Analyzing and Predicting Sentiment of Images on the Social Web

Stefan Siersdorfer,
Enrico Minack, Fan Deng
L3S Research Center
Leibniz Universität Hannover
30167 Hannover, Germany
{siersdorfer,minack,deng}@L3S.de

Jonathon Hare
Electronics & Computer Science
University of Southampton
SO17 1BJ Southampton, UK
jsh2@ecs.soton.ac.uk

ABSTRACT

In this paper we study the connection between sentiment of images expressed in metadata and their visual content in the social photo sharing environment Flickr. To this end, we consider the bag-of-visual words representation as well as the color distribution of images, and make use of the SentiWordNet thesaurus to extract numerical values for their sentiment from accompanying textual metadata. We then perform a discriminative feature analysis based on information theoretic methods, and apply machine learning techniques to predict the sentiment of images. Our large-scale empirical study on a set of over half a million Flickr images shows a considerable correlation between sentiment and visual features, and promising results towards estimating the polarity of sentiment in images.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: INFORMATION STORAGE AND RETRIEVAL; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Experimentation, Measurement

Keywords

Color features, Classification, Sentiment Analysis, Automatic Annotation, Visual Terms

1. INTRODUCTION

The social web has opened many avenues for exploration of large multimedia datasets that have previously been unavailable. Popular social websites, such as Flickr, contain massive amounts of visual information in the form of photographs. Many of these photographs have been collectively tagged and annotated by members of the respective community. In this work, we exploit a large dataset of images crawled from Flickr in order to investigate links between visual features and sentiment values extracted from the images' textual metadata.

Emotional semantic image retrieval is the heading given to the research area in which low-level image features are associated with emotions for the purposes of retrieval or categorization. Schmidt and Stock [7] investigated emotional image retrieval from the library science perspective and performed a study based on 763 human-tagged emotions on a set of 25 images from Flickr. Their study indicated that for at least some images in their test collection inter-tagger consensus was very high. This finding is a justification for there being a repeatable underlying link between visual content and evoked emotion/sentiment. Colombo et al. [1] proposed one of the first automatic emotional image retrieval systems. In their system, a novel technique to obtain a high-level representation of art images was implemented, which allowed the derivation of emotional semantics such as action, relaxation, joy and uneasiness. Finally, Wang et al. [9] demonstrated a novel scheme to automatically annotate emotional image semantics and provide emotional image retrieval. Wang et al.'s approach is based on a set of global color and sharpness descriptors, based on theories of human visual perception, and designed especially for their experiments. The evaluation on a small dataset showed promising results. For an entirely different problem setting, where tags and visual features in combination with favorite assignments in Flickr are used to classify and rank photos according to their attractiveness, [6] can be seen as an example of using metadata of images to obtain training data and ground truth for classification and regression.

Compared to previous work, our paper is the first to apply and evaluate automatic classification of sentiment using a large-scale collection of photos and annotations from Flickr. Furthermore, we are the first to automatically generate sets of discriminative, sentiment-related visual features from a large Web 2.0 dataset.

2. VISUAL FEATURES

Digital images are represented as two dimensional arrays of pixels. Typically, images are processed in order to extract *feature vectors* that represent the images' content. The intuition behind the feature vector representation is that images with similar visual content should have similar vectors in the feature space.

Recently, it has become popular to transform image features into discrete elements or *terms*. These so-called "visual

MM'10, October 25–29, 2010, Firenze, Italy.

terms" are elegant because they enable image content to be described in almost the same way as a text document. Typically, an image is represented by a histogram of the number of occurrences of each distinct visual term [8]. This kind of approach is often called a "bag of visual terms" model, as the terms are treated completely independently of each other, regardless of their relative or absolute positioning in the image. It is worth noting that traditional histogram-based image features can also be considered to represent a visual bag of words as will be seen in the following sections. In this study, different types of progressively more powerful visual bag of word based image descriptions have been used.

Global and Local RGB Histogram.

The first type of feature selected is a 64-bin (4×4×4) global RGB histogram. This simple feature is also popular in content-based retrieval applications. Each bin of the histogram can be considered to represent a single visual term that in turn represents a certain range of similar colors. The size of a histogram bin represents the number of times a visual term occurs in the image.

The global color histogram completely discards all information about the layout of color in the image. If we first segment the image into blocks, and then calculate a color histogram for each block, it is possible to develop a rudimentary descriptor that describes the color at a rough location in the image. For our experiments, we split each image into 16 blocks (four evenly sized intervals along each axis) and calculated a 64-bin RGB histogram for each block. Each histogram bin at each of the 16 different locations was taken to be a different visual term, and, as before, the size of the bin represented the number of occurrences of the respective term.

SIFT-based Bag of Visual Terms.

Rather than globally using all of the pixels in the image, recent advancements in low-level image description have lead to approaches that are based on the detection and description of locally interesting regions. Whilst there have been many region detection approaches proposed, Lowe's SIFT [4] remains the most popular descriptor currently. Sivic and Zisserman [8] demonstrated how SIFT descriptors could be quantized into visual words. In their approach, the k-means clustering algorithm was used to find clusters of SIFT descriptors. The centroids of these clusters then became the 'visual' words representing the entire possible vocabulary. The vector quantizer then worked by assigning local descriptors to the closest cluster. The biggest problem of the k-means based approach is that it is computationally very expensive to create large numbers of clusters in high dimensional spaces. More recently, Nistér and Stewénius [5] proposed the use of hierarchical k-means to enable them to build visual vocabularies with over 1 million SIFT-based terms. In this work, we detect interest regions in each image by detecting peaks in a difference-of-Gaussian pyramid [4]. Hierarchical k-means is used to learn various sized codebooks for vector quantization. Because the datasets are so large, we cannot use all the SIFT features in the clustering, so we use uniform random sampling to select one million SIFT features from which we learn the clusters. Once the vector quantization has been applied and discrete visual terms have been extracted from an image, the whole image can be represented by a sparse histogram of visual term occurrences.

3. SENTIMENT CLASSIFICATION OF PHOTOS

In order to automatically classify the sentiment of images, we first assign numerical sentiment values to photos based on their textual metadata. We then make use of these sentiment values and the visual features described in the previous section to build machine learning models (SVMs) for the sentiment classification of photos. Furthermore, we show how discriminative, sentiment-related, visual features can be automatically obtained from a large set of images.

Sentiment Assignment.

SentiWordNet [2] is a lexical resource built on top of WordNet. WordNet is a thesaurus containing textual descriptions of terms and relationships between terms (for examples hypernyms and synonyms). WordNet distinguishes between different part-of-speech types (verb, noun, adjective, etc.) A *synset* in WordNet comprises all terms referring to the same concept (e.g. $\{car, automobile\}$). In SentiWordNet a triple of three *sentiment values* (pos, neg, obj) corresponding to positiveness, negativeness, or objectiveness are assigned to each WordNet synset. The sentiment values are in the range of $[0, 1]$ and sum up to 1 for each triple. For instance $(pos, neg, obj) = (0.875, 0.0, 0.125)$ for the term "good" or $(0.25, 0.375, 0.375)$ for the term "ill". Senti-values were partly created by human assessors and partly automatically assigned using an ensemble of different classifiers.

As the simplest sentiment value computation approach we use a dictionary of clearly positive and negative SentiWords (SW). With that, we can assign a positive (+1) sentiment value if the text representation only contains positive sentiment terms, and a negative (-1) sentiment value if it only contains negative sentiment terms (discarding all other cases). In a more complex computation, we consider the synsets of each term obtained from SentiWordNet that clearly conveys a sentiment. To this end, we filter out all synsets having a sentiment value below a certain *threshold* τ. The average of the remaining synsets represents the sentiment value of the respective term. Again, we filter out all terms that do not clearly express sentiment, i.e., do not yield threshold τ. The average over all remaining sentiment terms constitutes the sentiment value of the image; we call this value SentiWordNet-avg-τ (SWN-avg-τ).

Note that we derived our notion of sentiment polarity directly from SentiWordNet; a more fine grained distinction between different types of mood, opinion, and emotions is subject of our future work.

Classifying Sentiments.

In order to classify the sentiment of photos into categories "positive" or "negative", we use a supervised learning paradigm which is based on training items (photos in our case) that need to be provided for each category. Both training and test items, which are later given to the classifier, are represented as feature vectors. We construct these vectors from the color and SIFT visual-bag-of-words features described in Section 2. Photos associated with "positive" or "negative" sentiment are used to train a classification model, using probabilistic (e.g., Naive Bayes) or discriminative models (e.g., SVMs).

To obtain training sets of photos with "positive" or "negative" sentiment we transform the sentiment value into binary categories. Formally, we obtain a set $\{(\vec{p_1}, l_1), \ldots, (\vec{p_n}, l_n)\}$ of photo vectors $\vec{p_i}$ labeled by l_i with $l_i = 1$ if the sentiment value is positive ("positive" examples), $l_i = -1$ if the sentiment value is negative ("negative" examples). Linear support vector machines (SVMs) construct a hyperplane $\vec{w} \cdot \vec{x} + b = 0$ that separates the set of positive training examples from a set of negative examples with maximum margin. For a new, previously unseen, photo $\vec{p}$ the SVM merely needs to test whether it lies on the "positive" or the "negative" side of the separating hyperplane.

Discriminative Analysis of Visual Features.

For identifying the most characterizing visual features related to sentiment, we used the Mutual Information (MI) measure [10] from information theory which can be interpreted as a measure of how much the joint distribution of features X_i (color or SIFT features in our case) deviate from a hypothetical distribution in which features and categories ("positive" and "negative" sentiment) are independent of each other. We will show the outcome of this analysis conducted on a large set of Flickr images in the next section.

4. EXPERIMENTS

Flickr Image Dataset.

In our experiments, we aimed to gather images that 1) were likely to express sentiments, 2) contained textual metadata reflecting these sentiments. To this end, we crawled for images using query terms conveying a strong positive or negative sentiment. We used the top-1,000 positive / negative terms of the SentiWordNet dictionary as query terms, and for each query, gathered the first 5,000 images if available. In this way, we gathered over 586,000 images with many of them having a resolution of 1024 pixels at the larger dimensions, and most of them having at least 500 pixels. The image files of this dataset sum up to approx. 45 Gbyte in size. For all images, we also retrieved metadata including url, resolution, title, description, and a list of tags, comprising about 200 Mbyte[1].

Image Features.

In a second step, we extracted the Global Color Histogram (**GCH**), the Local Color Histogram (**LCH**), and the SIFT bag-of-visual-term features (**SIFT**) from all our images, and also generated any possible combination of the three. As a simple baseline, we further generated random features (**RND**) for each image (with feature values uniformly distributed among $[0, 1]$) to crosscheck classification on features that are clearly uncorrelated to the sentiment values of the images.

The GCH features were 64-dimensional, LCH features 1024-dimensional, SIFT features 3125-dimensional (based on a 5-level quantization tree with 5 splits per level), and for the random features we chose 1000 dimensions. All features except for the random ones were sparse, i.e., they contained many zero values. Color histograms and the random features contained continuous values in $[0, 1]$, whereas the SIFT bag-of-visual-term features were positive integer values.

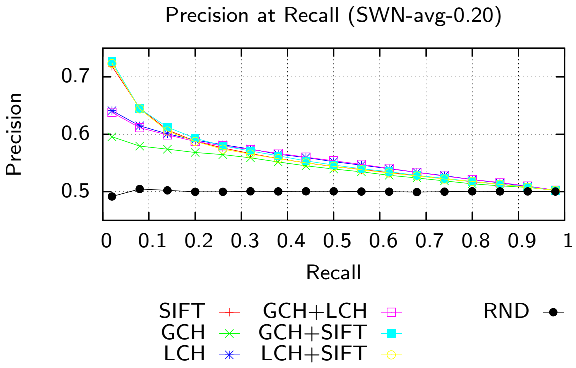

Figure 1: Classification results for sentiment assignments SW and SWN-avg-τ with $\tau = 0.20$ for training with 50,000 photos per category

Classification.

We performed different series of binary classification experiments of Flickr photos into the classes "positive sentiment" and "negative sentiment". From the labeled images and image features, we created training and test sets for classification. We randomly picked 50,000 images for each category (positive and negative sentiment). We trained an SVM model on these labeled data and tested on the remaining labeled data. For testing, we chose an equal number of positive and negative test images, with at least 35,000 of each kind. We used the SVMlight [3] implementation of linear support vector machines (SVMs) with standard parameterization in our experiments, as this has been shown to perform well for various classification tasks.

Our quality measure is the precision-recall curve with the positive category corresponding to positive sentences. The characteristic precision-recall curve is shown in Figure 1.

We can observe that for small recall values, precision values of up to 70% can be reached. Due to the challenging character of this task, for high recall values, the precision degrades down to the random baseline. With increasing number of training examples, we see that the precision in the lower regions increases as expected. The SIFT features perform best for SWN-avg with increasing τ; for the simple SW sentiment value computation, the color histogram features perform best. A combination of SIFT and color histogram features provides best performance for the SWN-avg sentiment values with high τ. This shows the sensitivity of the SIFT features for training with strongly positive or negative sentiment values.

Sentiment Correlated Features.

We used Mutual Information to discover image features that are most correlated with sentiments. For each feature, we computed the MI value with respect to the both positive and negative sentiment category. Figure 2 illustrates the 16 most discriminative visual features based on their MI value for the positive and negative categories (in rank order; top-to-bottom). Overall, the selected features mirror the features we inferred for the sample of classified images described in the previous section.

The GCH features for positive sentiment are dominated by earthy colors and skin tones. Conversely, the features for

[1] `http://www.l3s.de/~minack/flickr-sentiment`

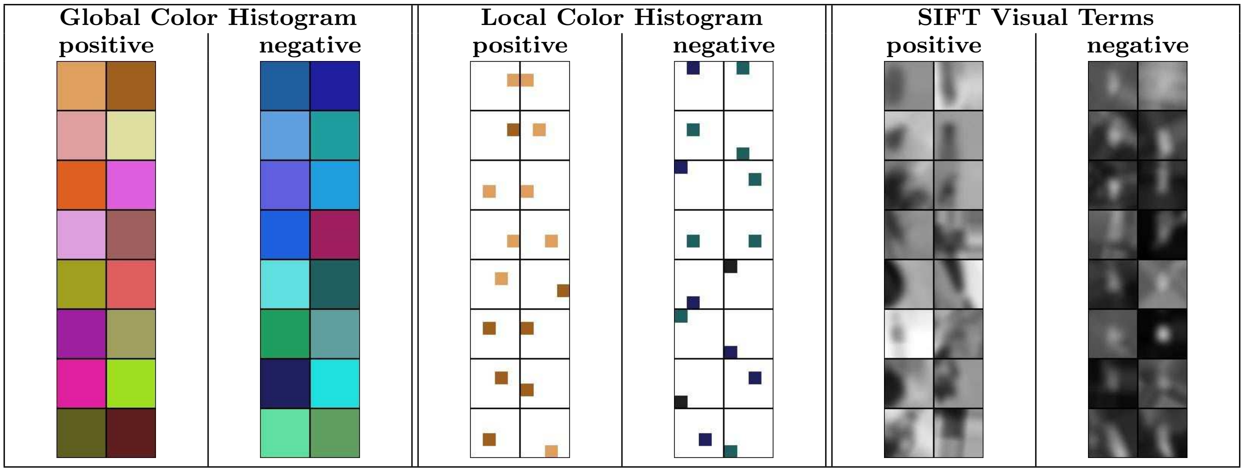

Figure 2: The 16 most predictive visual features for positive and negative sentiment, calculated using mutual information. The visualizations are ranked by decreasing MI score (shown left-right, top-bottom), so the patches depicted at the top have more predictive power than those further down. Depictions of the SIFT visual features have been extracted from interest regions of images in the dataset, and normalized for scale and rotation.

negative sentiment are dominated by blue and green tones. Interestingly, this association can intuitively be hypothesized because it mirrors human perception of warm (positive) and cold (negative) colors. The LCH features show the same trend as the GCH features — blue tones associated with negative sentiment, and skin tones associated with positive sentiment. In addition, the LCH features indicate that there is no bias to the spatial location in which pixels of the respective colors occur for positive sentiment. Negative features appear to be biased away from the far right of the image plane.

Results based on SIFT visual terms are difficult to interpret directly, but we can make some general observations. Looking at the most discriminative SIFT visual term features, the first observation is that the features within the two classes are remarkably similar, but there is a clear difference between the classes. The negative features seem dominated by a very light central *blob* surrounded by a much darker background. The positive features are dominated by a dark *blob* on the side of the patch (the patches have been normalized for rotation, so the dark blob could occur in any orientation in the image).

5. CONCLUSION

We conducted an in-depth analysis of the connection between different image features and sentiment on a sample consisting of more than half a million images from the social sharing site Flickr. Our studies revealed strong and intuitive dependencies between the sentiment values extracted from metadata and visual features based on color histograms and SIFT visual term representations. In our classification experiments, we further confirmed that visual features can, to a certain degree, help predicting the polarity of sentiment. We are aware that this work is just one of many steps; in order to make results applicable for real systems, a combination with additional information obtained through advanced text analysis techniques, considering complementary domain knowledge, and focusing on specific problems domains is of high practical importance.

Acknowledgments.

This work was supported by EU FP7 integration project LivingKnowledge (Contract No. 231126), and the LiveMemories project, graciously funded by the Autonomous Province of Trento (Italy).

6. REFERENCES

[1] C. Colombo, A. D. Bimbo, and P. Pala. Semantics in visual information retrieval. *IEEE MultiMedia*, 6:38–53, 1999.

[2] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. *LREC*, 6, 2006.

[3] T. Joachims. Making large-scale support vector machine learning practical. *Advances in kernel methods: support vector learning*, pages 169–184, 1999.

[4] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[5] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, pages 2161–2168, 2006.

[6] J. San Pedro and S. Siersdorfer. Ranking and classifying attractiveness of photos in folksonomies. In *WWW*, pages 771–780, New York, USA, 2009. ACM.

[7] S. Schmidt and W. G. Stock. Collective indexing of emotions in images. a study in emotional information retrieval. *Journal of the American Society for Information Sci. and Tech.*, 60(5):863–876, 2009.

[8] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, October 2003.

[9] W. Wei-ning, Y. Ying-lin, and J. Sheng-ming. Image retrieval by emotional semantics: A study of emotional space and feature extraction. In *SMC*, volume 4, pages 3534–3539, 2006.

[10] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *ICML*, pages 412–420, 1997.