Image and Collateral Text in Support of Auto-annotation and Sentiment Analysis

Pamela Zontone and **Giulia Boato**
University of Trento
Trento, Italy.
{zontone|boato}@disi.unitn.it

Jonathon Hare and **Paul Lewis**
University of Southampton
Southampton, United Kingdom
{jsh2|phl}@ecs.soton.ac.uk

Stefan Siersdorfer and **Enrico Minack**
L3S Research Centre
Hannover, Germany
{siersdorfer|minack}@l3s.de

Abstract

We present a brief overview of the way in which image analysis, coupled with associated collateral text, is being used for auto-annotation and sentiment analysis. In particular, we describe our approach to auto-annotation using the graph-theoretic dominant set clustering algorithm and the annotation of images with sentiment scores from SentiWordNet. Preliminary results are given for both, and our planned work aims to explore synergies between the two approaches.

1 Automatic annotation of images using graph-theoretic clustering

Recently, graph-theoretic approaches have become popular in the computer vision field. There exist different graph-theoretic clustering algorithms such as minimum cut, spectral clustering, dominant set clustering. Among all these algorithms, the Dominant Set Clustering (DSC) is a promising graph-theoretic approach based on the notion of a *dominant set* that has been proposed for different applications, such as image segmentation (Pavan and Pelillo, 2003), video summarization (Besiris et al., 2009), etc. Here we describe the application of DSC to image annotation.

1.1 Dominant Set Clustering

The definition of Dominant Set (DS) was introduced in (Pavan and Pelillo, 2003). Let us consider a set of data samples that have to be clustered. These samples can be represented as an undirected edge-weighted (similarity) graph with no self-loops $G = (V, E, w)$, where $V = 1, \ldots, n$ is the vertex set, $E \subseteq V \times V$ is the edge set, and $w : E \rightarrow \mathbb{R}^*_+$ is the (positive) weight function. Vertices in G represent the data points, whereas edges represent neighborhood relationships, and finally edge-weights reflect similarity between pairs of linked vertices. An $n \times n$ symmetric matrix $A = (a_{ij})$, called affinity (or similarity) matrix, can be used to represent the graph G, where $a_{ij} = w(i, j)$ if $(i, j) \in E$, and $a_{ij} = 0$ if $i = j$. To define formally a Dominant Set, other parameters have to be introduced. Let S be a non-empty subset of vertices, with $S \subseteq V$, and $i \in S$. The (average) weighted degree of i relative to S is defined as:

$$\text{awdeg}_S(i) = \frac{1}{|S|} \sum_{j \in S} a_{ij}$$

where $|S|$ denotes the number of elements in S. It can be observed that $\text{awdeg}_{\{i\}}(i) = 0$ for any $i \in V$. If $j \notin S$ we can define the parameter $\phi_S(i, j) = a_{ij} - \text{awdeg}_S(i)$ that is the similarity between nodes j and i with respect to the average similarity between node i and its neighbors in S. It can be noted that $\phi_{\{i\}}(i, j) = a_{ij}$, for all $i, j \in V$ with $i \neq j$. Now, if $i \in S$, the weight $w_S(i)$ of i relative to S is:

$$w_S(i) = \begin{cases} 1 & \text{if } |S| = 1 \\ \sum_{j \in S \setminus \{j\}} \phi_{S \setminus \{i\}}(j, i) w_{S \setminus \{i\}}(j) & \text{otherwise.} \end{cases}$$

This is a recursive equation where to calculate $w_S(i)$ the weights of the set $S \setminus \{i\}$ are needed. We can deduce that $w_S(i)$ is a measure of the overall similarity between the node i and the other nodes in $S \setminus \{i\}$, considering the overall similarity among the nodes in $S \setminus \{i\}$. So, the total weight of S can be defined as:

$$W(S) = \sum_{i \in S} w_S(i).$$

A non-empty subset of vertices $S \subseteq V$ such that $W(T) > 0$ for any non-empty $T \subseteq S$ is defined as a *dominant set* if the following two conditions

are satisfied: 1. $\forall i \in S$, $w_S(i) > 0$; and 2.$\forall i \notin S$, $w_{S\cup\{i\}}(i) < 0$. These conditions characterize the internal homogeneity of the cluster and the external inhomogeneity of S. As a consequence of this definition, a dominant set cluster can be derived from a graph by means of a quadratic program (Pavan and Pelillo, 2003). Let $\mathbf{x}$ be an n-dimensional vector, where n is the number of vertices of the graph and its components indicate the presence of nodes in the cluster. Let A be the affinity matrix of the graph. Let us consider the following standard quadratic program:

$$\begin{array}{c} \max f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} \\ s.t. \quad \mathbf{x} \in \Delta \end{array} \tag{1}$$

where $\Delta = \{\mathbf{x} \geq 0 \text{ and } e^T\mathbf{x} = 1\}$ is the standard simplex of $\mathbb{R}^n$. If a point $\mathbf{x}^* \in \Delta$ is a local maximum of f, and $\sigma(\mathbf{x}*) = \{i \in V : x_i^* > 0\}$ is the support of $\mathbf{x}^*$, it can be shown that the support $\sigma(\mathbf{x}*)$ is a dominant set for the graph. So, a dominant set can be derived by solving the equation (1). The following iterative equation can be used to solve (1):

$$x_i(t+1) = x_i(t)\frac{(A\mathbf{x}(t))_i}{\mathbf{x}(t)^T A\mathbf{x}(t)}$$

where t denotes the number of iterations. To summarize the algorithm, a dominant set is found and removed from the graph. A second dominant cluster is extracted from the remaining part of the graph, and so on. This procedure finishes when all the elements in the graph have been assigned to a cluster.

1.2 Image annotation using DSC

Here we present an approach to automatically annotate images using the DSC algorithm. In the initialization phase (training) the image database is split into L smaller subsets, corresponding to the different image categories or visual concepts that characterize the images in the database. In this process only tags are exploited: an image is included in all subsets corresponding to its tags. Given a subset l, the corresponding affinity matrix A_l is calculated and used by the DSC algorithm. Following (Wang et al., 2008), the elements of the affinity matrix $A_l = (a_{ij})$ are defined as $a_{ij} = \mathrm{e}^{-w(i,j)/r^2}$ where $w(i,j)$ represents the similarity function between images i and j in the considered subset l, and $r > 0$ is the scaling factor used as an adjustment function that allows the control of clustering sensitivity. We use the MPEG.7 descriptors (Sikora, 2001) as features for computing the similarity between images. Following the DSC approach, we can construct all clusters of subset l with similar images, and associate them with the tag of subset l.

In the test phase, a new image is annotated associating to it the tag of the cluster that best matches the image. To do this, we use a decision algorithm based on the computation of the MSE (Mean Square Error), where for each cluster we derive a feature vector that represents all the images in that cluster (e.g., the average of all the feature vectors). The tag of the cluster with smaller MSE is used for the annotation.

For our experiments, we consider a subset of the Corel database, that consists of 4287 images in 49 categories ($L = 49$). The 10% of images in each category have been randomly selected from the database and used only for testing. In Figure 1 we report the annotation accuracy results obtained on 15 different classes with optimal parameter $r = 0.2$. For some classes the accuracy is very high, whereas for others the accuracy is very low (under 30%). The total annotation accuracy considering all the 49 classes is roughly 69%.

In a second set of experiments we consider a set of 6531 images from the MIR Flickr database (Huiskes and Lew, 2008), where each image is tagged with at least one of the chosen 30 visual concepts ($L = 30$). Images are characterized by multiple tags associated to them, thus an image is included in all the corresponding subsets. For testing we use 875 images. To evaluate the annotation accuracy we compare the automatically associated tag with the user defined tags of that image. In Figure 1 we report the annotation accuracy obtained for the 30 different categories, with the optimal parameter $r = 0.2$. The total annotation accuracy is about 87%.

Further simulations are in progress to evaluate the accuracy of multiple tags that can be associated to the test set in the MIR Flickr database. Indeed, our idea is to annotate the images considering the other common tags of the images belonging to each cluster.

2 Annotating Sentiment

In the previous section we were concerned with annotating images with visual concepts, typically object names or descriptors. A separate strand of

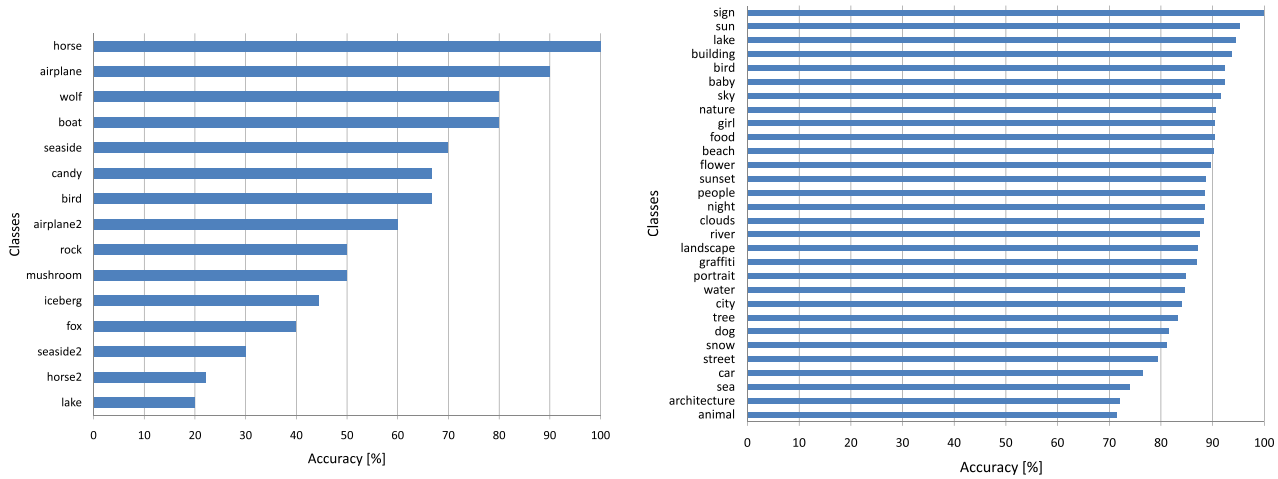

Figure 1: Annotation accuracy for 15 classes of the Corel database (left) and for 30 classes of the MIR Flickr database (right).

our work is concerned with opinion analysis in multimedia information and the automatic identification of sentiment. The study of image indexing and retrieval in the library and information science fields has long recognized the importance of sentiment in image retrieval (Jörgensen, 2003; Neal, 2006). It is only recently however, that researchers interested in automated image analysis and retrieval have become interested in the sentiment associated with images (Wang and He, 2008).

To date, investigations that have looked at the association between sentiment and image content have been limited to small datasets (typically much less than 1000) and rather specific, specially designed image features. Recently, we have started to explore how sentiment is related to image content using much more generic visual-term based features and much larger datasets collected with the aid of lexical resources such as SentiWordNet.

2.1 SentiWordNet and Image Databases

SentiWordNet (Esuli and Sebastiani, 2006) is a lexical resource built on top of WordNet. WordNet (Fellbaum, 1998) is a thesaurus containing textual descriptions of terms and relationships between terms (examples are hypernyms: "car" is a subconcept of "vehicle" or synonyms: "car" describes the same concept as "automobile"). 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 *senti-values* (pos, neg, obj) (corresponding to positive, negative, or rather neutral sentiment flavor of a word respectively) are assigned to each WordNet synset (and, thus, to each term in the synset). The senti-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 (see (Esuli, 2008) for an evaluation of these methods).

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. Recently in the image analysis community it has become popular to use Flickr as a resource for building datasets to experiment with. We have been exploring how we can crawl Flickr for images that have a strong (positive or negative) sentiment associated with them. Our initial explorations have been based around crawling Flickr for images tagged with words that have very high positive or negative sentiment according to their SentiWordNet classification.

Our image dataset has been refined by assigning an overall sentiment value to each image based on its textual metadata and discarding images with low overall sentiment. At the simplest level we use a dictionary of clearly positive and negative SentiWords, with which we assign a positive (+1) sentiment value if the text representation only con-

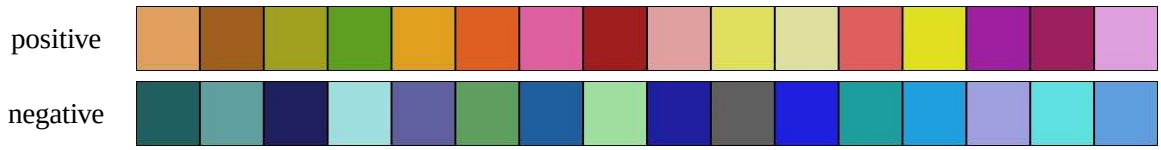

Figure 2: Top 16 most discriminative colours (from left to right) for positive and negative sentiment classes.

tains positive sentiment terms, and a negative (-1) sentiment value if it only contains negative sentiment terms. We discarded images with neither a positive nor negative score. Currently we are also exploring more powerful ways to assign sentiment values to images.

2.2 Combining Senti-values and Visual Terms

In the future we intend to exploit the use of techniques such as the one described in Section 1.2 in order to develop systems that are able to predict sentiment from image features. However, as a preliminary study, we have performed some small-scale experiments on a collection of 10000 images crawled from Flickr in order to try and see whether a primitive visual-bag-of-terms (Sivic and Zisserman, 2003; Hare and Lewis, 2005) can be associated with positive and negative sentiment values using a linear Support Vector Machine and Support Vector Regression. The visual-term bag-of-words for the study was based upon a quantisation of each pixel in the images into a set of 64 discrete colours (i.e., each pixel corresponds to one of 64 possible visual terms). Our initial results look promising and indicate a considerable correlation between the visual bag-of-words and the sentiment scores.

Discriminative Analysis of Visual Features. In our small-scale study we have also performed some analysis in order to investigate which visual-term features are most predictive of the positive and negative sentiment classes. For this analysis we have used the Mutual Information (MI) measure (Manning and Schuetze, 1999; Yang and Pedersen, 1997) from information theory which can be interpreted as a measure of how much the joint distribution of features (colour-based visual-terms in our case) deviate from a hypothetical distribution in which features and categories ("positive" and "negative" sentiment) are independent of each other.

Figure 2 illustrates the 16 most discriminative colours for the positive and negative classes. The dominant visual-term features for positive sentiment are dominated by earthy colours and skin tones. Conversely, the features for negative sentiment are dominated by blue and green tones. Interestingly, this association can be explained through intuition because it mirrors human perception of warm (positive) and cold (negative) colours.

Currently we are working on expanding our preliminary experiments to a much larger image dataset of over half a million images and incorporating more powerful visual-term based image features. In addition to seeking improved ways of determining image sentiment for the training set we are planning to combine the dominant set clustering approach to annotation presented in Section 1.2 with the sentiment annotation task of this section and compare the combined approach with other state of the art approaches as a step towards achieving robust image sentiment annotation.

3 Conclusions

The use of dominant set clustering as a basis for auto-annotation has shown promise on image collections from both Corel and from Flickr. We have also shown how that visual-term feature representations show some promise as indicators of sentiment in images. In future work we plan to combine these approaches to provide better support for opinion analysis of multimedia web documents.

Acknowledgments

This work was supported by the European Union under the Seventh Framework Programme (FP7/2007-2013) project LivingKnowledge (FP7-IST-231126), and the LiveMemories project, graciously funded by the Autonomous Province of Trento (Italy). The authors are also grateful to the creators of Flickr for providing an API that can be used in scientific evaluations and the broader Flickr community for making images and metadata available.

References

D. Besiris, A. Makedonas, G. Economou, and S. Fotopoulos. 2009. Combining graph connectivity and dominant set clustering for video summarization. *Multimedia Tools and Applications*, 44 (2):161–186.

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

Andrea Esuli. 2008. *Automatic Generation of Lexical Resources for Opinion Mining: Models, Algorithms and Applications*. PhD in Information Engineering, PhD School "Leonardo da Vinci", University of Pisa.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Jonathon S. Hare and Paul H. Lewis. 2005. On image retrieval using salient regions with vector-spaces and latent semantics. In Wee Kheng Leow, Michael S. Lew, Tat-Seng Chua, Wei-Ying Ma, Lekha Chaisorn, and Erwin M. Bakker, editors, *CIVR*, volume 3568 of *LNCS*, pages 540–549, Singapore. Springer.

Mark J. Huiskes and Michael S. Lew. 2008. The MIR Flickr Retrieval Evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA. ACM.

Corinne Jörgensen. 2003. *Image Retrieval: Theory and Research*. Scarecrow Press, Lanham, MD.

C. Manning and H. Schuetze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Diane Neal. 2006. *News Photography Image Retrieval Practices: Locus of Control in Two Contexts*. Ph.D. thesis, University of North Texas, Denton, TX.

M. Pavan and M. Pelillo. 2003. A new graph-theoretic approach to clustering and segmentation. *IEEE Conf. Computer Vision and Pattern Recognition*, 1:145–152.

Thomas Sikora. 2001. The mpeg-7 visual standard for content description - an overview. *IEEE Trans. Circuits and Systems for Video Technology*, 11 (6):262–282.

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

Weining Wang and Qianhua He. 2008. A survey on emotional semantic image retrieval. In *ICIP*, pages 117–120, San Diego, USA. IEEE.

M. Wang, Z. Ye, Y. Wang, and S. Wang. 2008. Dominant sets clustering for image retrieval. *Signal Processing*, 88 (11):2843–2849.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *ICML*, pages 412–420.