movie information summary platform, IMDb, and construct a movie concept ontology with affectively biased audiovisual concepts, subsequently mine associated film trailers, acquire temporally localized annotations, and develop computational machinery to recognize both the movie concepts and the affective states in the video clips.

The main goals of this work include: (1) the development of an affectively biased movie concept ontology suitable as a mid-level representation of affective states in film, (2) the proposal of a three-shot composite called "shot-triplets" that sit between fine-grained shots and coherent scenes, which we argue is well-suited for Affective Computing in the context of audiovisual streams, (3) a movie trailer dataset with concept and perceived emotions annotations temporally localized at the "shot-triplet"-level, and (4) visual detector banks useful for movie concept detection as well as perceived emotion prediction.

8.2 Related Work

Our particular focus on movies content comes from the many prior works that have investigated the application of affect to movies in the past, given film's clearly defined use case as a media and well-defined added business value. Among the earliest, [Hanjalic and Xu, 2005] developed affective trajectories of movie clips, plotted in the valence-arousal space. In [Wang and Cheong, 2006], the correlation between genre of movies was investigated in seven emotion states. Work in [Teixeira *et al.*, 2011] and [Ellis *et al.*, 2014b] continued along the

same lines but focused on fusing audio and visual features. Then in [Canini *et al.*, 2013], with some semblance to our work, mid-level features were proposed that included style distinctives which they called "film grammar" features, encompassing elements like shot length, distance-to-camera and lighting conditions. And recently, in [Baveye *et al.*, 2015a], fine-tuned deep networks were used to predict valence-arousal states. All these form a strong precedent of seminal work, but still largely rely on low-level features for affect state prediction, where even in [Canini *et al.*, 2013], the features are not as semantically interpretable and perceptually identifiable as one would prefer, e.g., 'fire' or 'humming'. The largest video-based affect study we are aware of is in [Vandal *et al.*, 2015] which we discussed in §4.2, but does not focus on films and relies exclusively on facial expressions.

Despite all this prior art, there are very few publicly available datasets for visual affect research in videos and films. In [Schaefer *et al.*, 2010], a small dataset called FilmStim of 64 short film clips and LIRIS-ACCEDE [Baveye *et al.*, 2015b], a dataset of 9,800 clips, are proposed using the PANAS and valence-arousal representations, respectively (mentioned briefly also in §3.2). The sparsity of movie datasets for affect has largely been due to licensing restrictions and so works like [Baveye *et al.*, 2015b] focus on niche and little-viewed films licensed under Creative Commons[1]. In the same spirit as MVSO (q.v. Chap. 4), we endeavor to develop a large-scale movie ontology with affective mid-level concepts associated with movie clips that are useful for modeling perceived emotional states.

8.3 IMDb Movie Concept Ontology

For a viable mid-level semantic representation, we desire movie concepts that are used by actual viewers to describe scenes. As a result, we take an approach similar to VSO and MVSO (q.v. Chap. 4) of mining these movie concepts from a social multimedia platform. To this end, we use the Internet Movie Database (IMDb)[2], and mine 'plot keywords' that have been entered by users on the website. The hope is that some subset of these keywords will directly occur in audiovisual form in the content of the movie itself, and because they are

[1] `https://creativecommons.org`

[2] `http://www.imdb.com`

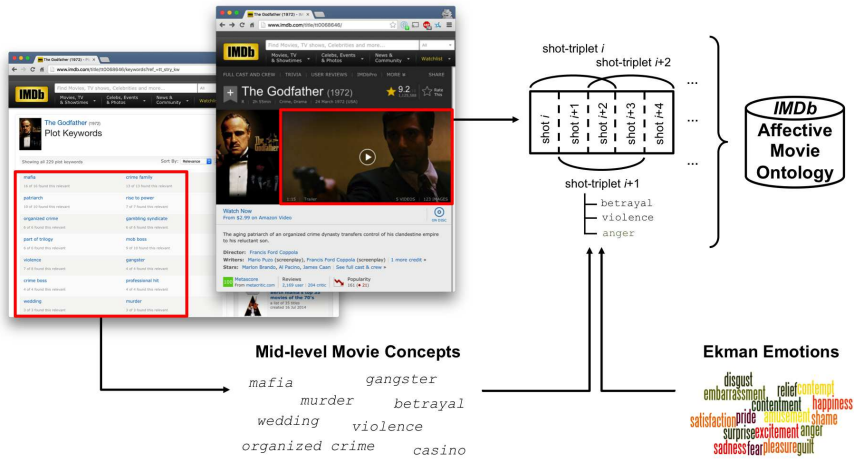


Figure 8.1: Overview of the IMDb Affective Movie Ontology Construction Process. We mine plot keywords and genre-organized movie trailers from IMDb and connect them to emotions with near-shot-level temporal localization.

entered by users they will be entities that have specifically been both *perceived* by viewers of the film and assessed as useful (and even, memorable) semantic descriptors.

To determine which films to mine these plot keywords from, we began by targeting a subset of the movie genre categorizations on IMDb. On IMDb, movies are grouped into multiple genres including: action, adventure, comedy, fantasy, thriller, sport, war, western, romance, horror, crime, documentary, musical, biography and film-noir. In addition, IMDb implements an in-house current popularity algorithm they call "movieMETER"[3] which ranks movies in each genre. In late October 2014, we downloaded movie trailers from the top-400 of this rank list in each genre given that a trailer existed. We omitted the musical, biography and film-noir genres due the natural of their tendency to ill-suited for audiovisual detection tasks. Note that we did not acquire 400 *unique* movies for each genre since a movie can fall under multiple genres. We also additionally filtered movie trailers based aspect ratio to remove poor quality trailers, removing anything greater than 16:9 and less than 4:3, resulting in 1,042 unique movie trailers across 10 movie genres mined from IMDb. We chose to focus on movie trailers here rather than full-length movies for their potential to scale with much greater ease as well as the reduced licensing restrictions given

[3] `http://www.imdb.com/chart/moviemeter`

their wide, public accessibility on the Web.

After mining these movie trailers, we parsed their associated metadata, e.g., release date, cast, plot, location and budget, as well as associated 'plot keywords'. Some examples of popular plot keywords we mined included *murder*, *blood*, *death*, *friendship*, *flashback*, *love*, *police*, *kiss*, *chase* and *female nudity*. More telling and interesting though were some of the less-popular movie concepts on the long-tail of the distribution like *bayonet*, *space travel*, *eye patch*, *ex-soldier* and *airplane accident*. In total, we collected 998 movie concepts using this mining process on IMDb, essentially matching the scale of the number of classes found in the ILSVRC [Russakovsky *et al.*, 2015]. We note that, in-post, we had three expert judges annotate all 998 concepts and determined that 876 movie concepts were likely to be machine "detectable" (87.78% coverage), e.g., concepts like 'father-daughter relationship' and 'based on book' were deemed undetectable. Every concept appeared in at least 20 movies, and 211 concepts appeared in at least 100 movies.

8.3.1 Movie Trailer Concept Annotation

Movie trailers are typically intentionally designed to create brief, but impressionable affect responses and serve as a preview, giving viewers a broad understanding of the storyline. So while the movie concepts we mined from IMDb are intended to describe the full movie, we can still reasonably expect that a sizable subset of the concepts will occur in the trailers. However, because these plot keywords are annotated per movie on IMDb, the labels are not only potentially noisy but also weak. As a result, we seek to develop an annotation task to temporally localize the occurrence of these movie concepts in our movie trailers to later train detectors on.

To localize these annotations, we first performed shot and scene detection on our movie trailers with [Sidiropoulos *et al.*, 2011], yielding approximately 75 shots per trailer. Shots lengths ranged from 0.5 seconds to 20 seconds. Most of the trailers consisted of as many as 165 detected shots with only 31 trailers with less than 20 shots. There were a total of 82,959 detected shots over 1,613 detected scenes across all our movie trailers.

Given some of the observations from §7.4 that there is a slight delay in the affect onset (although usually still faster than cognitive processes) and also to alleviate the annotation

Emotion	#Pos. Labels	Emotion	#Pos. Labels
Anger	20,208	Sadness	6,802
Disgust	2,596	Surprise	29,238
Fear	28,877	Neutral	54,040
Happiness	16,910		

Table 8.1: Perceived Emotion Annotation Counts in Movie Trailers Mined from IMDb at the Shot-Triplet Granularity. The unaggregated number of shot-triplets positively labeled (pos. label) on MTurk for a given Ekman emotion [Ekman, 1999] is shown.

burden, we form what we call "shot-triplets" from the shots of a trailer. Much like our choice of six second windows in §7.4, a sufficient trade-off between contextual information and a minimum affect onset period is to allow for several visually consistent shots. We define a shot-triplet as a contiguous audiovisual stream composited by three consecutive shots, so for example, given five shots in a trailer, there are three shot-triplets where the triplets have some temporal overlap with each other. The shot-triplet composition allows for a convenient, psychologically-inspired middle ground between fine-grained audiovisual slices like frames and shots and coarse-grained groupings like scenes or an entire video.

Given these shot-triplets, we developed an annotation task on Amazon's Mechanical Turk (MTurk) to label the occurrence of our movie concepts. In addition, we also had workers annotate their perception of seven discrete semantic emotions from [Ekman, 1999] in the shot-triplets: anger, disgust, fear, happiness, sadness, surprise and neutral. Note that based on our earlier research in Chapter 3 (i.e., §3.3), we particularly targeted *perceived* emotion given its greater affinity to objective affect labels. To date, we annotated a subset of our full movie trailer corpus, including 198 annotated trailers for 5,303 non-overlapping shot-triplets.

In Table 8.1, we show the unaggregated number of shot-triplets positively annotated for each emotion. Interestingly, we observe that the 'disgust' and 'sadness' emotions occur significantly fewer times than the other emotions. Intuitively, this may be due to the fact that we chose to work on movie trailers and director's make a cinematic choice to avoid such scene clips in these short preview to maximize viewer engagement and curiosity. On the

Anger	Disgust	Fear	Happiness	Sadness	Surprise
drug addict	wristwatch	key	text messaging	grief	body landing on car
impalement	vomit	dream sequence	engagement ring	fishing	critically bashed
punched in the chest	ex-convict	killing an animal	drug use	bloody nose	engineer
kicked	self mutilation	hiding in a closet	marriage proposal	apology	target practice
shot in eye	barefoot	death of child	christmas tree	tear	sabotage
shot in stomach	rape	torso cut in half	factory	sleeping	firework
kicked in head	dead body	blindness	gift	cross	rat
stabbed in stomach	eaten alive	blood on camera lens	hairy chest	death of husband	waitress
shot in leg	champagne	death of friend	male underwear	sadness	shot in face
decapitation	dismemberment	suicide	wine	looking in mirror	crashing through window

Table 8.2: Example of Plot Keywords from IMDb Matched to Emotions. Matches are determined by annotation co-occurrence at the shot-triplet level.

other hand, emotions like 'surprise' and 'fear' were highly annotated which give a sense of suspense to drive a potential audience to the theater for the full-length motion film. Also, it is unsurprising that a considerable number of annotations of shot-triplets were neutral since many triplets were likely contextualizing content that lead up to a shot-triplet with more affective intensity.

8.3.2 Movie Emotion-to-Concept Matching

While our movie concepts are already hierarchically nested with respect to genres in the ontology, we also investigated the relationship between our movie concepts and emotions via our gathered annotations. We computed the co-occurrence of emotion annotations and movie concepts and ranked the concepts by frequency. Naturally, some movie concepts did not co-occur with some emotions. We omitted the 'neutral' emotion since it would have too wide of a variance in matched semantic concepts and would generally yield uninteresting matched concepts.

In Table 8.2, we show the emotion-to-concept matching based on co-occurrences in our annotation data. Many of the matched concepts agree with our intuition for what we expect from a given emotion. For example, 'marriage proposal' and 'gift' have a clear relationship to 'happiness' and likewise, concepts like 'rape' or 'vomit' to 'disgust'. Others, on the surface, are cause for pause, but make sense after consideration; for example, 'drug

addict[s]' are likely portrayed in films as 'angry' or 'anger'-inducing, and the 'cross' concept is likely shown in movies during the death of a character, giving rise to its high co-occurrence with 'sadness'. However, some matched concepts are also surprising, e.g., 'engineer' with 'surprise' or 'text messaging' with 'happiness,' where we can only guess that either there are some associations that do tend to occur often enough in films because there is indeed an affective connection or simply co-occur often by chance in cinematography. In addition, it is worth noting that, just as we argued in Chapter 3 (q.v. §3.6), we observe that many of the matched concepts across emotions are semantically similar, reinforcing our argument that these emotions are not necessarily mutually exclusive from each other.

8.4 Affective Movie Concept Detection

To detect our affective mid-level movie concepts, we formulate a multiple-instance learning (MIL) problem. Briefly, in MIL, given an instance $\mathbf{x}_{ij} \in \mathbb{R}^d$ and associated label $y_{ij} \in \mathbb{R}$, we consider a grouping of instances called "bags" $\mathbf{x}_i = \{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iN_i}\}$. Given a label for the i-th bag $Y_i \in \{-1, 1\}$, we have a positive bag of instances when *any* one of the instances are positive and a negative bag when *all* of the instances in the bag are negative; this relationship can be encoded as $Y_i = \max\{y_{ij}\}$. As an initial proof-of-concept, in our experiments, instances are movie shots and bags correspond to movie trailers, where plot keywords are bag labels.

All the trailers in our dataset together constitute 3,410,986 frames, but we subsample three representative frames output by [Sidiropoulos *et al.*, 2011] within each shot for our experiments. For our initial experiments, we extracted features from an AlexNet network [Krizhevsky *et al.*, 2012] trained on ILSVRC12 [Russakovsky *et al.*, 2015] from our representative frames in each shot-triplet taking the 1000-dimensional softmax outputs and averaging them across the three representative frames. In addition, we computed audio features using openSMILE [Eyben *et al.*, 2013] for *emobase* features as we did in §7.4.1. For learning, we use an approach that adapts the soft-margin classifier to the MIL setting

called multiple-instance SVM (MI-SVM) [Andrews *et al.*, 2002], which solves the problem

$$\begin{aligned} \max_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i, \\ \text{s.t.} \quad & Y_i \max_{\{j\in i\}}(\langle \mathbf{w}, \mathbf{x}_j\rangle + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \quad \forall\, i \end{aligned} \tag{8.1}$$

where bag-level predictions are given by $\hat{Y}_i = \text{sgn}\left(\max_{\{j\in i\}}(\langle \mathbf{w}, \mathbf{x}_j\rangle + b)\right)$.

Given that our current annotations contain considerable class imbalance at the triplet-level, for our preliminary experiments, we manually balanced the dataset and treated audio-based and visual movie concepts for classification separately as well as focused on weak trailer labels. We used a 70/30% train/test split and selected the MI-SVM slack parameter by grid search with cross-validation. We performed experiments over 218 visual concepts and 291 audio-based concepts. Using the openSMILE audio features, we achieved an average classification accuracy of ∼67.3% across all audio-based concepts. The visual features, on the other hand, performed significantly worse with only ∼25.6% of the visual concepts performing better than chance. We suspect this is due to the high visual variance within a trailer and a weak label at the trailer/bag level gives a frame-averaged feature little chance to learn distinctive elements about a given class; meanwhile, for the audio features, since they are extracted over the duration of the entire shot, much more contextual information is given to the representation, allowing the learner to discriminate.

8.5 Open Issues

A primary challenge with the current work is that while we based our intuition and motivations for affective mid-level movie concepts on the successes of MVSO and in other similar vision works, there is not yet sufficient psychology evidence that concepts like those we mined tightly correlate with human emotions (though there is some very preliminary psychology research now on low-level visual features, like the presence of orange and blue tones, in film). Additionally, the number of videos, both at the trailer and shot triplet level, are still not as large as we might desire, as this would essentially require a greater volume of dense crowd annotations. Denser and larger volumes of temporally localized affect and concept annotations will be necessary to pave the way forward for reliable affect detection

at the mid-level in videos. Alternatively, existing datasets could be used to bolster the size of our own, e.g., LIRIS-ACCEDE [Baveye *et al.*, 2015b], where we would compromise on professional high-end content to aid in achieving larger scales. Regardless, we believe that coupling mid-level concept detection with affect state understanding in films will prove particularly useful in affective psychology and sociology as it provides a clean, familiar medium for both researchers and subjects along with explainable, semantic factors to better qualify emotional elements.

In addition, there are many more experiments that need to be done once we can acquire a larger volume of data to improve mid-level movie concept detection as well as experiments toward predicting perceived emotion labels. Specifically, one direction may be to integrate work with cross-residual learning (q.v. Chap. 6) with multiple-instance learning and possibly multimodal features, i.e., multi-view learning. In the preliminary experiments, we performed MIL using trailers as bags in the learning process, but using triplets as bags would yield greater temporal localization provided such fine-grained annotations exist. In future, affect detection in videos, mid-level or otherwise, is also likely to benefit from investigating memory- and attention-based NNs, e.g., LSTMs [Hochreiter and Schmidhuber, 1997].

Chapter 9

Future Directions in Visual Affective Computing

There are many exciting and unexplored challenges in Visual Affective Computing at scale. While we have discussed several areas for future work and improvements at the end of each respective chapter's conclusion, we highlight three specific high-impact areas we believe Visual Affective Computing is likely to benefit from dedicated research in the coming years.

9.1 Other Visual Affect Oracles

In this thesis, we proposed and explored several ways to partition or divide the "affect oracle" (q.v. §2.2 and Fig. 2.2) which result in multiple label-generating distributions rather than just one, including induced vs. intended vs. perceived (q.v. Chap. 3), person-specific (c.f. works like [Koelstra *et al.*, 2011]), and language and geography toward culture (q.v. Chap. 4). As hinted briefly in §5.6, there are naturally many more ways in which the affect oracle can be partitioned. Age group, gender and profession are exemplary and intuitive choices for demographics that would likely bias the affective states of individuals, and thus the labels used in computational learning. However, we believe that there are also far more faceted and nuanced partitionings of the affect oracle to be explored.

Personal interests are one example of such a nuanced partitioning that leads to multiple visual affect oracles. And in fact, there is some psychology research that suggest these

preferential biases influence our judgment of things like aesthetics [Vessel *et al.*, 2014]. For example, though not to the degree of specific individuals, consider the fans of football team A compared to that of team B or C. Certainly, the manner in which such fans will rate certain visual content such as video livestreams of games or images of sports paraphernalia will be biased by the nature of their team affiliations. This idea naturally extends to many other areas like political affiliations, positions on ethics, etc. and even to seemingly far more abstract dimensions like musical aptitude, medical conditions or dietary practices. Further still, composite affect partitionings or even explicitly modeling the partitionings hierarchically are exciting future directions.

9.2 Multimodality for Visual Affective Computing

We have primarily looked at models trained using a single type of feature, e.g., color histograms, local binary patterns, MFCCs, deep learned representations, in this thesis. However, visual affect is likely to benefit from multiple feature representations, i.e., multiple visual features and/or features from multiple modalities like audio, visual and temporal. In fact, some works have already begun to explore the use of multimodal features in Affective Computing [Morency *et al.*, 2011; Koelstra *et al.*, 2011; Soleymani *et al.*, 2012; Perez-Rosas *et al.*, 2013; Jiang *et al.*, 2014; Pang and Ngo, 2015; Wang *et al.*, 2015b], although not all deal with affect in visual multimedia [Soleymani *et al.*, 2016].

Given the success of deep neural networks, even for Visual Affective Computing tasks [You *et al.*, 2014; Xu *et al.*, 2014; Jiang *et al.*, 2014; Pang and Ngo, 2015; Xu *et al.*, 2015; Campos *et al.*, 2015; Jou *et al.*, 2015; Campos *et al.*, 2016; Jou and Chang, 2016a], and recently, multimodal networks, e.g., [Srivastava and Salakhutdinov, 2012; Wu *et al.*, 2014], one promising direction is to explore multimodal networks for multi-oracle affective computing tasks in visual multimedia. Complementary to multitask learning and fan-out network architectures discussed in Chapter 6, fan-in network architectures can be used to implement a set of problems knowns as multi-view learning, i.e., where a single instance with a single label now has multiple input representations. In designs like [Wu *et al.*, 2014], networks can then be developed that are simultaneously multi-view and multitask. Such an approach

would match our intuition that visual affect is a complex objective that requires multiple vantage points to understand as well as a decomposable problem made up of smaller, but related affective subproblems.

9.3 Visualizing Affective States

While we have showed that, despite the ambitious challenge, visual affective states can be predicted with reasonable fidelity in this thesis, a critical missing piece in the greater field of Affective Computing is the ability to visualize and understand what learned models are representing and how they trickle down into machine judgments. This need will becoming increasingly pressing as the community continues to scale-up research. One noteworthy attempt made in this area was in [Hanjalic and Xu, 2005], where affective states in the valence-arousal space for input video streams were plotted as "trajectories" in the 2-D plane. There is a spectrum of existing visualization works in the broader multimedia and vision community from stage-wise visualization, e.g., [Zeiler and Fergus, 2014], to end-to-end generation, e.g., DeepDream[1], that are all viable approaches that should be explored in the context of Visual Affective Computing.

One other promising direction is the use of visual pattern mining [Li *et al.*, 2016a] to localize specific pixel-level regions in visual inputs to identify what image patches may have triggered a certain concept detector. Due to visual variance this method is likely to be less useful for affective representations like discrete emotions or valence-arousal-dominance space (q.v. §2.1.2), but far more effective for mid-level affective states like adjective-noun pairs [Borth *et al.*, 2013b] or movie concepts (q.v. Chap. 8). Being able to identify such image patches, for example in MVSO (i.e., Chap. 4), would allow us to visually compare different cultures, e.g., after a round of image-based clustering.

[1] `http://googleresearch.blogspot.com/2015/06/inceptionism-going-deeper-into-neural.html`

Part IV

Conclusion

Chapter 10

Conclusions

In this thesis, we have developed principles, methods and computational machinery for large-scale affective computing in the context of visual multimedia. We began by proposing the idea that the ground truth generating mechanism in affective computing, i.e., the affect oracle, should be understood as a separable entity (q.v. §2.2 and Fig. 2.2). The principle is intuitive, but has been largely unexplored in affective computing problems, and it has proven effective in this thesis for opening the way for large-scale computable affect in vision and multimedia (q.v. Chap. 3 and 4). Given "large-scale" problems have both a depth *and* breadth component, one interpretation of this "partitioning" of the affect oracle is that we are "scaling out" (i.e., versus "scaling up") affective computing. In this thesis, while investigating this 'scale out' view of affect, we did not neglect the 'scale up' component, but instead simultaneously pushed the limits of visual affect also along the dimensions of volume and veracity. In fact, we integrated annotations from 2.5M users in Chapter 3 and 15.6K concepts across 7.3M social media images from 12 languages and 237 countries in Chapter 4 and 5. It is our belief that as we continue scaling up and out Visual Affective Computing problems, it will reveal new insights in how we can teach machines to computational model and integrate human affections.

10.1 Summary of Contributions

In Part I of this thesis, we illustrated content-driven affect detection in a novel visual domain for a well-known psychology-grounded affect representation called Ekman emotions. Specfically, in Chapter 3, we delineated between induced vs. intended vs. perceived emotion and made a case for *perceived* emotion because, given that it allows for a cognitive assessment of an affective entity, the traditional vision and multimedia approaches to scaling up datasets like crowdsourcing become valid, e.g., majority voting annotations (which otherwise, would be wrong since affect is subjective). We studied this in the context of animated GIF image sequences which are particularly well-suited for emotion analysis given their use in popular social media and encompassed the first computational multimedia assessment of its kind.

In Part II, we explored the very recent area of Visual Affective Computing, poised to enable even larger scales of visual affect data: mid-level representations. Specifically, in Chapter 4 we looked at a mid-level semantic representation called adjective-noun pairs (ANPs) and proposed another division of the affect oracle along cultural lines. We investigated this multicultural partitioning along language and geographic dimensions, resulting in the largest, public visual affect image dataset to-date called multilingual visual sentiment concept ontology (MVSO) consisting of 7.3M images from 12 languages across 237 countries with over 15.6K ANP visual concepts. In Chapter 5, we developed convolutional neural networks for constructing multilingual detector banks across this massive MVSO corpus and showed their ability to enable impactful applications like culture-diversified image-based query expansions as well as investigated their use in cross-lingual sentiment prediction. And in Chapter 6, we proposed a new learning paradigm called cross-residual learning that extends residual learning to jointly learning from multiple related tasks and showed its particular usefulness in the context of simultaneously predicting affective states (adjectives), visual objects (nouns), and affective mid-level representations (ANPs).

Finally, in Part III, we discussed two ongoing works and future challenges in large-scale Visual Affective Computing. In Chapter 7, we proposed a completely new computing paradigm for affective computing which we call implicit affective computing where signals used in more traditional affective computing like biometrics and other physiological signals actually become the target labels rather than the input data; we presented preliminary

experiments that seek to accomplish this by binarizing these affective physiological signals to create temporally-localized physiological markers which we seek to learn from visual multimedia features. And in Chapter 8, we discussed ongoing work in trying to integrate perceived emotion and mid-level concepts for visual affect detection in movie trailers. Lastly, in Chapter 9, we discussed several areas which we are excited to see progress in the next several years for Visual Affective Computing. These included other affect oracle partitionings, multimodality for affect detection at scale, and improved visualization strategies for understanding affective states.

10.2 Concluding Remarks on Affective Applications

There are a wide range of applications for Visual Affective Computing ahead, some of which are now just realizations of early suggestions in [Picard, 1997]. While we have highlighted some fo these in this thesis already, it is worth briefly noting that visual affective computing poised to have high impact in end-consumer products and business services. Very early on in works like [Ou *et al.*, 2004] the relations between color, emotions and preferential attachment were studied. Today, it is unsurprising that engineering color and many other visual elements in digital advertisements, films, and public relations on social media have are critical to revenue payouts, audience engagement and experience, etc. Motivated by this, some works like [Wang *et al.*, 2013a; Peng *et al.*, 2015] have begun to investigate the problem of "affective image adjustment" where input images can be automatically adjusted by changing attributes like color palettes given a selected output image 'mood'. In addition, there are likely to be benefits yet to be seen with the application of visual affective computing to summarization tasks which assist in *generating* trailers or animated GIFs like those seen in Google's Motion Stills[1] and similar products, where say automatically detected high-arousal content is surfaced, or in visual captioning like in [Mathews *et al.*, 2015] or question-answering like in [Tapaswi *et al.*, 2016].

Naturally, automated affective understanding of the natural visual world will also be critical in robotics and human-computer interactions in general, e.g., [Hoque *et al.*, 2013;

[1] `https://research.googleblog.com/2016/06/motion-stills-create-beautiful-gifs.html`

Chen *et al.*, 2014c]. One challenge with such applications though is that false detections can have far more adverse consequences to their users and additional care will need to be taken to build in administrative and engineering controls. In a different line of application, recently, [Szekely *et al.*, 2015] showed that systems for monitoring and combatting human trafficking can benefit from affective mid-level concept detection. Likewise, visual affective computing is related to a number of other multimedia topics gaining similar traction, including aesthetics [Ke *et al.*, 2006; Bhattacharya *et al.*, 2013], memorability [Isola *et al.*, 2011], interestingness [Gygli *et al.*, 2013], popularity [Khosla *et al.*, 2014; Bakhshi *et al.*, 2014], creativity [Redi *et al.*, 2014], and fashionability [Simo-Serra *et al.*, 2015]. We believe that all these applications also will only benefit as we become better at teaching machines to detect and integrate human affections.

10.3 Concluding Remarks on Ethics in Affective Computing

In closing, it is helpful to also consider the *ethics* of building affective systems like those in this thesis as well as their applications, at least in brief. As seen in recent years, systems that learn from data can go awry and when deployed in real-world environments the consequences may offend[2], introduce injustice and social stereotypes[3], infringe on privacy, etc [Barocas and Selbst, 2016]. Alan M. Turing himself addressed some of these ethics in his classic question on "Can machines think?" [Turing, 1950], and the discussion is still hot as ever today [Pistono and Yampolskiy, 2016]. Even though such model biases are usually not intentionally malicious by design, often resulting from negligence of data sources and distributions [Wagstaff, 2012], it is still important to acknowledge that no technology is ever inherently neutral [Dyer, 2011], thus making Affective Computing both susceptible *and not* somehow unique compared to other technologies with regard to ethics. In Affective Computing, some have observed that even in the most simple representation of affect, sentiment (q.v. §2.1.2), stereotype biases can be unknowingly be introduced in situations where such

[2]E.g., see vision-based examples at `https://www.flickr.com/help/forum/en-us/72157653088504775`, `https://youtu.be/t4DT3tQqgRM` and `http://www.bbc.com/news/technology-33347866`

[3]E.g., see `https://youtu.be/gdCJYsKlX_Y`

sentiment analyses drive, for example, marketing strategy choices that favor males over females, or a specific race of individuals[4]. In the extreme, some may point to the infamous HAL 9000 of the science fiction classic "2001: A Space Odyssey," or any number of other modern day variant, as additional cases against Affective Computing research.

Being no small issue, Picard actually highlighted several areas of ethical caution in several of her seminal writings [Picard, 1997; Picard, 2003], at one point suggesting distinguishing between machine capacities to perceive affect, express affect, induce affect and act based on affect. This delineation she suggests allows for a more thoughtful and principled approach choosing which of these capabilities to engineer into an affective system and to what degree. We believe some of our work on partitioning the affect oracle actually helps to further cement this framework of thinking about ethics in Affective Computing by providing some additional dimensions of granularity. For example, our work in Chapter 4 and 5 around cultural diversification of affect may help identify misrepresentations from certain cultures to take action against model biases (note that we even intentionally tried to mitigate effects like portion biases in our own work by properly normalizing). And yet, we acknowledge such an approaches could also be used inversely by malicious individuals to target such cultures. We believe the specific consideration of affect as a divisible entity also actually allows for a trade-off in ease of scaling but greater susceptibility to ethical abuse at coarser levels of partitioning, down to the finest level where systems are less susceptible to biases like stereotypes because they are highly personalized affects, but scale much more poorly since it is difficult to get high-volume per-person data.

Ultimately, there may be many potential ethical challenges and debates ahead for Affective Computing, but like Turing's "Head in the Sand" objection [Turing, 1950] this does not therefore mean that progress in understanding affect along computational lines is better off avoided altogether. We believe that like its very foundation, the computer, the depth and span of the ethical shadow of Affective Computing will only be an indicator of the potential for good casting that shadow.

[4]E.g., see `https://vimeo.com/163292139`

Part V

Bibliography

Bibliography

[Ahn and Picard, 2014] Hyung-Il Ahn and Rosalind W. Picard. Measuring affective-cognitive experience and predicting market success. *IEEE Transactions on Affective Computing (TAC)*, 5(2), 2014.

[Andrews *et al.*, 2002] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofman. Support vector machines for multiple-instance learning. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2002.

[Argyriou *et al.*, 2008] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. In *Machine Learning*, 2008.

[Baccianella *et al.*, 2010] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Language Resources and Evaluation Conference (LREC)*, 2010.

[Backstrom *et al.*, 2010] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: Improving geographical prediction with social and spatial proximity. In *International World Wide Web Conference (WWW)*, 2010.

[Bakhshi *et al.*, 2014] Saeideh Bakhshi, David A. Shamma, and Eric Gilbert. Faces engage us: Photos with faces attract more likes and comments on Instagram. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2014.

[Balahur and Turchi, 2012] Alexandra Balahur and Marco Turchi. Multilingual sentiment analysis using machine translation? In *Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, 2012.

[Banea *et al.*, 2008] Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. Multilingual subjectivity analysis using machine translation. In *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, 2008.

[Barocas and Selbst, 2016] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104, 2016.

[Barrett and Satpute, 2013] Lisa Feldman Barrett and Ajay B. Satpute. Large-scale brain networks in affective and social neuroscience: Towards an integrative architecture of the human brain. *Current Opinion in Neurobiology*, 23(3), 2013.

[Barrett, 2006] Lisa Feldman Barrett. Are emotions natural kinds? *Perspectives on Psychological Science*, 1(1), 2006.

[Bartlett *et al.*, 2005] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Recognizing facial expression: Machine learning and application to spontaneous behavior. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[Bautin *et al.*, 2008] Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena. International sentiment analysis for news and blogs. In *International Conference on Weblogs and Social Media (ICWSM)*, 2008.

[Baveye *et al.*, 2015a] Yoann Baveye, Emmanuel Dellandréa, Christel Chamaret, and Liming Chen. Deep learning vs. kernel methods: Performance for emotion prediction in videos. In *International Conference of the Association for the Advancement of Affective Computing (ACII)*, 2015.

[Baveye *et al.*, 2015b] Yoann Baveye, Emmanuel Dellandréa, Christel Chamaret, and Liming Chen. LIRIS-ACCEDE: A video database for affective content analysis. *IEEE Transactions on Affective Computing (TAC)*, 6(1), 2015.

[Benevenuto *et al.*, 2009] Fabrício Benevenuto, Tiago Rodrigues, Virgílio Almeida, Jussara Almeida, and Marcos Gonçalves. Detecting spammers and content promoters in online

video social networks. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2009.

[Bhattacharya *et al.*, 2013] Subhabrata Bhattacharya, Behnaz Nojavanasghari, Tao Chen, Dong Liu, Shih-Fu Chang, and Mubarak Shah. Towards a comprehensive computational model for aesthetic assessment of videos. In *ACM International Conference on Multimedia (MM)*, Grand Challenge, 2013.

[Borth *et al.*, 2013a] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. SentiBank: Large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *ACM International Conference on Multimedia (MM)*, Technical Demonstration, 2013.

[Borth *et al.*, 2013b] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM International Conference on Multimedia (MM)*, Brave New Ideas, 2013.

[Boyd-Graber and Resnik, 2010] Jordan Boyd-Graber and Philip Resnik. Holistic sentiment analysis across languages: Multilingual supervised latent dirichlet allocation. In *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, 2010.

[Bradley *et al.*, 1992] Margaret M. Bradley, Mark K. Greenwald, Margaret C. Petry, and Peter J. Lang. Remembering pictures: Pleasure and arousal in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(2), 1992.

[Bretzner *et al.*, 2002] Lars Bretzner, Ivan Laptev, and Tony Lindeberg. Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2002.

[Brooke *et al.*, 2009] Julian Brooke, Milan Tofiloski, and Maite Taboada. Cross-linguistic sentiment analysis: From english to spanish. In *International Conference on Recent Advances in Natural Language Processing (RANLP)*, 2009.

[Calvo and D'Mello, 2010] Rafael A. Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods and their applications. *IEEE Transactions on Affective Computing (TAC)*, 1(1), 2010.

[Campos *et al.*, 2015] Víctor Campos, Amaia Salvador, Xavier Giró-i-Nieto, and Brendan Jou. Diving deep into sentiment: Understanding fine-tuned CNNs for visual sentiment prediction. In *ACM Multimedia Workshop on Affect and Sentiment in Multimedia (ASM)*, 2015.

[Campos *et al.*, 2016] Víctor Campos, Brendan Jou, and Xavier Giró-i-Nieto. From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction. *arXiv preprint arXiv:1604.03489*, 2016. In review.

[Canini *et al.*, 2013] Luca Canini, Sergio Benini, and Riccardo Leonardi. Affective recommendation of movies based on selected connotative features. *IEEE Transactions on Circuits and Systems for Video Technology (CSVT)*, 23(4), 2013.

[Caruana, 1997] Rich Caruana. Multitask learning. *Machine Learning*, 28(1), 1997.

[Castellano *et al.*, 2007] Ginevra Castellano, Santiago D. Villalba, and Antonio Camurri. Recognising human emotions from body movement and gesture dynamics. In *International Conference of the Association for the Advancement of Affective Computing (ACII)*, 2007.

[Cha *et al.*, 2007] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I Tube, You Tube, Everybody Tubes: Analyzing the world's largest user generated content video system. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2007.

[Chanel *et al.*, 2005] Guillaume Chanel, Julien Kronegg, Didier Grandjean, and Thierry Pun. Emotional assessment: Arousal evaluation using EEG's and peripheral physiological signals. Technical report, University of Geneva, 2005.

[Chang *et al.*, 2006] Ya Chang, Changbo Hu, Rogerio Feris, and Matthew Turk. Manifold-based analysis of facial expression. *Journal of Image and Vision Computing (IMAVIS)*, 24(6), 2006.

[Chen *et al.*, 2011] Jianhui Chen, Jiayu Zhou, and Jieping Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2011.

[Chen *et al.*, 2014a] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014.

[Chen *et al.*, 2014b] Tao Chen, Felix X. Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and Shih-Fu Chang. Object-based visual sentiment concept analysis and application. In *ACM International Conference on Multimedia (MM)*, 2014.

[Chen *et al.*, 2014c] Yan-Ying Chen, Tao Chen, Winston H. Hsu, Hong-Yuan M. Liao, and Shih-Fu Chang. Predicting viewer affective comments based on image content in social media. In *ACM International Conference on Multimedia Retrieval (ICMR)*, 2014.

[Chen *et al.*, 2015] Yan-Ying Chen, Tao Chen, Taikun Liu, Hong-Yuan Mark Liao, and Shih-Fu Chang. Assistive image comment robot - A novel mid-level concept-based representation. *IEEE Transactions on Affective Computing*, 6(3), 2015.

[Chikazoe *et al.*, 2014] Junichi Chikazoe, Daniel H. Lee, Nikolaus Kriegeskorte, and Adam K. Anderson. Population coding of affect across stimuli, modalities and individuals. *Nature Neuroscience*, 17, 2014.

[Collobert and Weston, 2008] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *International Conference of Machine Learning (ICML)*, 2008.

[Cortes and Vapnik, 1995] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3), 1995.

[Cunningham *et al.*, 2004] William A. Cunningham, Carol L. Raye, and Marcia K. Johnson. Implicit and explicit evaluation: fMRI correlates of valence, emotional intensity, and control in the processing of attitudes. *Journal of Cognitive Neuroscience*, 16(10), 2004.

[Dahl *et al.*, 2014] George E. Dahl, Navdeep Jaitly, and Ruslan Salakhutdinov. Multi-task neural networks for QSAR predictions. *arXiv preprint arXiv:1406.1231*, 2014.

[Dai *et al.*, 2016] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[Dan-Glauser and Scherer, 2011] Elise S. Dan-Glauser and Klaus Scherer. The Geneva affective picture database: A new 730-picture database focusing on valence and normative significance. *Behavior Research Methods*, 43(2), 2011.

[Darwin, 1872] Charles Darwin. *The Expression of Emotions in Man and Animals*. John Murray, 1872.

[Datta *et al.*, 2006a] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision (ECCV)*, 2006.

[Datta *et al.*, 2006b] Ritendra Datta, Jia Li, and James Z. Wang. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *IEEE International Conference on Image Processing (ICIP)*, 2006.

[Davis and Lang, 2003] Michael Davis and Peter J. Lang. Emotion. In M. Gallagher, R. J. Nelson, and I. B. Weiner, editors, *Comprehensive Handbook of Psychology: Vol. 3. Biological Psychology*. John Wiley & Sons, Inc, 2003.

[Deng *et al.*, 2014] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *European Conference on Computer Vision (ECCV)*, 2014.

[Dietz and Lang, 1999] Richard B. Dietz and Annie Lang. Aefective agents: Effects of agent affect on arousal, attention, liking and learning. In *Cognitive Technology Conference*, 1999.

[Divvala *et al.*, 2014] Santosh Divvala, Ali Farhadi, and Carlos Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[Dodds *et al.*, 2015] Peter Sheridan Dodds, Eric M. Clark, Suma Desu, Morgan R. Frank, Andrew J. Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M. Kloumann, James P. Bagrow, Karine Megerdoomian, Matthew T. McMahon, Brian F. Tivnan, and Christopher M. Danforth. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences (PNAS)*, 112(8), 2015.

[Donahue *et al.*, 2014] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *International Conference of Machine Learning (ICML)*, 2014.

[Dryer and Haspelmath, 2013] Matthew S. Dryer and Martin Haspelmath, editors. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, 2013. `http://wals.info/chapter/87`.

[Dyer, 2011] John Dyer. *From the Garden to the City: The Redeeming and Corrupting Power of Technology*. Kregel, Grand Rapids, 2011. 198.

[Ekman *et al.*, 1980] Paul Ekman, Wallace V. Friesen, and Sonia Ancoli. Facial signs of emotional experience. *Journal of Personality and Social Psychology*, 39(6), 1980.

[Ekman *et al.*, 1987] Paul Ekman, Wallace V. Friesen, Maureen O'Sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E. Ricci-Bitti, Klaus Scherer, Masatoshi Tomita, and Athanase Tzavaras. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 4(53), 1987.

[Ekman, 1992] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3/4), 1992.

[Ekman, 1999] Paul Ekman. Basic emotions. In *Handbook of Cognition and Emotion*. John Wiley & Sons Ltd., 1999. 54.

[Ellis *et al.*, 2014a] Joseph G. Ellis, Brendan Jou, and Shih-Fu Chang. Why we watch the news: A dataset for exploring sentiment in broadcast video news. In *ACM International Conference on Multimodal Interaction (ICMI)*, 2014.

[Ellis *et al.*, 2014b] Joseph G. Ellis, W. Sabrina Lin, Ching-Yung Lin, and Shih-Fu Chang. Predicting evoked emotions in video. In *IEEE International Symposium on Multimedia (ISM)*, 2014.

[Essa and Pentland, 1997] Irfan A. Essa and Alex P. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19(7), 1997.

[Eyben *et al.*, 2013] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *ACM International Conference on Multimedia (MM)*, 2013.

[Fan *et al.*, 2008] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research (JMLR)*, 9, 2008.

[Fasel and Luettin, 2003] B. Fasel and Juergen Luettin. Automatic facial expression analysis: A survey. *Pattern Recognition*, 36(1), 2003.

[Finkel *et al.*, 2005] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2005.

[Fleureau *et al.*, 2012] Julien Fleureau, Philippe Guillotel, and Quan Huynh-Thu. Physiological-based affect event detector for entertainment video applications. *IEEE Transactions on Affective Computing (TAC)*, 3(3), 2012.

[Fontaine *et al.*, 2007] Johnny R. J. Fontaine, Klaus R. Scherer, Etienne B. Roesch, and Phoebe C. Ellsworth. The world of emotions is not two-dimensional. *Psychological Science*, 18(12), 2007.

[Freeman and Roth, 1995] William T. Freeman and Michal Roth. Orientation histograms for hand gesture recognition. In *IEEE International Workshop on Automatic Face and Gesture Recognition*, 1995.

[Gabrielsson, 2002] Alf Gabrielsson. Emotion perceived and emotion felt: Same or different? *Musicae Scientiae*, 5(1), 2002.

[Gazzaniga, 2009] Michael S. Gazzaniga, editor. *The Cognitive Neurosciences*. MIT Press, 2009.

[Gelli *et al.*, 2015] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang. Image popularity prediction in social media using sentiment and context features. In *ACM International Conference on Multimedia (MM)*, 2015.

[Gers *et al.*, 2002] Felix A. Gers, Nicol N. Schraudolph, and J"urgen Schmidhuber. Learning precise timing with LSTM recurrent networks. *Journal of Machine Learning Research (JMLR)*, 3, 2002.

[Ghifary *et al.*, 2015] Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *International Conference on Computer Vision (ICCV)*, 2015.

[Glorot *et al.*, 2011] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *International Conference of Machine Learning (ICML)*, 2011.

[Goffman, 1967] Erving Goffman. *Interaction Ritual: Essays on Face-to-Face Behavior*. Anchor/Doubleday, 1967.

[Gonzalez *et al.*, 2010] Hector Gonzalez, Alon Halevy, Christian S. Jensen, Anno Langen, Jayant Madhavan, Rebecca Shapley, and Warren Shen. Google Fusion Tables: Data

management, integration, and collaboration in the cloud. In *ACM Symposium on Cloud Computing (SoCC)*, 2010.

[Gross *et al.*, 2010] Ralph Gross, Iain Matthews, Jeffrey F. Cohn, Takeo Kanade, and Simon Baker. Multi-PIE. *Journal of Image and Vision Computing (IMAVIS)*, 28(5), 2010.

[Gunes and Pantic, 2010] Hatice Gunes and Maja Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, 1(1), 2010.

[Güngördü and Oflazer, 1994] Zelal Güngördü and Kemal Oflazer. Parsing Turkish using the lexical functional grammar formalism. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 1994.

[Gygli *et al.*, 2013] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. The interestingness of images. In *International Conference on Computer Vision (ICCV)*, 2013.

[Haag *et al.*, 2004] Andreas Haag, Silke Goronzy, Peter Schaich, and Jason Williams. Emotion recognition using bio-sensors: First steps towards an automatic system. *Affective Dialogue Systems*, 3068, 2004.

[Halácsy *et al.*, 2007] Péter Halácsy, András Kornai, and Csaba Oravecz. HunPos: An open source trigram tagger. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007.

[Hanjalic and Xu, 2005] Alan Hanjalic and Li-Qun Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia (TMM)*, 7(1), 2005.

[Hao *et al.*, 2011] Ming Hao, Christian Rohrdantz, Halldór Janetzko, Umeshwar Dayal, Daniel A. Keim, Lars-Erik Haug, and Mei-Chun Hsu. Visual sentiment analysis on Twitter data streams. In *IEEE Visual Analytics Science and Technology (VAST)*, 2011.

[Haselton and Ketelaar, 2006] Martie G. Haselton and Timothy Ketelaar. Irrational emotions or emotional wisdom? The evolutionary psychology of affect and social behavior. *Affect in Social Thinking and Behavior*, 8(21), 2006.

[He *et al.*, 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imageNet classification. In *International Conference on Computer Vision (ICCV)*, 2015.

[He *et al.*, 2016a] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[He *et al.*, 2016b] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *arXiv preprint arXiv:1603.05027*, 2016.

[Herbrich *et al.*, 2006] Ralf Herbrich, Tom Minka, and Thore Graepel. TrueSkill$^{\text{TM}}$: A Bayesian skill rating system. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2006.

[Hipps, 2009] Shane Hipps. *Flickering Pixels: How Technology Shapes Your Faith*. Zondervan, Grand Rapids, 2009. 76.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8), 1997.

[Hochschild, 1983] Arlie Russell Hochschild. *The Managed Heart: Commercialization of Human Feeling*. University of California Press, 1983.

[Hoque *et al.*, 2013] Mohammed Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W. Picard. MACH: My automated conversation coach. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2013.

[Hu and Yang, 2014] Xiao Hu and Yi-Hsuan Yang. Cross-cultural mood regression for music digital libraries. In *Joint Conference on Digital Libraries (JCDL)*, 2014.

[Huang *et al.*, 2015] Junshi Huang, Rogerio Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *International Conference on Computer Vision (ICCV)*, 2015.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference of Machine Learning (ICML)*, 2015.

[Isola *et al.*, 2011] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[Jack *et al.*, 2012] Rachael E. Jack, Oliver G. B. Garrod, Hui Yu, Roberto Caldar, and Philippe G. Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences (PNAS)*, 109(19), 2012.

[James, 1883] William James. What is an emotion? *Mind Association*, 9, 1883.

[Ji and Ye, 2009] Shuiwang Ji and Jieping Ye. An accelerated gradient method for trace norm minimization. In *International Conference of Machine Learning (ICML)*, 2009.

[Ji *et al.*, 2015] Rongrong Ji, Yue Gao, Wei Liu, Xing Xie, Qi Tian, and Xuelong Li. When location meets social multimedia: A survey on vision-based recognition and mining for geo-social multimedia analytics. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(1), 2015.

[Jia *et al.*, 2012] Jia Jia, Sen Wu, Xiaohui Wang, Peiyun Hu, Lianhong Cai, and Jie Tang. Can we understand van Gogh's mood?: Learning to infer affects from images in social networks. In *ACM International Conference on Multimedia (MM)*, 2012.

[Jia *et al.*, 2014] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia (MM)*, 2014.

[Jiang *et al.*, 2008] Wei Jiang, Eric Zavesky, Shih-Fu Chang, and Alex Loui. Cross-domain learning methods for high-level visual concept classification. In *IEEE International Conference on Image Processing (ICIP)*, 2008.

[Jiang *et al.*, 2009] Yu-Gang Jiang, Jun Wang, Shih-Fu Chang, and Chong-Wah Ngo. Domain adaptive semantic diffusion for large scale context-based video annotation. In *International Conference on Computer Vision (ICCV)*, 2009.

[Jiang *et al.*, 2014] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. Predicting emotions in user-generated videos. In *Conference of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2014.

[Jin *et al.*, 2010] Xin Jin, Andrew Gallagher, Liangliang Cao, Jiebo Luo, and Jiawei Han. The wisdom of social multimedia: Using Flickr for prediction and forecast. In *ACM International Conference on Multimedia (MM)*, 2010.

[Joachims, 2003] Thorsten Joachims. Optimizing search engines using clickthrough data. In *ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2003.

[Jou and Chang, 2016a] Brendan Jou and Shih-Fu Chang. Deep cross residual learning for multitask visual recognition. In *ACM International Conference on Multimedia (MM)*, 2016.

[Jou and Chang, 2016b] Brendan Jou and Shih-Fu Chang. Going deeper for multilingual visual sentiment detection. *arXiv preprint arXiv:1605.09211*, 2016. Technical Report, Department of Electrical Engineering, Columbia University.

[Jou *et al.*, 2013] Brendan Jou, Hongzhi Li, Joseph G. Ellis, Daniel Morozoff-Abegauz, and Shih-Fu Chang. Structured exploration of who, what, when, and where in heterogeneous multimedia news sources. In *ACM International Conference on Multimedia (MM)*, Grand Challenge, 2013.

[Jou *et al.*, 2014] Brendan Jou, Subhabrata Bhattacharya, and Shih-Fu Chang. Predicting viewer perceived emotions in animated GIFs. In *ACM International Conference on Multimedia (MM)*, Grand Challenge, 2014.

[Jou *et al.*, 2015] Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *ACM International Conference on Multimedia (MM)*, 2015.

[Jou *et al.*, 2016] Brendan Jou, Margaret Yuying Qian, and Shih-Fu Chang. SentiCart: Cartography and geo-contextualization for multilingual visual sentiment. In *ACM International Conference on Multimedia Retrieval (ICMR)*, Technical Demonstration, 2016.

[Kaiser and Roessler, 1970] Charles Kaiser and Robert Roessler. Galvanic skin responses to motion pictures. *Perceptual and Motor Skills*, 30(4), 1970.

[Kallinen and Ravaja, 2006] Kari Kallinen and Niklas Ravaja. Emotion perceived and emotion felt: Same and different. *Musicae Scientiae*, 10(2), 2006.

[Kanade *et al.*, 2000] Takeo Kanade, Jeffrey F. Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2000.

[Katz, 1980] Judith Milstein Katz. Discrepancy, arousal and labeling: Towards a psychosocial theory of emotion. *Sociological Inquiry*, 50(2), 1980.

[Ke *et al.*, 2006] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[Khosla *et al.*, 2014] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. What makes an image popular? In *International World Wide Web Conference (WWW)*, 2014.

[Kiesler, 1983] Donald J. Kiesler. The 1982 interpersonal circle: A taxonomy for complementarity in human transactions. *Psychological Review*, 90(3), 1983.

[Kim *et al.*, 2013] Yelin Kim, Honglak Lee, and Emily Mower Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[Kipp and Martin, 2009] M. Kipp and J. C. Martin. Gesture and emotion: Can basic gestural form features discriminate emotions? In *International Conference of the Association for the Advancement of Affective Computing (ACII)*, 2009.

[Koelstra *et al.*, 2011] Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing (TAC)*, 3(1), 2011.

[Kovashka and Grauman, 2015] Adriana Kovashka and Kristen Grauman. Discovering attribute shades of meaning with the crowd. *International Journal of Computer Vision (IJCV)*, 114(1), 2015.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2012.

[Krizhevsky, 2009] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Univerity of Toronto, 2009.

[Kunst-Wilson and Zajonc, 1980] William Raft Kunst-Wilson and Robert B. Zajonc. Affective discrimination of stimuli that cannot be recognized. *Science*, 207(4430), 1980.

[Kuo *et al.*, 2014] Tzu-Ming Kuo, Ching-Pei Lee, and Chih-Jen Lin. Large-scale kernel RankSVM. In *SIAM International Conference on Data Mining (SDM)*, 2014.

[Lang *et al.*, 1997] Peter J. Lang, Margaret M. Bradley, and Bruce N. Cuthbert. International Affective Picture System (IAPS): Technical manual and affective ratings. Technical report, National Institute of Mental Health (NIMH) Center for the Study of Emotion and Attention (CSEA), 1997.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 1998.

[Lee *et al.*, 2005] Jin Ha Lee, J. Stephen Downie, and Sally Jo Cunningham. Challenges in cross-cultural/multilingual music information seeking. In *Conference of the International Society of Music Information Retrieval (ISMIR)*, 2005.

[Li *et al.*, 2010] Li-Jia Li, Hao Su, Eric P. Xing, and Li Fei-Fei. Object Bank: A high-level image representation for scene classification & semantic feature sparsification. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2010.

[Li *et al.*, 2013] Hongzhi Li, Brendan Jou, Joseph G. Ellis, Dan Morozoff, and Shih-Fu Chang. News Rover: Exploring topical structures and serendipity in heterogeneous multimedia news. In *ACM International Conference on Multimedia (MM)*, Technical Demonstration, 2013.

[Li *et al.*, 2016a] Hongzhi Li, Joseph G. Ellis, and Shih-Fu Chang. Event specific multimodal pattern mining with image-caption pairs. *arXiv preprint arXiv:1601.00022*, 2016.

[Li *et al.*, 2016b] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. TGIF: A new dataset and benchmark on animated GIF description. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[Liao and Poggio, 2016] Qianli Liao and Tomaso Poggio. Bridging the gaps between residual learning, recurrent neural networks and visual cortex. *arXiv preprint arXiv:1604.03640*, 2016.

[Lieberman, 2007] Matthew D. Lieberman. Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology*, 58, 2007.

[Lienhart and Maydt, 2002] Rainer Lienhart and Jochen Maydt. An extended set of haar-like features for rapid object detection. In *IEEE International Conference on Image Processing (ICIP)*, 2002.

[Lin *et al.*, 2014] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2014.

[Lisetti and Nasoz, 2004] Christine L. Lisetti and Fatma Nasoz. Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP Journal on Applied Signal Processing*, 11, 2004.

[Liu *et al.*, 2015] Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2015.

[Liu *et al.*, 2016] Hongyi Liu, Brendan Jou, Tao Chen, Mercan Topkara, Nikolaos Pappas, Miriam Redi, and Shih-Fu Chang. Complura: Exploring and leveraging a large-scale multilingual visual sentiment ontology. In *ACM International Conference on Multimedia Retrieval (ICMR)*, Technical Demonstration, 2016.

[Liu, 2012] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1), 2012.

[Lomas, 2016] Tim Lomas. Towards a positive cross-cultural lexicography: Enriching our emotional landscape through 216 'untranslatable' words pertaining to well-being. *Journal of Positive Psychology*, 2016.

[Lowe, 2004] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2), 2004.

[Lucey *et al.*, 2010] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010.

[Luo *et al.*, 2011] Jiebo Luo, Dhiraj Joshi, Jie Yu, and Andrew Gallagher. Geotagging in multimedia and computer vision - A survey. *Multimedia Tools and Applications*, 51(1), 2011.

[Luong *et al.*, 2016] Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task sequence to sequence learning. In *International Conference on Learning Representations (ICLR)*, 2016.

[Lyons *et al.*, 1998] Michael J. Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 1998.

[Machajdik and Hanbury, 2010] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM International Conference on Multimedia (MM)*, 2010.

[Marsella and Gratch, 2014] Stacy Marsella and Jonathan Gratch. Computationally modeling human emotion. *Communications of the ACM*, 57(12), 2014.

[Martínez *et al.*, 2014] Héctor P. Martínez, Georgios N. Yannakakis, and John Hallam. Don't classify ratings of affect; Rank them! *IEEE Transactions on Affective Computing (TAC)*, 5(3), 2014.

[Mathews *et al.*, 2015] Alexander Mathews, Lexing Xie, and Xuming He. SentiCap: Generating image descriptions with sentiments. In *Conference of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2015.

[McCarthy, 1994] E. Doyle McCarthy. The social construction of emotions: New directions from culture theory. *Social Perspectives on Emotion*, 2, 1994.

[McDuff *et al.*, 2013] Daniel McDuff, Rana el Kaliouby, Thibaud Senechal, May Amr, Jeffrey F. Cohn, and Rosalind W. Picard. Affectiva-MIT facial expression dataset (AM-FED): Naturalistic and spontaneous facial expressions collected in-the-wild. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013.

[Mediratta *et al.*, 2013] Namita Mediratta, Rana el Kaliouby, Evan Kodra, and Pankaj Jha. Does facial coding generalize across cultures? In *European Society for Opinion and Marketing Research (ESOMAR) - Asia Pacific*, 2013.

[Mehrabian, 1980] Albert Mehrabian, editor. *Basic Dimensions for a General Psychological Theory.* Oelgeschlager, Gunn & Hain, 1980.

[Mihalcea *et al.*, 2007] Rada Mihalcea, Carmen Banea, and Janyce Wiebe. Learning multilingual subjective language via cross-lingual projections. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007.

[Mikels *et al.*, 2005] Joseph A. Mikels, Barbara L. Fredrickson, Gregory R. Larkin, Casey M. Lindbery, Sam J. Maglio, and Patricia A. Reuter-Lorenz. Emotional category data on images from the International Affective Picture System. *Behavior Research Methods*, 37(4), 2005.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2013.

[Mitra and Acharya, 2007] S. Mitra and T. Acharya. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics (SMC)*, 37(3), 2007.

[Moore *et al.*, 2014] Joshua L. Moore, Thorsten Joachims, and Douglas Turnbull. Taste space versus the world: An embedding analysis of listening habits and geography. In *International Society for Music Information Retrieval (ISMIR)*, 2014.

[Morency *et al.*, 2011] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *ACM International Conference on Multimodal Interaction (ICMI)*, 2011.

[Nair and Hinton, 2010] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference of Machine Learning (ICML)*, 2010.

[Narihira *et al.*, 2015] Takuya Narihira, Damian Borth, Stella X. Yu, Karl Ni, and Trevor Darrell. Mapping images to sentiment adjective noun pairs with factorized neural nets. *arXiv preprint arXiv:1511.06838*, 2015.

[Öhman, 2002] Arne Öhman. Automaticity and the amygdala: Nonconscious responses to emotional faces. *Current Directions in Psychological Science*, 11(2), 2002.

[Oliva and Torralba, 2001] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal on Computer Vision (IJCV)*, 42(3), 2001.

[Ou *et al.*, 2004] Li-Chen Ou, M. Ronnier Luo, Andrée Woodcock, and Angela Wright. A study of colour emotion and colour preference. Part I-III. *Color Research & Application*, 29(3-5), 2004.

[Oviatt and Cohen, 2000] Sharon Oviatt and Philip Cohen. Perceptual user interfaces: Multimodal interfaces that process what comes naturally. *Communications of the ACM (CACM)*, 43(3), 2000.

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(10), 2010.

[Pang and Ngo, 2015] Lei Pang and Chong-Wah Ngo. Multimodal learning with deep boltzmann machine for emotion prediction in user generated videos. In *ACM International Conference on Multimedia Retrieval (ICMR)*, 2015.

[Pappas *et al.*, 2016] Nikolaos Pappas, Miriam Redi, Mercan Topkara, Brendan Jou, Hongyi Liu, Tao Chen, and Shih-Fu Chang. Multilingual visual sentiment concept matching. In *ACM International Conference on Multimedia Retrieval (ICMR)*, 2016.

[Parikh and Grauman, 2011] Devi Parikh and Kristen Grauman. Relative attributes. In *International Conference on Computer Vision (ICCV)*, 2011.

[Patterson and Hays, 2012] Genevieve Patterson and James Hays. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[Peng *et al.*, 2015] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[Perez-Rosas *et al.*, 2013] Veronica Perez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. Utterance-level multimodal sentiment analysis. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2013.

[Picard, 1997] Rosalind W. Picard. *Affective Computing*. MIT Press, 1997.

[Picard, 2003] Rosalind W. Picard. Affective Computing: Challenges, 2003.

[Pistono and Yampolskiy, 2016] Federico Pistono and Roman V. Yampolskiy. Unethical research: How to create a malevolent artificial intelligence. *arXiv preprint arXiv:1605.02817*, 2016.

[Plutchik, 1980] Robert Plutchik. *Emotion: A Psychoevolutionary Synthesis*. Harper & Row, 1980.

[Raiko *et al.*, 2012] Tapani Raiko, Harri Valpola, and Yann LeCun. Deep learning made easier by linear transformations in perceptrons. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[Ramsundar *et al.*, 2015] Bharath Ramsundar, Steven Kearnes, Patrick Riley, Dale Webster, David Konerding, and Vijay Pande. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*, 2015.

[Rasmus *et al.*, 2015] Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. Semi-supervised learning with ladder networks. *arXiv preprint arXiv:1507.02672*, 2015.

[Razavian *et al.*, 2014] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014.

[Redi *et al.*, 2014] Miriam Redi, Neil O'Hare, Rossano Schifanella, Michele Trevisiol, and Alejandro Jaimes. 6 Seconds of sound and vision: Creativity in micro-videos. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[Rudd *et al.*, 2016] Ethan Rudd, Manuel Günther, and Terrance Boult. MOON: A mixed objective optimization network for the recognition of facial attributes. *arXiv preprint arXiv:1603.07027*, 2016.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3), 2015.

[Russell, 1990] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1990.

[Russell, 1991] James A. Russell. Culture and the categorization of emotions. *Psychological Bulletin*, 110(3), 1991.

[Samal and Iyengar, 1992] Ashok Samal and Prasana A. Iyengar. Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern Recognition*, 25(1), 1992.

[Schaefer *et al.*, 2010] Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion*, 24(7), 2010.

[Schein *et al.*, 2002] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and metrics for cold-start recommendations. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002.

[Schmid, 1994] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, 1994.

[Schuller *et al.*, 2011] Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. AVEC 2011 – The First International Audio/Visual Emotion Challenge. In *International Conference of the Association for the Advancement of Affective Computing (ACII)*, 2011.

[Shan *et al.*, 2009] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image and Vision Computing*, 27(6), 2009.

[Shin and Kim, 2010] Yunhee Shin and Eun Yi Kim. Affective prediction in photographic images using probabilistic affective model. In *ACM International Conference on Image and Video Retrieval (CIVR)*, 2010.

[Sidiropoulos *et al.*, 2011] Panagiotis Sidiropoulos, Vasileios Mezaris, Ioannis Kompatsiaris, Hugo Meinedo, Miguel Bugalho, and Isabel Trancoso. Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Transactions on Circuits and Systems for Video Technology (CSVT)*, 21(8), 2011.

[Siersdorfer *et al.*, 2010] Stefan Siersdorfer, Enrico Minack, Fan Deng, and Jonathon Hare. Analyzing and predicting sentiment of images on the social web. In *ACM International Conference on Multimedia (MM)*, 2010.

[Silveira *et al.*, 2013] Fernando Silveira, Brian Eriksson, Anmol Sheth, and Adam Sheppard. Predicting audience responses to movie content from electro-dermal activity signals. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2013.

[Sim *et al.*, 2002] Terence Sim, Simon Baker, and Maan Bsat. The CMU pose, illumination, and expression (PIE) database. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2002.

[Simo-Serra *et al.*, 2015] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[Simon, 1967] Hebert A. Simon. Motivational and emotional controls of cognition. *Psychological Review*, 74(1), 1967.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[Singh *et al.*, 2010] Vivek K. Singh, Mingyan Gao, and Ramesh Jain. Social pixels: Genesis and evaluation. In *ACM International Conference on Multimedia (MM)*, 2010.

[Soleymani *et al.*, 2012] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing (TAC)*, 3(1), 2012.

[Soleymani *et al.*, 2016] Mohammad Soleymani, Björn Schuller, David Garcia, Brendan Jou, Shih-Fu Chang, and Maja Pantic. A survey of multimodal sentiment analysis. *Journal of Image and Vision Computing (IMAVIS)*, 2016. In preparation.

[Solli and Lenz, 2008] Martin Solli and Reiner Lenz. Color emotions for image classification and retrieval. In *European Conference on Colour in Graphics, Imaging, and Vision (CGIV)*, 2008.

[Solli and Lenz, 2010] Martin Solli and Reiner Lenz. Emotion related structures in large image databases. In *ACM International Conference on Image and Video Retrieval (CIVR)*, 2010.

[Srivastava and Salakhutdinov, 2012] Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2012.

[Srivastava *et al.*, 2014] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR)*, 15(1), 2014.

[Srivastava *et al.*, 2015] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. In *International Conference of Machine Learning (ICML) Workshop on Deep Learning*, 2015.

[Sudowe *et al.*, 2015] Patrick Sudowe, Hannah Spitzer, and Bastian Leibe. Person attribute recognition with a jointly-trained holistic CNN model. In *International Conference on Computer Vision (ICCV)*, 2015.

[Szegedy *et al.*, 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[Szekely *et al.*, 2015] Pedro Szekely, Craig A. Knoblock, Jason Slepicka, Andrew Philpot, Amandeep Singh, Chengye Yin, Dipsy Kapoor, Prem Natarajan, Daniel Marcu, Kevin Knight, David Stallard, Subessware S. Karunamoorthy, Rajagopal Bojanapalli, Steven Minton, Brian Amanatullah, Todd Hughes, Mike Tamayo, David Flynt, Rachel Artiss, Shih-Fu Chang, Tao Chen, Gerald Hiebel, and Lidia Ferreira. Building and using a knowledge graph to combat human trafficking. In *International Semantic Web Conference (ISWC)*, 2015.

[Tang, 2013] Yichuan Tang. Deep learning using linear support vector machines. In *International Conference of Machine Learning (ICML) Workshop on Challenges in Representation Learning*, 2013.

[Tao and Tan, 2005] Jianhua Tao and Tieniu Tan. Affective Computing: A review. In *International Conference of the Association for the Advancement of Affective Computing (ACII)*, 2005.

[Tapaswi *et al.*, 2016] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding stories in movies through question-answering. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[Teixeira *et al.*, 2011] René Marcelino Abritta Teixeira, Toshihiko Yamasaki, and Kiyoharu Aizawa. Determination of emotional content of video clips by low-level audiovisual features. *Multimedia Tools and Applications*, 61(1), 2011.

[Thelwall *et al.*, 2010] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, and Di Cai. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12), 2010.

[Tian *et al.*, 2001] Ying-li Tian, Takeo Kanade, and Jeffrey F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(2), 2001.

[Torresani *et al.*, 2010] Lorenzo Torresani, Martin Szummer, and Andrew Fitzgibbon. Efficient object category recognition using classemes. In *European Conference on Computer Vision (ECCV)*, 2010.

[Toutanova *et al.*, 2003] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Annual Meeting of the North American Association for Computational Linguistics (NAACL)*, 2003.

[Toyama *et al.*, 2003] Kentaro Toyama, Ron Logan, Asta Roseway, and P. Anandan. Geographic location tags on digital images. In *ACM International Conference on Multimedia (MM)*, 2003.

[Turing, 1950] Alan M. Turing. Computing machinery and intelligence. *Mind*, 49, 1950.

[Turney, 2002] Peter D. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.

[van der Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research (JMLR)*, 9, 2008.

[Vandal *et al.*, 2015] Thomas Vandal, Daniel McDuff, and Rana El Kaliouby. Event detection: Ultra large-scale clustering of facial expressions. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2015.

[Vandenberghe and Boyd, 1996] Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *Society for Industrial and Applied Mathematics (SIAM) Review*, 38(1), 1996.

[Vessel *et al.*, 2014] Edward A. Vessel, Jonathan Stahl, Natalia Maurer, Alexander Denker, and G. Gabrielle Starr. Personalized visual aesthetics. In *Proceedings of the SPIE 9014*, Human Vision and Electronic Imaging XIX, 90140S, 2014.

[Viola and Jones, 2001] Paul Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[Wagstaff, 2012] Kiri L. Wagstaff. Machine learning that matters. In *International Conference of Machine Learning (ICML)*, 2012.

[Wan, 2009] Xiaojun Wan. Co-training for cross-lingual sentiment classification. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2009.

[Wang and Cheong, 2006] Hee Lin Wang and Loong-Fah Cheong. Affective understanding in film. *IEEE Transactions on Circuits and Systems for Video Technology (CSVT)*, 16(6), 2006.

[Wang *et al.*, 2006] Wei-Ning Wang, Ying-Lin Yu, and Sheng-Ming Jiang. Image retrieval by emotional semantics: A study of emotional space and feature extraction. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2006.

[Wang *et al.*, 2013a] Xiaohui Wang, Jia Jia, and Lianhong Cai. Affective image adjustment with a single word. *The Visual Computer: International Journal of Computer Graphics*, 29(11), 2013.

[Wang *et al.*, 2013b] Xiaohui Wang, Jia Jia, Jiaming Yin, and Lianhong Cai. Interpretable aesthetic features for affective image classification. In *IEEE International Conference on Image Processing (ICIP)*, 2013.

[Wang *et al.*, 2015a] Xiaohui Wang, Jia Jia, Lianhong Cai, and Jie Tang. Modeling emotion influence in image social networks. *IEEE Transactions on Affective Computing (TAC)*, PP(99), 2015.

[Wang *et al.*, 2015b] Yilin Wang, Yuheng Hu, Subbarao Kambhampati, and Baoxin Li. Inferring sentiment from web images with joint inference on visual and social cues: A regulated matrix factorization approach. In *Conference of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2015.

[Wang *et al.*, 2015c] Yilin Wang, Suhang Wang, Jiliang Tang, Huan Liu, and Baoxin Li. Unsupervised sentiment analysis for social media images. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

[Wang *et al.*, 2016] Xi Wang, Zhenfeng Sun, Wenqiang Zhang, and Yu-Gang Jiang. Matching user photos to online products with robust deep features. In *ACM International Conference on Multimedia Retrieval (ICMR)*, 2016.

[Watson *et al.*, 1988] David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6), 1988.

[Wu and Huang, 1999] Ying Wu and Thomas S. Huang. Vision-based gesture recognition: A review. In Annelies Braffort, Rachid Gherbi, Sylvie Gibet, Daniel Teil, and James Richardson, editors, *Gesture-Based Communication in Human-Computer Interaction*, volume 1739. Springer Berlin Heidelberg, 1999.

[Wu *et al.*, 2014] Zuxuan Wu, Yu-Gang Jiang, Jun Wang, Jian Pu, and Xiangyang Xue. Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In *ACM International Conference on Multimedia (MM)*, 2014.

[Xu *et al.*, 2012] Min Xu, Changsheng Xu, Xiangjian He, Jesse S. Jin, Suhuai Luo, and Yong Rui. Hierarchical affective content analysis in arousal and valence dimensions. *Signal Processing: Indexing of Large-Scale Multimedia Signals*, 93(8), 2012.

[Xu *et al.*, 2014] Can Xu, Suleyman Cetintas, Kuang-Chih Lee, and Li-Jia Li. Visual sentiment prediction with deep convolutional neural networks. *arXiv preprint arXiv:1411.5731*, 2014.

[Xu *et al.*, 2015] Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li, and Leonid Sigal. Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. *arXiv preprint arXiv:1511.04798*, 2015.

[Yacoob and Davis, 1996] Yaser Yacoob and Larry S. Davis. Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 18(6), 1996.

[Yang *et al.*, 2006] Yi-Hsuan Yang, Chia-Chu Liu, and Homer H. Chen. Music emotion classification: A fuzzy approach. In *ACM International Conference on Multimedia (MM)*, 2006.

[Yang *et al.*, 2008] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H. Chen. A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 16(2), 2008.

[Yang *et al.*, 2014] Yang Yang, Jia Jia, Shumei Zhang, Boya Wu, Qicong Chen, Juanzi Li, Chunxiao Xing, and Jie Tang. How do your friends on social media disclose your emotions? In *Conference of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2014.

[Yanulevskaya *et al.*, 2008] Victoria Yanulevskaya, Jan van Gemert, Katharina Roth, Ann-Katrin Herbold, Nicu Sebe, and Jan-Mark Geusebroek. Emotional valence categorization using holistic image features. In *IEEE International Conference on Image Processing (ICIP)*, 2008.

[Yim *et al.*, 2015] Junho Yim, Heechul Jung, Byung-In Yoo, Changkyu Choi, Dusik Park, and Junmo Kim. Rotating your face using multi-task deep neural network. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[Yin *et al.*, 2000] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato. A 3D facial expression database for facial behavior research. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2000.

[You *et al.*, 2014] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Conference of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2014.

[Yuan *et al.*, 2013] Jianbo Yuan, Sean Mcdonough, Quanzeng You, and Jiebo Luo. Sentribute: Image sentiment analysis from a mid-level perspective. In *ACM Knowledge Discovery and Data Mining (KDD) Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)*, 2013.

[Zajonc, 1980] Robert B. Zajonc. Feeling and thinking: Preferences need no inferences. *American Psychologist*, 35(2), 1980.

[Zeiler and Fergus, 2014] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, 2014.

[Zeng *et al.*, 2009] Zhihong Zeng, Maja Pantic, Glenn I. Poisman, and Thomas S. Huang. A survey of affect recognition: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 31(1), 2009.

[Zhang and Kender, 2013] John R. Zhang and John R. Kender. Recognizing and tracking clasping and occluded hands. In *IEEE International Conference on Image Processing (ICIP)*, 2013.

[Zhao *et al.*, 2014] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art features for image emotion recognition. In *ACM International Conference on Multimedia (MM)*, 2014.

[Zheng *et al.*, 2009] Yan-Tao Zheng, Ming Zhao, Yang Song, Hartwig Adam, Ulrich Buddemeier, Alessandro Bissacco, Fernando Brucher, Tat-Seng Chua, and Hartmut Neven. Tour the world: Building a web-scale landmark recognition engine. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[Zhou *et al.*, 2012] Jiayu Zhou, Jianhui Chen, and Jieping Ye. MALSAR: Multi-tAsk Learning via StructurAl Regularization. Technical report, Arizona State University, 2012.