Object-Based Visual Sentiment Concept Analysis and Application

Tao Chen, Felix X. Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, Shih-Fu Chang

Columbia University, USA
{tc2650,xy2154,jc3960,yc2776,sc250}@columbia.edu yanying@gmail.com

ABSTRACT

This paper studies the problem of modeling object-based visual concepts such as "crazy car" and "shy dog" with a goal to extract emotion related information from social multimedia content. We focus on detecting such adjective-noun pairs because of their strong co-occurrence relation with image tags about emotions. This problem is very challenging due to the highly subjective nature of the adjectives like "crazy" and "shy" and the ambiguity associated with the annotations. However, associating adjectives with concrete physical nouns makes the combined visual concepts more detectable and tractable. We propose a hierarchical system to handle the concept classification in an object specific manner and decompose the hard problem into object localization and sentiment related concept modeling. In order to resolve the ambiguity of concepts we propose a novel classification approach by modeling the concept similarity, leveraging on online commonsense knowledgebase. The proposed framework also allows us to interpret the classifiers by discovering discriminative features. The comparisons between our method and several baselines show great improvement in classification performance. We further demonstrate the power of the proposed system with a few novel applications such as sentiment-aware music slide shows of personal albums.

Categories and Subject Descriptors

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

Keywords

social multimedia; visual sentiment; affective computing

1. INTRODUCTION

The explosive growth of social media and the ever-growing volume of online visual content have greatly motivated the research on large-scale social multimedia analysis. Among these research efforts, understanding the emotion and sentiment in visual media content shows its emerging importance. Images and videos embedding strong sentiments can strengthen the opinion conveyed in the content and influence the public viewer with viral effects. Understanding sentiment expressed in visual content will greatly benefit social media communication and enable broad applications in education, advertisement and entertainment.

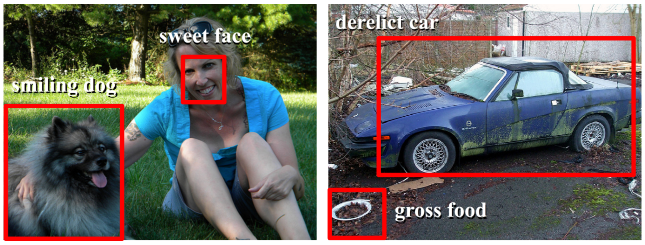


Figure 1: We propose a novel framework to detect visual concepts corresponding to adjective-noun pairs that statistically correlate with sentiments or emotions expressed in images. The proposed framework focuses on detection of objects and associated attributes. (photos from throughpan*flashing@Flickr and allenthepostman@Flickr)

While the modeling of generic visual concepts (nouns) such as "sky" and "dog" has been studied extensively in computer vision, modeling adjectives correlated with visual sentiments like "amazing" and "shy" directly is very difficult, if not impossible, due to the big "affective gap" between the low-level visual features and the high-level sentiment. Therefore, we follow a more tractable approach proposed by Borth et al. [2] which models sentiment related visual concepts as a mid-level representation to fill the gap. Those concepts are Adjective Noun Pairs (ANPs), such as "happy dog" and "beautiful sky", which combine the sentimental strength of adjectives and detectability of nouns. Though these ANP concepts do not directly express emotions or sentiments, they were discovered based on strong co-occurrence relationships with emotion tags of web photos, and thus are useful as effective statistical cues for detecting emotions depicted in the images. In this work, we systemically study the modeling of these sentiment related visual concepts (termed visual sentiment concepts in this paper) by addressing two important challenges:

Many concepts are object based and need to be localized. These include popular content in social media

MM'14, November 3–7, 2014, Orlando, Florida, USA.
Copyright 2014 ACM 978-1-4503-3063-3/14/11 ...$15.00.
http://dx.doi.org/10.1145/2647868.2654935.

such as “face”, “dog” and “car”. Modeling concepts using the features extracted from the whole image as in [2] will lose the specific characteristics of the objects and unavoidably encompass noise from the background. However, it is very difficult and time consuming to localize each concept because of the lack of bounding box annotation of the ANPs, and the prohibitive computational complexity required in training the local concept detectors.

The ambiguity of the visual sentiment annotation. The existing annotation of visual sentiment concepts are highly ambiguous and they are often hard to distinguish since semantically related adjectives often have similar visual reflexes in images. For example, “cute dog” and “adorable dog”. Traditional multiclass classification using hard labels will not work. In contrast, the nouns representing objects are much less ambiguous.

Therefore a natural and efficient approach is to model the object (noun) first, and then model the object based sentiment attributes (adjective) with a specific aim to tackle the ambiguity. The requirement of the hierarchical scheme can be further justified by several observations. First, visual sentiment concepts are inconsistent across objects. For example, the visual characteristics that indicate a girl is beautiful could be very different from those indicate a beautiful sunset scene. Second, the features that effectively detect sentiment concepts is usually different from those effective for object detection. For example, Machajdik et al. [20] proposed to use features inspired by psychology and art to model affect in images. Third, to extract the features such as the composition of the image and the color distribution inside the object, we also need to first localize the object. Moreover, instead of modeling visual sentiment concepts directly with an object detection method such as Deformable Part Model (DPM) [10], the hierarchical method allows us to apply special approaches like soft weight SVM to address the concept overlap among multiple adjectives of the same noun. Traditional object detection methods do not allow this.

In this work we first build object detection models to recognize six frequent objects in social media including “car”, “dog”, “dress”, “face”, “flower” and “food” to model the 135 ANP concepts associated with those nouns. We then build object-based sentiment concept models, based on the detected objects, with location and composition information. In order to resolve the visual ambiguity among the adjectives, we propose a novel method exploiting rich semantic similarity information computed from an online commonsense knowledgebase, conceptNet [19]. The method models the ambiguity by two kinds of visual similarity, namely attributive and proportional similarity. The implementation is based on a fusion of weighted SVM and a recently proposed machine learning model, called proportion-SVM or pSVM. Several visual sentiment concepts detected by our system are shown in Figure 1. The very significant performance improvement (up to 50% relatively) compared to the previous work proves the success of our model. To understand how these high-level models work, we interpret the classifiers by discovering the discriminative features for each visual sentiment concept classifier.

Additionally, we demonstrate an innovative application that benefit from our model. The application is to automatically generate sentiment-aware personalized music slide show. We also show our model can improve the existing affective application such as an image commenting assistant. The user studies and comparisons show that our hierarchical classification greatly improves the quality of these applications.

To the best of our knowledge, this is the first work systematically studying the modeling of sentiment concepts related to objects. Our technical contributions include a complete hierarchical system for object based visual sentiment modeling in which we i) exploit an online commonsense knowledgebase to model the visual sentiment relations from ambiguous and incomplete Internet labels, ii) propose a unique classification model to handle attributive and proportional similarity between visual sentiment concepts. We prove the sentiment modeling technique is a highly useful tool for solving various real-world problems by demonstrating novel applications and improving existing ones.

The rest of the paper is organized as follows. Section 2 summarizes the related work. Section 3 includes an overview of the problem and our solution. In Section 4 and Section 5, we present the proposed object-based visual sentiment models in detail. Section 6 evaluates our results by various comparisons and further explains and justifies our hierarchical classification framework. Section 7 shows applications that greatly benefit from our new visual sentiment concept detection.

2. RELATED WORK

2.1 Modeling Sentiment

Most work on sentiment analysis so far has been based on textual information [31, 8, 28]. Sentiment models have been demonstrated to be useful in various applications including human behavior prediction [8], business [23], and political science [29].

Compared to text-based sentiment analysis, modeling sentiment based on images is much less studied. Perhaps the most relevant work is [2], which proposed to *design* a large-scale visual sentiment ontology based on Adjective-Noun Pairs (the sentiment modeling is then based on one-vs-all SVMs). Our work is clearly distinct from [2], as we focus on the fundamental understanding of visual sentiment concepts, rather than ontology construction and basic detector training. Such understanding is important for identifying successful components and open issues in designing features and detection models. This work is also necessary as the performance of the modeling has been shown to be critical for various applications of visual sentiment analysis [2, 3].

2.2 Modeling Visual Concepts

Concept modeling has been widely studied in multimedia [22, 27], and computer vision (often referred as “attributes”) [11]. The concepts being modeled are mostly objects [27], scenes [24], or activities [12]. There is work trying to solve the “fine grained recognition” task, where the categories are usually organized in a hierarchical structure. [6, 7, 5]. There is also work trying to model “non-conventional” concepts or properties of the images, such as image aesthetic and quality [17, 21], memorability [15], interestingness [15], and affection/emotions [20, 30, 16, 20, 30, 32]. Our work is different from the above as our objective is to model mid-level sentiment concepts, e.g., “beautiful flower”, and

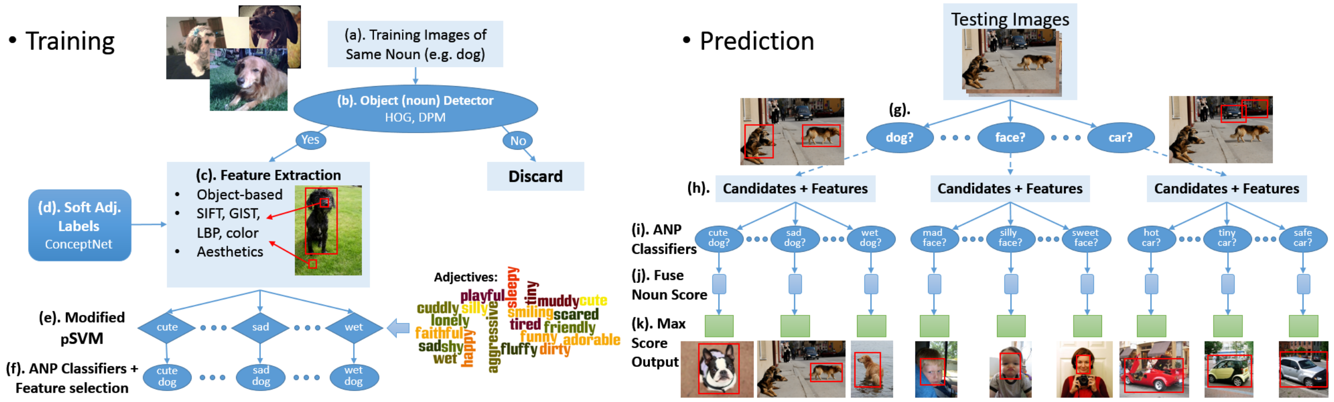
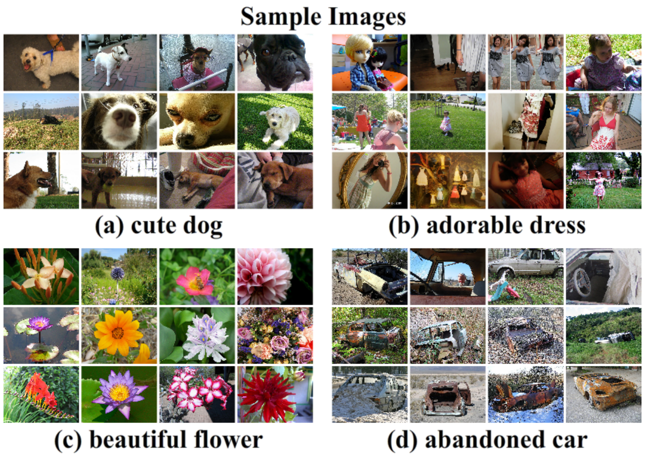


Figure 2: The pipeline of our hierarchical system (details in Section 3).

"happy dog", which have been shown promising in supporting several exciting applications [2, 3], but have not been analyzed with sufficient rigors and robustness.

We propose to use a hierarchical approach to model object-based sentiment concepts by firstly localizing the objects, and then modeling the object-specific adjectives. An alternative approach is to train a detection model for each object-based sentiment concept separately. One piece of related work is [25], where a DPM model is trained for each "visual phrase" such as "person riding a horse", and "dog laying on sofa". Different from the a visual phrase, which is a combination of two objects, an object-based sentiment concept is an adjective noun pair. As mentioned in the introduction, the hierarchical approach is necessary due to the ambiguity of the ANP labels and efficiency considerations.



(a) cute dog

(b) adorable dress

(c) beautiful flower

(d) abandoned car

Figure 3: Example images with ANP pseudo label.

3. OVERVIEW

Our goal is to detect visual sentiment concepts from images. Following [2], visual sentiment concepts are defined as Adjective Noun Pairs (ANPs), such as "beautiful girl" and "happy dog". In [2], more than 3,000 ANPs were discovered by a data-driven method and form the Visual Sentiment Ontology (VSO). Image based classifiers were trained for these ANPs with standard SVM. 1,200 ANPs among VSO were considered detectable based on the classifier performance and their classifiers are known as SentiBank. The 1,200 dimensional detector response of SentiBank has been proven effective as mid-level features in tasks such as predicting the sentiments of photo tweets.

However we find the individual concept detector of SentiBank less reliable and interpretable, especially for object-based concepts such as "adorable dress" and "abandoned car". The ANPs containing the same noun are often confused with each other. This drawback limits the application and reliability of SentiBank as mentioned in [3]. This is due to the fact that the difference of the sentiment concepts of these objects are usually reflected in the subtle difference *within* the image region of the object in addition to the context in the surrounding background. Because the classifiers trained in SentiBank use features extracted from the whole image without considering object localization, the inability to separate foreground and background features as expected limits the detection accuracy and interpretability of ANP concept classifiers. To address this problem, we propose a hierarchical classification system that detects and localizes the object first, and then classify the sentiment concepts.

3.1 Data Collection

Our data is from the VSO image dataset mentioned in [2]. For each ANP, it contains URLs of about 500 Flickr images, which are tagged with the ANP by the image uploader. Note that as the tags may not be fully reliable, such labels are considered as pseudo labels, rather than the ground truth. Figure 3 shows example images of several ANPs. Without losing generality, we focus on the aforementioned "detectable" ANPs and our experiment in this paper is limited to the six most popular nouns, namely "car", "dog", "dress", "face", "flower" and "food". These nouns are not only frequently tagged in the social multimedia, but also associated with diverse adjectives to form a large set of ANPs. In addition, if we can successfully model such ANPs, we can easily extend the proposed approach to cover more nouns. The total number of ANPs associated with the 6 nouns is 135, as shown in Table 1. We download the images of these ANPs as our training and test data. The data partition will be discussed in Section 6.1.

Table 1: ANPs used in our experiment (categorized by nouns).

Adjective	Noun
abandoned, amazing, awesome, bad, broken, classic, clean, crazy, damaged, derelict, dirty, expensive, famous, fancy, fantastic, filthy, hot, lonely, powerful, safe, shiny, super, tiny, ugly	car
adorable, aggressive, cuddly, cute, dirty, faithful, fluffy, friendly, funny, happy, lonely, muddy, playful, sad, scared, shy, silly, sleepy, smiling, tiny, tired, wet	dog
adorable, colorful, cute, elegant, fancy, favorite, gorgeous, pretty, sexy, stunning, traditional	dress
adorable, angry, attractive, bright, chubby, clean, crazy, crying, cute, dirty, dumb, excited, funny, grumpy, handsome, hilarious, innocent, mad, pensive, pretty, scared, scary, sick, silly, sleepy, stupid, sweet, tired, tough, ugly, weird, worried	face
amazing, attractive, awesome, beautiful, cheerful, colorful, delightful, dry, dying, fantastic, fascinating, favorite, fragile, fresh, gentle, golden, little, magical, magnificent, pretty, prickly, smelly, strange, stunning, sunny, tiny, wild	flower
amazing, awesome, colorful, delicious, dry, excellent, fancy, favorite, greasy, great, gross, healthy, hot, natural, super, tasty, traditional, weird, yummy	food

3.2 Pipeline

Figure 2 shows the pipeline of our hierarchical classification. The left part of the figure shows the training pipeline. Given the training images of the same noun Figure 2(a), we first train an object (noun) detector and apply it to those images (b). Only the images with relatively high detection response will be kept on into the next level of the hierarchy. Each training image passed contains at least one object bounding box. Next, for each bounding box a group of comprehensive features including low-level visual features and aesthetic features are extracted (c) for the sentiment concept classification. The details of the features will be described in Section 5. To train each classifier, the ANP pseudo labels of each object passed from the image is treated as soft labels to take into account the semantic overlap of multiple ANPs. We use semantic similarities between concepts to compute the weights (d). Fusion of weighted SVM and proportional SVM [34] (e) is specifically used to learn from the overlapped class labels. The fusion employs a rank minimization technique, which also provides us the ability to select the most discriminative features (f).

The predicting pipeline is illustrated on the right part of Figure 2. In the first level of the hierarchy, we apply all the noun detectors to the test image (g). All the detected bounding box candidates will be considered in the second level. For each candidate, we extract the aforementioned comprehensive features (h) and apply the ANP classifiers corresponding to (e.g., “shy dog” after “dog” detector) detected nouns (i). The ANP predicted probability score is fused with the corresponding noun detection score so that the confidence of noun detection is taken into consideration (j). The maximum score among all the candidates is then chosen as the ANP score (k).

4. NOUN DETECTION

Our object dependent visual sentiment classifiers are designed as hierarchical classifiers. The goal of the first level of the hierarchy is to detect the object. There are several reasons for such a design. First, the objects or nouns are much easier to detect than the visual sentiment attributes described in the adjectives. Second, the features suitable for object detection and sentiment attribute detection are usually different. For example, Histogram of Oriented Gradients (HOG) is known to be efficient for object detection, but sentiment attributes may be more sensitive to aesthetic based features [20]. Third, object detectors can localize an object and enable the extraction of object-based features for training the sentiment attributes. Removal of the interference from the background improves the possibility of success for the difficult task of sentiment attribute classification.

For objects that have clear parts (car, dog, and dress), we adopt the 5th release [13] of Deformable Part Model (DPM) [10], one of state-of-the-art object detection methods. The method models the object as several movable parts attached to a root and uses latent SVM to train the model. Specifically, for car and dog, we use the pre-trained models in the release. For dress, we train our own models using manually labeled bounding boxes on 500 randomly selected images from all the training images associated with the noun “dress”. For the objects which has no clear parts (flower and food), we train SVM models for them (with the same features used for DPM). For “face”, instead of using a general object detector, we apply a rotation invariant multiview face detection method [14] for face images.

In the prediction step, we apply the noun detector on test images with a loose threshold, so that more object candidates can be detected to improve the recall. The detection scores are also fused with the subsequent ANP classifier step.

5. MODELING VISUAL ADJECTIVES

The second step in the hierarchy is the ANP classification. The goal is to train classifiers for different ANPs that contain the same noun, e.g., “happy dog” vs. “sad dog” vs. “lonely dog”. In this subsection, when we mention ANPs, we are referring to ANPs of the same noun by default.

Training set and pseudo labels.

For each noun, every instance in the training set corresponds to a detected object bounding box passed from the noun detection step. This implies multiple instances may come from same training image and some training images are discarded. Each instance inherits the pseudo ANP labels from its original training image.

Feature extraction.

Intuitively, the best set of features is different for different adjectives. Instead of manually designing features for each adjective, we target extracting a comprehensive set of features for all images, and we then rely on automatic feature selection to choose the most discriminative features. The features we used can be classified as low-level features and aesthetic features.

The *low-level features* include BoW features of SIFT, Gist, LBP and Color. The dictionary size of BoW is 500. These features are extracted in an “object-based” manner based on three different settings: whole image, inside the object bounding box, and background only. For the features of the whole image, we further apply two-level Spatial Pyramid Matching. Thus each BoW feature has 3,500 dimensions and the overall low-level feature has 14,000 dimensions. We further extract *aesthetic features* including dark channel, luminosity feature, sharpness, symmetry, low depth of field, white balance, colorfulness, color harmony, eye sensitivity as well as the relative position and size of the bounding box. Details of such features can be found in [1]. These features are also computed for both the whole image and inside the bounding box if applicable. Overall the aesthetic features have 301 dimensions. By extracting object-based features we are able to utilize the object localization in the ANP classification.

Leveraging ConceptNet.

With the extracted features and pseudo labels, one naive way of modeling the ANPs is to train a one-vs-all classifier for each ANP. However, this is not a good approach as different adjectives with related meanings can be applied to the same images. For example, an image of "cute dog" can also be an image of "lovely dog". Therefore, the adjective modeling process is a multi-label learning problem rather than a multi-class learning problem. Unfortunately, although there might be multiple labels for each image, due to the image collection process, one image is usually associated with a single adjective label. The incomplete labels create difficulties in training the ANP classifiers.

Labeling training images manually is infeasible due to the large number of categories and instances. While there exist abundant accurately labeled image datasets for various objects and even fine grained categories, there are no annotated datasets for ANPs.

We propose to learn ANP from incomplete labels by incorporating an additional similarity measure between ANPs. We observed two types of similarity between adjectives related to visual sentiment, which we call "attributive similarity" and "proportional similarity". The former refers to cases in which every instance of one ANP shares certain visual similarity with another. For example, every "cute dog" is usually considered as "adorable dog" too. The later refers to cases in which a certain proportion of the instances of one ANP can be labeled as another. For example, some "dirty dogs" are also "wet dogs", but not all. Modeling these similarities are very challenging. Our hypothesis is that both types are correlated with the *semantic similarity* between adjectives in the domain of language. Thus we build a $n \times n$ matrix of semantic similarities S for n ANPs. Each entry $S_{i,j}$ indicates the semantic similarity between the adjectives of the ith and the jth ANPs. Note we also normalize each row of S so that $S_{i,i} = 1$ and the entry of the least similar ANP in each row equals to -1.

We compute the semantic similarities between adjectives by ConceptNet [19]. The ConceptNet is a hypergraph that links a large amount of concepts (words) by the knowledge discovered from Wikipedia and WordNet, or provided by community contributors. The concept similarity can be computed by searching for the shortest path in this graph. We apply the association function of the Web API of ConceptNet 5 [1] to get the semantic similarity of two adjectives. If an adjective does not exist as a concept in ConceptNet, we manually select one of its synonyms or derivatives instead. The similarity value generated by ConceptNet is between -1 and 1, where 1 corresponds to the similarity to itself. We will explain how we approximate the attributive and proportional similarity by semantic similarity in the next paragraph.

Training the Classifiers.

By considering the two types of similarity between visual sentiment attributes, we design our classification methods as follows.

For "attributive similarity" it is natural to use weighted SVM (wSVM), where each instance in the training data has a weight between 0 and 1 to regulate its contribution to the classifier. Given training samples of multiple ANPs of the same noun, when training a classifier for an ANP$_i$, an image of the ANP$_j$ will be treated as a training sample with a label equaling the sign of $S_{i,j}$ and a sample weight equaling the absolute value of $S_{i,j}$. We subsample from the training instances so that the ratio of positive to negative is 1:1. We employ weighted linear SVM to train the classifiers for each ANP.

To deal with "proportional similarity", our solution is pSVM ($\propto$SVM) [34]. The input of pSVM is a "bag" of the training instances. Each bag is assigned a proportion value indicating the proportion of positive instances in the bag. In our case each bag consists of the instances with the same ANP pseudo labels, and while training the classifier for the ith ANP the proportion value of the jth bag is set as follows:

$$p_{i,j} = \begin{cases} 0 & \text{if } S_{i,j} \leq 0 \\ S_{i,j} & \text{if } S_{i,j} > 0 \end{cases}$$

We use the alter-$\propto$SVM algorithm [34]. It iteratively trains the linear SVM model and updates the inferred labels until convergence.

Model fusion and feature selection.

To combine multiple features and select the most discriminative features for each ANP model, we employ the low-rank fusion proposed by Ye et al. in [33], which has been shown to work much better than naive late fusion. Multiple models trained from wSVM and pSVM with different low-level features and aesthetic features, different regions of the image (whole, object, background) are used for prediction and their confidence scores are fused by rank minimization as mentioned above. The process also provides us individual feature weights that can be used to find the most discriminative features for each ANP classifier.

Following the approach used in detecting noun phrase [25], we further combine the noun detection score with the ANP detection score of each candidate bounding box. If a test image has multiple candidates for an ANP, the maximal score is chosen as the final ANP score.

6. EXPERIMENTS AND DISCUSSIONS

6.1 Test Set and Ground Truth

To evaluate the performance of hierarchical classification, we prepare two test sets from the dataset. Test set A evaluates the performance of distinguishing multiple related ANPs of the same noun. For each ANP, we choose 240 images from the dataset; 40 of them are randomly sampled

[1] http://conceptnet5.media.mit.edu/

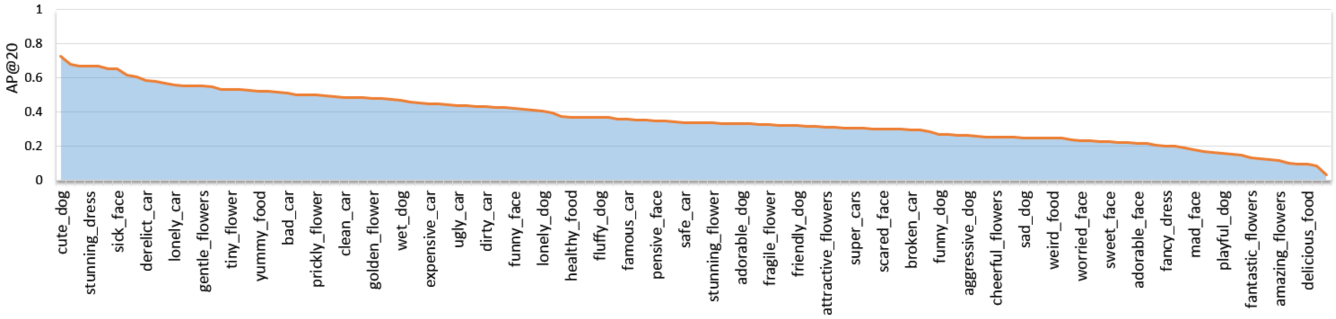


Figure 4: The curves show AP@20 (avg. over 5 runs) for test set A. Only a subset of ANP names are shown due to space limit.

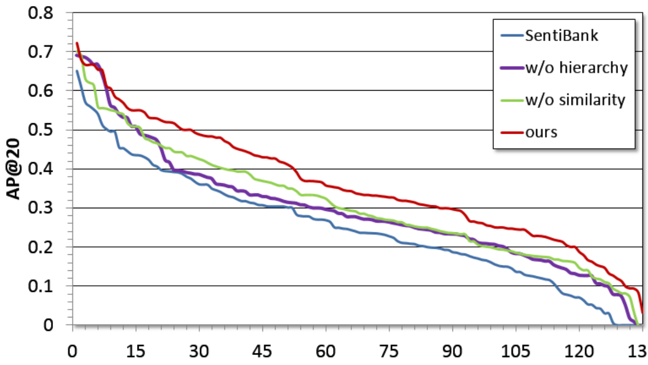


Figure 5: Comparison of ANP detectors with different baselines. ANPs are sorted on the horizontal axis based on the detection performance. Performance is computed by averaging over 5 runs on test set A.

from pseudo-labeled images of each ANP class. The other 200 are randomly sampled from the pseudo-labeled images of other ANPs containing the same noun. Since we cannot rely on the pseudo labels from testing, each image is manually labeled by three individuals (graduate students without detailed knowledge of our work). Each individual answers yes or no to the question "Does the image contain the ANP". The ANP ground truth label of the image is decided by the majority vote of the individual labels. All 240 images are excluded from the training set of the ANP; 20 positives and 40 negatives are randomly sampled from the 240 images to form the test set A of the ANP.

Test set B evaluates the overall visual sentiment classification performance on the real world data. For each ANP (whose noun is denoted by N), 20 positives are sampled in the same way as test set A. To get the negatives, we first randomly select one image from each of the 1,200 SentiBank ANPs by using the pseudo labels from the web except those containing the same noun N. Then we sample additional negatives from the manually labeled ground truth set of the ANPs sharing the same noun. The total number of negative examples in the test set B for each ANP is 1,200.

Our test sets and the trained ANP detectors will be made available to the community at http://visual-sentiment-ontology.appspot.com/.

6.2 Performance and Comparisons

We conduct our experiments on a eight-core CPU workstation. In the training step, training each noun detector takes about six hours including feature extraction, which is the most time consuming step. Fortunately we only need to train once for each noun. In the second level of the hierarchy, feature extraction takes two seconds for each image. Training each ANP classifier in the second level of the hierarchy only takes ten seconds on average using LibLinear library [9]. In the prediction step, it takes about three seconds to compute all 135 ANP scores for each image including feature extraction.

In our experiments all the parameters are chosen by 5-fold cross validation on the training set with pseudo labels. For both test sets, we evaluate the classification performance in terms of AP@20 averaged over five runs, and in each run we randomly sample the test images as mentioned above. The result for test set A is shown in Figure 4. It shows two thirds of ANPs have a AP@20 score above 0.3. The hierarchical classification results are also compared with following baseline and variations:

- SentiBank. A straightforward linear SVM classification in a one-against-all setting using pseudo labels and features extracted from the whole images.
- Without hierarchy. Noun detection is skipped. Features are extracted only from the whole image. Weighted SVM and pSVM are used to handle similarity over multiple related ANPs
- Without similarity among concepts. One-vs-all classification is directly trained based on pseudo labels in the second level of the hierarchy.

The comparison results on test set A are shown in Figure 5. The mean AP on test set A and B for each noun and all nouns along with comparison are also shown in Figure 6.

Figure 7 further shows the top results detected by ANP classifiers using our method. The corresponding object bounding box is shown as blue rectangle. Misclassified images are surrounded by red frames.

For both test set A and test set B, our proposed approach improves over the SentiBank classifiers reported in [2] by a very large margin. The improvement can be as high as 50% as shown in Fig 6: the AP for all nouns for test set A increases from 0.24 to 0.36. The results also provide justification for both the hierarchical framework in Section 3

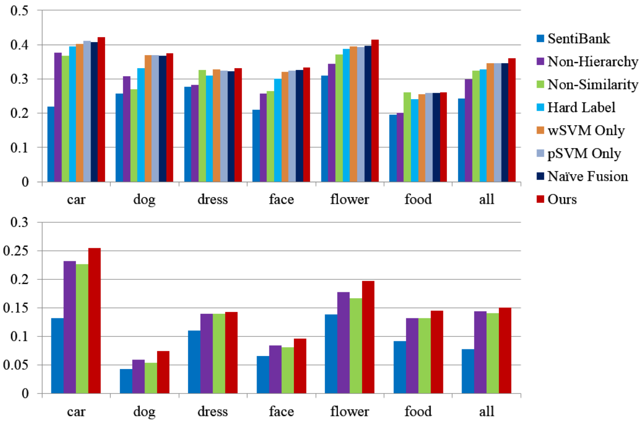


Figure 6: The mean AP on test set A (top) and B (bottom) for each noun and all nouns.

and the similarity modeling technique in Section 5 proposed in this paper.

6.3 Evaluation of Components

We further justify each component of our system. These experiments are conducted on test set A and measured by the mean AP@20 of all nouns averaged over 5 runs. We first evaluate the multi-concept similarity model. In the second level of the hierarchy, if we only use semantic similarity to generate hard labels for each instance and use standard SVM for classification, the AP@20 will be reduced from 0.36 to 0.33. The result for each noun and all nouns are shown as the "Hard Label" in Figure 6. It indicates that leveraging ConceptNet similarity as soft labels is more effective. Next we evaluate the classification model. If we only use weighted SVM or pSVM, the mean AP will both decline to around 0.35. The two groups of results are shown as "wSVM Only" and "pSVM Only" in Figure 6 respectively, and thus confirm the benefit of fusing the two models. At last, we test the classification model with naive late fusion instead of the low-rank fusion, the mean AP is also reduced to 0.35. The result is shown as "Naive Fusion" in Figure 6. It verifies our assumption that it is useful to fuse models of different visual features. Although the improvement due to low-rank fusion is not significant, the more important benefit of this fusion method is generating a sparse error matrix for each feature so that we can achieve feature selection, as well as classifier interpretation, which will be discussed in Section 6.4.

6.4 Discussion

Further justification of the hierarchical framework.

Previous research suggests that phrase detection or a combination of phrase and object detection performs better than general object detection [25]. For example, "jumping dog" can be better detected using a combination of "jumping dog" detector and "dog" detector. Therefore, one may argue against our choice of general object detector and suggest using specific object detector for each ANP. Several observations lead to our decision. First, most ANPs are different from the visual phrases mentioned in [25]. The additional visual cues involving many global and aesthetic related features extracted from localized regions usually do not contribute much to the typical object detection features such as HoG. However, there are exceptions such as "sleepy dog", which has a very different pose from normal dogs. Second, it is very difficult and inefficient to integrate our comprehensive features into the object detection method such as DPM. Third, it is too time consuming and impracticable to label object bounding boxes for each ANP. Therefore, our hierarchical classification offers the best trade-off between accuracy and efficiency.

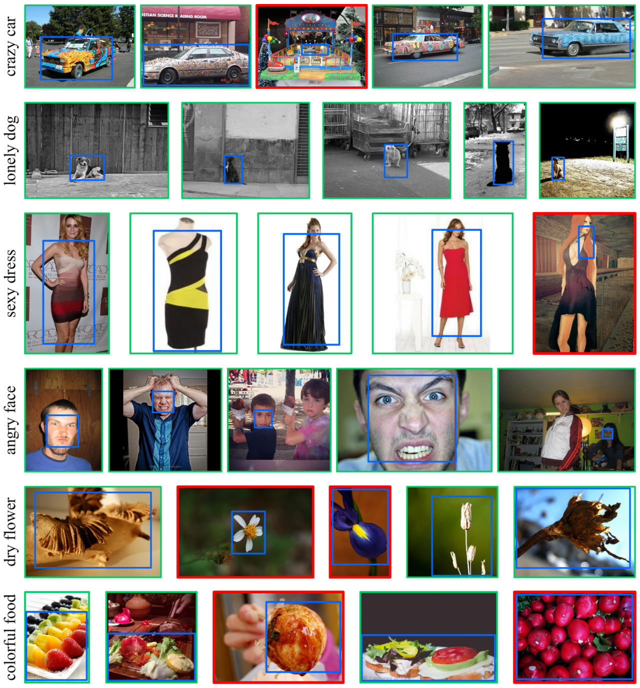


Figure 7: Top 5 detection results of ANP classifiers trained by our hierarchical model. Misclassified images are red framed. Detected concept regions are shown as blue bounding boxes. For image credits see footnote 3.

Extension to scenes or settings.

According to Figure 6, even without object localization our method still outperforms the baseline, thanks to the soft labels that exploit the concept similarity and the deliberately designed model. This means that we can apply our method to "global nouns" such as "sunset" and "sky". These concepts do not need localization.

The factors for classification accuracy.

Some of our ANP classifiers cannot achieve good classification accuracy. An important reason is that some ANP categories may still be ambiguous. Some adjectives, such as "favourite", are subjective because the interpretations may vary significantly among individuals. This causes greater diversity in the image labels and also the visual features. Other reasons include the lack of visual cues of the adjective such as "smelly", low performance of noun detection such as "food" and the lack of more efficient features such as the lack of facial marker features for "face".

Table 2: The examples of the discriminative features of ANP.

ANP	Discriminative features
abandoned car	object SIFT, white balance, background color histogram, background Gist
crazy car	object LBP, object colorfulness, color harmony, object color histogram
lonely dog	background Gist, composition, depth of field, colorfulness
shy dog	object luminosity, dark channel, object symmetry, object white balance
ugly face	object color harmony, object symmetry, object Gist, color histogram

Discriminative features.

Other than improving the classification accuracy, it is also interesting to understand how the classification works by studying the contributions of different features. Specifically we try to interpret the classifiers by discovering discriminative features.

Low-rank fusion [33] provides us with statistics on the discrimination ability of each feature during ANP classification. For each ANP we rank the features according to the average over the fusion weights of all the test examples. Table 2 shows the top features for some ANPs. The result shows the features extracted from the object are more discriminative than those extracted from the whole image, and the background features are also useful. For example, images of "abandoned car" are usually shot in the wild and with a dim tone, thus the white balance and background features are important. Many crazy cars modeled by our method are with exaggerated color combinations (Figure 7). Images of "lonely dog" contain more uniform background, and are mostly black and white with relatively small object (dog). "Shy dogs" images usually show half face, thus object symmetry is important. The similar reasons can be found for "ugly face".

Limitations and future work.

Our work has several limitations. First, object detection may not be always reliable, especially when applied to ANPs associated with unique postures of the corresponding noun such as "sleepy dog". Training specific object detectors for such ANPs may be a solution. Second, the semantic similarity derived from ConceptNet might be incorrect. For such cases manually provided similarity can be helpful. Third, our feature set can be more comprehensive, for example including the shape based and "Wang histogram" features that are proved efficient for aesthetics and emotion classification [4, 20], or even the recently proposed "OverFeat" feature based on deep learning [26]. And finally, our discriminative features are limited to our feature set, further work can be done to discover new features for better interpretation of sentiment concept model.

7. APPLICATIONS

We demonstrate two applications that benefit from our visual sentiment concept detection. Note since our experiments are limited to six nouns, the following applications also focus on the images associated with these nouns. More nouns can be easily incorporated by training more ANP detectors using our system.

7.1 Mood-Aware Personalized Music Slide show

Music slide shows are popular for personal entertainment and photo sharing. Modern music playing software is capable of showing visual effects, album covers or photo albums when playing music. Recently, GettyImages presents Moodstream[2], a web-based application that is able to show photos according to the mood of the music. However the photos are from a pre-selected set where the mood labels are already available. Given the ability of our system to detect visual sentiment concepts, we can generate such mood-aware music slide shows directly from users' personal photo albums.

To generate a slide show, we first compute the real-time mood of the music. We adopt Mood Cloud [18] to detect the mood of the music. This method predicts a five dimensional mood vector for any given time stamp of the music in real-time using SVM model trained from manually labeled data. The five dimensions correspond to aggressive, happy, relaxed, sad and party. Next we classify every candidate photo to one of these five moods. We build a 135×5 matrix to map each of the 135 ANPs to the mood vector, where each entry of the matrix is the mean of the semantic similarity and co-occurrence rate between corresponding ANP and name of the mood. The semantic similarity is computed from ConceptNet. The co-occurrence rate is a normalized value of the returned image count while searching the ANP and mood together on Flickr. The mood vector of the photo can be estimated by the mean mood vector of its top three detected ANPs, and we classify the photo to the mood class with the highest score.

Given a music track, we repeatedly decide the slide show photo of each segment from the beginning of the track. Each time we take three seconds of music to compute a dominant mood. A photo of same mood is randomly selected for this segment. If the mood is sad or relaxed, the segment is extended to three seconds. The next segment starts right after the previous one. This process repeats until the end of the track. We also apply several rules for photo transition. The length of each transition is 0.5 seconds. The type includes fade, zoom in and out, and flying in from eight surrounding directions. If the detected mood is "aggressive" and "party", the transition type is restricted to zoom in and flying in, randomly selected otherwise. Since our system also provides ANP bounding boxes, we generate the zooming path according to the center of the top ANP bounding box.

We conduct the experiment on five music tracks and three Creative Common licensed personal photo albums with a total of 100 images downloaded from Flickr. The music tracks are within the length of 15 to 20 seconds. Generating all the slide shows on the machine mentioned in Section 6 takes 4.5 minutes, where 99% of the time is spent on ANP detection of the photos. This indicates that our method can generate a slide show of a real-time music stream with preprocessed photo albums. Figure 8 shows several screenshots from the music slide shows. Please refer to our supplemental video for better illustration of the sidle show results.

We evaluated our result by a preliminary user study. We generate another five slide shows by randomly selecting photos, and ask evaluators to choose the preferred one.

[2] http://moodstream.gettyimages.com/

sad happy aggressive

Figure 8: Screenshots of the music slide show with detected moods. Please refer to the supplemental material for the video. For image credits see footnote 3.

Each pair of the five music slide shows is shown to ten evaluators. These evaluators are graduate students majoring in computer science, engineering, art and physics without detailed knowledge of our project. In 42 of 50 evaluations, the evaluator preferred the mood-aware music slide show.

7.2 Image Commenting Assistant

A novel application of visual sentiment concept detection presented in [3] is the automatic image commenting assistant. In their work, the authors first compute the ANP response of the image by SentiBank, then select proper elements from a comment database as the comment on the image. The selection is based on the correlation between the top three detected visual sentiment concepts and affective concepts extracted from large-scale user comments. While showing promising results, this method suffers from low detection accuracy of the object-based ANPs. We follow this pipeline and replace the sentiment concept predictors with our hierarchical predictors to generate comments. The quality of the comments generated by the two methods is compared. For fair comparison, we only consider the response of the 135 object based ANPs for both SentiBank and our method. We randomly select 20 images from test set A to test the commenting assistants.

Figure 9 shows the comments of five images generated by the two commenting assistants. The top three detected concepts are listed for both methods. The concepts detected by our method are constantly better than [3]. The comparison of top three detected concepts clearly shows two advantages of our method. Our approach can more reliably detect the objects (nouns), especially when they are relatively small in the image, as shown in the first example. Second, our method provides better classifiers for the sentiment attributes (adjective) since ambiguity is considered. We further evaluate the comment quality with a user study. The two comments generated from each of the 20 images are shown to five evaluators. These evaluators are graduate students majoring in engineering and art without detailed knowledge of our project. For each comparison, they are asked "which comment is from the real user". In 78 of the 100 evaluations, the evaluator prefers the comment generated by our approach.

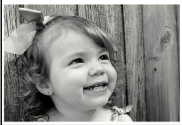
	Ours: Cool idea! (Hope you did something nice for him to make up for all his suffering :-) Top ANPs: worried face, crying face, angry face Chen2014: Gorgeous waterlily, very nice capture. Have a good weekend. Top ANPs: playful dog, muddy dog, delicious food
	Ours: haha, awesome! you have beautiful eyes even when they're psycho Top ANPs: weird face, angry face, mad face Chen2014: That cute, a beautiful picture. A very funny moment. Top ANPs: excited face, silly face, sleepy face
	Ours: That is too funny. And a little scary. Very nice work, Diana. Top ANPs: ugly face, cute dress, funny face Chen2014: beautiful lady with nice legs Top ANPs: dirty dog, sexy dress, stupid face
	Ours: Good god. My dad would think that was awesome... lol. Top ANPs: ugly car, classic cars, awesome cars Chen2014: very nice shot !! amazing cars, perfect !! Top ANPs: classic cars, awesome cars, super cars
	Ours: OMG I just melt every time I see this perfect beautiful sweet little face with those ever so cute chubby baby cheeks and awesome stunning huge eyes! She really is so perfect!!!!!! Top ANPs: sweet face, adorable face, chubby face Chen2014: Happy and beautiful capture! Very nice! Top ANPs: dirty face, scared face, cute dress

Figure 9: Image comments generated by our method and [3]. Top 3 detected ANPs are listed for both. For image credits see footnote 3.

8. CONCLUSIONS

This paper presents a novel hierarchical system to model object-based visual sentiment concepts. The system is designed to handle sentiment concept classification in an object specific manner and tackle the challenge of concept localization and resolving sentiment attribute ambiguity. By leveraging an online commonsense knowledgebase and proposing novel classification to model concept similarity, our system has greatly improved the classification performance over previous work (by up to 50%). The proposed framework also allows us to interpret the classifiers by discovering discriminative features. We demonstrate the power of the proposed system with a few novel applications such as sentiment-aware music slide show of personal albums. Although our experiments currently focus on a visual sentiment dataset, the proposed framework of discovering visual concepts as mid-level representation may be extended to handle other important yet challenging topics such as in aesthetics, style, credibility and intent.

9. ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers and Dr. Martha Larson for their invaluable input. 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.

[3] Credits of the images used in Figure 7, 8, 9 can be found at http://www.ee.columbia.edu/dvmm/vso/credits.html

10. REFERENCES

[1] Subhabrata Bhattacharya, Behnaz Nojavanasghari, Tao Chen, Dong Liu, Shih-Fu Chang, and Mubarak Shah. Towards a comprehensive computational model for aesthetic assessment of videos. In *Proceedings of the 21st ACM International Conference on Multimedia*. ACM, 2013.

[2] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia*. ACM, 2013.

[3] Yan-Ying Chen, Tao Chen, Winston H. Hsu, Hong-Yuan Mark Liao, and Shih-Fu Chang. Predicting viewer affective comments based on image content in social media. In *Proceedings of the International Conference on Multimedia Retrieval*. ACM, 2014.

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

[5] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Computer Vision and Pattern Recognition*, 2009.

[6] Jia Deng, Jonathan Krause, and Li Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 580–587. IEEE, 2013.

[7] Kun Duan, Devi Parikh, David Crandall, and Kristen Grauman. Discovering localized attributes for fine-grained recognition. In *Computer Vision and Pattern Recognition*. IEEE, 2012.

[8] Andrea Esuli and Fabrizio Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the Conference on Language Resources and Evaluation*, volume 6, 2006.

[9] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[11] V. Ferrari and A. Zisserman. Learning visual attributes. In *Neural Information Processing Systems*, 2007.

[12] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. Attribute learning for understanding unstructured social activity. In *European Conference on Computer Vision*. Springer, 2012.

[13] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/~rbg/latent-release5/.

[14] Chang Huang, Haizhou Ai, Yuan Li, and Shihong Lao. High-performance rotation invariant multiview face detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(4):671–686, 2007.

[15] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *Computer Vision and Pattern Recognition*, 2011.

[16] Jia Jia, Sen Wu, Xiaohui Wang, Peiyun Hu, Lianhong Cai, and Jie Tang. Can we understand van Gogh's mood?: Learning to infer affects from images in social networks. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 857–860. ACM, 2012.

[17] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z Wang, Jia Li, and Jiebo Luo. Aesthetics and emotions in images. *Signal Processing Magazine, IEEE*, 28(5):94–115, 2011.

[18] Cyril Laurier and Perfecto Herrera. Mood Cloud: A real-time music mood visualization tool. In *Proceedings of the 2008 Computers in Music Modeling and Retrieval Conference*, 2008.

[19] Hugo Liu and Push Singh. ConceptNet - a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004.

[20] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of ACM Multimedia*, pages 83–92, 2010.

[21] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *Proceedings of the International Conference on Computer Vision*, 2011.

[22] M. Naphade, J. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and Curtis J. Large-scale concept ontology for multimedia. In *IEEE Multimedia*, 2006.

[23] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Information Retrieval*, 2(1-2):1–135, 2008.

[24] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Computer Vision and Pattern Recognition*. IEEE, 2012.

[25] Mohammad Amin Sadeghi and Ali Farhadi. Recognition using visual phrases. In *Computer Vision and Pattern Recognition*. IEEE, 2011.

[26] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv:1312.6229*, 2013.

[27] J.R. Smith, M. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. In *International Conference on Multimedia and Expo*, 2003.

[28] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[29] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, 2010.

[30] Weining Wang and Qianhua He. A survey on emotional semantic image retrieval. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 117–120. IEEE, 2008.

[31] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005.

[32] V. Yanulevskaya, J. van Gemert, K. Roth, A. Herbold, N. Sebe, and J.M. Geusebroek. Emotional valence categorization using holistic image features. In *Proceedings of the IEEE International Conference on Image Processing*, pages 101–104, 2008.

[33] Guangnan Ye, Dong Liu, I-Hong Jhuo, and Shih-Fu Chang. Robust late fusion with rank minimization. In *Computer Vision and Pattern Recognition*. IEEE, 2012.

[34] Felix X. Yu, Dong Liu, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. $\propto$SVM for learning with label proportions. In *International Conference on Machine Learning*, 2013.