Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs

Damian Borth [1] Rongrong Ji [2] Tao Chen [2] Thomas Breuel [1] Shih-Fu Chang [2]

[1]University of Kaiserslautern, Germany [2]Columbia University, USA
{d_borth, tmb}@cs.uni-kl.de {rrji, taochen, sfchang}@ee.columbia.edu

ABSTRACT

We address the challenge of sentiment analysis from visual content. In contrast to existing methods which infer sentiment or emotion directly from visual low-level features, we propose a novel approach based on understanding of the visual concepts that are strongly related to sentiments. Our key contribution is two-fold: first, we present a method built upon psychological theories and web mining to automatically construct a large-scale Visual Sentiment Ontology (VSO) consisting of more than 3,000 Adjective Noun Pairs (ANP). Second, we propose SentiBank, a novel visual concept detector library that can be used to detect the presence of 1,200 ANPs in an image. The VSO and SentiBank are distinct from existing work and will open a gate towards various applications enabled by automatic sentiment analysis. Experiments on detecting sentiment of image tweets demonstrate significant improvement in detection accuracy when comparing the proposed SentiBank based predictors with the text-based approaches. The effort also leads to a large publicly available resource consisting of a visual sentiment ontology, a large detector library, and the training/testing benchmark for visual sentiment analysis.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Retrieval and Indexing

Keywords

Sentiment Prediction, Concept Detection, Ontology, Social Multimedia

1. INTRODUCTION

Nowadays the Internet, as a major platform for communication and information exchange, provides a rich repository of people's opinion and sentiment about a vast spectrum of topics. Such knowledge is embedded in multiple facets, such as comments, tags, browsing actions, as well as shared media objects. The analysis of such information either in the area of opinion mining, affective computing or sentiment analysis plays an important role in behavior sciences, which aims to understand and predict human decision making [26] and enables applications such as brand monitoring, stock market prediction, or political voting forecasts.

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502282.

Figure 1: Tweets from the "2012 Year on Twitter" collection: Barack Obamas reelection tweet (left) and a tweet capturing the destruction caused by Hurricane Sandy (right). Both tweets are characterized by a short text ("four more years" and "rollercoaster at sea" respectively) and conveying the sentiment visually.

So far, the computational analysis of sentiment mostly concentrates on the textual content [26]. Limited efforts have been conducted to analyze sentiments from visual content such as images and videos, which is becoming a pervasive media type on the web. For example, two of the most popular tweets in 2012 (see Fig. 1) conveying valuable sentiment information primarily visually[1]. Thus, an open issue in sentiment analysis research is the need of visual content analysis.

This problem poses a set of unique challenges as it addresses abstract human concepts in the sense of emotion and affect. Typically, semantic concept detection in images is concerned with the physical presence of objects or scenes like "car" or "building". On the other hand sentiment could differ among persons as the stimuli evoked human responses are naturally subjective. In some sense, this is analogous to the differentiation between content-based image retrieval (CBIR) and emotional semantic image retrieval (ESIR) [33]. There exists an *affective gap* in ESIR [21] between low-level features and the emotional content of an image reflecting a particular sentiment, similar to the well-known *semantic gap* in CBIR between low-level features and image semantics. To fill the semantic gap, mid-level representations based on visual concepts have been proposed. In this paper, we propose to discover and detect a set of visual concepts that can be used to fill the affective gap and automatically infer the sentiments reflected in an image. Note, that our mid-

[1]Please note that, throughout the paper we will define sentiment similarly to [26], as the polarity of an opinion item which either can be *positive, neutral or negative*

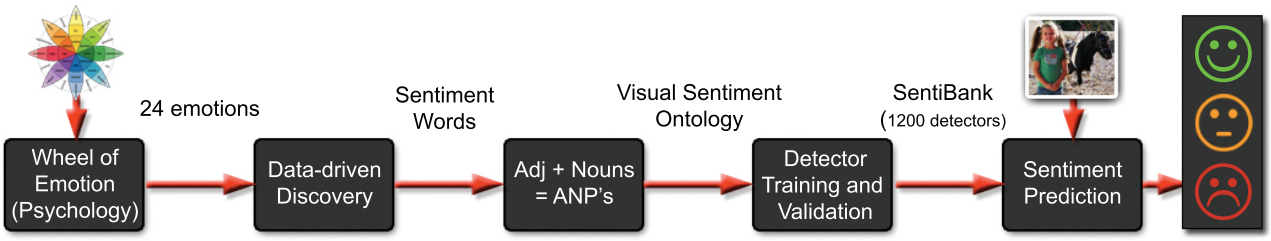

Figure 2: Overview of the proposed framework for constructing the visual sentiment ontology and SentiBank. Applications in multimodal sentiment prediction is also shown.

level representation is much expressive than the ones (e.g., color schemes or geometric shapes) described in [33] and has a better capability for explaining sentiment prediction results.

We apply the psychological theory, Plutchik's Wheel of Emotions [27], as the guiding principle to construct a large-scale **visual sentiment ontology** (VSO) that consists of more than 3,000 semantic concepts. Our construction criteria ensure that each selected concept **(1)** reflects a strong sentiment, **(2)** has a link to emotions, **(3)**is frequently used in practice, and **(4)** has a reasonable detection accuracy. To satisfy these conditions, we introduce **Adjective Noun Pairs (ANP)** such as "beautiful flower" or "disgusting food". The advantage of using ANPs, compared to nouns or adjectives only, is its capability to turn a neutral noun like "dog" into an ANP with strong sentiment like "cute dog" by adding an adjective with a strong sentiment. Such combined phrases also make the concepts more detectable than adjectives (like "beautiful"), which are typically abstract and difficult to detect. Building upon the VSO we introduce **SentiBank**, a library of trained concept detectors providing a mid-level visual representation. We show - through extensive experiments - that useful detector performance can be achieved for 1,200 ANP concepts, which form the released detector library SentiBank. Further, we demonstrate the usefulness of the proposed approach towards sentiment prediction on image tweets as it outperforms text-based prediction approaches by a very large margin. In summary, our contributions are - **first**, a systematic, data-driven methodology for the construction of a visual sentiment ontology from user-generated content and folksonomies on the web; **second**, the large-scale Visual Sentiment Ontology founded by a well-known psychological model; **third**, a mid-level representation built on automatic detectors of the discovered concepts in order to bridge the *affective gap* mentioned earlier; and, **forth**, the public release of the VSO including its large-scale dataset, the SentiBank detector library, and the benchmark for visual sentiment analysis.

In the rest of the paper, we first discuss related work (Sec.2) and show an overview of the proposed framework (Sec. 3). Then, the design and construction methodology of the VSO (Sec.4) and SentiBank, the proposed mid-level visual concept representation (Sec.5) are discussed. Finally, application in image tweet sentiment prediction is presented (Sec.6).

2. RELATED WORK

The challenge of automatically detecting semantic concepts such as objects, locations, and activities in visual data, referred to as video annotation [1], concept detection [28], semantic indexing [25] or multimedia event detection [20], has been studied extensively over the last decade. In benchmarks like TRECVID [25] or the PASCAL visual object challenge [10], the research community has investigated a variety of features and statistical models. In addition, there also has been much work in creating large ontologies and datasets [7, 14, 29]. Typically, such vocabularies are defined according to utility for retrieval, coverage, diversity, availability of training material, and its detectability by automatic detection systems [23, 25]. Recent approaches have also turned towards web portals like Flickr and YouTube as information sources for visual learning, employing user-generated tags as an alternative to manual labels [16, 31]. Aligned with the aforementioned trend, our approach also exploits large-scale image tags available on the web. Our focus, however, is less on concept detection itself but rather on the construction of an ontology of visually detectable ANPs serving as mid-level representation of sentiment attributes of visual content. In contrast, the prior works focus on physical concepts corresponding to objects, scenes, location but not concepts that characterize sentiment visually.

With respect to sentiment analysis, much progress has been achieved in text analysis [9, 30] and textual dictionary creation [9, 35]. However, efforts for visual analysis fall far behind. The closest that comes to sentiment analysis for visual content is the analysis of aesthetics [6, 15, 22], interestingness [12], and affect or emotions [13, 21, 37, 36]. To this end, either low-level features are directly taken to predict emotion [18, 13], or indirectly by facial expressions detection [32] or user intent [11]. Similarly [34], which introduced a high-level representation of emotions, is limited to low-level features such as color based schemes. Please refer to [15, 33] for a comprehensive study of aesthetics and emotions in images. Compared to the above works, our proposed approach is novel and ambitious in two ways. First, we build a large-scale ontology of semantic concepts correlated with strong sentiments like "beautiful landscape" or "dark clouds" as a complement to a textual sentiment dictionary [9, 35]. Such an ontology is the first of its kind and would open new research opportunities for the multimedia community and beyond. Second, from such an ontology and a publicly shared detector library a mid-level visual representation can be learned for the purpose of robust sentiment prediction.

Only a few small datasets exist today for affect / emotion analysis on visual content. A prominent one is the *International Affective Picture System* (IAPS) [17] providing normative ratings of emotion (pleasure, arousal, dominance) for a set of color photographs. The dataset consists of 369 photos covering various scenes showing insects, puppies, children, poverty, diseases and portraits, which are rated by 60 participants using affective words. Similarly, the *Geneva Affective Picture Database* (GAPED) [4] provides

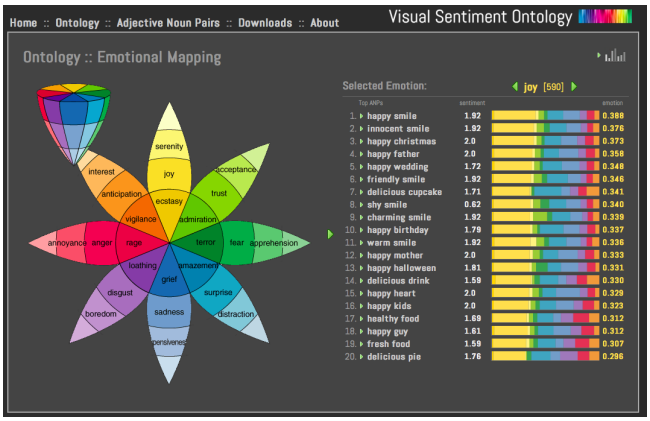

Figure 3: Plutchik's Wheel of Emotions and the visualization interface of the ontology based on the wheel.

730 pictures including negative (e.g., spiders, snakes, scenes containing human rights violation), positive (e.g., human and animal babies, nature sceneries) and neutral pictures. All pictures were rated according to valence, arousal, and the consistency of the represented scenes. In [21], the *Affective Image Classification Dataset* includes two separate datasets of abstract painting (228 paintings) and artistic photos (807 photos), which are labeled with 8 basic emotions through a crowd-sourcing procedure. In contrast to the above mentioned datasets our work provides a significantly larger dataset (about 0.5 million) of images crawled from social media and labeled with thousands of ANP concepts. In addition, we created a separate image dataset from Twitter for a sentiment prediction benchmark.

3. FRAMEWORK OVERVIEW

An overview of the proposed framework is shown in Fig. 2. The construction process is founded on psychological principles such as *Plutchik's Wheel of Emotions* [27]. During the first step, we use each of the 24 emotions defined in Plutchik's theory to derive search keywords and retrieve images and videos from Flickr and YouTube. Tags associated with the retrieved images and videos are extracted - for example "joy" leads to "happy", "beautiful", and "flower". These tags are then analyzed to assign sentiment values and to identify adjectives, verbs, and nouns. The set of all adjectives with strong sentiment values and all nouns is then used to form adjective noun combinations or **Adjective Noun Pairs (ANP)** such as "beautiful flowers" or "sad eyes". Those ANPs are then ranked by their frequency on Flickr and sampled to form a diverse and comprehensive ontology containing more than 3,000 ANP concepts. We then train individual detectors using Flickr images that are tagged with an ANP and keep only detectors with reasonable performance to form **SentiBank**. This detector library consists of 1,200 ANP concept detectors providing a 1,200 dimension ANP detector response for a given image. As a sample application, we apply SentiBank and train classifiers to predict sentiment values of image tweets and demonstrate a superior performance over conventional sentiment prediction using text only.

4. VISUAL SENTIMENT ONTOLOGY

In this section we outline the design and systematic construction of the proposed Visual Sentiment Ontology (VSO). Here we focus on sentiment or emotion expressed by the content owner shared on social media such as Twitter. We assume the sentiments of the receivers (i.e., viewers of the visual content), though not directly addressed in this paper, are strongly related to those of the content owners. Our goal is to construct a large-scale ontology of semantic concepts, which **(1)** reflect a strong sentiment, **(2)** have a link to an emotion, **(3)** are frequently used and **(4)** have reasonable detection accuracy. Additionally, the VSO is intended to be comprehensive and diverse enough to cover a broad range of different concept classes such as *people, animals, objects, natural or man-made places*, and so on.

joy	terror	amazement	disgust
joy	terror	amazing	disgusting
happy	horror	beautiful	gross
love	zombie	nature	food
smile	fear	wonder	nasty
beautiful	dark	light	sick
flowers	street	love	dirty
light	halloween	sky	dead
nature	war	eyes	face
kids	undead	clouds	blood
christmas	bomb	landscape	insect

Figure 4: Example top tags for different emotions. Colors of the boxes (green, grey, red) indicate different sentiments (positive, neutral, negative).

4.1 Psychological Foundation

To establish a solid foundation for the construction of the VSO we utilize a well-known emotion model derived from psychological studies. There are several well-known early works such as Darwin's evolutionary motivation of emotions [5], Ekman's facial expression system [8] and Osgood's appraisal and valence model [24]. Here, focus on *Plutchnik's Wheel of Emotions* [27] as seen in Fig. 3 is organized into 8 basic emotions, each with 3 valences. Beginning from the top we have:

1. ecstasy → joy → serenity
2. admiration → trust → acceptance
3. terror → fear → apprehension
4. amazement → surprise → distraction
5. grief → sadness → pensiveness
6. loathing → disgust → boredom
7. rage → anger → annoyance
8. vigilance → anticipation → interest

Why Plutchnik's Emotion Model? The model is inspired by chromatics in which emotions elements are arranged along a wheel and bi-polar emotions are opposite to each other - a useful property for the construction of a sentiment ontology. Further, it maps well to psychological theories such as Ekman, where 5 basic emotions are the same (anger, disgust, fear, sadness, surprise) while Ekman's "happiness" maps well to Plutchnik's "joy". Compared to the emotional model utilized in [21], Plutchnik basic emotions correspond to all 4 negative emotions but have slightly different positive emotions. In contrast, Plutchnik introduced two additional basic emotions (interest, trust) and organizes each of them into 3 intensities providing a richer set of different emotional valences. Statistics of our crawled

Table 1: Statistics of the Visual Sentiment Ontology construction process

(a)	Flickr	YouTube
# of emotions	24	24
images or videos	150,034	166,342
tags	3,138,795	3,079,526
distinct top 100 tags	**1,146**	**1,047**
(b)	**Sentiment Words**	
distinct top 100 tags	**1,771**	
pos+neg adjectives	268	
neutral adjectives	0	
total adjectives	**268**	
pos+neg nouns	576	
neutral nouns	611	
total nouns	**1,187**	
(c)	**VSO Statistics**	
ANP concept candidates	320k	
ANPs (with non-empty images)	47k	
ANPs included in VSO	**3k**	
top pos. adjectives	beautiful, amazing, cute	
top neg. adjectives	sad, angry, dark	
top nouns	face, eyes, sky	

dataset confirm useful contribution of each emotion group in Plutchik to the final VSO.

4.2 Sentiment Word Discovery

Initial Image & Video Retrieval: For each of the 24 emotions we retrieve images and videos from Flickr and YouTube respectively and then extract their distinct associated tags by the Lookapp tool [2]. In total we retrieve about 310k images and videos and about 6M tags, which are made of a set of 55k distinct tags. An overview of this step can be seen in Tab. 1 (a).

Tags Analysis: For tag analysis we first remove stopwords and perform stemming. For each emotion, we perform tag frequency analysis to obtain the top 100 tags. Examples of such top tags can be seen in Fig. 4. Finally, the sentiment value of each tag is computed using two popular linguistics based sentiment models, SentiWordNet [9] and SentiStrength [30]. In this work, each word is assigned a sentiment value ranging from -1 (negative) to +1 (positive). Overall, as shown Tab. 1 (b), we are able to retrieve 1146 distinct tags from Flickr and 1,047 distinct tags from YouTube forming the final set of 1,771 distinct tags with 1,187 nouns (576 positive and negative ones and 611 neutral ones) and 268 positive or negative adjectives. Note that we ignore verbs in this work because of the current focus on still images.

4.3 Adjective Noun Pair (ANP) Construction

Looking closer at the results of the previous step we can see that the 576 discovered nouns with positive or negative sentiment would satisfy the first concept selection condition mentioned above for ontology construction (reflecting strong sentiment), but in this case we would not be able to include the 611 neutral nouns. As for the adjectives, all 268 have either a positive or negative sentiment value (satisfying condition **(1)**) but probably we would not be able to satisfy condition **(4)**: reasonable detection accuracy. Visual learning

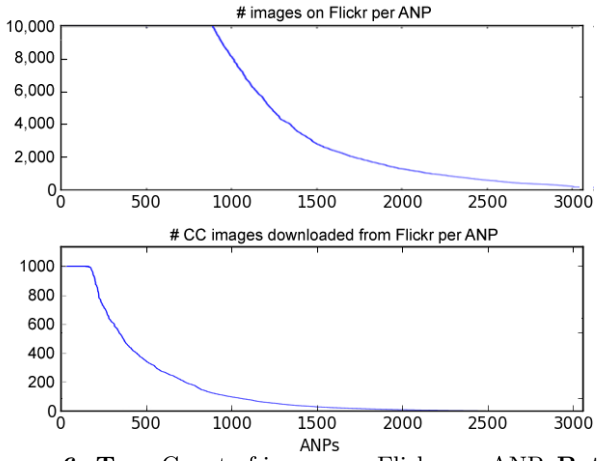

Figure 6: Top: Count of images on Flickr per ANP. **Bottom:** count of CC images downloaded per ANP (limited to max of 1000 images per ANP).

of adjectives is understandably difficult due to its abstract nature and high variability. Therefore, we propose adjective nouns combinations or **Adjective Noun Pairs (ANP)** to be the main semantic concept elements of the VSO. The advantage of using ANPs, as compared to nouns or adjectives only, is the feasibility of turning a neutral noun into a strong sentiment ANP. Such combined concepts also make the concepts more detectable, compared to adjectives only. The above described ANP structure shares certain similarity to the recent trend in computer vision and multimedia concept detection, i.e. *bi-concepts* [19] or TRECVID's *concept pairs* [25].

Candidate Selection: The set of all strong sentiment value adjectives and the set of all nouns are now used to form ANPs such as "beautiful flower" or "disgusting food". After ANP concepts are formed, an extra text analysis step is employed to avoid ANPs that correspond to named entities with meaning changed (e.g., "hot" + "dog" leads to a named entity instead of a generic concept). Obviously, during the construction of ANPs we also have to fuse the sentiment value of the adjective and the noun. This is done by applying a simple model to sum up the corresponding sentiment values $s(ANP) = s(adj) + s(noun)$ where the sentiment value $s(ANP)$ is between -2 and +2. Obviously, with this model we have to be careful with cases like "abused" being negative and "child" being positive forming the ANP "abused child", which reflects definitely a strong negative sentiment. We address this issue, by identifying ANPs that include an adjective and a noun with opposite sentiment values. We observed that in such cases the adjective usually has a stronger impact on the overall ANP sentiment than the noun and thus let the ANP inherits the sentiment value of the adjective.

Candidate Ranking: Those ANPs candidates (about 320k) are then ranked by their frequency on Flickr to remove meaningless or extremely rare constructions like e.g. "frightened hat" or "happy happiness". Having this ranked list of ANP frequencies (characterized by a long tail as seen in Fig. 6 (top)), we dismiss all ANPs with no images found on Flickr. This leads to a remaining set 47k ANP candidates. In this step we also eliminate cases where both, the singular and plural forms of an ANPs exists in the VSO. In such a case we take the more frequent one.

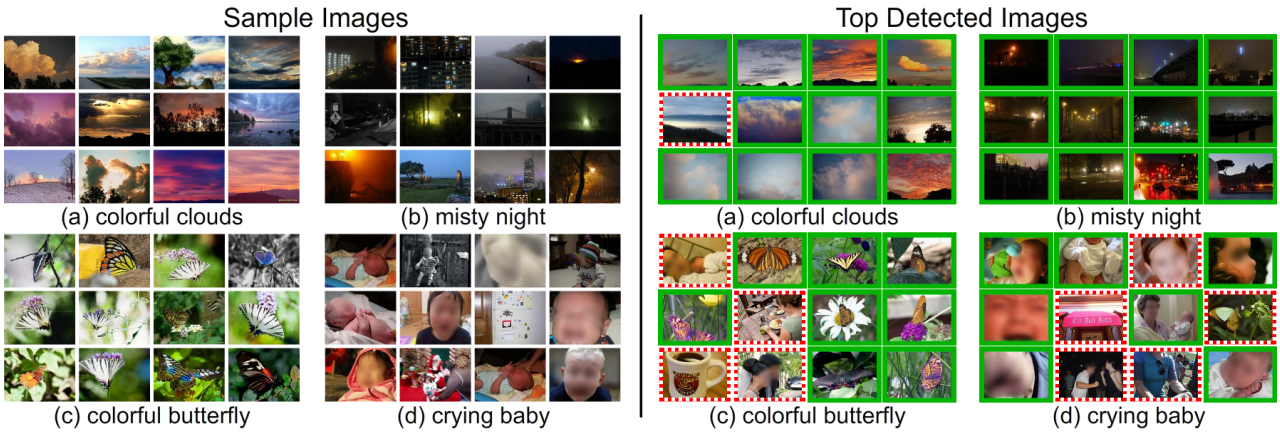

Figure 5: Left: Selected images for four sample ANPs, **(a)**,**(c)** reflecting a positive sentiment and **(b)**, **(d)**, a negative one. **Right:** top detected images by SentiBank ANPs with high detection accuracy (top) and low accuracy (bottom). Correct detections are surrounded by green and thick frames and incorrect ones by red and dashed frames. Faces in the images are blurred.

Table 2: Top 3 ANPs for basic emotions.

Emotion	Top ANPs
joy	happy smile, innocent smile, happy christmas
trust	christian faith, rich history, nutritious food
fear	dangerous road, scary spider, scary ghost
surprise	pleasant surprise, nice surprise, precious gift
sadness	sad goodbye, sad scene, sad eyes
disgust	nasty bugs, dirty feet, ugly bug
anger	angry bull, angry chicken, angry eyes
anticipation	magical garden, tame bird, curious bird

Ontology Sampling: The final step is to subsample the concepts in a diverse and balanced way and include those with a high frequency only. To avoid dominance by just a few popular adjectives, we partition candidate concepts into individual adjective sets and sample from each adjective a subset of ANPs. Further we only take ANPs if they have sufficient (currently > 125) images found on Flickr.

Link back to Emotions: We are interested in how the discovered ANPs are related to the basic emotions used in the very first retrieval step. Here, we measure the counts of images that have both the emotion term and the ANP string in their meta-data and normalize the resulting 24 dimension histogram to unit sum. This way a two-directional connection between an emotion and an ANP can be established. For example, the most dominant emotion for "happy smile" ANP is "joy" and for the emotion "disgust" the popular ANP is "nasty bugs". More examples can be seen in Table 2.

The final VSO contains more than 3,000 ANP concepts with 268 adjectives and their corresponding ANPs. Some of top ranked APNs are: "happy birthday", "beautiful flower", and "little girl" being positive and "dark night", "heavy rain", and "broken window" being the negative counterpart.

4.4 Flickr CC Dataset & Visualization Tool

An essential part of the VSO is its image dataset representing each ANP. The images are used for SentiBank detector training (Sec. 5). We used the Flickr API to retrieve and download Creative Common (CC) images for each ANP (limited to 1000 images by the API service) and include only images that contain the ANP string either in the title, tag or description of the image. With this we were able to download a sufficient amount of CC images for 1,553 of the 3,000 ANPs (in total about 500k images). The distribution of image count per ANP can be seen in Fig. 6 (bottom). Selected images of four example ANPs are show in Fig. 5 (left).

To help visualize the VSO and the associated dataset, we have developed two novel visualization techniques, one based on Wheel of Emotion (shown in Fig. 3) and the other the well-known TreeMap hierarchical visualization method (Fig. 8). The Emotion Wheel interface allows users to view and interact with the Plutchik 24 emotions directly and then zoom in to explore specific ANP concepts and associated images. The TreeMap interface offers a complementary way of navigating through different levels of VSO - emotion, adjective, noun, and ANPs. At each level, the map shows statistics and summaries of information from the level below. Interactive demos of these tools are available online[2].

5. SENTIBANK

Given the Visual Sentiment Ontology constructed above, we propose *SentiBank*, a novel visual sentiment analysis framework using the output of ANP detectors as a mid-level concept representation for each image. Its objective is to detect ANP concept presence and to characterize the sentiment reflected in visual content. To this end, we address several key issues, namely ANP label reliability, design of individual ANP detectors, detector performance, and coverage.

5.1 Reliability of ANP labels

It is known that web labels may not be reliable [7, 31]. Using Flickr tags directly as pseudo labels of ANPs might incur either *false positive*, i.e. an image is labeled by an ANP but actually does not show the ANP, or *false negative*, i.e. if an image is not labeled with an ANP it does not imply the ANP is not present in the image.

Dealing with Pseudo Labels: Considering the potential of *false positive*, we further evaluate the ANP labels by an Amazon Mechanical Turk (AMT) experiment. We ran-

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

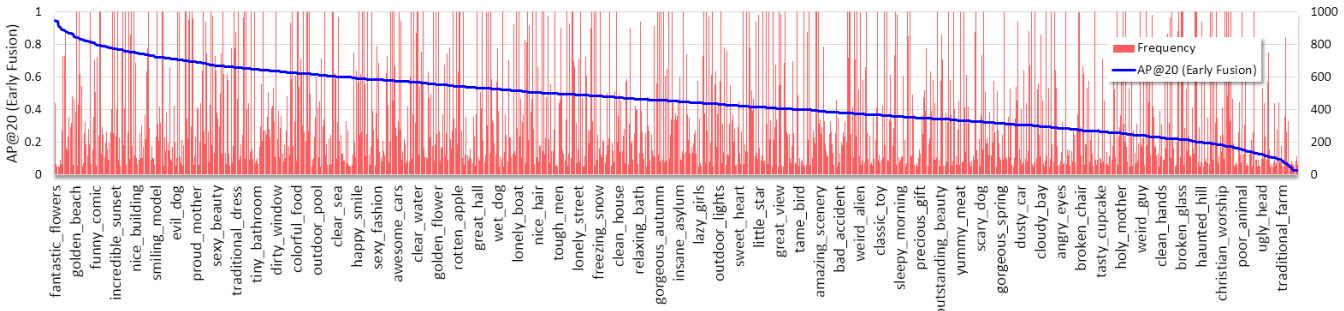

Figure 7: AP@20 (avg. over 5 runs of the reduced testsets) vs. frequency of 1,553 ANP detectors ranked by detector performance. Note only a subset of ANP names are shown due to space limit.

Figure 8: VSO visualization interface using Treemap.

domly sample images of 200 ANP concepts to manually validate their actual presence, namely using AMT crowdsource to check whether an image indeed contains the corresponding ANP. Each image label is validated by 3 Turkers and is treated as "correct" only if at least 2 Turkers agree that an image contains the given ANP label. Results of this experiment show that 97% of AMP image labels are actually "correct", which indicates that false positive is not a major issue.

Unfortunately, for the *false negative* issue, such a label validation procedure is prohibitive since it would require to fully label all images for all ANPs. This is also an open issue for existing crowdsourced visual recognition benchmarks such as ImageNet LISVRC2010-2012 and ObjectBank. In this work, we resolve this by randomly sampling positives of other ANPs (except those containing the same adjective or noun) when forming the negative set for each ANP class. This way we can minimize the probability of false negative, while avoiding the prohibitive task of labeling the entire dataset over every ANP concept.

Training and Testing Partitions: For training we sample 80% of pseudo positive images of each ANP and twice as many negative images using the subsampling scheme described above. For testing, we prepare two different testsets, denoted as the *full* and *reduced* testsets. Both use the remaining 20% of pseudo positive samples of a given ANP as positive test samples. But the negative samples are different - the full testset includes 20% pseudo positive samples from each of the other ANPs (except those with the same adjective or noun). This leads to a balanced training set and a large and diverse testset for individual detector performance evaluation. However, the prior of the positive in each testset is very low, only about 1/1,553. The reduced testset, intended for fast implementations and balanced test sample distributions, includes much less negative samples - the number of negative test samples for each ANP is just twice as many the positive samples. To avoid testset bias, we also generate 5 runs of the reduced testset, each of which includes different negative samples while keeping the positive samples fixed. We will use performance average over these 5 runs in later experiments (Fig. 7, 9, 10) in the paper and leave the performance details over the full testset in supplementary materials.

5.2 Training ANP Detectors

With the partition of training and test data for each ANP, detectors for individual ANPs can be readily trained.

Visual Feature Design: Following the feature design for state-of-the-art visual classification systems such as ObjectBank [38], we first include generic visual features: a 3×256 dimension Color Histogram extracted from the RGB color channels, a 512 dimension GIST descriptor [40] that has been shown to be useful for detecting scenes like "beautiful landscape", a 53 dimension Local Binary Pattern (LBP) descriptor suitable for detecting textures and faces, a Bag-of-Words quantized descriptor using a $1,000$ word dictionary with a 2-layer spatial pyramid and max pooling, and finally a $2,000$ dimensional attribute [39] useful for characterizing abstract ANPs. Additional features specialized for detecting objects, faces, or aesthetics will be presented later in Sec. 5.5.

ANP Detector Training: Due to the large amount of ANPs in the ontology, we employ Linear SVMs to train ANP detectors in order to ensure high efficiency. Parameter tuning of SVM was performed by cross-validation optimizing Average Precision at rank 20 (AP@20), a performance measure focusing on the accuracy of the top ranked samples. Detector performance was also measured by the Area Under Curve (AUC), estimating the probability of ranking a random positive sample higher than a random negative one. Finally, our third measure is F-Score, describing the harmonic mean between precision and recall. All three measures are standard metrics for detector evaluation.

Detector performance using various features can be seen in Fig. 9. Here we can observe a clear dominance by the attribute features followed by Bag-of-Words (BOW). Considering feature fusion, both early and late fusions are evaluated. The former refers to merging and normalizing different feature vectors into a single vector. The latter refers to the fusion of detector scores after classification. We have evaluated different fusion methods including *Early*

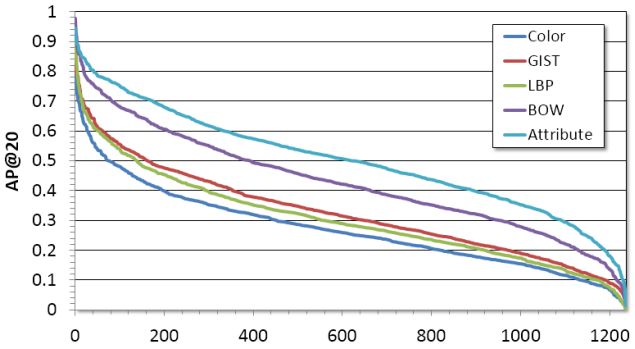

Figure 9: Comparison of ANP detectors using different features. Performance computed by avg. over 5 runs of the reduced testsets.

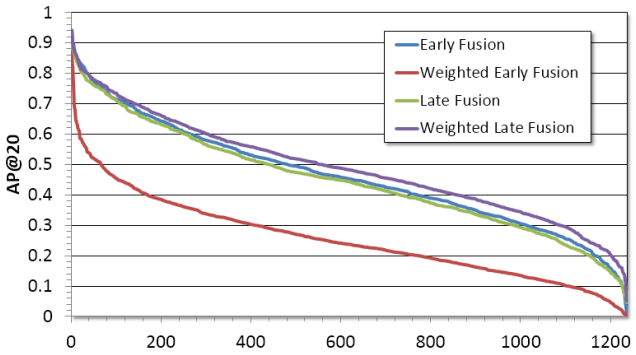

Figure 10: AP@20 of ANP detectors using different fusion approaches. Performance computed by avg. over 5 runs of the reduced testsets.

Fusion, Weighted Early Fusion, Late Fusion, and Weighted Late Fusion[3]. As shown in Fig. 10, Weighted Late Fusion out-performs other fusion schemes by a small margin, while the performance of early and late fusion is quite close. For implementation simplicity, we use the early fusion approach in the released SentiBank detectors.

5.3 Ontology Structure

An ontology consists of not only concepts but also relations that can be used for browsing and reasoning about concepts within a domain. To construct such ontological structures, we have conducted an interactive process in which multiple subjects were asked to combine the 1,200 ANP concepts in SentiBank into distinct groups, so that each group shares coherent semantics among group members. Consensus among subjects was reached through result comparison and discussion. We perform grouping processes for adjectives and nouns in the VSO separately. Such separate processes allow exploitation of relations unique for adjectives or nouns. For example, a hierarchical structure for nouns (a total of about 520) was found to include six levels and 15 nodes at the top level, such as person, place, object, food, and abstract concept. The adjectives (a total of about 260) were grouped to two levels with 6 nodes at the top level. The standard hyponym-hypernym ("is-a") relations were found in the noun hierarchy, while special relations like exclusive ("sad" vs. "happy") and strength order ("nice" vs. "great" vs. "awesome") were found among adjectives. We also found special conditions for combining adjectives and nouns into ANPs. For example, some adjectives are only applicable to certain noun groups, such as people, place, food, and objects. In other words, adjective groups can be considered as facets of specific types of nouns - a practice often used in ontology construction.

Comparison of the constructed noun taxonomy and the well-known ImageNet shows 59% of the VSO nouns being mapped to ImageNet synsets. This leads to 41% of VSO nouns not covered by ImageNet, although they can still be found in WordNet. These concepts unique to VSO are mainly related to abstract concepts such as "violence" or "religion", which reflect strong emotions or sentiments. This confirm the unique focus on emotions and sentiments in the concept discovery process of VSO, as described earlier in this section. Due to the space limit, we refer for more details to the technical report [3]

[3]weights are also tuned by cross-validation

5.4 ANP Detectability and Coverage

ANP Detectability Overview: An important step in building SentiBank is to select only ANPs with reasonable detection accuracy. To this end, we rank the ANP detectors based on the previously described performance measures such as F-Score, AUC, or AP@20. It is worth noting that using different performance metrics only slightly affect the relative orders of the ANP detectors. At the end, we choose $1,200$ ANP detectors, all of which have non-zero AP@20 and most have F-score greater than 0.6, when evaluated over the reduced testset. It's interesting to see (as shown in Fig. 7) that there is no correlation between the detectability of an ANP and its occurrence frequency. Instead, the difficulty in detecting an ANP depends on the content diversity and the abstract level of concept. We show some examples of the best and worst ANPs based on AP@20 in Figure 5.

Generalizability and Coverage: Are the selected SentiBank ANP detectors generalizable to different data domains? It would be useful to check the performance degradation if we apply SentiBank detectors to new data domains, such as TRECVID or PASCAL VOC. In the next section, we answer this question indirectly by predicting the sentiment of image tweets which have different characteristics than the Flickr images. In addition, as discussed in Sec. 5.3, the scope of the constructed VSO is broad, covering many categories in other well-known visual ontologies, such as LSCOM [23] and ImageNet [7]. This will help to address a common out-of-vocabulary problem when applying small-sized detectors to general domains.

5.5 Special Visual Features

Other than the generic features, we also test several special features for training the SentiBank detectors. First, since many of the ANP concepts are associated with objects, we utilize object detection techniques to localize the concept within an image. We choose 210 ANPs that are associated with detectable objects such as people, dog, or cars. We apply the object detection tools from [38] and combine multi-scale detection results to form a spatial map for constraining the image region from which the visual features described in Sec. 5.2 are extracted. Another option is to take the object detection response scores directly as features. Secondly, we evaluate facial features on 99 ANPs with nouns like face, smile, tears, etc. These include the detected face count, relative face size, relative facial marker position

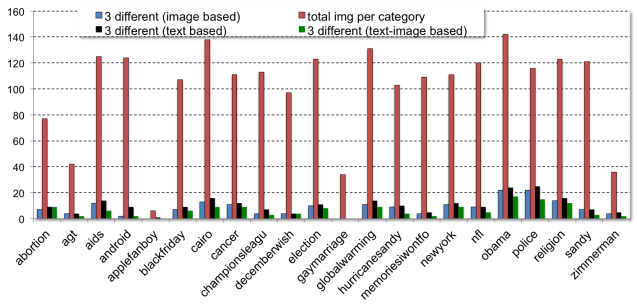

Figure 11: The volumes and label disagreements for different hashtags. For each hashtag, the total number of images is shown, in addition to the number of images receiving complete disagreement among Turkers (i.e., 3 different sentiment labels: positive, negative and neural), while labeling is done using text only, image only, and joint image-text combined.

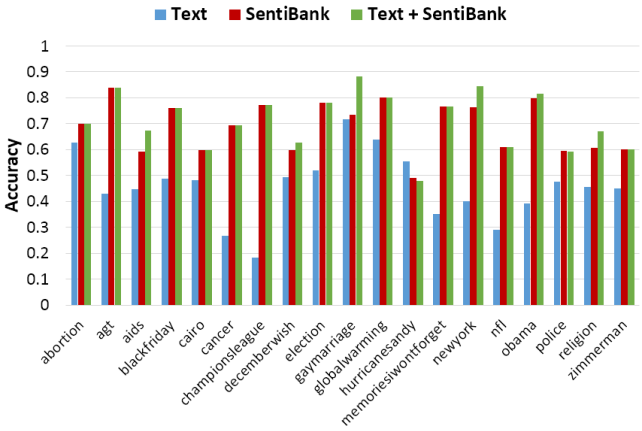

Figure 12: Phototweet sentiment prediction accuracy over different hashtags by using text only (SentiStrength), visual only (SentiBank), and combination. Accuracy averaged over 5 runs.

and Haar features at the markers. Thirdly, we test aesthetics related features on all of the ANPs. These features from [41] include dark channel and luminosity feature, sharpness, symmetry, low depth of field, white balance, etc. The above three groups of features increase the mean AP@20 score of selected ANP sets by 10.5%, 13.5% and 9.0% (relative gains) respectively on the reduced testset and the mean AP@100 by 39.1%, 15.8% and 30.7% on the full testset. Based on the above comparisons, we conjecture that the generic features offer a competitive solution for detecting ANP visual sentiment concepts, while special features offer great potential for further improvements.

6. SENTIBANK APPLICATIONS

In this section we demonstrate potential applications of SentiBank. Since our initial motivation to construct the V-SO and create SentiBank is to capture the sentiment reflected in visual content, our first application focuses on sentiment prediction in image tweets[4]. Additionally, we evaluate

[4] Sentiment is often measured in a particular domain such as movies, food, or politics. For this study we adopt a generic definition across different tweet topics without topic-specific differentiation

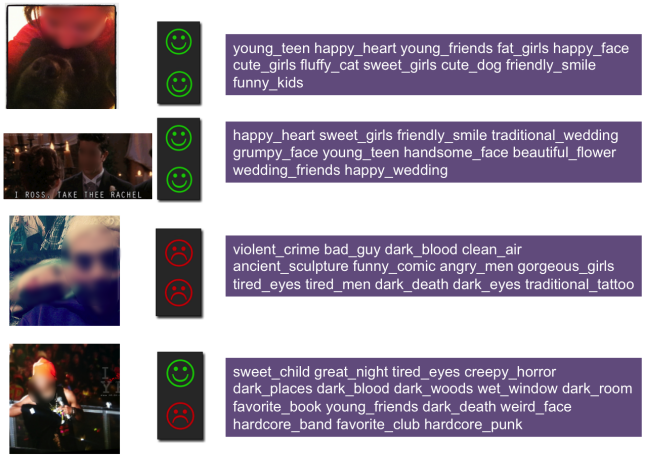

Figure 13: Sentiment prediction results using SentiBank as features (top icon in each box in the middle: ground truth sentiment, bottom icon: predicted sentiment). On the right the top responding ANPs found in the prediction.

SentiBank against a well-known emotion dataset of art photos [21].

6.1 Sentiment Prediction

For sentiment prediction, state-of-the-art approaches typically rely on text-based tools such as SentiWordNet [9] or SentiStrength [30]. However, due to the length restriction of 140 characters in tweets, such approach is limited and often unable to correctly discern the sentiment of the text content. To overcome this issue, we use the proposed visual sentiment analysis tool, SentiBank, to complement and augment the text features in sentiment prediction.

Hashtag Selection: We collect tweets containing images from the PeopleBrowsr API using the following popular hashtags: *Human:* #abortion, #religion, #cancer, #aids, #memoriesiwontforget, *Social:* #gaymarriage, #police, #nuclearpower, #globalwarming, *Event:* #election, #hurricanesandy, #occupywallstreet, #agt (america got talent), #nfl, #blackfriday, #championsleague, #decemberwish, *People:* #obama, #zimmerman, *Location:* #cairo, #newyork, *Technology:* #android, #iphonefan, #kodak, #androidgame, #applefan. The resulting dataset consists of 20 to 150 images per hashtag, crawled during November 2012.

Ground Truth Labeling: To obtain sentiment ground truth for the collected image tweets, we conduct three labeling runs using AMT, namely *image-based*, *text-based*, and joint *text-image based*. They correspond to image only inspection, text only inspection, and full inspection using both image and text contained in the tweet. Each run is assigned to 3 randomly assigned Turkers, but no Turkers are asked to annotate the same tweet under different modality settings. Fig. 11 shows the labeling statistics, where we define an image as "agreed", if more than 2 Turkers assign the same label (either positive, negative or neutral). From the results, we clearly see that, joint *text-image based* labels are the most consistent ones, following by *image-based* labels and then the *text-based* labels. This indicates the limitation of text-based sentiment analysis for Twitter and highlights the potential for a holistic sentiment analysis using both the image and text analysis. At the end, we include only the image tweets that receive unanimously agreed labels among three Turkers from the image-text annotation as the final benchmark set.

Table 3: Tweet Sentiment Prediction Accuracy (Visual Based Methods)

	Linear SVM	Logistic Regr.
Low-level Features	0.55	0.57
SentiBank	0.67	**0.70**

It includes 470 positive tweets and 133 negative tweets over 21 hashtags, among which 19 hashtags each with more than 10 samples are shown in Fig. 12.

Visual-based Classifier: As mentioned before, SentiBank serves as an expressive mid-level representation of visual concept. For each image, SentiBank provides a 1,200 dimension ANP response, which is used as an input feature for the sentiment classification. Here, we employ linear classifiers such as Linear SVM and Logistic Regression. To this end, we are not only aiming to predict the sentiment being reflected in images but also to provide an explanation of the prediction result. This is achieved by providing a list of top responding ANP detectors in addition to the sentiment prediction label.

We first compare the proposed SentiBank mid-level representation with low-level features, using two different classification models, LinearSVM and Logistic Regression. For low-level features, we use the same set as those described in Sec. 5.2 (color histogram, GIST, LBP, BoW, and attributes). Prediction accuracy is shown in Table 3 - confirming the significant performance improvement (more than 20% relatively) achieved by the SentiBank features. The logistic regression model is also found to be better than Linear SVM.

In a separate experiment, we have also confirmed the superiority of SentiBank using Adjective-Noun Pairs over the concepts of nouns only, adjectives only, or their union. This again verifies the advantage of using ANP concepts to build out SentiBank representations.

Text Based Classification: We adopt two text-based sentiment predictors:

`(1) Naive Bayesian text-based Sentiment Classifier:`

$$Score = \frac{1}{M}\sum_{m=1}^{M} Frequency_m \times Score_m \quad (1)$$

in which *Score* is the sentiment prediction score normalized to [-1,1], M the number of unique words after stemming and stop words removal, $Frequency_m$ the frequency of word m, and $Score_m$ is the individual sentiment score of word m obtained from SentiStrength.

`(2) SentiStrengh API:` We directly use the sentiment prediction by the publicly available SentiStrength API[5] to compute the sentiment score for the entire tweet text. Our experiment has shown that SentiStrenght API prediction accuracy based on the entire tweet text is higher than the one combining scores of individual words using the Naive Bayesian method.

Joint text-image Classification Performance: Finally, we compare the accuracy using text only (SentiStrength), visual only (SentiBank), and their combination. From Tables 3 and 4, we find visual based methods using SentiBank concepts are significantly better than the text only (70% vs. 43%). By further analyzing the results, we find most of the text contents in the tweets are short and neutral, explaining the low accuracy of text-based methods in predicting the sentiment. In such cases, the sentiment values of the visual content predicted by the SentiBank-based classifiers play a much more important role in predicting the overall sentiment of the tweet.

[5]http://sentistrength.wlv.ac.uk/

Table 4: Comparison of Tweet Sentiment Prediction Accuracy

text only (SentiStrength)	text + visual (SentiBank LinearSVM)	text + visual (SentiBank Logistic Regr.)
0.43	0.68	**0.72**

Fig. 12 shows comparison of sentiment prediction accuracy for each individual hashtag. Here, we find the visual-based approach using SentiBank concepts consistently outperforms the text-based method using SentiStrength API, except only one hashtag ("hurricanesandy"). It's also very encouraging to see combining text and SentiBank features further improves accuracy for several hashtags, despite the low accuracy of the text-based method.

Results of a few sample images can be seen in Fig. 13. Here, SentiBank's capability in explaining predictions is illustrated by showing a list of top ANPs detected in each test image.

6.2 Emotion Classification

Although our initial motivation was to predict sentiment reflected in images, an evaluation of the proposed method in emotion classification (similar to the task in [21]) might be of interest. Especially since our VSO construction process starts with web search using emotions as search terms. The dataset is based on ArtPhotos retrieved from DeviantArt.com and contains 807 images from 8 emotion categories. Such evaluation poses a set of challenges such as domain change. SentiBank is trained on a different set of images than the testset. Further, our emotion categories are slightly different, not mentioning our focus is on a framework with generic visual features rather than the specialized affective features used in [21]. We follow a similar process to select SentiBank ANPs as features for each emotion category, combined with the Naive Bayesian classifier. Results are reported in Fig. 14. Even in such a challenging setting, SentiBank compares well to [21] when using the same classification model (Naive Bayesian) and even slightly outperforms the best results in [21] when using the Logistic Regression model. This demonstrates the potential of SentiBank for applications in different domains.

7. DISCUSSION

In this paper we have presented an approach towards the prediction of sentiment reflected in visual content. To reach this goal we propose a systematic, data-driven methodology to construct a large-scale sentiment ontology built upon psychology and web crawled folksonomies. Further, we present SentiBank, a concept detector library based on the constructed ontology to establish a novel mid-level representation for bridging the *affective gap*. Finally, we release the concept ontology, dataset, the detector library, and the benchmark for tweet sentiment analysis to stimulate research in this direction.

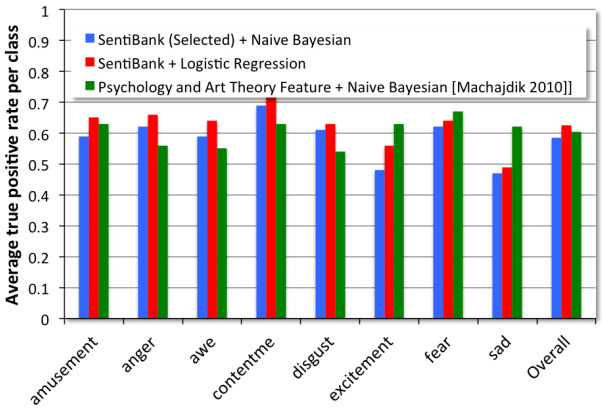

Figure 14: Emotion classification performance.

Several exciting directions are open for investigation. First, the cultural influence on expressing sentiment and emotion is of interest. Further, as shown earlier the application of special features such as aesthetic features used in [21] and face expression features offers interesting potential for further improvement. Additionally, other applications such as advertising, games, virtual reality are promising when the cross-domain performance of the detectors is studied in more depth. Another important task is the extension of VSO and SentiBank to video applications.

8. ACKNOWLEDGEMENT

Research was sponsored in part by the U.S. Defense Advanced Research Projects Agency (DARPA) under the Social Media in Strategic Communication (SMISC) program, Agreement Number W911NF-12-C-0028. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Defense Advanced Research Projects Agency or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

9. REFERENCES

[1] H. Aradhye, G. Toderici, and J. Yagnik. Video2Text: Learning to Annotate Video Content. *Internet Multimedia Mining*, 2009.

[2] D. Borth, A. Ulges, and T.M. Breuel. Lookapp - Interactive Construction of web-based Concept Detectors. *ICMR*, 2011.

[3] D. Borth and S-F. Chang. Constructing Structures and Relations in SentiBank Visual Sentiment Ontology. *Technical Report #CUCS-020-13, Columbia University, Computer Science Dep.*, 2013.

[4] E. Dan-Glauser et al. The Geneva Affective Picture Database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior Research Methods*, 2011.

[5] Charles Darwin. *The Expression of the Emotions in Man and Animals*. Oxford University Press, USA, 1872 / 1998.

[6] R. Datta, D. Joshi, J. Li, and J. Wang. Studying Aesthetics in Photographic Images using a Computational Approach. *ECCV*, 2006.

[7] J. Deng et al. ImageNet: A Large-Scale Hierarchical Image Database. *CVPR*, 2009.

[8] P. Ekman et al. Facial Expression and Emotion. *American Psychologist*, 48:384-384, 1993.

[9] A. Esuli and F. Sebastiani. SentiWordnet: A publicly available Lexical Resource for Opinion Mining. *LREC*, 2006.

[10] M. Everingham, et al. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. of Computer Vision*, 88(2):303-338, 2010.

[11] A. Hanjalic, C. Kofler, and M. Larson. Intent and its Discontents: the User at the Wheel of the Online Video Search Engine. *ACM MM*, 2012.

[12] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an Image Memorable? *CVPR*, 2011.

[13] 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. *ACM MM*, 2012.

[14] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. Loui. Consumer Video Understand.: Benchmark Database and an Eval. of Human and Machine Performance. *ICMR*, 2011.

[15] D. Joshi, R. Datta, E. Fedorovskaya, Q. Luong, J. Wang, J. Li, and J. Luo. Aesthetics and Emotions in Images. *Signal Processing Magazine*, 28(5):94-115, 2011.

[16] L. Kennedy, S.-F. Chang, and I. Kozintsev. To Search or to Label?: Predicting the Performance of Search-based Automatic Image Classifiers. *MIR Workshop*, 2006.

[17] P. Lang, M. Bradley, and B. Cuthbert. International Affective Picture System (IAPS): Technical Manual and Affective Ratings, 1999.

[18] B. Li, et al. Scaring or Pleasing: Exploit Emotional Impact of an Image. *ACM MM*, 2012.

[19] X. Li, C. Snoek, M. Worring, and A. Smeulders. Harvesting Social Images for Bi-Concept Search. *IEEE Transactions on Multimedia*, 14(4):1091-1104, 2012.

[20] N. Codella et al. IBM Research and Columbia University TRECVID-2011 Multimedia Event Detection (med) System. *NIST TRECVID Workshop*, 2011.

[21] J. Machajdik and A. Hanbury. Affective Image Classification using Features inspired by Psychology and Art Theory. *ACM MM*, 2010.

[22] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the Aesthetic Quality of Photographs using Generic Image Descriptors. *ICCV*, 2011.

[23] M. Naphade, J. Smith, J. Tesic, S. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-Scale Concept Ontology for Multimedia. *IEEE MultiMedia*, 13(3):86-91, 2006.

[24] C. Osgood, G. Suci, and P. Tannenbaum. *The Measurement of Meaning*, volume 47. University of Illinois Press, 1957.

[25] P. Over et al. Trecvid 2012 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. *TRECVID Workshop*, 2012.

[26] B. Pang and L. Lee. Opinion Mining and Sentiment Analysis. *Information Retrieval*, 2(1-2):1-135, 2008.

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

[28] C. Snoek and M. Worring. Concept-based Video Retrieval. *Foundations and Trends in Inf. Retrieval*, 4(2), 2009.

[29] S. Strassel et al. Creating HAVIC: Heterogeneous Audio Visual Internet Collection. *LREC*, 2012.

[30] M. Thelwall et al. Sentiment Strength Detection in Short Informal Text. *J. of the American Soc. for Information Science and Tech.*, 61(12):2544-2558, 2010.

[31] A. Ulges, C. Schulze, M. Koch, and T. Breuel. Learning Automatic Concept Detectors from Online Video. *Journal on Comp. Vis. Img. Underst.*, 114(4):429-438, 2010.

[32] V. Vonikakis and S. Winkler. Emotion-based Sequence of Family Photos. *ACM MM*, 2012.

[33] W. Wang and Q. He. A Survey on Emotional Semantic Image Retrieval. *IEEE ICIP*, 2008.

[34] X. Wang, J. Jia, P. Hu, S. Wu, J. Tang, and L. Cai. Understanding the Emotional Impact of Images. *ACM MM*, 2012.

[35] T. Wilson et al. Recognizing Contextual Polarity in phrase-level Sentiment Analysis. *HLT/EMNLP*, 2005.

[36] V. Yanulevskaya et al. In the Eye of the Beholder: Employing Statistical Analysis and Eye Tracking for Analyzing Abstract Paintings. *ACM MM*, 2012.

[37] V. Yanulevskaya et al. Emotional Valence Categorization using Holistic Image Features. *IEEE ICIP*, 2008.

[38] Li et al. ObjectBank: A high-level Image Rep. for Scene Classification and Semantic Feature Sparsification. *NIPS*, 2010.

[39] F. Yu, et al Designing Category-level Attributes for Discriminative Visual Recognition. *CVPR*, 2013.

[40] A. Olivai and A. Torralba. Modeling the Shape of the Scene: a Holistic Representation of the Spatial Envelope. *Int. J. of Computer Vision*, 42(3):145-175, 2001.

[41] S. Bhattacharya and B. Nojavanasghari and T. Chen and D. Liu and S.-F. Chang and M. Shah. Towards a Comprehensive Computational Model for Aesthetic Assessment of Videos. *ACM MM*, Grand Challenge, 2013.