Visual Affect Around the World: A Large-scale Multilingual Visual Sentiment Ontology

Anonymous

Paper ID: 170

ABSTRACT

Every culture and language is unique. Our work expressly focuses on the uniqueness of culture and language in relation to human affect, specifically sentiment and emotion semantics, and how they manifest in social multimedia. We develop sets of sentiment- and emotion-polarized visual concepts by adapting semantic structures called adjective-noun pairs, originally introduced by Borth et al. [6], but in a multilingual context. We propose a new language-dependent method for automatic discovery of these adjective-noun constructs. And 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 structure in [6], our unified ontology is organized hierarchically by multilingual clusters of visually detectable nouns and subclusters of emotionally biased versions of these nouns. In addition, we present an image-based prediction task to show how generalizable language-specific models are in a multilingual context. A new, publicly available dataset over 12 languages, >15.6K sentiment-biased visual concepts with language-specific detector banks, >7.36M images and metadata are also released to the research community.

Categories and Subject Descriptors

H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding

Keywords

Multilingual; Language; Cultures; Cross-cultural; Emotion; Sentiment; Ontology; Concept Detection; Social Multimedia

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 languages used to express affective visual concepts, e.g. *beautiful place* or *delicious food*? And (2) how are such affective visual concepts used to convey different emotions and sentiment across languages?

In Psychology, there are two commonly held schools-of-thought on the connection between cultural context and human affect, i.e. our experiential feelings via our sentiments and emotions. Some believe emotion to be culture-specific [34], that is, emotion is dependent on one's cultural context, while others believe emotion to be universal [17], that is, emotion and cultural are independent mechanisms. For example, while this paper is written in English, there are emotion words/phrases in other languages for which there is no exact translation in English, e.g., *Schadenfreude* in German

Figure 1: Example images from "around the world" organized by affective visual concepts. Top set shows images of *old market* concept from three different cultures/languages; and the bottom, images of *good food*. Even though conceptual reference is the same, each culture's sentimental expression of these concept may be adversely different.

ACM Multimedia '15 Brisbane, Australia

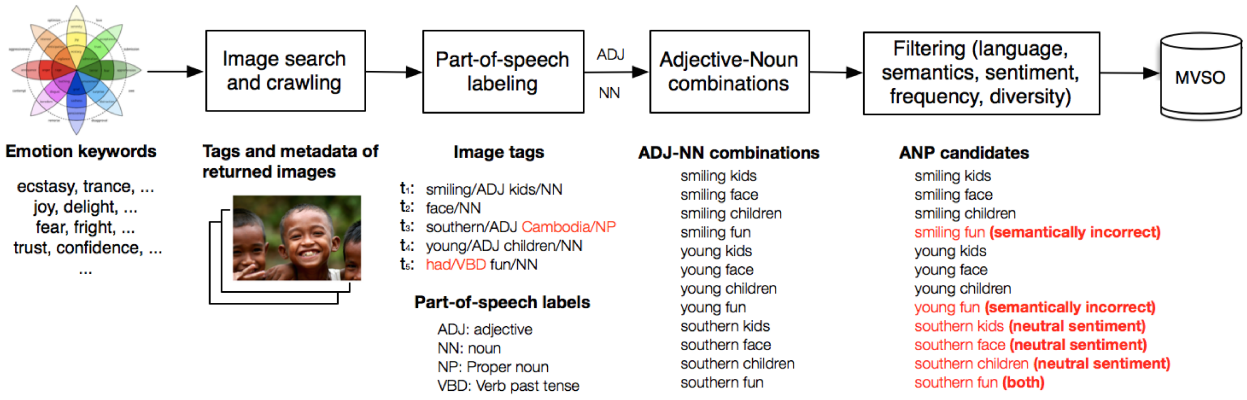

Figure 2: The construction process of our multilingual visual sentiment ontology (MVSO) begins with crawling images and metadata based on emotion keywords. Image tags ($t_1, \ldots, t_5$) are labeled with part-of-speech tags, and adjectives and nouns are used to form candidate adjective-noun pair (ANP) combinations [6], while others are ignored (in red). Finally, these candidate ANPs are filtered based on various criteria (Sec. 3.2) which help remove incorrect pairs (in red), forming a final MVSO with diversity and coverage.

refers to the 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 reference is the same, perhaps the underlying emotion is different?

In Affective Computing [38] and Multimedia, 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*. In one attempt to bridge sentiment and visual media, Borth et al. [6] developed a *visual sentiment ontology* (VSO), a set of 1200 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*. Many works like this that have built algorithms, models and datasets on the assumption of the psychology theory that emotions are universal. However, while such works provide great research contributions in that native language, their applicability and generalization to other languages and cultures remains largely unexplored.

We present a large-scale multilingual visual sentiment ontology (MVSO) and dataset including adjective-noun pairs from 12 languages of diverse origins: Arabic, Chinese, Dutch, English, French, German, Italian, Persian, Polish, Russian, Spanish, and Turkish. We make the following contributions: **(1)** a principled, content-aware pipeline for designing a multilingual visual sentiment ontology, **(2)** a Multilingual Visual Sentiment Ontology mined from social multimedia data end-to-end, **(3)** a MVSO organized hierarchically into noun-based clusters and sentiment-biased adjective-noun pair subclusters, **(4)** a multilingual, sentiment-driven visual concept detector bank, and **(5)** the release of a dataset containing MVSO and large-scale image dataset with benchmark cross-lingual sentiment prediction[1].

2. RELATED WORK

We address the general challenge of *affective image understanding*, aiming at both recognition and analysis of sentiment and emotions in visual data, but from a multilingual and cross-cultural perspective. Our work is closely related to Multimedia and Vision research that focus on visual aesthetics [4, 25], interestingness [15], popularity [26], memorability [19] and creativity [40]. Our work also relates to research in Cognitive and Social Psychology, especially emotion and culture research [12, 35, 39], but also neuroaesthetics [47], visual preference [47, 50], and social interaction [33].

Progressive research in "visual affect" recognition was done in [48] and [32] where image features were designed based on art and psychology principles for emotion prediction. And such works were later improved in [20] by adding social media data in semi-supervised frameworks. From this research effort in visual affect understanding, several affective image datasets were released to the public. The International Affective Picture System (IAPS) dataset [29] is a seminal dataset of ~1,000 images, focused on induced emotions in humans for biometric measurement. The Geneva Affective PicturE Database (GAPED) [9] consists of 730 pictures meant to supplement IAPS and tries to narrow the themes across images. And recently, in [6], a *visual sentiment ontology* (VSO) and dataset was created from Flickr image data, resulting in a collection of adjective-noun pairs along with corresponding images, tags and sentiment. It is also worth mentioning that there are several affective video-based datasets as well, e.g. see the DEAP [27] and MAHNOB-HCI Tagging [43] datasets. One major issue with these datasets and existing methods is that they do not consider the *context* in which emotions are felt and perceived. Instead they assume that visual affect is universal, and do not take in account the influence of culture or language. We explicitly tackle visual affect understanding from a multi-cultural and multilingual perspective. In addition, while existing works often use handpicked data, we gather our data "in the wild" on a popular, multilingual social multimedia platform.

The study of emotions across language and culture has long been a topic of research in Psychology. The main controversy among psychologists in the field is with regard to whether emotions are culture-specific [34], i.e. their perception and elicitation varies with the context, or universal [17]. In [41], a survey of cross-cultural work on semantics surrounding emotion elicitation and perception is given, show-

[1] *[Website link withheld for double-blind submission]*

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 1: Most representative keywords for several basic emotions according to native/proficient speakers for 7 of our 12 languages, chosen and shown top-to-bottom in decreasing no. of discovered visual affect concepts, or adjective-noun pairs.

ing 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[2] from a multimedia-driven and computational perspective, as far as we know.

Other work in cross-lingual research come from text sentiment analysis and music information retrieval. In [3] and [36], they developed multilingual methods for international text sentiment analysis in online blogs and news articles, respectively. In [31] and [18], they presented approaches to indexing digital music libraries with music from multiple languages. Specific to emotion, [18] 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.

It is important to distinguish our work from that of Borth et al. on VSO [6] and its associated detector bank, SentiBank [5]. Their mid-level approach has recently proven effective for affective understanding in images [5] and video [22] as well as finding a wide range of applications in emotion prediction [5, 22, 24], social media commenting [8], etc. However, in addition to lack of multilingual and there are several technical challenges with VSO that we seek to address. We improve on VSO [5, 6] by: (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 criterion based on image tag counts instead of metadata count. Our proposed MVSO discovery method can be easily extended to any language, while achieving greater coverage and diversity than VSO.

3. ONTOLOGY CONSTRUCTION

An overview of the proposed method for multilingual visual sentiment concept ontology construction is shown in Figure 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 [39] or [12]. 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 language, semantics, sentiment, frequency and uploader 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 speakers 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. We settled on 12 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. We selected Flickr because there is an existing body of multimedia research using it in the past, and in particular, [23] 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 [23], but note that in addition those benefits, Flickr has multilingual support and the use of Flickr provides an natural comparison to the seminal VSO [6] work.

3.1 Adjective-Noun Pair Discovery

As our seed emotion ontology, we selected the *Plutchik's Wheel of Emotions* [39]. 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 [12]; it has also been shown to be useful for VSO [6]. The Plutchik emotions are organized in 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.

Multilingual Query Construction: To obtain seeds

[2]Note that we use *language* and *culture* interchangeably often. We define language as the "lens" through which we can observe culture. So while the two can be distinguished, for simplicity, we use them interchangeably.

[3]www.flickr.com

	#images	#tags	#cand	#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
#total	5,940,610	107,782,684	8,154,766	15,630

Table 2: 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 #cand. Through a series of filters – frequency, language, semantics filter, sentiment filter and diversity – and after crowdsourcing, we got our final visual sentiment concepts #anps.

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. 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[4] or other resources to enrich their emotion keywords. Table 1 lists top 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 query on tags on Flickr API to retrieve images and their related tags. Like [6], 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.

Part-of-speech Labeling: To identify the type of each word in a Flickr tag, we performed automatic part-of-speech labeling using pre-trained language-specific taggers which achieve high accuracy (>95% for most languages), namely TreeTagger [42], Stanford tagger [45], HunPos tagger [16] and a morphological analyzer for Turkish [14]. 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 is almost always taken into account. From the full set of part-of-speech labels, we retain those that 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.

Discovery Strategy: We 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 together 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 [6] overlap with the English ANPs discovered by our method.

[4] `translate.google.com`

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.

Language & Semantics: We used a combination of language dictionaries[5] instead of language classifiers to verify the correctness of the ANP as the performance of using the latter alone was low on short-length text, especially for Romance languages which share characters. All of the English ANPs were classified as indeed English by the dictionary, and 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 all other languages were mixed primarily with English. We removed candidate pairs which referred to named entities or technical terms, where named entities were detected using several publicly available knowledge bases such as Wikipedia[6] and dictionaries for names[7], cities, regions and countries[8], and technical terms were removed with a manual created listing of words specific to our source domain, Flickr, containing photography-related (e.g. *macro*, *exposure*) and camera-related words (e.g. *DSLR*, *Canon*).

Non-neutral Sentiment: To filter out neutral candidate adjective-noun pairs, each ANP was scored in sentiment using two publicly available sentiment ontologies: SentiStrength [44] and SentiWordnet [13]. SentiStrength ontology supports all the languages consider, 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 [1, 2]. We computed the ANP sentiment score $S(anp) \in [-2, +2]$ as:

$$S(anp) = \begin{cases} S(a) & : sgn\{S(a)\} \neq sgn\{S(n)\} \\ S(a) + S(n) & : \text{otherwise} \end{cases} \quad (1)$$

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 adjective 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.

Frequency: Good ANPs are those which are actually used together, that is, the adjective and noun co-occur as a

[5] `www.winedt.org`
[6] `www.wikipedia.org`
[7] `www.ssa.gov`
[8] `www.geobytes.com`

pair. Here, we loosely defined an ANP's "frequency" of usage as its number of occurrences as a 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 [11], 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 (non-zero frequency); but for English, since Flickr's most dominant number of users are English-speaking, we set a conservatively higher frequency threshold of 40.

Diversity: The shear frequency of an adjective-noun pair occurrence alone was not sufficient to ensure the 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 ANPs whose associated images by searching metadata were uploaded by a small number of users, and observed that a power law distribution occurs 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.

To further ensure the diversity of MVSO, we subsampled nouns in every language by limiting no more than 100 nouns per adjective so that we do not have, for example, the adjective *surprising* modifying every possible noun in the corpus. In addition, we performed stem unification by checking what inflected form (e.g. singular/plural) of an ANP was most popular in their usage as a tag on Flickr. This unification did also filter some candidate ANPs as some "duplicates" were present but simply in different inflected forms.

3.3 Crowdsourcing Validation

A further inspection of the corpus after the automatic filtering process showed that some frequent 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 ontology, we crowdsourced a validation task. For each language, we asked native speaking workers to evaluate the correctness of the post automatic filtering ANPs. We collected judgements using CrowdFlower[9], a crowdsourcing platform that distributes small tasks to a large number of workers, where we limited workers by their language expertise. We note that while we elected to perform this additional stage of crowdsourcing, other researchers may find a fully automatic pipeline more desirable, so in our public release, we also release the pre-crowdsourced version of our MVSO.

To help workers assess the correctness of the ANPs, we developed separate crowdsourcing jobs for each language. For each validation job, we selected about 5,000 ANPs by ranking ANPs by their number of crawled images by tag search. The choice of 5,000 was largely based simply on our monetary budget.

[9] `www.crowdflower.com`

	#cand	#users	#coun	%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 3: Crowdsourcing results via no. of input candidate ANPs #cand, #users, countries #coun, and perc. of ANPs accepted %correct and annotator agreement %agree.

3.3.1 Crowdsourcing Setup

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 has 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 reasons, for three languages (Persian, Polish and Dutch), the CrowdFlower platform does not to evaluate workers based on their language expertise, so we filtered them by provenience, selecting the countries according the official language spoken (e.g. Netherlands, Belgium, Aruba and Suriname for Dutch).

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 truth were carefully judged and selected by four independent expert annotators. In the interface, aside from instructions, workers were shown the five ANPs and simply chose between "yes" or "no" to validate ANPs.

Quality Control: Like some other crowdsourcing platforms, CrowdFlower provides 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 question 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 for the test questions.

3.3.2 *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 ANP 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 proportion of incorrect ANPs according to the crowd was actually greater than the number of correct ANPs. In Table 3, we summarize the statistics from our crowdsourcing experiments relative to the number of ANPs, percentages of correct/incorrect ANPs according to worker majority vote, and average agreement.

4. DATASET ANALYSIS & STATISTICS

Having acquired a final set of adjective-noun pairs for each of the 12 languages, we downloaded images from Flickr API querying by 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 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 images which Flickr or uploaders removed or changed privacy settings on midway were removed in post. 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.

4.1 Comparison with VSO [6]

To verify and test the efficacy of our MVSO, we provide a comparision of our extracted English visual sentiment ontology with that of VSO [6] along dimensions of size (number of ANPs) and diversity of nouns and adjectives (Figure 3). In Figure 3a, the overlap of English MVSO with VSO is compared with VSO alone after applying all filtering criteria except from subsampling filter which might exclude ANPs belonging to VSO. As mentioned previously, about 86% overlaps between them. As we vary a frequency threshold t over image tag counts, the overlap converges to 100%.

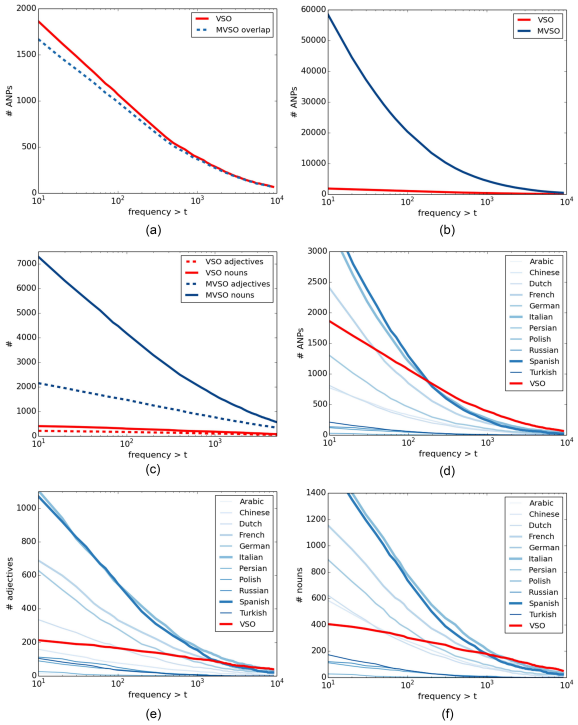

Figure 3: Comparison of our English MVSO and VSO [6] in Figures (a), (b) and (c), in terms of ANP overlap, no. of ANPs, adjectives and nouns; and with all other languages in Figures (d), (e) and (f), in terms of the no. of ANPs, adjectives and nouns when varying the frequency threshold t from 0 to 10,000 (on log-scale), respectively.

In Figure 3b, we show that there are far greater number of ANPs in our English MVSO compared to VSO ANPs throughout all the possible values of a frequency threshold, after applying all filtering criteria. Similarly, as shown in Figure 3c, given there are more adjectives and nouns in our English MVSO, we also achieve greater diversity than VSO.

In Figure 3d, 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 3e and 3d, we observe that 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.

4.2 Sentiment Distributions

Returning to our research motivation from the Introduction, an interesting question to ask is which languages tend to be more positive or negative in their visual content. To answer this question, we computed the median sentiment value across all ANPs and ranked languages as in Figure 4. Here, we used a weighted sentiment per ANP $S(anp)(1 + \frac{1}{N}\text{count}(anp))$, where $S(anp)$ is from Eq. (1) but is now weighted by count(anp), which denotes the number of im-

ages tagged with the ANP and N is the total number of images in the given language. The median was used instead of the mean because the latter deviates a lot from the center of the distribution due to biasing by extreme values; the median generally remain closer to the center (50^{th} percentile), and allows us to better compare across languages.

Overall, we observed that there is a tendency toward positive sentiment across all languages, where Spanish demonstrates the highest positive sentiment, followed by Chinese and Dutch. This surprising observation is in fact compatible with previous research showing that there is a universal positivity bias over languages with Spanish being the most relatively positive language [10]. The languages with the lowest sentiment were Arabic and French. The sentiment distributions (Figure 4 - right) were concentrated in low sentiment score ranges around 0, for example -0.05 to 0.25 for Spanish and -1.5 to 0.18 for Persian. The most variant sentiment distributions were from Spanish, Persian, Russian and Arabic, while the least ones were Polish, Dutch and Turkish. The higher variance, the wider spectrum of sentiment of ANPs in that language, and likely a greater variety in the expression of visual sentiment in such concepts.

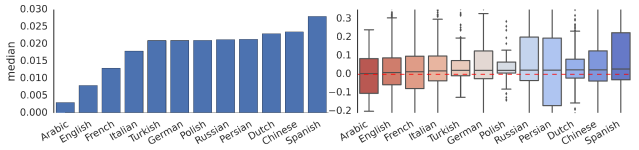

Figure 4: Median sentiment computed over all ANPs per language is shown on left, and the sentiment distribution using box plots on the right (zoomed at 90% of the distributions). On right, languages are sorted by median sentiment in ascending order (left to right).

4.3 Emotion Distributions

Another interesting question arises when considering co-occurrence of ANPs with the emotions in different languages. What emotions are the most frequently occurring across languages? Aside from sentiment, which focuses on only positivity/negativity, what are probable mappings of ANPs to emotions for each language? 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:

$$\text{emo}^i(x) = \frac{\frac{1}{n_i}\sum_{j=1}^{n_i} c_{ij}^{(x)}}{\sum_{i=1}^{24}\frac{1}{n_i}\sum_{j=1}^{n_i} c_{ij}^{(x)}} \in [0,1] \qquad (2)$$

Note the model in (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 (2),

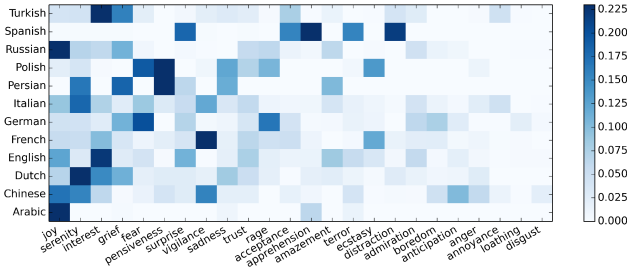

Figure 5: Probabilities of emotions per language with respect to their visual sentiment content. Emotions are ordered by the sum of their probabilities across languages (left to right) and clipped for better visualization.

we compute a normalized emotion score per language l and emotion i as:

$$\text{score}^i(l) = \frac{\sum_{x=1}^{|X^{(l)}|}\text{emo}^i(x)\cdot\text{count}(x)}{\sum_{i=1}^{24}\sum_{x=1}^{|X^{(l)}|}\text{emo}^i(x)\cdot\text{count}(x)} \in [0,1] \qquad (3)$$

Figure 5 shows these scores per language and Plutchik emotion [39] on a heatmap diagram. Scores in each row sum to 1 (over 24 emotions). The emotions 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.

5. CROSS-LINGUAL MATCHING

To get a gauge on the 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. We ensured that translations of ANPs also passed all our validation filters for this analysis.

Exact Alignment: We grouped ANPs from each language that have the exact same translation. 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 that covered the greatest number of languages was *beautiful girl* with ANPs from ten languages. Figure 6 shows a correlation matrix of the number of times ANPs from pairs of languages appeared together in a set of ANPs with exact same translation, e.g. out of all the translations that German ANPs were translated to (782), more of them 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).

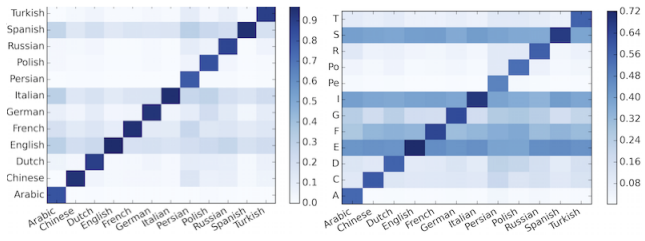

Figure 6: No. of times ANPs from two languages were translated to the same phrase (Left: Exact Alignment) or phrase in the same cluster (Right: Approximate Alignment). Read columnwise to compare a language with others.

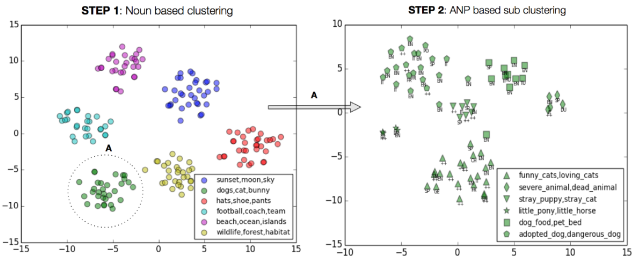

Figure 7: Examples of noun clusters (left) and ANP sub-clusters (right) from our two-stage clustering for cross-lingual matching. For visualization, word2vec [37] vectors were projected to $\mathbb{R}^2$ using t-SNE [46].

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 Stanford tagger [45] from the list of translated phrases and discovered 3,315 total clusters, or nouns. We then extracted word2vec [37] features, a word representation trained on a Google News[10] corpus, for these translated nouns (112 nouns were out-of-vocabulary), and performed k-means clustering (k=200) to get groups of nouns with similar meaning. The number of clusters was 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 [37] features for this full translated phrase in each cluster and used them to run one more round of k-means clustering (adjusting k based on the number of phrases in each cluster, where clusters from first stage varied from 8 to 414). 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), unlike the flat structure in VSO [6]. We discovered 3,545 sub-clusters of ANP concepts, e.g. resulting in clusters containing *little pony* and *little horse* as in Figure 7. This approach also yielded a larger intersection between languages, where German and Dutch share 153 clusters, and German and Chinese intersect over 128 ANP clusters.

The correlation matrix from this approximate matching is shown in Figure 6, along with one subtree from our ontology by hierarchical clustering in Figure 7. For Figure 7, we projected data to $\mathbb{R}^2$ using t-SNE dimensionality reduction [46]. On the left, six clusters composed of different sets of nouns are shown with clusters of *sunset-moon-sky* and *dogs-cat-bunny*. On the right, we show the sub-clustering of ANPs for the *dogs-cat-bunny* cluster in **A**, giving us noun groupings modified by sentiment-biasing adjectives to get ANPs like *funny cats-loving cats* and *adopted dog-dangerous dog*.

6. VISUAL SENTIMENT PREDICTION

To test the effectiveness of a vision-based approach for visual affet understanding when crossing languages, we designed and built language-specific sentiment predictors using the data collected with MVSO. Inspired by work in [18], 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 there is a large body of work expressly focused on sentiment (e.g. [6, 48, 49]) for its simplicity, compared to emotion prediction. More importantly, we wanted to reduce the number of variables to be analyzed since our primary goal is to uncover cross-lingual differences.

We first constructed a bank of visual concept detectors like in [5] for our final MVSO adjective-noun pairs. For simplicity, we focused on the six languages with the most ANPs and associated images in our dataset: in descreasing 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 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.

6.1 Visual Sentiment Concept Detectors

To construct our bank of visual concept detectors of ANPs, we used convolutional neural networks (CNNs) [30], in particular, adopting the AlexNet architecture [28] for its good performance on large-scale vision recognition and detection tasks. To train our detector bank, we fined-tuned six different models for each language, where network weights were initialized with DeepSentiBank [7], an AlexNet model trained on VSO [6] dataset, which we obtained from its authors. This fine-tuning approach ensures that each network begins

	#ANPs	#train	#test	lrs (K)	time (hr)	top-1	top-5
English	4,342	3,236,728	807,447	50	40	10.1%	21.7%
Spanish	2,382	1,085,678	270,400	40	35	12.4%	25.4%
Italian	1,561	602,424	149,901	30	30	17.0%	30.9%
French	1,115	462,522	115,112	30	26	17.7%	35.5%
German	275	108,744	27,048	20	12	30.1%	52.8%
Chinese	243	102,740	25,575	20	15	27.1%	45.0%
DSB [7]	2,089	826,806	41,113	-	-	8.2%	19.1%

Table 4: Adjective-noun pair (ANP) classification performance on Flickr images for six major languages in MVSO and compared to DeepSentiBank (DSB) [7]. No. of visual sentiment concepts #ANPs, #train and #test images along with learning rate step size (lrs, in thousands) are shown with training times (in hours), top-1 and top-5 accuracies.

[10] news.google.com

Figure 8: Example top-5 classification results from our multilingual visual sentiment detector bank. Translations to English provided for convenience.

with weights that are already somewhat "affectively" biased. The base learning rates were set to 0.001 and 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 in the second stage were set proportional to number of training images per language. For a single language, fine-tuning took between 12 and 40 hours for convergence on a single NVIDIA GTX 980 GPU implemented with Caffe [21]. From Table 4, as expected we achieve higher top-1 and top-5 accuracies than DeepSentiBank [7], but also even when having more output neurons as in English and Spanish. Top-k accuracy refers to the normalized number of correct classifications given that true class is in the top k predicted ranks.

6.2 Sentiment Prediction on Flickr

Once the deep models were fine-tuned, we extracted features from the last three layers, two fully-connected and softmax output layer, from running the feedforward network on the validation set of Sec. 6.1 and applied them as inputs for image-based sentiment prediction. Training and testing was split from this same validation set but pooled from across all ANPs and by binning the sentiment labels from Eq. (1) into positive and negative classes. We rejected all samples from ANPs with a sentiment score from Eq. (2) of less than 0.05; and for fair comparison in cross-lingual experiments, we stratified our training and testing sets across all languages so that the amount of training and testing for positive and negative sentiment classes was the same. We trained sentiment prediction models using standard linear SVMs with grid search for selecting the slack.

We found that the softmax output features, i.e. probabilistic outputs of each language's deep network over the presence of ANPs, performed the best for all languages, and show resulting sentiment prediction results in Figure 9. Each language expectedly did better in predicting test samples from its own language, but in addition, Chinese generally was the most difficult to predict by models trained from other languages; and using a Chinese-specific sentiment model to predict the sentiment in other languages was also the worst in average. We speculate that this is due to the difference

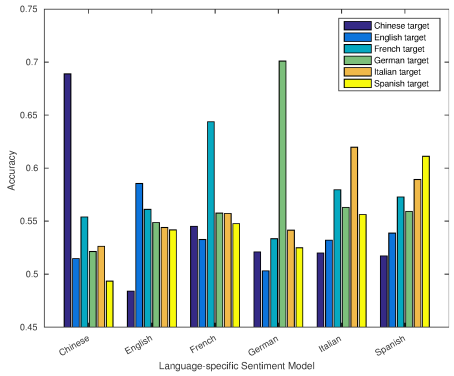

Figure 9: Image-based, cross-lingual domain transfer sentiment prediction results with language-specific models applied on cross-lingual examples.

in the visual sentiment portrayal from Eastern and Western cultures. Interestingly, the classification of French and Italian sentiment was the most consistently high using models from all languages. We observed good performance in cross-lingual prediction for Latin languages, i.e. Spanish, Italian and French. For example, Italian was the best cross-lingual classifier for Spanish and French sentiment, and Spanish was best cross-lingual classifier for Italian sentiment, followed by French. Despite not being a very strong sentiment classifier, the English-specific sentiment model had the least variance in its accuracy across all languages, likely a result of the pervasive use of English all over the world.

Figure 10: 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.

In Figure 10, 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 and binning, 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 that an image from the English test set of *foggy morning* as positive, possibly for its resemblence to a Chinese 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 experiment highlight some similarities and differences in how visual sentiment is expressed and perceived by various cultures.

7. CONCLUSION & FUTURE WORK

We proposed a new multilingual discovery method for vi-

sual sentiment concepts and showed its efficacy on a social multimedia platform for 12 languages. We based our approach on the 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. In addition, we presented a two-stage hierarchical clustering approach to unify our ontology across languages. We make our Multilingual Visual Sentiment Ontology (MVSO), pre-crowdsourcing as well as post, and image dataset, available to the public. A cross-lingual analysis of our large-scale MVSO and image dataset using semantic matching and visual sentiment prediction hint that emotions are not necessarily culturally universal. Our preliminary results show that there are indeed commonalities, but also distinct separations, in how visual affect is expressed and perceived, where other works assumed only commonalities. And yet, we believe these point to the colorful diversity of our world, rather than our inability to understand one another.

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 will also adopt our approach to other language-specific social multimedia platforms to counter the insufficient data for some languages like Arabic, Persian and Chinese. In addition, while we discussed culture and languages in this work, we have not yet performed an in-depth study on geo-location data in MVSO, often provided along with uploaded images on Flickr. While such information could be useful to distinguish between sub-cultures speaking the same languages (e.g. Spanish vs. South-Americans), we omitted such a study here because of the noise that geo-location data can add. For example, an American traveling in China uploading pictures is still more likely to use his native tongue to tag and sentimentally describe his content. The trade-off is that while his semantics are culturally American, his uploaded visual content is now from another culture, so there is still much to be explored from geo-location and user metadata.

8. REFERENCES

[1] A. Balahur and M. Turchi. Multilingual sentiment analysis using machine translation? In *WASSA*, 2012.

[2] C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan. Multilingual subjectivity analysis using machine translation. In *EMNLP*, 2008.

[3] M. Bautin, L. Vijayarenu, and S. Skiena. International sentiment analysis for news and blogs. In *ICWSM*, 2008.

[4] S. Bhattacharya, B. Nojavanasghari, T. Chen, D. Liu, S.-F. Chang, and M. Shah. Towards a comprehensive computational model for aesthetic assessment of videos. In *ACM MM*, 2013.

[5] D. Borth, T. Chen, R. Ji, and S.-F. Chang. SentiBank: Large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *ACM MM*, 2013.

[6] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, 2013.

[7] T. Chen, D. Borth, T. Darrell, and S.-F. Chang. DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014.

[8] Y.-Y. Chen, T. Chen, W. H. Hsu, H.-Y. M. Liao, and S.-F. Chang. Predicting viewer affective comments based on image content in social media. In *ICMR*, 2014.

[9] E. S. Dan-Glauser and K. Scherer. The Geneva affective picture database: A new 730-picture database focusing on valence and normative significance. *Behav. Res. Meth.*, 43(2), 2011.

[10] S. Dodds and et al. Human language reveals a universal positivity bias. *PNAS*, 112(8), 2015.

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

[12] P. Ekman. Facial expression and emotion. *American Psychologist*, 48(4), 1993.

[13] A. Esuli and F. Sebastiani. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *LREC*, 2006.

[14] Z. Güngördü and K. Oflazer. Parsing Turkish using the lexical functional grammar formalism. In *ACL*, 1994.

[15] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. V. Gool. The interestingness of images. In *ICCV*, 2013.

[16] P. Halácsy, A. Kornai, and C. Oravecz. HunPos: An open source trigram tagger. In *ACL*, 2007.

[17] M. G. Haselton and T. Ketelaar. Irrational emotions or emotional wisdom? The evolutionary psychology of affect and social behavior. *Affect in Soc. Think. and Behav.*, 8(21), 2006.

[18] X. Hu and Y.-H. Yang. Cross-cultural mood regression for music digital libraries. In *JCDL*, 2014.

[19] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *CVPR*, 2011.

[20] J. Jia, S. Wu, X. Wang, P. Hu, L. Cai, and J. Tang. Can we understand van Gogh's mood?: Learning to infer affects from images in social networks. In *ACM MM*, 2012.

[21] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014.

[22] Y.-G. Jiang, B. Xu, and X. Xue. Predicting emotions in user-generated videos. In *AAAI*, 2014.

[23] X. Jin, A. Gallagher, L. Cao, J. Luo, and J. Han. The wisdom of social multimedia: Using Flickr for prediction and forecast. In *ACM MM*, 2010.

[24] B. Jou, S. Bhattacharya, and S.-F. Chang. Predicting viewer perceived emotions in animated GIFs. In *ACM MM*, 2014.

[25] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR*, 2006.

[26] A. Khosla, A. Das Sarma, and R. Hamid. What makes an image popular? In *WWW*, 2014.

[27] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. DEAP: A database for emotion analysis using physiological signals. *IEEE TAC*, 3(1), 2011.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

[29] P. Lang, M. Bradley, and B. Cuthbert. International Affective Picture System (IAPS): Technical manual and affective ratings. Technical report, NIMH CSEA, 1997.

[30] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proc. of the IEEE*, 1998.

[31] J. H. Lee, J. S. Downie, and S. J. Cunningham. Challenges in cross-cultural/multilingual music information seeking. In *ISMIR*, 2005.

[32] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM MM*, 2010.

[33] H. R. Markus and S. Kitayama. Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98(2), 1991.

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

[35] B. Mesquita, N. H. Frijda, and K. Scherer. Culture and emotion. In J. W. Berry, P. R. Dasen, and T. S. Saraswathi, editors, *Handbook of Cross-cultural Psychology*, volume 2. Allyn & Bacon, 1997.

[36] R. Mihalcea, C. Banea, and J. Wiebe. Learning multilingual subjective language via cross-lingual projections. In *ACL*, 2007.

[37] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

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

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

[40] M. Redi, N. O'Hare, R. Schifanella, M. Trevisiol, and A. Jaimes. 6 Seconds of sound and vision: Creativity in micro-videos. In *CVPR*, 2014.

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

[42] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Intl Conf. on New Methods in Language Proc.*, 1994.

[43] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimedia database for affect recognition and implicit tagging. *IEEE TAC*, 3(1), 2012.

[44] M. Thelwall, K. Buckley, G. Paltoglou, and D. Cai. Sentiment strength detection in short informal text. *Jour. Ameri. Soci. for Info. Sci. & Tech.*, 61(12), 2010.

[45] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL*, 2003.

[46] L. van der Maaten and G. E. Hinton. Visualizing high-dimensional data using t-SNE. *JMLR*, 9, 2008.

[47] E. A. Vessel, J. Stahl, N. Maurer, A. Denker, and G. G. Starr. Personalized visual aesthetics. In *SPIE-IS&T Electronic Imaging*, 2014.

[48] V. Yanulevskaya, J. van Gemert, K. Roth, A. Herbold, N. Sebe, and J. M. Geusebroek. Emotional valence categorization using holistic image features. In *ICIP*, 2008.

[49] Q. You, J. Luo, H. Jin, and J. Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*, 2014.

[50] R. B. Zajonc. Feeling and thinking: Preferences need no inferences. *American Psychologist*, 35(2), 1980.