Probabilistic Reverse Annotation for Large Scale Image Retrieval

Pramod Sankar K., C. V. Jawahar
Center for Visual Information Technology
International Institute of Information Technology
Hyderabad, India.
jawahar@iiit.ac.in

Abstract

Automatic annotation is an elegant alternative to explicit recognition in images. In annotation, the image is matched with keyword models, and the most relevant keywords are assigned to the image. Using existing techniques, the annotation time for large collections is very high, while the annotation performance degrades with increase in number of keywords. Towards the goal of large scale annotation, we present an approach called "Reverse Annotation". Unlike traditional annotation where keywords are identified for a given image, in Reverse Annotation, the relevant images are identified for each keyword. With this seemingly simple shift in perspective, the annotation time is reduced significantly. To be able to rank relevant images, the approach is extended to Probabilistic Reverse Annotation. Our framework is applicable to a wide variety of multimedia documents, and scalable to large collections. Here, we demonstrate the framework over a large collection of 75,000 document images, containing 21 million word segments, annotated by 35000 keywords. Our image retrieval system replicates text-based search engines, in response time.

1. Introduction

Since the advent of economical imaging devices, the number of digital images and videos has grown exponentially. Large collections of images and videos are now available and shared online. Efficient retrieval from such large collections of multimedia data, is becoming an important problem.

In the early years of image retrieval, images were annotated manually. Since manual effort was costly, this was affordable for mostly military and medical domains [5]. Content Based Image Retrieval (CBIR) [16] systems have shown ample promise for querying-by-example, but the image matching techniques are often computationally intensive and thus time consuming. Transcriptions of images through object recognition, scene analysis, etc. has not been effective, partly due to the restricted applicability of present day recognition techniques. Recently, recognition-free approaches have been demonstrated for image retrieval, where a search index is built in a feature space [11, 13, 15]. However, the popular image or video retrieval systems are those based on text, such as Google Images, which indexes multimedia with the surrounding text. The popularity is mostly due to the interactive retrieval times for text based systems, as well as being able to query-by-text. Consequently, there has been a growing interest in automatic annotation of images [6, 7].

In this paper, we present what we call, the *Reverse Annotation* framework. Unlike traditional annotation where keywords are identified for a given image, in Reverse Annotation, the relevant images are identified for each keyword. This converts the annotation problem from *classification* to *verification*. This improves the annotation performance as well as reduces the annotation time. This framework is primarily designed for situations where the number of images to be annotated is much greater than the number of keywords to be annotated with.

We shall explain our framework in Section 3, which is further extended to a probabilistic setting in Section 4. We demonstrate the utility of this annotation framework by building an interactive retrieval system over a collection of 75,000 document images. Implementation details are presented in Section 5. Before proceeding further, we shall look at existing annotation techniques and the issues to be addressed for building retrieval systems in the next section.

2. Traditional Auto-Annotation

Recent work towards automatic annotation concentrates on learning a mapping between keywords and the image features. The image is generally segmented into regions using a rectangular grid or by using Normalized Cuts or Blobworld ([4]) [6, 7]. A hierarchy of regions at multiple scales could also be used to represent the image [3, 8]. Suitable

1-4244-1180-7/07/$25.00 ©2007 IEEE

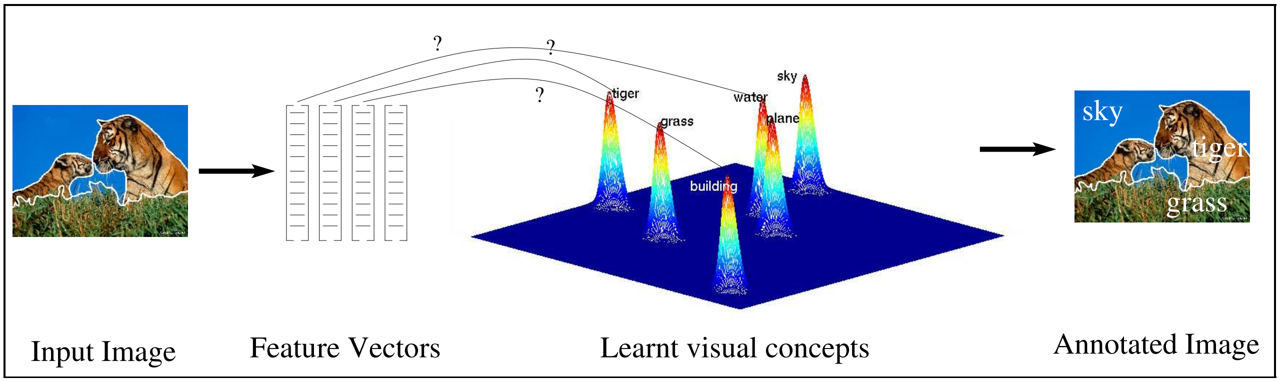

Figure 1. Depiction of traditional annotation approaches. Given a test image, features are extracted from its regions. These features are compared against the learnt concept models, and the keyword of the closest match is given to the region

features for the color, texture and region location are extracted from these regions. The features are clustered (using K-means) to discretize the feature space.

A training set of annotated images is obtained from either the Corel database, a hand annotated dataset, news photographs with captions, or surrounding text in a webpage. A relationship between the image features and the annotations is learnt from the training examples. This relationship could be a *co-occurrence model*, a *translational model* of probabilities [6], a generative probabilistic *cross media relevance model* [7], statistical linguistic indexing [8] or a *latent semantic* model [9]. Depending on the level of ambiguity in the annotations, the sophistication of the corresponding learnable model is improved.

Traditional annotation process is depicted in Figure 1. The keyword models in the feature space, are learnt from training data. To annotate a test image, its regions are identified, and features are extracted from each region. The features are compared against the keyword models and the labels are assigned using a suitable classifier, such as nearest neighbor or asymmetrical-SVM [18]. Annotations could also be propagated across images, as successfully demonstrated by Rath *et. al.* [12], over a collection of 1000 handwritten document images. The image retrieval system, is built on individually annotated images, considering each image as a document of its annotations.

The current state-of-the-art approaches have shown promising results in representing meaningful concepts using visual features. However, the problem of image annotation for building large scale retrieval systems, has not been well addressed. For image retrieval, large collections of images are required to be annotated with a large set of keywords. Towards this end, the issues that need to be addressed are:

- Scalability: Previous approaches have learnt feature spaces and their relationships with only a small set of keywords. The scalability of features to represent thousands of keywords has to be explored. With increasing number of keywords, the representative features from one would begin to overlap with others.
- Loss of Variety: The variety present in the training data, is lost with vector quantization. Points at the edge of the clusters, are generally misclassified, even though similar exemplars are provided during training.
- Computational complexity: Suppose we are given N images, with m regions, to be annotated by n keywords. The annotation complexity would be $O(N \cdot m \cdot n)$. If $N = 100,000$, $m = 10$ and $n = 30,000$, and the comparison between two feature vectors takes 0.1 seconds, the annotation time would be close to 100 years.
- Ranked Retrieval: When a retrieval system is built on annotated images, it is necessary to rank the images based on relevance [7]. The annotation procedure should allow for ranking relevant documents.

3. Reverse Annotation

The issues toward large scale annotation are addressed by a novel *Reverse Annotation* framework. Reverse Annotation takes-off from the fact that, for any retrieval system, the number of items in the index is *limited*, while the images that are indexed could be unlimited. For example, in the annotation of images of animals, the number of animals is limited by the variety of fauna, while the number of images of animals is not limited. Similarly, the number of distinct people in the news is a small set, while the number of detected faces across news videos is very large. With a careful selection of the keywords/concepts for annotation, a large percentage of documents can be indexed, as well as a large number of user queries can be retrieved.

In traditional annotation, the image features are compared with keyword models, where the annotation problem was that of classification. A multi class classifier or a hierarchy of classifiers is required to be learnt for this purpose.

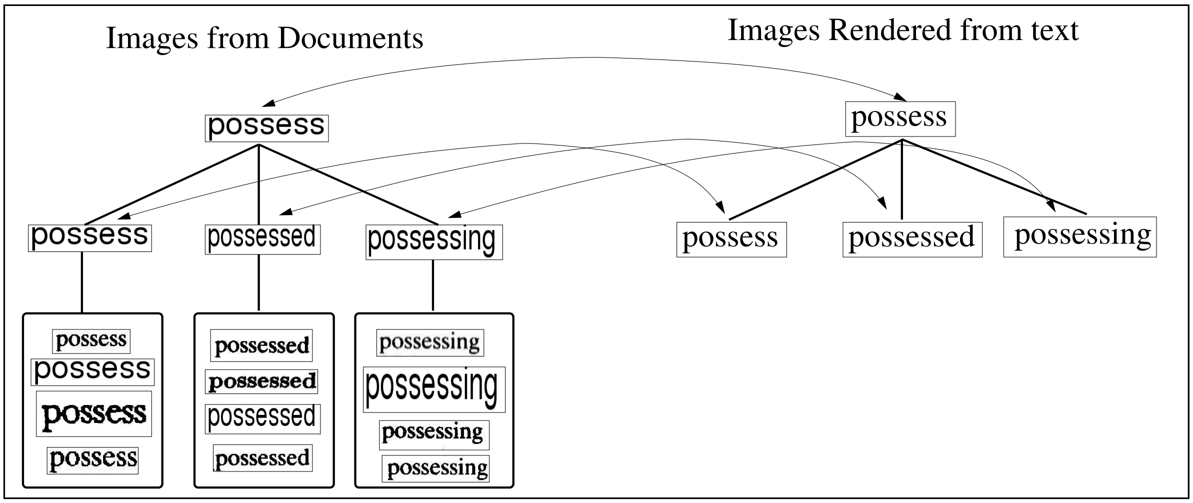

Figure 2. Hierarchical comparisons in the Reverse Annotation framework. The keywords are converted to word images, whose cluster tree is shown on the right, while that of images from documents is shown on the left. The lower levels of hierarchy are compared only if the higher levels match. It suffices to annotate one level above the leaf nodes.

Given N image regions to annotate, and n labels to annotate with, each of the N features are classified against the n classes. This reduces the annotation performance and increases annotation time.

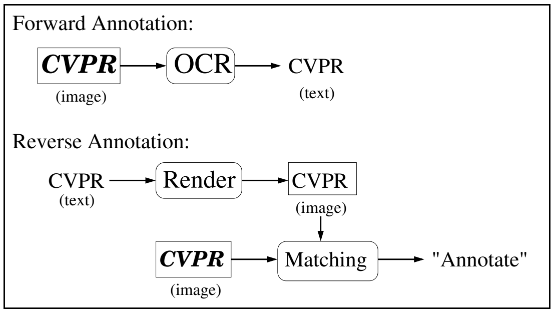

Figure 3. Comparison between Forward and Reverse Annotation schemes for Word recognition

Conversely, in Reverse Annotation, the keyword/label is matched with the images, and the matches are annotated with the keyword. This effectively is a verification problem, which involves only a two-class classification. Evidence from the biometrics community suggests that the verification problem has better performance than a classification problem [17]. Thus, the annotation performance is expected to improve. The comparison between forward and reverse annotations for word image annotation is depicted in Figure 3.

The images annotated in one step of verification, need not be included for verification in the subsequent steps. If $k_1, k_2, ..., k_n$ are the keywords, the images annotated by k_1 would not match the keywords $k_2, ..., k_n$, and can thus be eliminated from future classifications. This reduces the number of computations required at each step. Further, the keywords can be reverse-annotated in an order that minimizes the annotation time. According to the Zipf's law [19], the frequency of keywords when ranked by the frequency, follows a power-law distribution. Assuming Zipf's law to hold for image collections, we could arrange the keywords in decreasing order of their frequency. With such an order, the number of images to be annotated at each reverse-annotation step, reduces exponentially.

Moreover, the fact that $N >> n$ implies that there is abundant repetition in the image collection. This allows us to cluster the image features, such that multiple occurrences of a particular concept are found in one cluster. A simple clustering procedure or a hashing scheme [14] could be used to obtain the clusters quickly. Given such clusters, it suffices to annotate one representative of the cluster, allowing for considerable annotation speed-up.

Further, the keywords themselves are not totally isolated. Keywords can be clustered based on semantic distances or the similarity in their representative features. The clusters can be built in a tree-like hierarchy to enable quicker comparisons between the keyword clusters and the feature clusters [10]. The procedure of annotation over clustering is depicted in Figure 2.

The framework is suitable for a large variety of annotation applications. Consider, for example, the problem of annotating faces in news videos. The keywords (names of people in news), could be collected from online news stories. The keyword examples can be obtained from a suitable knowledge base (Google Images, Flickr etc.), or from a hand labeled album of images. Face images in the news videos, can be clustered such that each cluster has all instances of one person. The Reverse Annotation framework can now be used to annotate the face images by matching

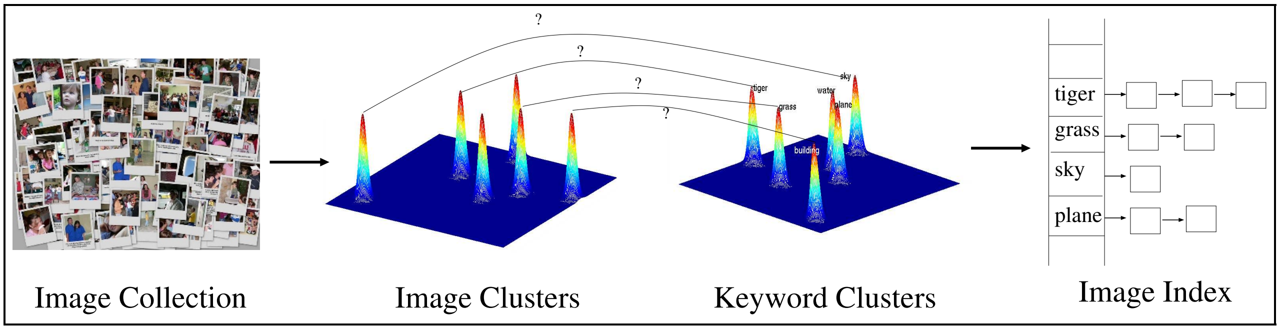

Figure 4. Depiction of Reverse Annotation. The features from the image collections are clustered into unknown concepts. To annotate the images, it suffices to find the correspondences between these clusters and that of the keywords. Once the correspondences are identified, the search index could be easily built.

the labeled face exemplars with the face clusters.

Once annotated, building a search index and a retrieval system over the annotated images is straightforward.

4. Probabilistic Reverse Annotation

The Reverse Annotation procedure has addressed the goal of identifying the relevant images for a given keyword. This is useful to build an index for search and retrieval. However, the indexed images cannot be ranked based on relevance, since the associations in the index are binary: either the image is relevant for a given keyword, or it is not. Ranking of the images is necessary to retrieve images that are more relevant to the given search query.

The Reverse Annotation framework provides a mechanism for probabilistic annotation. For this purpose, we extend the framework to *Probabilistic Reverse Annotation*. In Reverse Annotation, the closest cluster was identified for a given keyword. In Probabilistic Reverse Annotation, we estimate the probability that each cluster belongs to the keyword. The procedure is shown schematically in Figure 4.

$k_1, k_2, ..., k_n$ $k_{i_1}, k_{i_2}, ..., k_{i_{n'_i}}$	keyword clusters points in cluster k_i
$I_1, I_2, ..., I_N$ $R_{j_1}, R_{j_2}, ..., R_{j_{m'_j}}$	images in the collection region descriptors for image I_j
$t_1, t_2, ..., t_m$ $t_{l_1}, t_{l_2}, ..., t_{l_{n''_l}}$	region clusters points in cluster t_l

Table 1. Naming convention for the Probabilistic Reverse Annotation framework

Let us follow the naming convention presented in Table 1. The probability $p1_{il}$ is estimated for the keyword k_i to be found at each of the cluster t_l, $l = 1, 2, ..., m$ as

$$p1_{il} = \frac{d(k_i||t_l)}{\sum_{l=1}^{m} d(k_i||t_l)}$$

where $d(x||y)$ is a similarity measure (inverse of a distance measure).

Within each cluster, the probability $p2_{lj}$ is estimated for the image region t_{l_j} to belong to t_l. This probability can be estimated as

$$p2_{lj} = \frac{d(t_l||t_{l_j})}{\sum_{j=1}^{n''_l} d(t_l||t_{l_j})}$$

The total probability p_{ij} that cluster t_{l_j} belongs to keyword k_i is given by

$$p_{ij} = p1_{il} * p2_{lj}$$

The cumulative weight of an image to a given keyword, is calculated by accumulating the total probabilities from each region of the image.

4.1. Ranking by tf-idf

When an image contains two instances of a given object, then the probability of the image belonging to the keyword is doubled, and so on for multiple instances of the object. Thus, the ranking by the above probability measure, corresponds to the *term-frequency* (TF) measure used in text retrieval. TF measures the importance of the keyword for the image. The relevant images for each keyword, are sorted according to the TF measure.

Similarly, the *inverse document frequency* (IDF) measure indicates the overall importance of the given keyword in the entire collection. The IDF measure is defined as the logarithm of the inverse of the sum of the cumulative probabilities of the images. The IDF is used to normalize the TF value across all the images. This normalization does not affect the ranking within the same keyword. However, IDF plays an important role when retrieving from multiple keywords such as “tiger in grass” etc.

4.2. Handling Off-Center Points

In an annotation scheme, the points away from the mean are generally penalized heavily. Especially when the keyword exemplars are vector quantized, the points that fall far from the center of this quantization tend to be misclassified. However, in the given training data, one could find an exemplar that closely matches such off-center points. If the training data is assumed to be reliable, the off-center points need not be penalized due to the averaging.

In such cases, there would be an exemplar in keyword cluster, $k_{i_{j'}}$ $(j' = 1, 2, ..., n'_i$), which is very similar to t_{l_j}. The probability estimation for $p2_{lj}$ is modified accordingly as

$$d_{lj_{min}} = \min_{j'} d(t_{l_j}||k_{i_{j'}})$$

$$p2_{lj} = \frac{d_{lj_{min}}}{\sum_{j=1}^{n''_l} d_{lj_{min}}}$$

4.3. Efficient Implementation

The index would contain all the images in the collection, ranked according to the tf-idf measure. However, for all practical purposes, for a given keyword, we could conveniently ignore all those images, with a tf-idf measure less than a particular threshold. This corresponds to a small set of neighboring clusters, which can be viewed as another level of clustering over the first-level clusters.

For an efficient implementation, we extend this concept to a hierarchy of clusters, over both the keyword exemplars and the image regions. The comparison is initialized at the coarsest level of clusters. Further comparisons at a finer level of clusters are performed only among the clusters that matched at the immediate coarser level. The assumption is that the clusters that do not match at a coarse level would not match at a finer level, similar to the approach of Nister and Stewenius [10]. Therefore, a large number of comparisons between dissimilar clusters are avoided.

5. Reverse Annotation over Document Images

We demonstrate the Reverse Annotation framework over the domain of document images. There is a rich variety of keywords in documents, which can be easily sampled from a text corpus. Keyword exemplars could be easily generated by rendering the text to word images. The scalability of the approach could be tested over large collections of images and it is comparatively easy to evaluate a retrieval system over document images.

Moreover, large collections of document images are now publicly available from the various digitization projects. The Million Book Project [1], has thus far digitized a monumental *one million* books. A large percentage of digitized content comes from a variety of scripts, including Indian, Chinese, Arabic etc. Conventional techniques for transcription of document images, using Optical Character Recognition (OCR), do not produce text that is suitable for search and retrieval. The recognizers fail due to the inherent complexity of the scripts and printing variations. Building a search system over a large collection of document images without OCRs, is an immediate real-world requirement.

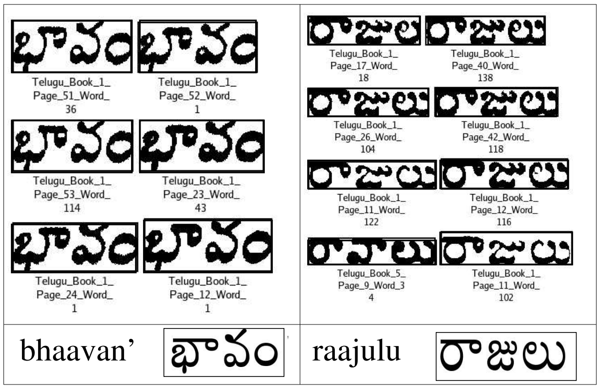
Telugu_Book_1_ Page_51_Word_ 36
Telugu_Book_1_ Page_52_Word_ 1
Telugu_Book_1_ Page_53_Word_ 114
Telugu_Book_1_ Page_23_Word_ 43
Telugu_Book_1_ Page_24_Word_ 1
Telugu_Book_1_ Page_12_Word_ 1
Telugu_Book_1_ Page_17_Word_ 18
Telugu_Book_1_ Page_40_Word_ 138
Telugu_Book_1_ Page_26_Word_ 104
Telugu_Book_1_ Page_42_Word_ 118
Telugu_Book_1_ Page_11_Word_ 122
Telugu_Book_1_ Page_12_Word_ 116
Telugu_Book_5_ Page_9_Word_3 4
Telugu_Book_1_ Page_11_Word_ 102
bhaavan'
భావం
raajulu
రాజులు

Figure 5. Examples of annotated clusters. The annotation of the cluster is shown below, along with the keyword exemplar that matched the cluster.

Dataset Our dataset comes from the Digital Library of India [1]. 500 books from the Telugu language, were randomly selected. The collection had 76,425 page images. The images were segmented to obtain individual word images. The total number of regions/segments in our collection was *21 Million*.

Keywords for Annotation Keywords were obtained from a text corpus of 9 million words. The ranks and the frequency of the words were computed, and the keywords were chosen from the middle of the power-law distribution. We obtained 35,000 keywords from the frequency range 10 to 2000. Exemplars for the keywords are generated by rendering the word to form a keyword-image.

Image Matching Matching of word images should be invariant to font type, style and size. The features extracted and the matching technique, should handle such variations, so that the matching score is invariant to these changes. Accordingly, word profile features such as upper word profile, lower word profile, projection profile and transition profile were chosen as the features [2]. Features are compared using a Dynamic Time Warping (DTW) approach since it inherently handles font type, and style variations [11]. This image matching was previously used for building search indexes in feature space, by forward annotation [2, 11].

Clustering The goal of clustering is to partition all the words into groups of individual words, with all instances of a given word occurring in one cluster. Our feature representation and similarity measure yield non-metric pairwise distances. In such cases, the popular choice of clustering is Hierarchical Agglomerative Clustering (HAC). HAC begins with individual clusters for each point and proceeds by merging the closest clusters until a stopping-criterion is met.

Two example clusters are shown in Figure 5. Clustering results were manually evaluated across 500 randomly chosen clusters. Precision is measured as the correctness of each cluster, while the recall measures the completeness. Precision was found to be 72.66%, while the recall was 75.45%.

Annotation The clusters were annotated using the Reverse Annotation framework, depicted in Figure 2. The closest word image cluster was searched for each keyword. The keywords that matched with the clusters in Figure 5, are given below the cluster. The annotation is performed in the *transliteration* scheme called "OmTrans", which is a Roman alphabet representation for Indian languages.

The annotation performance was assessed from 500 randomly picked cluster annotations. *The accuracy of annotation was found to be 73.2%.*

Performance of Retrieval System Search indexes were built separately over text documents from ground truth data and annotated images. In case of the ground truth collection, all words were indexed, ensuring a near-perfect precision-recall. The two search engines were evaluated against 20 queries picked at random from the keyword set. The retrieval results are evaluated using the R-precision measure, which is the precision of the system at R documents retrieved, R being the number of known relevant documents for the given query in the collection. R is obtained from the result of the ground truth search system.

The top R results from the search system were evaluated for retrieval performance and the overlap in the retrieved documents was found to be 77.38%. Since the chosen words were not stop-words, we have total recall from the baseline system. Thus the annotated documents are able to replicate text retrieval performance to upto an accuracy of 77%. The difference between the accuracies of the two systems comes from the inaccuracies in the image processing domain. The errors in segmentation, clustering and annotation propogate from one stage to the next and contribute to this mismatch.

6. Conclusions

We have presented a novel framework of Probabilistic Reverse Annotation towards enabling search over large collections of images. The performance of the framework over document image collections was found to be satisfactory. The approach is shown to be scalable to large multimedia collections. In the present work, annotation was performed by matching the centroids of clusters between keywords and test visual words. One could explore the possibility of matching clusters based on the distributions of points in them. Another possible extension could be to match a large set of clusters in one domain to those in another domain using graph matching algorithms.

References

[1] V. Ambati, N.Balakrishnan, Raj Reddy, L. Pratha, and C. V. Jawahar. The Digital Library of India Project: Process, Policies and Architecture. In *2nd International Conference on Digital Libraries(ICDL)*, 2006.

[2] A. Balasubramanian, Million Meshesha, and C. V. Jawahar. Retrieval from document image collections. In *7th International Workshop on Document Analysis Systems, DAS*, pages 1–12. LNCS, Springer-Verlag, 2006.

[3] D. Barnard, K.; Forsyth. Learning the semantics of words and pictures. In *ICCV*, volume 2, pages 408–415, 2001.

[4] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik. Blobworld: A system for region-based image indexing and retrieval. In *International Conference on Visual Information Systems*, pages 509–516. Springer, 1999.

[5] S. K. Chang and A. Hsu. Image information systems: Where do we go from here? *IEEE Trans. on Knowledge and Data Engineering*, 4(5):431–442, 1992.

[6] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, page 97, 2002.

[7] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *ACM SIGIR*, pages 119–126, 2003.

[8] J. Li and J. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE PAMI*, 25(10):1075–1088, 2003.

[9] F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 275–278, 2003.

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

[11] T. Rath and R. Manmatha. Word image matching using dynamic time warping. *CVPR*, 2:521–527, 2003.

[12] T. M. Rath, R. Manmatha, and V. Lavrenko. A search engine for historical manuscript images. In *SIGIR*, pages 369–376, 2004.

[13] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE PAMI*, 19(5):530–534, 1997.

[14] G. Shakhnarovich, T. Darrell, and P. Indyk, editors. *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice*. MIT Press, 2006.

[15] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, volume 2, pages 1470–1477, Oct. 2003.

[16] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE PAMI*, 22(12):1348, 2000.

[17] C. Vielhauer. *"Biometric User Authentication for IT Security"*. Springer, 2006.

[18] C. Yang, M. Dong, and J. Hua. Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. In *CVPR*, pages 2057–2063, 2006.

[19] G. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA, 1949.