Automatic Image Annotation Using a Visual Dictionary Based on Reliable Image Segmentation

Christian Hentschel[1], Sebastian Stober[1], Andreas Nürnberger[1], and Marcin Detyniecki[2]

[1] Otto-von-Guericke-University, Magdeburg
[2] Laboratoire d'Informatique de Paris 6
chentsch@student.uni-magdeburg.de,
{nuernb,stober}@iws.cs.uni-magdeburg.de,
marcin.detyniecki@lip6.fr

Abstract. Recent approaches in Automatic Image Annotation (AIA) try to combine the expressiveness of natural language queries with approaches to minimize the manual effort for image annotation. The main idea is to infer the annotations of unseen images using a small set of manually annotated training examples. However, typically these approaches suffer from low correlation between the globally assigned annotations and the local features used to obtain annotations automatically. In this paper we propose a framework to support image annotations based on a visual dictionary that is created automatically using a set of locally annotated training images. We designed a segmentation and annotation interface to allow for easy annotation of the traing data. In order to provide a framework that is easily extendable and reusable we make broad use of the MPEG-7 standard.

1 Introduction

The increased use of digital technologies for production, processing and distribution of digital images within the last decade has led to a sudden rise of valuable information now stored within pictorial data. In order to be able to efficiently retrieve sought information from large-scale image collections two major approaches have been meanwhile established. They mainly differ in the way a query is formulated. *Content-based Image Retrieval* approaches try to find images semantically similar to a given query image example by comparing them on a low-level basis. The requirement of an initial query image, however, disqualifies these approaches in any retrieval scenario, since the availability of such an image would most likely already solve the retrieval task. Therefore, in *Annotation-based Image Retrieval* an image collection is searched based on a textual description of the depicted content. While this approach is best-suited in scenarios where the desired pictorial information can be efficiently described by means of keywords, it demands for translation of the depicted contents into a textual representation (*annotation*). Manual image annotation is a tedious and time-consuming task. Hence, *Automatic Image Annotation* attempts to automatically infer the textual description of an image based on a small set of manually annotated training images.

N. Boujemaa, M. Detyniecki, and A. Nürnberger (Eds.): AMR 2007, LNCS 4918, pp. 45–56, 2008.

Automatic Image Annotation (AIA) describes a supervised classification of pictorial data. Each image class contains images, which are semantically similar and thus have at least one annotation in common. Typically these annotations are simple keywords such as "tree" or "sky". Since an image usually can be provided with more than one annotation most images belong to more than one image class at the same time. A classifier trained on a small set of manually annotated images tries to assign an image to one or more classes. A training set should preferably provide a unique mapping between a textual annotation and the described semantic entities within the image. The mapping is represented by a *visual dictionary* (also: *visual codebook*) [3,11,25,30]. A classifier compares the entries or visual words with an unknown image. A successfully rediscovered visual word leads to a corresponding image classification (and thus annotation).

In [9] we introduced Safire – an annotation framework to integrate semantic and feature based information about the content of images that is based on the MPEG-7 Multimedia Content Description Interface[1]. Safire enables the combined use of low-level features and annotations to be assigned to arbitrary hierarchically organized image segments. In this paper we exploit this framework in the field of AIA: Based on the extraction of MPEG-7-compliant low-level features and manually assigned textual descriptions a visual dictionary is assembled, which is later used for automatic image annotation. The extensive use of the MPEG-7 standard allows for a straightforward capability of extending of the visual dictionary, even by using external tools. Moreover, the annotated images can be easily sought by other MPEG-7-compliant applications.

In the following sections, we first offer a brief overview on the field of Automatic Image Annotation and present to what extent our approach is different from these. The subsequent section will present the AIA extension of Safire. Section 4 will present some results we were able to achieve with our prototypical implementation. Section 5 finally summarizes this work and gives some ideas for future research.

2 Related Work

According to [10], Automatic Image Annotation can be regarded as a type of multi-class image classification. The major characteristics are a large number of classes (as large as the annotation vocabulary size) and a possibly small number of example images per class. Recent research concentrates on applying machine learning techniques to identify correlations of low-level image features (typically color and texture) and annotations used for training. Classification of new images is later based on the visual dictionary that we obtain at this learning stage.

Learning a correlation between global low-level image features computed on a per-image level and annotation data has been successfully applied in general scene classification (see e.g. [6,27,28,29]). These approaches provide good results for classifying images when applied to image classes whose discriminative visual

[1] The Moving Picture Experts Group (MPEG), `http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm`

properties are spread equally over the whole image surface. The classification of “city” and “landscape” images for instance provides good classification results since city scenes typically show strong vertical and horizontal edges, whereas landscape scenes tend to have edges randomly distributed in various directions.

When applied to visual object detection, however, global visual features often insufficiently represent the prominent objects that have been used to annotate the images. Hence, more recent approaches (e.g. [7, 26] propose an automatic segmentation step before the actual learning stage to identify real-world objects within the images. The general assumption is that feature computation based on a potentially strong segmentation better describes the visual objects, depicted in the image, than global features. However, since up to today no general and robust automatic segmentation algorithm has been presented, these approaches suffer from the typically low segmentation accuracy current algorithms provide on low-level images.

Partition-based approaches try to overcome this obstacle by decomposing the images into multiple regions of equal shape (patches). This can be seen as a weak segmentation, which tries not to capture the shape of a visual object, but to produce multiple regions per image each corresponding to a single depicted object. This will result in more redundancy (depending on the patch size), which will help to statistically detect a correlation between global labels and local patches. Lipson et al. [16] applied this concept to general scene classification by modelling templates of spatially aligned colored patches, to describe a specific scene. Each template is assigned a descriptive label and automatic image annotation is done by matching these templates with unclassified images.

Other approaches [13, 14, 18] apply probabilistic methods to capture the relationship between images and globally assigned labels. In their approach, images are divided into a grid of patches. Color and texture features are computed for each patch. A patch inherits all labels from the original image. Using data clustering, groups within the set of all patches are computed. Then, for each cluster center, the probability for the occurrence of all labels assigned to a patch in this cluster is derived. Labeling of an unknown image is performed on patch level. For each patch, color and texture features are computed. The probabilities for the words in the nearest cluster of each patch are combined. The most plausible words for the global image annotation are those with the highest overall probability.

A problem of all the presented approaches is a typically high correlation between different annotations. Various labels that often appear together within the training images can not be distinguished. For example, if the training images always depict sky- and tree-regions within an image together, those objects are hardly distinguishable using the presented statistical methods. Furthermore, in [29] it is argued that global annotations are more general than a pure region labeling and thus a semantic correspondence between labels and image regions does not necessarily exist (e.g. an image globally labeled “wild life” might depict regions for elephants as well as regions for tigers – deducing from both regions to a global label “wild life” is impossible with approaches that are based on color and texture computation only).

Among the first approaches of AIA was the FourEyes system presented in [21] and [17]. It divides all images in the database into patches of equal quadratic shape. By selecting a label name from a limited set of annotations the user indicates a patch to be a positive example for the chosen label, from which the system immediately infers annotations for other patches in the database. In contrast to the aforementioned approaches, learning an association between low-level features and global annotation data is not necessary here since the training images are annotated *locally*. This concept avoids the aforementioned problems that arise from label correlation and implies a strong correlation between annotations and image regions, however, at the expense of frequent user interactions.

A major drawback of all the presented approaches is that all of them apply different low-level descriptors as well as different segmentation and annotation schemes, which makes it inherently difficult to extend the applied visual dictionaries as well as reuse the (manually or automatically) annotated image data. Moreover, the presented approaches do not respect the subjectivity of human perception. Two different beholders of an image may come to different image descriptions and may identify different prominent visual objects.

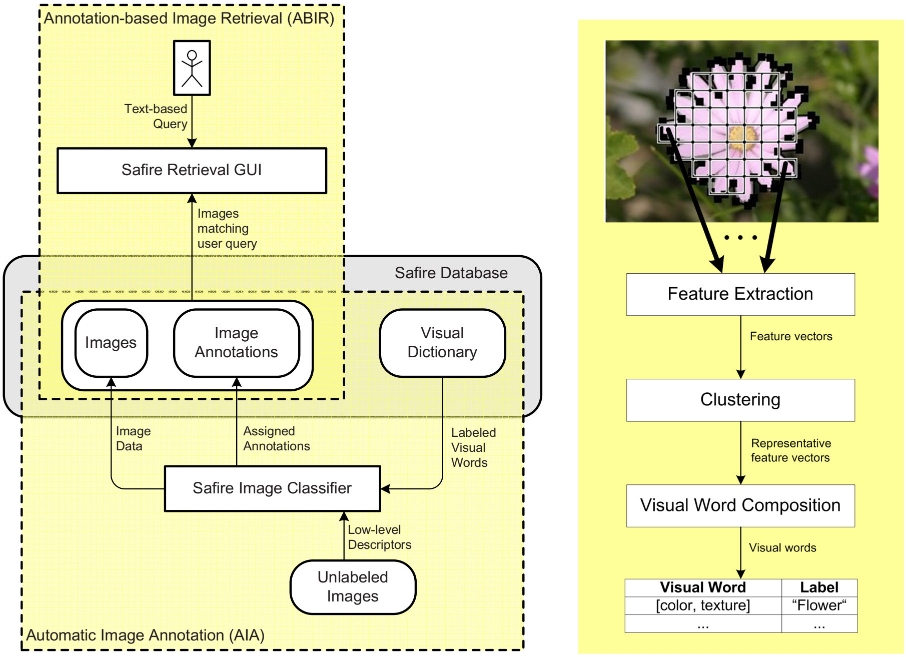
The system we present here is designed to allow for multiple image segmentations and annotations depending on the beholder. Moreover, the application of the MPEG-7 standard that was initially limited in [9] to region and annotation description has been extended and now covers all aspects of image description including the applied low-level image features. We therefore gain a high degree of extensibility and reusability

Moreover, in order to avoid a possible correlation between various annotations, we follow an approach similar the one briefly discussed above (for details see [21]). However, instead of designing an highly interactive system, we introduce a separate *manual* image segmentation step to construct a reliable visual dictionary while at the same time reducing the amount of user interactions required, once the dictionary has been created.

3 Automatic Image Annotation in Safire

As mentioned in the introduction, the prototypical implementation of the Safire framework as presented in [9] was extended by a component to support AIA and an Annotation-based Image Retrieval component to search digital color pictures based on textual descriptions. A general overview of the framework is given in Figure 1(a).

Generation of the visual dictionary in Safire is based on a locally annotated training set. An interface has been designed, to support manual, sloppy segmentation of a training image into regions of interest. These regions can be later assigned descriptive keywords by the user. Thus, each pixel in a training image can be unambiguously associated with a specific annotation. Segmentation and annotation data is stored in MPEG-7 compliant documents. These documents are used as training data for creating the visual dictionary.



(a) (b)

Fig. 1. Automatic Image Annotation in Safire: 1(a) shows the general system overview of the proposed framework. 1(b) depicts the process of creating the visual dictionary.

In order to respect the subjectivity of the human perception the annotation interface supports different "views" for the same image. Each user can create a specific image segmentation, with specific region annotations. The training dataset and thus the visual dictionary and the annotation vocabulary can be tailored to a specific user group.

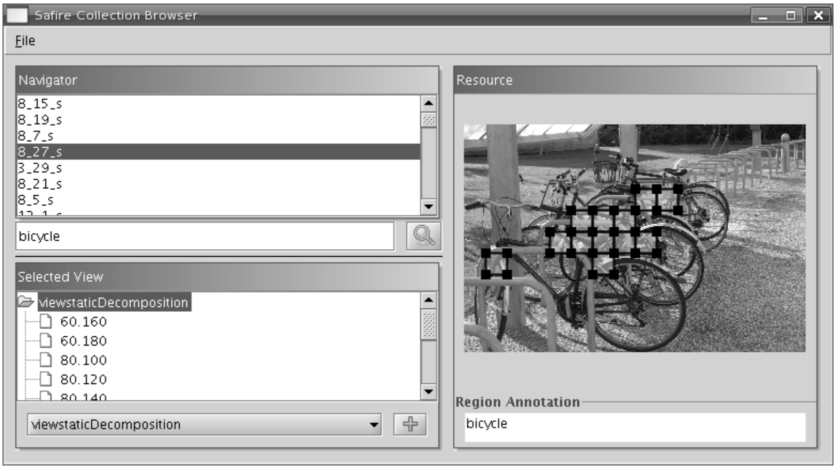
As manual image segmentation is a tedious task it should be performed only once to create a reliable dictionary. For labeling new images by matching visual words with image regions, manual segmentation should not be necessary. It was therefore decided to partition the manually derived segments into patches of equal size based on which the visual dictionary shall be created. Each patch is assigned directly the label of the surrounding image segment. The achieved local annotation can therefore be considered as highly reliable.

Image content is represented in a feature space that assembles the computed color and texture features of the derived patches. Color and texture are commonly applied characteristics of pictorial data in image retrieval [1]. Their strong advantage is a relatively low computational effort and full automatic extractability. The MPEG-7 standard comprises a number of different color and texture descriptors – each tailored to a specific application scenario. For a detailed overview see [19] and [2]. Among the color descriptors the Scalable Color Descriptor (SCD) was chosen. The SCD is a histogram-based descriptor and is recommended for retrieval of color pictures. The Homogeneous Texture Descriptor (HTD) was chosen as it has shown superior retrieval results in empirical evaluations for texture based image retrieval (see [20]). Both descriptors are computed in HSV space. While the SCD quantizes the HSV space for histogram computation, the HTD applies a Gabor filtering on the value component of the images.

Each derived feature vector describes a specific patch within the image and could be used directly for classifying unseen images. However, following [31], in order to reduce the amount of comparisons necessary for classification and to render the dictionary more robust to minor visual variances, the feature vectors of each annotation class are clustered using a simple k-means approach. The result of the clustering step can be seen as a reduction of all extracted feature vectors to a set of discriminative representatives for a specific annotation class. The computed cluster centers determine the visual words (see Fig. 1(b)). For each annotation entry in the visual dictionary a list of k describing representatives are assembled. The currently implemented classifier assumes a consistent number of visual words per annotation class. However, a varying parameter k is likewise imaginable. For example, for very homogenous classes a smaller set of representatives (a smaller values for k) might be sufficient, while very diverse classes where groups of similar objects are scattered in data space might require a bigger value for k.

New images to be labeled are statically decomposed into a regular grid of patches of the same size as the patches used in the dictionary. The automatic classification of new images in SAFIRE follows the k-nearest-neighbor approach. For each patch of the image to be classified, SCD and HTD feature vectors are computed and compared to the visual dictionary. Those k visual words are chosen, whose distance (Euclidean) is smallest to the patch to be classified. The class to which the new patch is to be assigned is derived based on a majority vote. The most frequent class is chosen. In order to avoid misclassification through visual objects that are not yet covered by the dictionary the classifier computes a confidence value for the proposed annotation. Annotation is performed only if more than 50% of the retrieved visual words agree on the same annotation class. Figure 2 shows the resulting sparse grid of patches that have been successfully labeled by the classifier. Please note that successfully here does not refer to a correct annotation in terms of human evaluation but rather to the fact that the computed confidence was above the threshold ($c(l) > 50\%$).

SAFIRE provides a simple user interface for retrieving images based on the automatically (and manually) assigned annotations. Fig. 2 shows the results of a request using the keywords "bicycle". Please note the small black squares representing the regions that have been successfully (s.a.) annotated by the classifier.



8_15_s
8_19_s
8_7_s
8_27_s
3_29_s
8_21_s
8_5_s

bicycle

viewstaticDecomposition
60.160
60.180
80.100
80.120
80.140

viewstaticDecomposition

Region Annotation: bicycle

Fig. 2. Results of a request in SAFIRE for images annotated by the keyword ”bicycle“. Black squares indicate credibly annotated region.

4 Results and Discussion

In order to minimize the effort required to create a dataset of locally annotated training data, we selected the Microsoft Research Cambridge (MSRC) dataset[2] to create the visual dictionary. The images in this collection have been manually segmented into visual objects and background and locally (that is objectwise) annotated. When compared to the typically used Corel or PASCAL datasets[3], the MSRC dataset content is rather small in size: it comprises 591 images depicting 23 different annotation classes. As a result, for some classes only very few example images were available (see Fig. 3). Two classes even had to be neglected due to the lack of sufficient example images. However, compared to the Corel or even the PASCAL dataset, the MSRC dataset provides a precise correlation between annotations and local visual features.

For each image a MPEG-7 compliant document containing region information, annotations and low-level image features was created. The documents were split into 60% training and 40% testing data used to estimate the annotation accuracy of the proposed system. The number of visual words per annotation class was restricted to 30 representatives in the clustering process. This decision is mainly due to the rather small dataset size and in order to reduce the computation effort required for image classification. The clustering was repeated to reduce the impact of the initial choice of cluster centers.

[2] Computer Vision at Microsoft Research Cambridge – Object class recognition, `http://research.microsoft.com/vision/cambridge/recognition/`

[3] The PASCAL Object Recognition Database Collection, `http://www.pascal-network.org/challenges/VOC/databases.html`

Annot. Class	# Patches
aeroplane	479
bicycle	761
bird	448
boat	320
body	691
book	2081
building	4231
car	1292
cat	583
chair	704
cow	1062
dog	666
face	713
flower	1258
grass	7215
road	3313
sheep	942
sign	963
sky	3795
tree	3342
water	2957

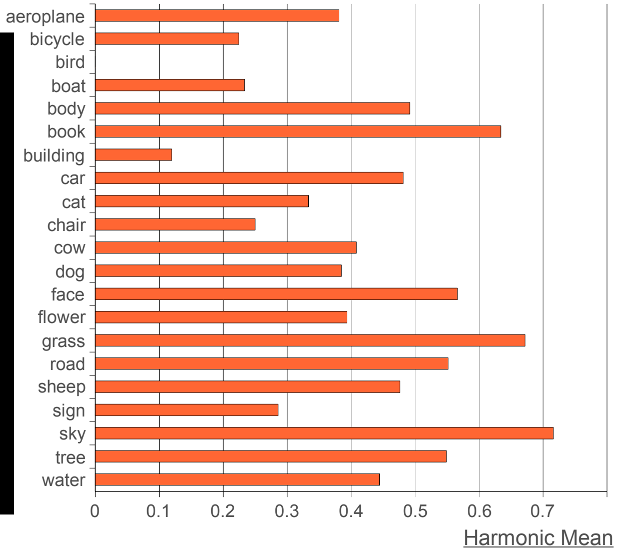


Fig. 3. Training Data Density and classification accuracy indicated by harmonic mean

Feng et al. [5] propose to measure the performance of an AIA system by computing the precision and recall values for each supported label. In image retrieval one typically seeks to optimize precision and recall at the same time. A measure that combines precision and recall into a single value is the weighted harmonic mean (or F-measure) [23], which averages the values of precision and recall with respect to a specific query q:

$$HM_q = \frac{2}{\frac{1}{R_q} + \frac{1}{P_q}} = \frac{2P_qR_q}{P_q + R_q} \tag{1}$$

The annotation-based retrieval process in Safire is predominantly based on the *global* annotations (that is the image-level annotations rather than the patch-level annotations). Global annotations determine whether an image appears in the result set or not. Fig. 3 lists the harmonic mean achieved for each of the 21 annotation classes. As can be seen, for several classes such as “sky”, “grass” and “tree” a rather high classification performance has been be achieved. For others, the classification performance is rather low (“bird”, “building”).

When analyzing the results we identified three major reasons for the low annotation accuracy of some classes. First of all, due to the chosen “careful classification” based on the described confidence computation, the system generates a large number of not classified patches (false dismissals). The high redundancy of the patch-based approach, however, slightly alleviates this effect at the global annotation level.

Second, when analyzing the training data density, we identified a strong correlation between the size of the annotation classes[4] and the classification result. On the one hand "Tree", "sky" and "grass" are among the four largest classes and show high class-wise accuracy. On the other hand, "bird", "dog", "body" and "chair" are rather small classes and also show a small classification accuracy. The fact that it seems more difficult to classify patches based on underrepresented classes is plausible when making aware that the number of patches available affects the representatives in the visual dictionary. A smaller number of patches results from fewer images representing a specific label class. Consequently, the computation of the most frequent representatives for each class to create the dictionary is based on fewer examples. Thus, the impact of outliers is much stronger whereas other, more prominent examples might not even be included in the training data.

Finally, the large visual variance of some classes also affects the classification accuracy. A larger in-class variance results in a more widespread instance space as the feature vectors computed for each training patch are much more different. As a result, the visual words, which have been computed using k-means clustering, are likewise much more widespread. The dictionary entries for different classes will tend to "overlap". An unclassified patch, which falls into an overlapping region is much more difficult to be classified with a high confidence. In compact classes, the computed visual words are densely populated, which makes classification more unambiguous. A class with a high in-class variance is, for example, the "building" class. This explains the rather low classification accuracy of the "building" class despite the large number of training examples.

The more annotation classes are represented in the dictionary, the more the computed visual words will tend to overlap in the described manner. This makes classification based on a majority vote as in the presented k-NN approach more and more difficult.

An important conclusion can be drawn from the classification results despite the discussed low accuracy for some classes. When reconsidering the label classes for whom a high accuracy has been achieved, it can be remarked, that all these classes can be considered as non-shaped structures. Clearly, "grass" and "sky" have no spatial dilation or shape. They rather exhibit a textured region with a specific color. The same holds to some extent for the "tree" class. Considering the training images for "flower" and "bicycle", one will notice that instead of showing a single representative for each class, most of the images actually depict a *bunch* of bicycles as well as flower*beds*. Multiple bicycles and flowers again possess no clear shape.

In other words, when exhibiting a low in-class variance, non-shaped regions can be efficiently described using solely color and texture descriptors as has been done within this work. On the other hand, the animal classes ("cow", "dog", "sheep", "cat", "bird") for instance cannot be distinguished based on color and texture only as suggested by the results. Shape seems to be an important discriminative characteristic to capture the visual appearance.

[4] In terms of patches used for dictionary creation, see Fig. 3.

5 Conclusion

In this paper we have proposed an extension to the Safire framework presented in [9] to support the user of an Annotation-based Image Retrieval system during the annotation process. Based on a small set of manually segmented and locally annotated training images a highly reliable, correlation-free visual dictionary is created, which is later used to infer the annotations of newly added images. The presented system fully relies on the MPEG-7 Multimedia Content Description Interface standard, which facilitates largely the extensibility and reusability of the presented solution. Our approach has shown good results in classifying non-shaped visual objects that can be sufficiently described by color and texture.

Our current research efforts concentrate on three main areas. First of all we intend to increase the classification accuracy by applying a more sophisticated classifier. Support-Vector-Machines [4,8] and Self-organizing Maps (SOM) [12,15,24] have been successfully applied in the domain of image retrieval and we intend to integrate these approaches in to Safire as well. Second, we are analyzing approaches to sample the patches on the image more sparsely based on so-called keypoints or salience measures as presented e.g. in [11]. This will allow the reduction of the sample space for dictionary creation to more discriminative examples. Likewise, integrating the spatial relation between sparsely computed patches might help to introduce shape as a characteristic in the classification process. Finally, we seek to increase the training data density for the currently sparsely populated classes. The *LableMe* project at the MIT [22] intends to build a large collection of images with a manually segmented and locally annotated ground truth to be used for object detection and recognition. The current size of the *LabelMe* database (183 annotation classes in 30369 images) outperforms the MSRC dataset by large and we are evaluating means to merge the two datasets into a single one within Safire, based on the standardized MPEG-7 description interface.

References

1. Bimbo, A.D.: Visual Information Retrieval. Morgan Kaufmann Publishers, Inc., San Francisco, CA (1999)
2. Choi, Y., Won, C.S., Ro, Y.M., Manjunath, B.S.: Texture Descriptors, Introduction to MPEG-7: Multimedia Content Description Interface, pp. 213–229. John Wiley & Sons, Ltd., Chichester (2002)
3. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3024. Springer, Heidelberg (2004)
4. Cusano, C., Ciocca, G., Schettini, R.: Image annotation using svm. In: Santini, S., Schettini, R. (eds.) Internet Imaging V, Proceedings of the SPIE, the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, December 2003, vol. 5304, pp. 330–338 (2003)
5. Feng, S., Manmatha, R., Lavrenko, V.: Multiple bernoulli relevance models for image and video annotation. In: Computer Vision and Pattern Recognition, 2004.

CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference, 27 June–2 July 2004, vol. 2, pp. II–1002–II–1009 (2004)

6. Feng, X., Fang, J., Qiu, G.: Color photo categorization using compressed histograms and support vector machines. In: Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference, 14-17 September, vol. 3, pp. III–753–6 (2003)
7. Frigui, H., Caudill, J.: Unsupervised image segmentation and annotation for content-based image retrieval. In: Fuzzy Systems, 2006 IEEE International Conference, July 16-21, pp. 72–77 (2006)
8. Goh, K.-S., Chang, E., Cheng, K.-T.: Support vector machine pairwise classifiers with error reduction for image classification. In: MULTIMEDIA 2001: Proceedings of the 2001 ACM workshops on Multimedia, pp. 32–37. ACM Press, New York, NY, USA (2001)
9. Hentschel, C., Nürnberger, A., Schmitt, I., Stober, S.: Safire: Towards standardized semantic rich image annotation. In: Marchand-Maillet, S., Bruno, E., Nürnberger, A., Detyniecki, M. (eds.) AMR 2006. LNCS, vol. 4398. Springer, Heidelberg (2007)
10. Inoue, M.: On the need for annotation-based image retrieval. In: Workshop on Information Retrieval in Context (IRiX), pp. 44–46 (2004)
11. Jurie, F., Triggs, B.: Creating efficient codebooks for visual recognition. In: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference, 17-21 October, vol. 1, pp. 604–610 (2005)
12. Laaksonen, J., Koskela, M., Oja, E.: PicSOM: Self-organizing maps for content-based image retrieval. In: Proc. of International Joint Conference on Neural Networks (IJCNN 1999), Washington, D.C., USA, July 10–16 (1999)
13. Lavrenko, V., Feng, S., Manmatha, R.: Statistical models for automatic video annotation and retrieval. In: Acoustics, Speech, and Signal Processing, 2004. Proceedings (ICASSP 2004). IEEE International Conference, 17-21 May, vol. 3, pp. iii–1044–7 (2004)
14. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. In: Proceedings of the 16th Conference on Advances in Neural Information Processing Systems NIPS (2003)
15. Lefebvre, G., Laurent, C., Ros, J., Garcia, C.: Supervised image classification by som activity map comparison. icpr 2, 728–731 (2006)
16. Lipson, P., Grimson, E., Sinha, P.: Configuration based scene classification and image indexing. In: Computer Vision and Pattern Recognition, 1997. Proceedings, 1997 IEEE Computer Society Conference, 17-19 June, pp. 1007–1013 (1997)
17. Minka, T.: An image database browser that learns from user interaction. Master's thesis, MIT Media Laboratory, Cambridge, MA (1996)
18. Mori, Y., Takahashi, H., Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words (1999)
19. Ohm, J.-R., Cieplinski, L., Kim, H.J., Krishnamachari, S., Manjunath, B.S., Messing, D.S., Yamada, A.: Color Descriptors, Introduction to MPEG-7: Multimedia Content Description Interface, pp. 187–212. John Wiley & Sons, Ltd., Chichester (2002)
20. Ojala, T., Mäenpää, T., Viertola, J., Kyllönen, J., Pietikäinen, M.: Empirical evaluation of mpeg-7 texture descriptors with a large-scale experiment. In: Proc. 2nd International Workshop on Texture Analysis and Synthesis, pp. 99–102 (2002)
21. Picard, R.W., Minka, T.P.: Vision texture for annotation. Multimedia Systems 3(1), 3–14 (1995)
22. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: LabelMe: A database and web-based tool for image annotation. MIT AI Lab Memo AIM-2005-025 (2005)

23. Schmitt, I.: Ähnlichkeitssuche in Multimedia-Datenbanken. Retrieval, Suchalgorithmen und Anfragebehandlung. Oldenbourg (2005)
24. Oh, K.s., Kaneko, K., Makinouchi, A.: Image classification and retrieval based on wavelet-som. dante 00, 164 (1999)
25. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference, 13-16 October, vol. 2, pp. 1470–1477 (2003)
26. Town, C., Sinclair, D.: Content based image retrieval using semantic visual categories (2000)
27. Vailaya, A., Figueiredo, M., Jain, A., Zhang, H.-J.: Image classification for content-based indexing. Image Processing, IEEE Transactions 10(1), 117–130 (2001)
28. Vailaya, A., Jain, A., Zhang, H.J.: On image classification: City vs. landscape. In: Content-Based Access of Image and Video Libraries, 1998. Proceedings. IEEE Workshop, 21 June, pp. 3–8 (1998)
29. Vogel, J.: Semantic Scene Modeling and Retrieval. In: Selected Readings in Vision and Graphics, vol. 33. Hartung-Gorre Verlag, Konstanz (2004)
30. Winn, J., Criminisi, A., Minka, T.: Object categorization by learned universal visual dictionary. In: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference, 17-21 October, vol. 2, pp. 1800–1807 (2005)
31. Zhang, R., Zhang, Z.: Hidden semantic concept discovery in region based image retrieval. In: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference, 27 June–2 July, vol. 2, pp. II–996–II–1001 (2004)