

Automatic Image Annotation using Visual Content and Folksonomies

Roland Mörzinger¹, Robert Sorschag¹, Georg Thallinger¹ and Stefanie Lindstaedt²

¹ Joanneum Research, Institute of Information Systems and Information Management
Steyrergasse 17, 8010 Graz, Austria

² Know-Center, Inffeldgasse 21a/II, 8010 Graz

Abstract. Automatic image annotation is an important and challenging task when managing large image collections. This paper describes techniques for automatic image annotation by taking advantage of collaboratively annotated image databases, so called visual folksonomies. Our approach applies two techniques based on image analysis: Classification annotates images with a controlled vocabulary while tag propagation uses user generated, folksonomic annotations and is therefore capable of dealing with an unlimited vocabulary. Experiments with a pool of Flickr images demonstrate the high accuracy and efficiency of the proposed methods in the task of automatic image annotation.

1 Introduction

With the prevalence of personal digital cameras, more and more images are hosted and shared on the Web. Systems organizing and locating images in these image databases heavily depend on textual annotations of images. For example, major existing image search engines like Google, Yahoo! and MSN rely on associated text in the web page, file names, etc. Naturally, the value of image databases immensely grows with rich textual annotations.

Image annotation or image tagging, is the process by which metadata is added to a digital image in the form of captioning or keywords. For this task humans interpret an image using their background knowledge and the capability of imagination. Therefore humans are able to annotate concepts which are not captured in an image itself. It is worth to note that the human labeling process is subjective and may therefore lead to ambiguous annotations especially in the absence of a fixed vocabulary. This issue can be addressed by making use of a shared annotation vocabulary. The practice of collaboratively creating and managing annotations - known as folksonomic tagging in the context of Web2.0 - aims at facilitating sharing of personal content between users.

A widely used website that relies on folksonomies as main method for the organization of their images is Flickr ¹. Flickr has a repository that is quickly approaching 1.5 billion images (as of August 2007) and growing. The large amount of information

¹ www.flickr.com

provided by such databases of annotated images is in the following referred to as *visual folksonomy*.

Every time a new image is added to an image database, it has to be annotated (or tagged) for the purpose of search. In order to avoid having the user to do expensive and time-consuming manual annotation of *untagged* images, automatic image annotation is highly desirable. In this paper, we present techniques for automatic image annotation exploiting visual content and existing folksonomies.

The following sections are organized as follows: In Section 2 we give an overview on related work and in Section 3 we describe our twofold approach for automatic image annotation, including a description of the content-based features we are using. Evaluation of experiments and results are shown in Section 4. A conclusion is in Section 5.

2 Related Work

Automatic image annotation is a challenging task which has not been solved satisfactorily for many real-world applications. Existing solutions cover only small domains and they usually work with a very limited vocabulary set. While there has been work on content-based image retrieval and object recognition over the last decades, see survey papers [2–4], work on automatic image annotation is a relatively new field.

One approach for automatic image annotation is the formulation as classification problem, where e.g. supervised learning by Support Vector Machines (SVMs) can be used to classify images and image parts to a number of concepts [6]. Another approach is to look at the probability of words associated with image features [5, 17].

The work in [10] proposes an image annotation technique that uses content-based image retrieval to find visually similar images from the Web and textual information that is associated with these images to annotate the query image. A similar approach is described in [12] where SIFT features [9] are utilized to retrieve similar images and to map keywords directly to these image descriptors. In [8] the Flickr image database is used to investigate collective annotations.

Some web-based image search engines use part of these techniques and operate on a pool of Flickr images. 'Flickr suggestions'³ combines regular text-based and content-based search, 'beholdsearch'⁴ includes a search for a number of predefined and trained high level concepts, 'retrievr'⁵ provides query-by-sketch using multiresolution wavelet decompositions. 'ALIPR'⁶[10] suggests automatic generated annotations for any on-line image specified by its URL.

3 Automatic Image Annotation

In our approach for automatic image annotation, we aim at taking advantage of the large amount of information provided by visual folksonomies. The visual content of untagged

³ <http://www.imgseek.net>

⁴ <http://www.beholdsearch.com>

⁵ <http://labs.systemone.at/retrievr>

⁶ <http://www.alipr.com>

images is a valuable source of information allowing cross-linking of these images to the images of the visual folksonomy. Given the high computational cost of content-based image analysis, efficient methods are required for enabling annotation in real-time.

In the context of automatic image annotation, we consider techniques that work in *real-time* to be fast enough to let the user work in an acceptable way, e.g. with a guaranteed response time of less than one second.

Our approach to automatic image annotation is split into two strategies

- automated image classification by off-line supervised learning of concepts from a folksonomy and
- tag propagation from visually similar images

Figure 1 shows the basic architecture incorporating the two strategies and related components. In the following the content-based features we are using and the two strategies are described in detail.

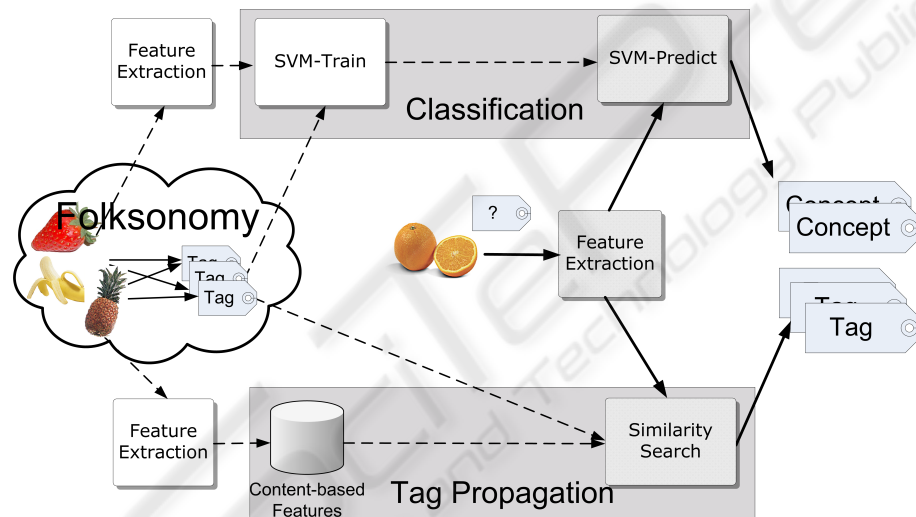


Fig. 1. Basic system architecture. Dashed lines indicate the semi-automatic off-line process; solid lines and shaded boxes depict components used on-line for automatic image annotation.

3.1 Content-based Features

We use three different MPEG-7 color features [7] and two texture features. They are computed globally to ensure that the image annotation is fast and scales with large visual folksonomies.

MPEG-7 color features. *Color Layout* describes the spatial distribution of colors. It is computed by dividing the image into 8x8 blocks and deriving the average value for each block. After computation of DCT and encoding, a set of low frequency DCT components is selected.

Dominant Color consists of a small number of representative colors, the fraction of the image represented by each color cluster and its variance. We use three dominant colors extracted by mean shift color clustering [15].

Color Structure captures both, color content and information about the spatial arrangement of the colors. Specifically, we compute a 32-bin histogram that counts the number of times a color is present in an 8x8 windowed neighborhood, as this window progresses over the image rows and columns.

Texture features. *Gabor Energy* is computed by filtering the image with a bank of orientation and scale sensitive filters and calculating the mean and standard deviation of the filter output in the frequency space. We applied a fast recursive gabor filtering [1] for 4 scales and 6 orientations.

Orientation Histograms are computed by dividing an image into a certain number of subregions and by accumulating a 1-D histogram of gradient directions of every subregion. Specifically, we use 4x4 slightly overlapping subregions with 8 histogram bins which makes it similar to SIFT features [9] and Histogram of Oriented Gradients [13].

3.2 Image Annotation by Classification

Automatic image annotation with concepts of a folksonomy is realized by supervised learning of classifiers. For this purpose, we selected a number of concepts by investigating high-level tags of the folksonomy. Afterwards a training set characteristic for the target concepts has to be generated. In our work this was accomplished in a *supervised* manner. Concepts are chosen manually based on the most frequent tags amongst the ones which are at the same semantic level and non-overlapping, i.e. tags which are not umbrella terms or subtopics of another tags. The training sets can initially be automatically compiled by text-search, however it is also necessary to manually exclude (visually) unrelated images for better classification results. Further input features have to be selected and their representation determined. Based on feature selection experiments, we set the feature vector as a concatenation of the feature values extracted for ColorLayout, DominantColor, ColorStructure and GaborEnergy. As learning algorithm we used a multi-class support vector machine (SVM). As argued in [16], early normalized fusion is a good choice for low- or intermediate-level features as above. We apply early feature fusion, statistical feature normalization (adjusting each feature dimension to have zero mean and unit standard deviation) and scaling of the feature value range to -1 and 1. The SVM model is trained in a one-against-rest fashion using a C++ implementation based on LibSVM [14].

3.3 Tag Propagation using CBIR

We use content-based image retrieval for automatic image annotation. In this approach untagged images are compared to the images of a visual folksonomy in order to obtain several tags of visually similar images.

Similar images are retrieved using the two image features ColorLayout and Orientation Histograms. These features capture both color and texture properties effectively

and they can be used for efficient image matching using the similarity search approach outlined below. In an off-line learning process the two features are computed for the images from a visual folksonomy. These image features are stored in a relational database along with the user tags. In order to produce tags for an untagged image, the features of ColorLayout and Orientation Histograms are computed and an image search is started to find visually similar images. The image search uses both features separately to find similar images before the resulting images are combined. As metric for ColorLayout similarity a non-linear distance measure [7] is used with the help of a hybrid tree to achieve fast computation times. The Orientation Histograms are compared with the Euclidean distance and break-up conditions to accelerate the process. After a merged set of visually similar images is generated, candidate tags for the untagged image could be selected from the tags of visually similar images, see Figure 2. We use tags that occur multiple times or tags with relationships to other tags (synonym, hypernyms, and hyponyms) in this set of tags from similar images. The available parameters for this method are the maximum number of tags to propagate and the number of visually similar images to use.







| Input | Similar images result | Propagated tags | Classified concept | Flickr tags |
|---|---|--|--------------------|--|
|  |  | fruit, food, fruits, vegetables, market, yellow, macro, 2006 | banana | FoodTrails, homestyle, horn,bananas, fruits, wet, market, Punggol 21, a favourite |
|  |  | red, fruits, fruit, food, market, strawberries, macro, tomato, summer, berries | strawberry | pike place market, seattle, june, 2007, fruits, produce, red, raspberries, perfect looking, fruit, sosio's, stand, food, m-p-g |
|  |  | fruits, food, fruit, apple, red, macro, black and white, vegetable, delicious, blueberries | blueberry | friendship, flickrfriends, comments, photos, holidays, happydays, summer, fruits, more, blackberries, fairytale, allthebest, Godblessyou, loveyouall, [+ 16 more] |

Fig. 2. Automatic image annotation examples. For three input images, the three visually most similar images resulting from similarity search, the propagated tags, the classified concept and the user generated tags of Flickr are shown.

4 Experiments, Results and Evaluation

This section describes the experimental setup, the used data sets and the results of our two strategies. The set of images used for our experiment is taken from the Flickr pool *Fruit&Veg*¹. The training set is composed of around 15,000 images and their tags from a snapshot of this pool from March, 2007. The corresponding test set is made up of 30 relevant images per concept (210 in total), which were added to Flickr afterwards (till August 2007). The performance of the two strategies is measured separately using precision and recall.

¹ <http://flickr.com/groups/fruitandveg>

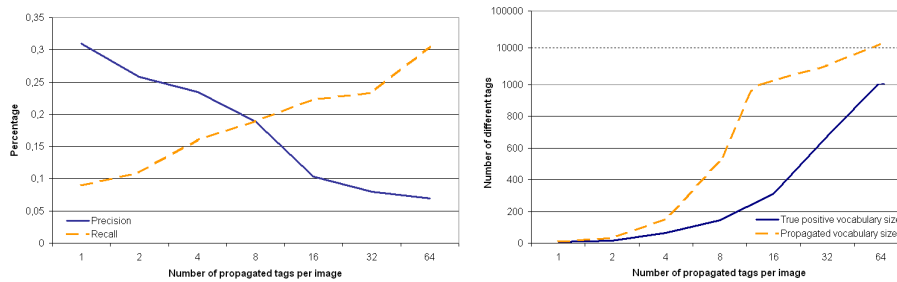


Fig. 3. Tag propagation results computed on a set of 15,000 images. Precision / recall (left) and propagated / matched vocabulary sizes (right) are plotted against the number of propagated tags (logarithmic scale).

For classification, we considered the following concepts: *banana*, *blueberry*, *kiwi*, *orange*, *strawberry*, *raspberry* and the negative class *other-fruits*. In this setup we intentionally included visually rather distinct concepts as well as visually similar concepts like raspberry and strawberry. For each of the concepts to be predicted, we subjectively selected a set of the 50 most relevant images from the training set, that is, images where the concept was most clearly predominant. After experimenting with different kernels and parameter tuning by applying a grid search using 5-fold cross-validation, we selected the RBF kernel with gamma parameter equal to 0.125 and the trade-off parameter C equal to 32.

Table 1. Classifier results computed on the test set.

| | precision | recall |
|--------------|-----------|--------|
| banana | 0.57 | 0.70 |
| blueberry | 0.96 | 0.80 |
| kiwi | 0.75 | 0.50 |
| orange | 0.85 | 0.77 |
| raspberry | 0.88 | 0.47 |
| strawberry | 0.64 | 0.90 |
| other-fruits | 0.58 | 0.83 |

Table 1 shows the classifier results applied on the test set. Generally, high precision was achieved, with up to 95% for the blueberry concept. The classifier produced fewer positive predictions for the concepts banana, kiwi and other-fruits. The precision for banana and kiwi is impaired by false positives for other-fruits. The high recall of 90% for strawberry sacrifices the recall of the visually similar raspberry. As can be seen in Table 2, 50% percent of images with strawberry were misclassified to raspberry. Issues with learning separations from the negative class other-fruits, reduced the precision for some concepts (13% to 23% false negatives). We observe that the classifier performed very well in discriminating blueberries against most of the other concepts, that is, there was no confusion between blueberry and the concepts of kiwi, orange, raspberry and strawberry in any sense. In total, we achieved 71% accuracy (percentage of predictions that are correct).

Table 2. Classifier confusion matrix computed on the test set.

| Predicted \ True | banana | blueberry | kiwi | orange | raspberry | strawberry | other-fruits |
|------------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------|
| banana | <u>0.70</u> | 0.00 | 0.03 | 0.03 | 0.00 | 0.00 | 0.23 |
| blueberry | 0.07 | <u>0.80</u> | 0.00 | 0.00 | 0.00 | 0.00 | 0.13 |
| kiwi | 0.23 | 0.00 | <u>0.50</u> | 0.07 | 0.00 | 0.00 | 0.20 |
| orange | 0.20 | 0.00 | 0.00 | <u>0.77</u> | 0.00 | 0.00 | 0.03 |
| raspberry | 0.00 | 0.00 | 0.03 | 0.00 | <u>0.47</u> | 0.50 | 0.00 |
| strawberry | 0.00 | 0.00 | 0.00 | 0.03 | <u>0.07</u> | <u>0.90</u> | 0.00 |
| other-fruits | 0.03 | 0.03 | 0.10 | 0.00 | 0.00 | 0.00 | <u>0.83</u> |

In the evaluation of the automatic tag propagation technique, the training set builds the reference database for similarity search. As in [10], we propagate tags for all 15,000 images using the (residual) images as database for similarity search and compare the propagated tags with the user tags. This evaluation shows the percentage of propagated tags which have also been generated by the user. Propagated tags are considered as false positives if they are not contained in the user annotations, even if they describe the image content correctly.

The two diagrams in Figure 3 present results of tag propagation using the 50 most similar images. The diagram on the left shows precision (percentage of propagated tags contained in user tags) and recall (percentage of user tags retrieved by the automatic tag propagation). These two measures behave complementary when the number of propagated tags is increased. The propagation of 8 tags, a number that corresponds to the average number of tags per image (7.85) in the *Fruit&Veg* pool, leads to 19% precision and recall. The diagram on the right shows the absolute number of different tags (vocabulary size) generated with our approach. Although this experiment uses a rather small image pool, more than 10,000 different tags can be propagated. The dashed curve indicates that several thousand of different tags are propagated from the system when propagating 32 tags per image. The solid line shows the number of different tags which are propagated and user generated (true positives).

Moreover we have manually evaluated the tag propagation system with the test set of 210 images described above. For that purpose, 8 tags per image were propagated using the 50 most similar images. A manual investigation of the relevance of these tags shows that about 70% of the propagated tags are useful annotations (precision) and only the residual 30% are wrong annotations. The reason for the difference between the results shown in Figure 3 and the manual evaluation is that many propagated tags are actually judged as correct by humans while considered as false positives when compared only to the image tags from a single user. Examples of this dilemma are shown in Figure 2.

It is worth to note that the classification and tag propagation of an untagged image of size 512x512 takes 0.9 seconds on average, mainly spent for feature extraction.

5 Conclusions

We presented two strategies for automatic image annotation that build on existing knowledge from labeled image databases, so called visual folksonomies. On the one hand, we implemented a supervised learning approach for classification using a SVM and a controlled vocabulary. On the other hand, we presented a tag propagation system that uses content-based image retrieval for automatic annotation that works with an uncontrolled folksonomic vocabulary. The proposed approaches use state-of-the-art image features and compute tags in less than 1 second per image which enables real-world applications on top of large-scale image databases.

Experiments on a set of images from Flickr's pool Fruit&Veg show that classification works with a high precision of 71% on a limited set of target concepts. The behaviour of the tag propagation approach heavily depends on the amount of propagated tags per image. When only few tags are propagated, high precision but low recall is achieved, while propagating many tags leads to low precision and high recall.

Although the classification setup presented in this work only implements the prediction of one concept per image, it is planned to extend the approach and learn several binary classifier for each concept individually for annotating one image with multiple concepts. Future work might include the automatic definition of concepts and experiments using more general and larger visual folksonomies.

Acknowledgements

The authors would like to thank their colleagues Marcus Thaler and Werner Haas for their support and feedback. The Know-Center is funded by the Austrian Competence Center program Kplus under the auspices of the Austrian Ministry of Transport, Innovation and Technology (<http://www.ffg.at>), by the State of Styria and by the City of Graz. The research leading to this paper was partially supported by the European Commission under contract FP6-027026 (K-Space), FP6-027122 (SALERO) and FP6-045032 (SEMEDIA).

References

1. I.T. Young, L.J. van Vliet, M. van Ginkel: Recursive Gabor Filtering. ICPR '00: Proceedings of the Int. Conf. on Pattern Recognition, Vol. 50, Issue 11, Nov 2002, 2798 - 2805
2. D. Forsyth, J. Ponce.: *Computer Vision: A Modern Approach*. Prentice Hall, 2002.
3. Datta, Ritendra, Li, Jia, Wang, James Z.: Content-Based Image Retrieval - Approaches and Trends of the New Age. MIR '05: Proc. of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval, 2005
4. A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain: Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000
5. A. Yavlinsky, E. Schofield, S. M. R uger: Automated Image Annotation Using Global Features and Robust Nonparametric Density Estimation. Proc. of the 4th Int. Conf. on Image and Video Retrieval (CIVR), Vol. 3568, 507-517, July 2005

6. C. Cusano, G. Ciocca, R. Schettini: Image annotation using SVM. Proceedings Of Internet imaging IV, SPIE, 2003
7. B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada: MPEG-7 color and texture descriptors. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, 703-715, June 2001
8. B. Shaw: Learning from a Visual Folksonomy. Automatically Annotating Images from Flickr. May 2006
9. D.G. Lowe: Distinctive Image Features from Scale-Invariant Keypoints. Int. Journal of Computer Vision, 60(2), 91-110, 2004
10. X. Li, L. Chen, L. Zhang, F. Lin, W.Y. Ma: Image Annotation by Large-Scale Content-based Image Retrieval. Proc. of ACM Int. Conf. on Multimedia, Santa Barbara, USA, 2006
11. X. Wang, L. Zhang, F. Jing, W.Y. Ma: AnnoSearch: Image Auto-Annotation by Search. Proc. of the International Conference on Computer Vision and Pattern Recognition, New York, 2006
12. D.R. Hardoon, C. Saunders, S. Szedmak, J. Shawe-Taylor: A Correlation Approach for Automatic Image Annotation. Int. Conf. on Advanced Data Mining and Applications, Springer LNAI 4093, 681-692, 2006
13. N. Dalai, B. Triggs: Histograms of oriented gradients for human detection. Conf. on Computer Vision and Pattern Recognition, Vol. 1, 886-893, June 2005
14. C. Chang, C.J. Lin: LIBSVM : a library for support vector machines. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
15. D. Comaniciu, P. Meer: Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002.
16. S. Ayache, G. Quenot, J. Gensel: Classifier Fusion for SVM-Based Multimedia Semantic Indexing. European Conf. on Information Retrieval, 2007
17. J. Jeon, V. Lavrenko, R. Manmatha: Automatic image annotation and retrieval using cross-media relevance models. ACM SIGIR Conference on Research and Development in Information Retrieval, 119-126, 2003.



SciTech Publications
Science and Technology Publications