# Inferring Sentiment from Web Images with Joint Inference on Visual and Social Cues: A Regulated Matrix Factorization Approach

**Yilin Wang**[1] **Yuheng Hu**[2] **Subbarao Kambhampati**[1] **Baoxin Li**[1]
[1]Department of Computer Science, Arizona State University, Tempe AZ
[2]IBM Almaden Research Center, San Jose CA
{ywang370,rao,baoxin.li}@asu.edu yuhenghu@us.ibm.com

**Abstract**

In this paper, we study the problem of understanding human sentiments from large scale collection of Internet images based on both image features and contextual social network information (such as friend comments and user description). Despite the great strides in analyzing user sentiment based on text information, the analysis of sentiment behind the image content has largely been ignored. Thus, we extend the significant advances in text-based sentiment prediction tasks to the higher-level challenge of predicting the underlying sentiments behind the images. We show that neither visual features nor the textual features are by themselves sufficient for accurate sentiment labeling. Thus, we provide a way of using both of them. We leverage the low-level visual features and mid-level attributes of an image, and formulate sentiment prediction problem as a non-negative matrix tri-factorization framework, which has the flexibility to incorporate multiple modalities of information and the capability to learn from heterogeneous features jointly. We develop an optimization algorithm for finding a local-optima solution under the proposed framework. With experiments on two large-scale datasets, we show that the proposed method improves significantly over existing state-of-the-art methods.

## 1 Introduction

**A picture is worth a thousand words.** It is surely worth even more when it comes to convey human emotions and sentiments. Examples that support this are abundant: great captivating photos often contain rich emotional cues that help viewers easily connect with those photos. With the advent of social media, an increasing number of people start to use photos to express their joy, grudge, and boredom on social media platforms like Flickr and Instagram. Automatic inference of the emotion and sentiment information from such ever-growing, massive amounts of user-generated photos is of increasing importance to many applications in health-care, anthropology, communication studies, marketing, and many sub-areas within computer science such as computer vision. Think about this: Emotional wellness impacts several aspects of people's lives. For example, it introduces self-empathy, giving an individual greater awareness of their feelings. It also improves one's self-esteem and resilience, allowing them to bounce back with ease, from poor emotional health, and physical stress and difficulty. As people are increasingly using photos to record their daily lives[1], we can assess a person's emotional wellness based on the emotion and sentiment inferred from her photos on social media platforms (in addition to existing emotion/sentiment analysis effort, e.g., see (De Choudhury, Counts, and Gamon 2012) on text-based social media).

In this paper, our goal is to automatically infer human sentiments (positive, neutral and negative) from photos shared on Flickr and Instagram. While sentiment analysis of photos is still in its infancy, a number of tools have been proposed during past two years(Yuan et al. 2013; Jia et al. 2012). A popular approach is to identify visual features from a photo that are related to human sentiments, such as objects (e.g., toys, birthday cakes, gun), human actions (e.g., crying or laughing), and many other features like color temperature. However, such an approach is often insufficient because the same objects/actions may convey different sentiments in different photo contexts. For example, consider Figure 1: one can easily detect the crying lady and girl (using computer vision algorithms such as face detection(Zhu and Ramanan 2012) and expression recognition(Song et al. 2010)). However, the same "crying" action conveys two clearly different sentiments: the "crying" in Figure 1a is obviously positive as the result of a successful marriage proposal. In contrast, the tearful girl in Figure 1b looks quite unhappy thus expresses negative sentiment. In other words, the so-called "visual affective gap" (Machajdik and Hanbury 2010) exists between rudimentary visual features and human sentiment embedded in a photo. On the other hand, one may also consider inferring the sentiment of a photo via its textual descriptions (e.g., titles) using existing off-the-shelf text-based sentiment analysis tools (Pang, Lee, and Vaithyanathan 2002). Although these descriptions can provide very helpful context information of the photos, solely relying on them while ignoring the visual features of the photos can lead to poor performance as well. Consider Figure 1 again: by analyzing only the text description, we can conclude that both Figure 1a and 1b convey negative sentiment as the keyword "crying" is often



[1]http://www.pewinternet.org/2015/01/09/social-media-update-2014/

(a) "Girlfriend crying a lot when I proposed to her". (b) Crying baby after her toy was taken

Figure 1: An example shows affective gap.

classified as negative sentiment in standard sentiment lexicon (Taboada et al. 2011). Last, both visual feature-based and text-based sentiment analysis approaches require massive amounts of training data in order to learn high quality models. However, manually annotating the sentiment of a vast amount of photos and/or their textual descriptions is time consuming and error-prone, presenting a bottleneck in learning good models.

The weaknesses discussed in the foregoing motivate the need for a more accurate automated framework to infer the sentiment of photos, with 1) considering the photo context to bridge the "visual affective gap", 2) considering a photo's visual features to augment text-based sentiment, and 3) considering the availability of textual information, thus a photo may have little or no social context (e.g., friend comments, user description). While such a framework does not exist, we can leverage some partial solutions. For example, we can learn the photo context by analyzing the photo's social context (text features). Similarly, we can extract visual features from a photo and map them to different sentiment meanings. Last, while manual annotation of all photos and their descriptions is infeasible, it is often possible to get sentiment labeling for small sets of photos and descriptions.

**Technical Contribution:** We propose an efficient and effective framework, named *RSAI* (Robust Sentiment Analysis for Images), for inferring human sentiment from photos that leverages these partial solutions. Figure 2 depicts the procedure of RSAI. Specifically, to fill the visual affective gap, we first extract visual features from a photo using low-level visual features (e.g., color histograms) and a large number of mid-level (e.g., objects) visual attribute/object detectors (Yuan et al. 2013; Tighe and Lazebnik 2013). Next, to add sentiment meaning to these extracted non-sentimental features, we construct Adjective Noun Pairs (ANPs)(Borth et al. 2013). Note that ANP is a visual representation that describes visual features by text pairs, such as "cloudy sky", "colorful flowers". It is formed by merging the low-level visual features to the detected mid-level objects and mapping them to a dictionary (more details on ANP are presented in Section 3). On the other hand, to learn the image's context, we analyze the image's textual description and capture its sentiment based on sentiment lexicons. Finally, with the help from ANPs and image context, RSAI infers the image's sentiment by factorizing an input image-features matrix into three factors corresponding to image-term, term-sentiment and sentiment-features. The ANPs here can be seen as providing the initial information ("prior knowledge") on sentiment-feature factors. Similarly, the learnt image context can be used to constrain image-term and term-sentiment factors. Last, the availability of labeled sentiment of the images can be used to regulate the product of image-term, term-sentiment factors. We pose this factorization as an optimization problem where, in addition to minimizing the reconstruction error, we also require that the factors respect the prior knowledge to the extent possible. We derive a set of multiplicative update rules that efficiently produce this factorization, and provide empirical comparisons with several competing methodologies on two real datasets of photos from Flickr and Instagram. We examine the results both quantitatively and qualitatively to demonstrate that our method improves significantly over baseline approaches.

The rest of this paper is organized as follows: first, we review the related work on sentiment prediction as well as work which utilizes the nonnegative matrix factorization. We then present a basic model for the problem and further improve the model by incorporating prior knowledge. The experimental results and a comprehensive analysis are presented in the experiment part. Last, we conclude by identifying future work.

## 2 Related Work

In this section, we review the related work on sentiment analysis and the methods for matrix factorization.

**Sentiment analysis on text and images**: Recently, sentiment analysis has shown its success in opinion mining on textual data, including product review(Liu 2012; Hu and Liu 2004), newspaper articles (Pang, Lee, and Vaithyanathan 2002), and movie rating (Pang and Lee 2004). Besides, there have been increasing interests in social media data (Borth et al. 2013; Yang et al. 2014; Jia et al. 2012; Yuan et al. 2013), such as Twitter and Weibo data. Unlike text-based sentiment prediction approaches, (Borth et al. 2013; Yuan et al. 2013) employed mid-level attributes of visual feature to model visual content for sentiment analysis. (Yang et al. 2014) provides a method based on low-level visual features and social information via a topic model. While (Jia et al. 2012) tries to solve the problem by a graphical model which is based on friend interactions. In contrast to our approach, all such methods restrict sentiment prediction to the specific data domain. For example, in Figure 1, we can see that approaches using pure visual information (Borth et al. 2013; Yuan et al. 2013) may be confused by the subtle sentiment embedded in the image. e.g., two crying people convey totally different sentiment. (Jia et al. 2012; Yang et al. 2014) assume that the images belong to the same sentiment share the same low-level visual features is often not true, because positive and negative images may have similar low-level visual features, e.g., two black-white images contain smiling and sad faces respectively. Recent, deep learning has shown its success in feature learning for many computer vision problem, (You et al. 2015) provides a

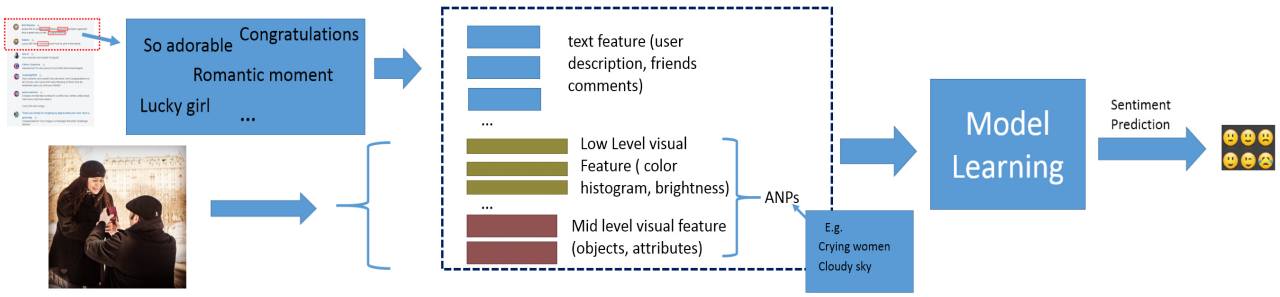


Figure 2: The framework of our proposed method. Comparing to conventional methods, which focus on single source/feature, the proposed method learns the heterogeneous features, including text features, low-level features and mid-level visual features, for sentiment analysis.

transfer deep neutral network structure for sentiment analysis. However, for deep learning framework, millions of images with associated sentiment labels are needed for network training. In real world, such label information is not available and how to deal with overfitting for small training data remains a challenging problem.

**Non-negative matrix factorization(NMF)**: Our proposed framework is also inspired by recent progress in matrix factorization algorithms. NMF has been shown to be useful in computer vision and data mining applications including face recognition(Wang et al. 2005), object detection (Lee and Seung 1999) and feature selection (Das Gupta and Xiao 2011), etc. Specifically, the work in (Lee and Seung 2001) brings more attention to NMF in the research community, where the author proposed a simple multiplicative rule to solve the problem and showed the factor coherence of original image data. (Ding, He, and Simon 2005) shows that if adding orthogonal constrains, the NMF is equivalent to $K$-means clustering. Further, (Ding et al. 2006) presents a work that shows, when incorporating freedom control factors, the non-negative factors will achieve a better performance on classification. In this paper, motivated by previous NMF framework for learning the latent factors, we extend these efforts significantly and propose a comprehensive formulation which incorporates more physically-meaningful constraints for regularizing the learning process in order to find a proper solution. In this respect, our work is similar in spirit to (Hu, Wang, and Kambhampati 2013) which develops a factorization approach for sentiment analysis of social media responses to public events.

## 3 The Proposed RSAI Framework

In this section, we first propose the basic model of our framework. Then we show the details of how to generate the ANPs. After that, we describe how to obtain and leverage the prior knowledge to extend the basic model. We also analyze the algorithm in terms of its correctness and convergence. Table 1 lists the mathematical notation used in this paper.

Table 1: Notations

| Notation | Dimension | Description |
|---|---|---|
| $X$ | $n \times m$ | Input data matrix |
| $T$ | $n \times t$ | Data-term matrix |
| $S$ | $t \times k$ | Term-sentiment matrix |
| $V$ | $m \times k$ | Feature-sentiment matrix |
| $T_0$ | $n \times t$ | Prior knowledge on $T$ |
| $S_0$ | $t \times k$ | Prior knowledge on $S$ |
| $V_0$ | $m \times k$ | Prior knowledge on $V$ |
| $R_0$ | $n \times k$ | Prior knowledge on the labels |

### 3.1 Basic Model

Assuming that all the images can be partitioned into $K$ sentiment ($K = 3$ in this paper as we focus on positive, neutral and negative. However, our framework can be easily extended to handle more fine-grained sentiment.) Our goal is to model the sentiment for each image based on visual features and available text features. Let $n$ be the number of images and the size of contextual vocabulary is $t$. We can then easily cluster the images with similar word frequencies and predict the cluster's sentiment based on its word sentiment. Meanwhile, for each image, which has $m$-dimensional visual features (ANPs, see below), we can cluster the images and predict the sentiment based on the feature probability. Accordingly, our basic framework takes these $n$ data points and decomposes them simultaneously into three factors: photo-text, text-sentiment and visual feature-sentiment. In other words, our basic model tries to solve the following optimization problem:

$$\begin{aligned} \min_{TSV} \quad & \left\|X - TSV^T\right\|_F^2 + \|T - T_0\|_F^2 \\ \text{subject to} \quad & T \geq 0, S \geq 0, V \geq 0;, \end{aligned} \tag{1}$$

where $X \in \mathbb{R}^{n \times m}$ represents input data matrix, and $T \in \mathbb{R}^{n \times t}$ indicates the text features. That is, the $i$th row of matrix $T$ corresponds to the posterior probability of the $i$th image's contextual social network information referring to the $t$ text terms (vocabulary). Similarly, $S \in \mathbb{R}^{t \times k}$ indicates the

posterior probability of a text belonging to $k$ sentiments. Finally, $V \in \mathbb{R}^{m \times k}$ represents the sentiment for each ANP. The regularization term $T_0$ is the term-frequency matrix for the whole word vocabulary (which is built based on textual descriptions of all photos). It is worth noting that the non-negativity makes the latent components easy to interpret.

As a result of this factorization, we can readily predict the image sentiment whether the contextual information (comments, user descriptions,etc.) is available or not. For example, if there is no social information associated with the image, then we can directly derive the image sentiment by applying non-negative matrix factorization for the input data $X$, when we characterize the sentiment of each image through a new matrix $R = T \times S$. Specifically, our basic model is similar to the probabilistic latent semantic indexing (PLSI) (Hofmann 1999) and the orthogonal nonnegative tri-matrix factorization (Ding et al. 2006). In their work, the factorization means the joint distribution of documents and words.

### 3.2 Extracting and Modeling Visual Features

In (Tighe and Lazebnik 2013; Tu et al. 2005; Yuan et al. 2013), visual content can be described by a set of mid-level visual attributes, however, most of the attributes such as "car", "sky","grass", etc., are nouns which make it difficult to represent high level sentiments. Thus, we followed a more tractable approach (Borth et al. 2013), which models the correlation between visual attributes and visual sentiment with adjectives, such as "beautiful" , "awesome", etc. The reason for employing such ANPs is intuitive: the detectable nouns (visual attributes) make the visual sentiment detection tractable, while the adjectives add the sentiment strength to these nouns. In (Borth et al. 2013), a large scale ANPs detectors are trained based on the features extracted from the images and the labeled tags with SVM. However, we find that such pre-defined ANPs are very hard to interpret. For example the pairs like "warm pool" , "abandoned hospital", and it is very difficult to find appropriate features to measure them. Moreover, in their work, during the training stage, the SVM is trained on the features extracted from the image directly, the inability of localizing the objects and scales bounds the detection accuracy. To address these problems, we have a two stage approach to detect ANPs based on the Visual Sentiment Ontology (Borth et al. 2013) and train a one vs all classifier for each ANP.

**Noun Detection**: The nouns in ANPs refer to the objects presented in the image. As one of fundamental tasks in computer vision, object detection has been studied for many years. One of most successful works is Deformable Part Model (DPM) (Felzenszwalb et al. 2010) with Histogram of Oriented Gradient (HOG) (Dalal and Triggs 2005) features. In (Felzenszwalb et al. 2010), the deformable part model has shown its capability to detect most common objects with rigid structure such as: car, bike and non-rigid objects such as pedestrian, dogs. (Pandey and Lazebnik 2011) further demonstrates that DPM can be used to detect and recognize scenes. Hence we adopt DPM to for nouns detection. The common objects(noun) are trained by the public dataset ImageNet(Deng et al. 2009). The scene detectors are trained on SUN dataset (Xiao et al. 2010). It is worth noting that selfie is one of most popular images on the web (Hu, Manikonda, and Kambhampati 2014) and face expression usually conveys strong sentiment, consequently, we also adopt one of state-of-the-art face detection methods proposed in (Zhu and Ramanan 2012).

**Adjective Detection**: Modeling the adjectives is more difficult than nouns due to the fact that there are no well defined features to describe them. Following (Borth et al. 2013), we collect 20,000 images associate with specific adjective tags from Web. The a set of discriminative global features, including Gist, color histogram and SIFT, are applied for feature extraction. Finally the adjective detection is formulated as a traditional image classification problem based on Bag of words(BOW)model. The dictionary size of BOW is 1,000 with the feature dimension size 1,500 after dimension reduction based on PCA.

### 3.3 Constructing Prior Knowledge

So far, our basic matrix factorization framework provides potential solution to infer the sentiment regarding the combination of social network information and visual features. However, it largely ignores the sentiment prior knowledge on the process of learning each component. In this part, we introduce three types of prior knowledge for model regularization: (1) sentiment-lexicon of textual words, (2) the normalized sentiment strength for each ANP, and (3) sentiment labels for each image.

**Sentiment Lexicon** The first prior knowledge is from a public sentiment lexicon named MPQA corpus [2]. In this sentiment lexicon, there are 7,504 human labeled words which are commonly used in the daily life. The number of positive words (e.g."happy", "terrific") is 2,721 and the number of negative words (e.g. "gloomy", "disappointed") is 4,783. Since this corpus is constructed without respect to any specific domain, it provides a domain independent prior on word-sentiment association. It should be noted that the English usage in social network is very casual and irregular, we employ a stemmer technique proposed in (Han and Baldwin 2011). As a result, the ill-formed words can be detected and corrected based on morphophonemic similarity, for example "good" is a correct version of "goooooooooooood". Besides some abbreviation of popular words such as "lol"(means laughing out loud) is also added as prior knowledge. We encode the prior knowledge in a word sentiment matrix $S_0$ where if the $i_{th}$ word belongs to $j_{th}$ sentiment, then $S_0(i, j) = 1$, otherwise it equals to zero.

**Visual Sentiment** In addition to the prior knowledge on lexicon, our second prior knowledge comes from the Visual Sentiment Ontology (VSO) (Borth et al. 2013), which is based on the well known previous researches on human emotions and sentiments (Darwin 1998; Plutchik 1980). It generates 3000 ANPs using Plutchnik emotion model and

[2] http://mpqa.cs.pitt.edu/

Table 2: Sentiment strength score examples

| ANP | Sentiment Strength |
|---|---|
| innocent smile | 1.92 |
| happy Halloween | 1.81 |
| delicious food | 1.52 |
| cloudy mountain | -0.4 |
| misty forest | -1.00 |
| ... | ... |

associates the sentiment strength (range in[-2:2] from negative to positive) by a wheel emotion interface[3]. The sample ANP sentiment scores are shown in Table 2. Similar to the word sentiment matrix $S_0$, the prior knowledge on ANPs $V_0$ is the sentiment indicator matrix.

**Sentiment labels of Photos** Our last prior knowledge focuses on the prior knowledge on the sentiment label associated with the image itself. As our framework essentially is a semi-supervised learning approach, this leads to a domain adapted model that has the capability to handle some domain specific data. The partial label is given by the image sentiment matrix $R_0$ where $R_0 \in \mathbb{R}^{n\times k}$. For example if the $i_{th}$ image belongs to $j_{th}$ sentiment, the $R_0(i,j) = 1$ otherwise $R_0(i,j) = 0$. The improvement by incorporating these label data is empirically verified in the experiment section.

### 3.4 Incorporating Prior Knowledge

After defining the three types of prior knowledge, we incorporate them into the basic model as regularization terms in following optimization problem:

$$\begin{aligned}\min_{TSV} &\left\|X - TSV^T\right\|_F^2 + \alpha \left\|V - V_0\right\|_F^2 \\ &+ \beta \left\|T - T_0\right\|_F^2 + \gamma \left\|S - S_0\right\|_F^2 \\ &+ \delta \left\|TS - R_0\right\|_F^2 \\ \text{subject to } & T \geq 0, S \geq 0, V \geq 0\end{aligned} \tag{2}$$

where $\alpha \geq 0$, $\beta \geq 0$, $\gamma \geq 0$ and $\delta \geq 0$ are parameters controlling the extent to which we enforced the prior knowledge on the respective components. The model above is generic and allows flexibility . For example, if there is no social information available for one image, we can simply set the corresponding row of $T_0$ to zeros. Moreover, the square loss function leads to an unsupervised problem for finding the solutions. Here, we re-write Eq (2) as :

$$\begin{aligned}L =& Tr(X^TX - 2X^TTSV^T + VS^TT^TTSV^T) \\ &+ \alpha Tr(V^TV - 2V^TV_0 + V_0^TV) \\ &+ \beta Tr(T^TT - 2T^TT_0 + T_0^TT_0) \\ &+ \gamma Tr(S^TS - 2S^TS_0 + S_0^TS_0) \\ &+ \delta Tr(S^TT^TTS - 2S^TT^TR_0 + R_0^TR_0)\end{aligned} \tag{3}$$

From Eq (3) we can find that it is very difficult to solve $T$, $S$ and $V$ simultaneously. Thus we employ the alternating multiplicative updating scheme shown in (Ding et al. 2006) to find the optimal solutions. First, we use fixed $V$ and $S$ to update $T$ as follows:

$$T_{ij} \leftarrow T_{ij}\sqrt{\frac{[XVS^T + \beta T_0 + \delta R_0S^T]_{ij}}{[TSV^TVS^T + \beta T + \delta TSS^T]_{ij}}} \tag{4}$$

Next, we use the similar update rule to update $S$ and $V$:

$$S_{ij} \leftarrow S_{ij}\sqrt{\frac{[T^TXV + \gamma S_0 + \delta T^TR_0]_{ij}}{[T^TTSV^TV + \gamma S + \delta T^TTS]_{ij}}} \tag{5}$$

$$V_{ij} \leftarrow V_{ij}\sqrt{\frac{[X^TTS + \alpha V_0]_{ij}}{[VS^TT^TTS + \alpha V]_{ij}}} \tag{6}$$

The learning process consists of an iterative procedure using Eq (3), Eq (4) and Eq (5) until convergence. The description of the process is shown in Algorithm 1.

**Algorithm 1** Multiplicative Updating Algorithm

**Input:** $X, T_0, S_0, V_0, R_0, \alpha, \beta, \gamma, \delta$
**Output:** $T, S, V$
**Initialization:** $T, S, V$
**while** Not Converge **do**
  Update $T$ using Eq(4) with fixed S,V
  Update $S$ using Eq(5) with fixed T,V
  Update $V$ using Eq(6) with fixed T,S
**End**

### 3.5 Algorithm Correctness and Convergence

In this part, we prove the guaranteed convergence and correctness for Algorithm 1 by the following two theorems.

**Theorem 1.** *When Algorithm 1 converges, the stationary point satisfies the Karush-Kuhn-Tuck(KKT) condition, i.e., Algorithm 1 converges correctly to a local optima.*

**Theorem 2.** *The objective function is nondecreasing under the multiplicative rules of Eq (4), Eq (5) and Eq (6), and it will converge to a stationary point.*

The detailed proof is presented in Appendix.

## 4 Empirical Evaluation

We now quantitatively and qualitatively compare the proposed model on image sentiment prediction with other candidate methods. We also evaluate the robustness of the proposed model with respect to various training samples and different combinations of prior knowledge. Finally, we perform a deeper analysis of our results.

### 4.1 Experiment Settings

We perform the evaluation on two large scale image datasets collected from Flickr and Instagram respectively. The collection of Flickr dataset is based on the image IDs provided by (Yang et al. 2014), which contains 3,504,192 images from 4,807 users. Because some images are unavailable now, and without loss of generality, we limit the number of

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

images from each user. Thus, we get 120,221 images from 3921 users. For the collection of the Instagram dataset, we randomly pick 10 users as seed nodes and collect images by traversing the social network based on breadth first search. The total number of images from Instagram is 130,230 from 3,451 users.

**Establishing Ground Truth**: For training and evaluating the proposed method, we need to know the sentiment labels. Thus, 20,000 Flickr images are labeled by three human subjects, the majority voting is employed. However, manually acquiring the labels for these two large scale datasets is expensive and time consuming. Consequently, the rest of more than 230,000 images are labeled by the tags, which was suggested by the previous works (Yang et al. 2014; Go, Bhayani, and Huang )[4]. Since labeling the images based on the tags may cause noise issue, and for better reliability we only label the images with primary sentiment labels, which include: positive, neutral and negative. It is worth noting that the human labeled images have both primary sentiment labels and fine grained sentiment labels. The fine grained labels, including: happiness, amusement, anger, fear, sad and disgust, are used to for fine grained sentiment prediction.

The comparison methods include: Senti API[5], SentiBank (Borth et al. 2013), EL(Yang et al. 2014) and the baseline method.

- Senti API is a text based sentiment prediction API, it measures the text sentiment by counting the sentiment strength for each text term.
- SentiBank is a state-of-the-art visual based sentiment prediction method. The method extracts a large number of visual attributes and associates them with a sentiment score. Similar to Senti API, the sentiment prediction is based on the sentiment of each visual attributes.
- EL is a graphical model based approach, it infers the sentiment based on the friend interactions and several low level visual features.
- Baseline: The baseline method comes from our basic model. To compare it fairly, we also introduce $R_0$ with the basic model which makes the baseline method have the ability to learn from training data.

## 4.2 Performance Evaluation

**Large scale image sentiment prediction**: As mentioned in Sec 3, the proposed model has the flexibility to incorporate the information and capability to jointly learn from the visual features and text features. For each image, the visual features are formed by the confidence score of each ANP detector, the feature dimension is 1200, which is as large as VSO (prior knowledge $V_0$). For the text feature, it is formed based on the term frequency and the dimension relies on the input data. To predict the label, the model input is unknown data $X \in \mathbb{R}^{n \times m}$ and its corresponding text feature matrix $T_0 \in \mathbb{R}^{n \times t}$, where $n$ is the number of images, $m = 1200$ and $t$ is the vocabulary size, we decompose it via Aglorithm 1 and get the label based on max pooling each row of $X * V$. *It is worth noting that in the proposed model,* ***tags are not included as input feature****.*

The results of comparison are shown in Table 3. We employ 30% data for training and remaining for testing. To verify the reliability of tags labeled images, we also included 20000 labeled Flickr images with primary sentiment label. Especially, the classifier setting for SentiBank and EL followed the original papers. The classifier of Sentibank is logistic regression and for EL it is SVM. From the results we can see that, the proposed method performs best in both datasets. Noting that proposed method improved 10% and 6% over state-of-the-art methods (Borth et al. 2013). Results from proposed method are shown in Figure 4. Noting that the number we reported in Table 3 is the prediction accuracy for each method.

From the table, we can see that, even though noise exists in the Flickr and Instagram dataset, the results are similar to the performance on human labeled dataset. Another interesting observation is that the performance of EL on Instagram is worse than on Flickr, one reason could be that the wide usage of "picture filters" lowers discriminative ability of the low level visual features, while the models based on the mid level attributes can easily avoid this filter ambiguity. Another interesting observation is that our basic model performs fairly well even if it does not incorporate the knowledge from sentiment strength of ANPs, which indicates that the object based ANPs by our method are more robust than the features used in (Borth et al. 2013).

**Fine Grained Sentiment Prediction**: Although our motivation is to predict the sentiment (positive, negative) on the visual data, to show the robustness and extension capability of the proposed model, we further evaluate the proposed model on a more challenging task in social media; predicting human emotions. Based on the definition of human emotion (Ekman 1992), our fine grained sentiment study labels the user posts with following human emotion categories including: happiness, amusement, disgust, anger, fear and sadness. The results on 20000 manually labeled flickr post are shown in Figure 5. Compared to sentiment prediction, fine grained sentiment prediction would give us more precise user behavior analysis and new insights on the proposed model.

As Figure 5 shows, compared to SentiBank and EL, the proposed method has the highest average classification accuracy and the variance of proposed method on these 6 categories is smaller than that of the baseline methods, which demonstrates the potential social media applications of the proposed method such as predicting social response. We noticed that the sad images have the highest prediction accuracy, and both disgust and anger are difficult to predict. Another observation is the average performance of positive categories, happiness and amusement, is similar to the negative categories. Explaining reason for this drives us to dig deeper into sentiment understanding in the following section.

[4]More details can be found in(Yang et al. 2014) and (Go, Bhayani, and Huang )

[5]http://sentistrength.wlv.ac.uk/,a text based sentiment prediction API


(a) Negative (b) Neutral (c) Positive

Figure 3: Sample tag labeled images from Flickr and Instagram.


(a) Negative (b) Neutral (c) Positive

Figure 4: Sample results from our proposed method. Photos with red bounding box are false positive predictions.

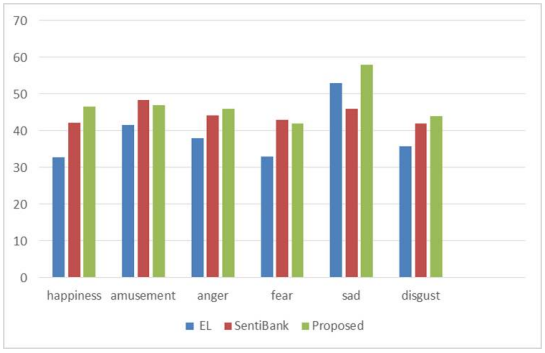


Figure 5: Fine grained sentiment prediction results (Y-axis represents the accuracy for each method).

### 4.3 Analysis and Discussion

In this section, we present an analysis of parameters for the proposed method and the results of the proposed method. Specifically, in last section we have studied the performance of different methods. In this part, our objective is to have deeper understanding on the datasets and the correlation between different features and the sentiments embedded in the images. Without loss of generality, we collected additional 20k images from Flickr and Instagram respectively (totally 40K) and we address the following research questions:

- **RQ1:**What is the relationship between visual features and visual sentiments?
- **RQ2:**Since the proposed method is better than pure visual feature based method, How does the model gain?

First, we start with RQ1 by extracting the visual features used in (Borth et al. 2013) and (Yang et al. 2014) for each image in the Flickr and Instagram datasets. Then we use k-means clustering to obtain 3 clusters of images for each dataset, where the image similarity is measured as Euclidean distance in the feature spaces. Based on each cluster center, we used the classfier trained in the previous experiment for cluster labeling. The results are shown in Figure 6. The x-axis is the different class label for each dataset and the y-axis is the number of images that belong to each cluster. From the

Table 3: Sentiment Prediction Results. The number means prediction accuracy, the higher the better.

| | Senti API | SentiBank | EL | Baseline | Proposed method |
|---|---|---|---|---|---|
| 20000 Flickr | 0.32 | 0.42 | 0.47 | 0.48 | **0.52** |
| Flickr | 0.34 | 0.47 | 0.45 | 0.48 | **0.57** |
| Instagram | 0.27 | 0.56 | 0.37 | 0.54 | **0.62** |

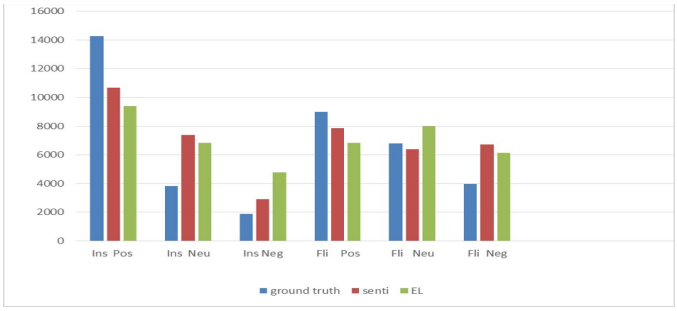


Figure 6: Sentiment distribution based on visual features. From left to rigth is number of positive, neutral, negative images in Instagram and Flickr, receptively. Y axis represents the number of images.

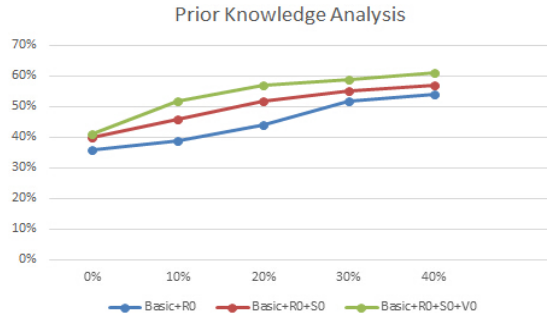


Figure 7: Performance gain by incorporating training data.

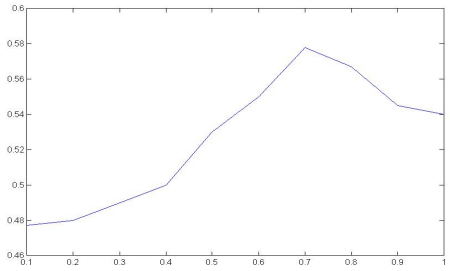


Figure 8: The value of $\beta$ versus model performance. X axis is $\beta$ value, y axis is value of model performance.

results, we notice that the "visual affective gap" does exist between human sentiment and visual features. For the state-of-the art method (Borth et al. 2013), the neural images are largely misclassified based on the visual features. While for (Yang et al. 2014), we observe t the low level features, e.g., color histogram, contrast and brightness, are not closely related to human sentiment as visual attributes.

We further analyze the performance of the proposed method based on these 40,000 images.

**Parameter study**: In the proposed model, we incorporate three types of prior knowledge: sentiment lexicon, sentiment labels of photos and visual sentiment for ANPs. It is important and interesting to explore the impact of each of them on the performance of the proposed model. Figure 7 presents the average results (y-axis) of two datasets on sentiment prediction with different amount of training data (x-axis)[6], where the judgment is on the same three sentiment labels with different combinations respectively. It should be noted that each combination is optimized by Algorithm 1, which has similar formulations. Moreover, we set the same parameter for $\alpha$,$\beta$, $\gamma$ and $\delta$ (0.9, 0.7, 0.8 and 0.7). Results give us two insights. First, employing more prior knowledge will make the model more effective than using only one type of prior knowledge. For our matrix factorization framework, $T$ and $V$ have independent clustering freedom by introducing $S$, thus it is natural to add more constraints for desired decomposed component. Second, when no training data, the basic model with $S_0$ performs much better than SentiAPI (refer Table 3), which means incorporating ANPs significantly improves image sentiment prediction. It is worth noting that there is no training stage for the proposed method. Thus when compared to fully supervised approaches, our method is more applicable in practice when the label information is unavailable.

**Bridging the Visual Affective Gap (RQ2)**: Figure 1 and Figure 7 demonstrate that a visual affective gap exists between visual features and human sentiments (i.e., the same visual feature may correspond to different sentiments in different context). To bridge this gap, we show that one possible solution is to utilize heterogeneous data and features available in social media to augment the visual feature-based sentiment. In the previous parameter study, we have studied the importance of the prior knowledge. Furthermore, we study importance of $\beta$ which contains the degree of contextual social information used in the proposed model. From Figure 8, we can observe that the performance of the proposed model increases along the value of $\beta$. However, when $\beta$ is greater than 0.8, the performance drops. This is because textual information in social media data is usually incomplete. Larger $\beta$ will cause negative effects on the prediction accuracy where there is none or little information available.

[6]The experiments setting is as same as discussed above.

## 5 Conclusion and Future Work

Can we learn human sentiments from the images on the web? In this paper, we proposed a novel approach for visual sentiment analysis by leveraging several types of prior knowledge including: sentiment lexicon, sentiment labels and visual sentiment strength. To bridge the "affective gap" between low-level image features and high-level image sentiment, we proposed a two-stage approach to general ANPs by detecting mid-level attributes. For model inference, we developed a multiplicative update algorithm to find the optimal solutions and proved the convergence property. Experiments on two large-scale datasets show that the proposed model is superior to other state-of-the-art models in both inferring sentiment and fine grained sentiment prediction.

In the future, we will employ crowdsourcing tools, such as AmazonTurk[7], to obtain high-quality, manually-labeled data to test the proposed method. Furthermore, inspired by the recent development of advanced deep learning algorithms and their success in image classification and detection tasks, we will follow this research direction to perform the sentiment analysis via deep learning. In order to have a robust trained architecture and network parameters, we will focus on the deep learning models that work for smaller dataset. Moreover, beyond sentiment analysis, we will study social event and social response (Hu et al. 2012; Hu, Manikonda, and Kambhampati 2014) via visual data in the social media.

## 6 Acknowledgment

Yilin Wang and Baoxin Li are supported in part by a grant (#1135616) from the National Science Foundation. Kambhampati's research is supported in part by the ARO grant W911NF-13-1- 0023, and the ONR grants N00014-13-1-0176, N00014-13-1-0519 and N00014-15-1-2027. Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

## References

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

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, 886–893. IEEE.

Darwin, C. 1998. *The expression of the emotions in man and animals*. Oxford University Press.

Das Gupta, M., and Xiao, J. 2011. Non-negative matrix factorization as a feature selection tool for maximum margin classifiers. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2841–2848. IEEE.

De Choudhury, M.; Counts, S.; and Gamon, M. 2012. Not all moods are created equal! exploring human emotional states in social media. In *Sixth International AAAI Conference on Weblogs and Social Media*.

[7]https://www.mturk.com/mturk/welcome

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 248–255. IEEE.

Ding, C.; Li, T.; Peng, W.; and Park, H. 2006. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 126–135. ACM.

Ding, C. H.; He, X.; and Simon, H. D. 2005. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SDM*, volume 5, 606–610. SIAM.

Ekman, P. 1992. An argument for basic emotions. *Cognition & Emotion* 6(3-4):169–200.

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

Go, A.; Bhayani, R.; and Huang, L. Twitter sentiment classification using distant supervision.

Han, B., and Baldwin, T. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 368–378. Association for Computational Linguistics.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 50–57. ACM.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177. ACM.

Hu, Y.; John, A.; Wang, F.; and Kambhampati, S. 2012. Et-lda: Joint topic modeling for aligning events and their twitter feedback. In *Proceedings of the 6th AAAI Conference*.

Hu, Y.; Manikonda, L.; and Kambhampati, S. 2014. What we instagram: A first analysis of instagram photo content and user types.

Hu, Y.; Wang, F.; and Kambhampati, S. 2013. Listening to the crowd: automated analysis of events via aggregated twitter sentiment. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, 2640–2646. AAAI Press.

Jia, J.; Wu, S.; Wang, X.; Hu, P.; Cai, L.; and Tang, J. 2012. 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*, 857–860. ACM.

Lee, D. D., and Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788–791.

Lee, D. D., and Seung, H. S. 2001. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, 556–562.

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies* 5(1):1–167.

Machajdik, J., and Hanbury, A. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the international conference on Multimedia*, 83–92. ACM.

Pandey, M., and Lazebnik, S. 2011. Scene recognition and weakly supervised object localization with deformable part-based models. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 1307–1314. IEEE.

Pang, B., and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, 271. Association for Computational Linguistics.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 79–86. Association for Computational Linguistics.

Plutchik, R. 1980. *Emotion: A psychoevolutionary synthesis*. Harper & Row New York.

Song, M.; Tao, D.; Liu, Z.; Li, X.; and Zhou, M. 2010. Image ratio features for facial expression recognition application. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 40(3):779–788.

Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; and Stede, M. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics* 37(2):267–307.

Tighe, J., and Lazebnik, S. 2013. Finding things: Image parsing with regions and per-exemplar detectors. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 3001–3008. IEEE.

Tu, Z.; Chen, X.; Yuille, A. L.; and Zhu, S.-C. 2005. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of Computer Vision* 63(2):113–140.

Wang, Y.; Jia, Y.; Hu, C.; and Turk, M. 2005. Non-negative matrix factorization framework for face recognition. *International Journal of Pattern Recognition and Artificial Intelligence* 19(04):495–511.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, 3485–3492. IEEE.

Yang, Y.; Jia, J.; Zhang, S.; Wu, B.; Li, J.; and Tang, J. 2014. How do your friends on social media disclose your emotions?

You, Q.; Luo, J.; Jin, H.; and Yang, J. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks.

Yuan, J.; Mcdonough, S.; You, Q.; and Luo, J. 2013. Sentribute: image sentiment analysis from a mid-level perspective. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, 10. ACM.

Zhu, X., and Ramanan, D. 2012. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2879–2886. IEEE.

# A Appendix

**Theorem 1.** *When Algorithm 1 converges, the stationary point satisfies the Karush-Kuhn-Tuck(KKT) condition, i.e., Algorithm 1 converges correctly to a local optima.*

**Proof of Theorem 1**. We prove the theorem when updating $V$ using Eq (6), similarly, all others can be proved in the same way. First we form the gradient of $L$ regards $V$ as Lagrangian form:

$$\frac{\partial L}{\partial V} = 2(VS^TT^TTS + \alpha V) - 2(X^TTS + \alpha V_0) - \mu \quad (7)$$

Where $\mu$ is Lagrangian multiplier $\mu_{ij}$ enforces the non-negativity constraint on $V_{ij}$. From the complementary slackness condition, we can obtain

$$(2(VS^TT^TTS + \alpha V) - 2(X^TTS + \alpha V_0))_{ij}V_{ij} = 0 \quad (8)$$

This is the fixed point relation that local minima for $V$ must hold. Given the Algorithm 1., we have the convergence point to the local minima when

$$V_{ij} = V_{ij}\sqrt{\frac{[X^TTS + \alpha V_0]_{ij}}{[VS^TT^TTS + \alpha V]_{ij}}} \quad (9)$$

Then the Eq (9) is equivalent to

$$(2(VS^TT^TTS + \alpha V) - 2(X^TTS + \alpha V_0))_{ij}V_{ij}^2 = 0 \quad (10)$$

This is same as the fixed point of Eq (8),i.e., either $V_{ij} = 0$ or the left factor is 0. Thus if Eq (10) holds the Eq (8) must hold and vice versa.

**Theorem 2.** *The objective function is nondecreasing under the multiplicative rules of Eq (4), Eq (5) and Eq (6), and it will converge to a stationary point.*

**Proof of Theorem 2**. First, let $H(V)$ be:

$$H(V) = Tr((VS^TT^TTS + \alpha V)V^T - (X^TTS + \alpha V_0 + \mu)V^T) \quad (11)$$

and it is very easy to verify that $H(V)$ is the Lagrangian function of Eq (3) with KKT condition. Moreover, if we can verify that the update rule of Eq (4) will monotonically decrease the value of $H(V)$, then it means that the update rule of Eq (4) will monotonically decrease the value of $L(V)$(recall Eq (3)). Here we complete the proof by constructing the following an auxiliary function $h(V, \widetilde{V})$.

$$h(V, \widetilde{V}) = \sum_{ik} \frac{(\widetilde{V}(VS^TT^TTS + \alpha V))_{ik}V_{ik}^2}{\widetilde{V_{ik}}} - \sum_{ik}(X^TTS + \alpha V_0 + \mu)_{ik}V_{ik}(1 + \log\frac{V_{ik}}{\widetilde{V_ik}}) \quad (12)$$

Since $z \geq (1 + \log z), \forall z > 0$ and similar in (Ding et al. 2006), the first term in $h(V, \widetilde{V})$ is always larger than that in $H(V)$, then the inequality holds $h(V, \widetilde{V}) \geq H(V)$. And it is easy to see $h(V, \widetilde{V}) = H(V)$, thus $h(V, \widehat{V})$ is an auxiliary function of $H(V)$. Then we have the following inequality chain:

$$H(V^0) = h(V^0, V^0) \geq h(V^0, V^1) = H(V^1).... \quad (13)$$

Thus, with the alternate updating rule of $V, S$ and $T$, we have the following inequality chain:

$$L(V^0, T^0, S^0) \geq L(V^1, T^0, S^0) \geq L(V^1, T^1, S^0).... \quad (14)$$

Since $L(V, S, T) \geq 0$. Thus $L(V, S, T)$ is bounded and the Algorithm 1 converges , which completes the proof.