I^{2}A: an Interactive Image Annotation system

Changbo Yang, Ming Dong and Farshad Fotouhi
Department of Computer Science
Wayne State University
Detroit, MI 48202, USA

Abstract

In this paper, we propose an interactive image annotation system. The proposed system has two connected components, the low-level feature space and the semantic space. Experiments show that our system is able to provide accurate annotation for images by learning the connections between the two spaces through statistical modelling, natural language processing, and users' interaction.

1 Introduction

With the rapid development of digital photography, digital image data in various formats has increased tremendously in recent years. Consequently image retrieval has drawn the attention of many researchers in the computer vision community. In many content-based image retrieval systems, image content is used in conjunction with user interaction to retrieve semantically meaningful information. However, current state-of-art computer vision technology lags far behind the human's ability to assimilate information at a semantic level. The retrieval based on the similarity of visual attributes when used arbitrarily cannot provide semantically meaningful information. Thus image annotation has become a problem of paramount importance for the further development of image retrieval.

Manual annotation and classification is one way to establish semantic indexing in image databases. However this method is very expensive when the volume of data is very large. An automatic computer program that can learn the relationship between the content of an image and its semantic meaning is highly desired to handle the massive digital image resources.

1.1 Related Work

Automatic image annotation is a highly challenging problem because of the semantic gap between low-level image content and high-level concepts. Recently Duygulu et al. [1, 2] thoroughly reviewed and studied this problem. In their work, images are described by a vocabulary of blobs. Several statistical machine translation models are developed to translate the set of blobs forming the image to a set of keywords. The reported annotation accuracy on a $5,000$ image database is very low, less than 10% in average. Consequently the retrieval performance in the semantic level is poor.

As emphasized in [3], computer vision researchers should identify features required for interactive image understanding, rather than their discipline's current emphasis on automatic techniques. User interaction is essential to accurately capture the semantic meaning of images [4]. A straightforward way of including the users in the loop is through relevance feedback. In this direction, Lu et al. [5] propose to form a semantic network on top of the keyword association on the images. In the semantic network, the relevance between an image and a keyword is explicitly memorized through relevance feedback and utilized for retrieval. However the learning of user knowledge is quite slow in this fashion. Also the retrieval performance of such a system is not high due to widespread synonymy and polysemy in natural languages.

In this paper, we propose an interactive image annotation system, in which image annotation is initialized based on statistical modelling and refined through relevance feedback and natural language processing. The remainder of the paper is organized as follows: Our system architecture is presented in Section 2. The proposed automatic image annotation algorithm is described in Section 3. In Section 4, we present our experiment and results. Conclusion is presented in Section 5.

2 System Architecture

The proposed system has two major components, the low-level feature space and the semantic space (see Figure 1). We assume that some images in our system have precise semantic meaning through manual annotation. We call them training images. Annotation information is not available for all other images in the system. We call them unlabelled images. The objective of our system is to index unlabelled images for later retrieval.

0-7803-9332-5/05/$20.00 ©2005 IEEE

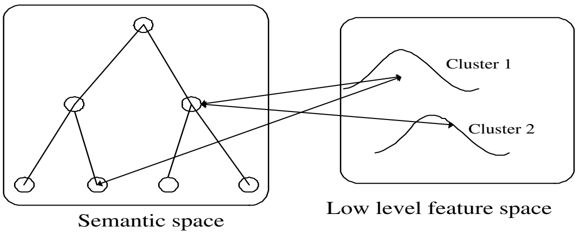

Figure 1: Semantic space and low level feature space

2.1 Low Level Feature Space

The training images in our system are grouped into clusters by K-means clustering based on their low-level features. Each cluster contains multiple images that are similar in the content. We then use annotation information from this collection of training images to generate the set of the keywords that could be used to describe the concept/concepts associated with the cluster of images.

A statistical keyword selection algorithm [6] is deployed to find the set of keywords that can effectively represent the concept/concepts of the cluster, and at the same time have the maximum discrimination power between different clusters. Statistically those keywords should appear frequently within the same cluster, but have low frequencies to show up in other clusters. After the keyword selection, we define the normalized frequency of a keyword in one cluster as its association weight to this cluster.

2.2 Semantic Space

To explore the semantic relationship among the keywords, a semantic hierarchy is built by the aid of WordNet. WordNet [7] is an online lexical reference system developed by the Cognitive Science Laboratory at Princeton University, which provides a comprehensive knowledge base for natural language processing. Although WordNet itself provides well-structured semantic hierarchies, it is not applicable to image retrieval due to its huge size and complex structure. Based on the annotation information from the training images, we select a set of sub-hierarchies in WordNet as our semantic hierarchy.

Initially, our semantic hierarchy is an empty one with only several root concepts. Each time a new keyword is identified in the available annotation, the semantic hierarchy will then be expanded by inserting it and related keywords into proper positions in the hierarchy. In this way, the semantic hierarchy grows progressively as more keywords are identified by the system. Since a vast majority of nouns in WordNet are very infrequently occurring words, our algorithm avoids inserting them by using a dictionary. Only the keywords appeared in the dictionary can be inserted into the hierarchy. In our implementation, we use a dictionary which includes all the words appeared in the annotation of the training images.

2.3 Semantic Expression

The set of the keywords and weights that is obtained in Section 2.1 only contains the statistical information for each cluster. To explore the semantic relationships among keywords, the keyword set has to be expanded with the aid of the semantic hierarchy. For example, if the keyword "coin" appears in the list, keyword "currency" should be added to the list since they are semantically related in our hierarchy. The weight of each expanded keyword is decided based on the similarity between it and the original keyword. Specifically the similarity is calculated as follows.

$$Sim(k_i, k_j) = \frac{depth(p)}{max(depth(k_i), depth(k_j))} \qquad (1)$$

where $depth(k_i)$ and $depth(k_j)$ are the depth of keywords k_i and k_j in the semantic hierarchy. p is the common ancestor of k_i and k_j.

The weight of expanded keyword is given by

$$S_n = \max(w_n, \sum_{i=1..M} w_i \cdot Sim(k_i, k_n)) \qquad (2)$$

where w_n is the weight of expanded keyword k_n and w_i is the weight of original keyword k_i, M is the number of total keywords in the hierarchy. Recall that we define normalized frequencies as the weight of the original keywords. After the keyword expansion, a longer keyword list is obtained for each cluster of images. We call it the *semantic expression*. The semantic information is embedded in the keyword weights. Based on the semantic expression, our system is able to handle natural language queries.

3 Interactive Image Annotation

The proposed two-step interactive image annotation algorithm is summarized in Figure 2. In the first step, we cluster the training images based on the low-level features and create semantic space using the method scribed in Section 2.2. Given an unlabelled image I, the weight of a keyword k in its semantic expression is given by,

$$w(k) = \sum_{i=1}^{K} w(k|C_i)p(C_i|I) \qquad (3)$$

where $w(k|C_i)$ is the weight of keyword k in the semantic expression of cluster C_i, $p(C_i|I)$ is the probability that image I belongs to the cluster C_i, and K is the total number of clusters.

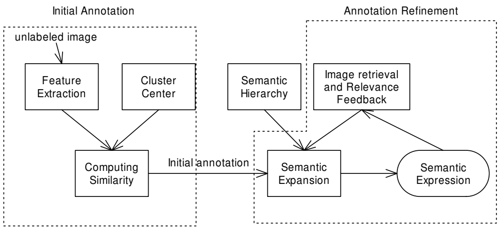

Figure 2: Two-step interactive image annotation algorithm

To simplify the initial annotation process, we compare the image I with each cluster center in the low-level feature space and annotate it only based on the semantic expression of its nearest cluster. In this case, $p(C_i|I)$ can only take binary values: 1 for the nearest cluster and 0 for other clusters.

Our system supports two kinds of query: query by an example and query by keywords. If a user provides an example image, we could convert the query to a semantic expression through the mapping between the low-level feature space and the semantic space. For keyword based search, we convert the keywords to a semantic expression through the keyword expansion. The retrieval then proceeds based on the similarity measure below:

$$Sim(q,d) = \frac{d^T q}{|d||q|} = \frac{\sum_k d_k q_k}{\sqrt{\sum_k d_k^2}\sqrt{\sum_k q_k^2}} \quad (4)$$

where q is the semantic expression of query image and d is the semantic expression of document image.

After presenting the retrieved images to the user, our system collects the user's positive feedback. We divide the positive feedback into two groups, group 1 containing training images and group 2 containing unlabelled images. Since the images in group 1 is manually annotated, we assume that they are semantically sound for user's searching objective. The semantic expression is then computed from group 1 and used to 1)modify the sematic expression of the query (short-term learning) and 2)refine the semantic expression of the unlabelled images in group 2. Specifically we employ Rocchio formula to both short-term and long-term learning,

$$Q' = \alpha Q + \beta \frac{\sum Q^+}{n} \quad (5)$$

where Q is the original semantic expression for either query image or a unlabelled image in group 2, Q^+ is the semantic expression of a positive example, n is the total number of positive examples, and Q' is the adjusted semantic expression. How to choose the optimal parameter in Rocchio's formula is a long debated problem. In our experiment, we set α=0.75 and β=0.25 as commonly used in many other applications.

After adjusting the semantic expression for the query, the retrieval is repeated for the next iteration. Through the short-term learning, relevant keywords in the semantic expression are assigned larger weights, and more images related to those keywords will be retrieved. On the other hand, we also move the semantic expression of those unlabelled images in group 2 towards the their true semantic meaning (identified by the user). The feedback accumulation will effectively lead us to the accurate semantic meaning of a unlabelled image in the long-run. Finally we can select some keywords with relative large weights in the semantic expression to annotate the image.

4 Experiment and Results

We collected 1625 images from Corel CD to conduct our preliminary experiment. These images are grouped into 8 concepts:“ animal”, “money”, “flower”, “snow mountain” “lantern”, “road”, “vegetable”, and “waterfall”. Each image comes with around 20 keywords annotated by Corel employees. We divide our collection into two groups. The first group has 525 randomly selected images. We use this group as the training image set. The other group has remaining 1100 images. We remove their annotation and treat them as unlabelled images. Based on our visual observation, the following set of visual features is extracted from every image in our collection: color histogram, color coherence histogram, edge histogram, and edge coherence histogram.

Notice that we always face the situation of having limited amount of annotated images and much more un-annotated images in real-world applications, for example, a large portion of images on the Internet are unlabelled. Our experiment is specially designed to simulate real-world scenarios by choosing a small training set and a large testing set.

We first group the training image set into 20 clusters based on the low-level features using K-means. Theoretically speaking, the number of the clusters reflects the balance between noise tolerance (larger K is preferred) and accuracy of the initial annotations (smaller K is preferred). We evaluated our system by setting K at 10, 20, 50, 100 respectively and find that a good balance is achieved at $K = 20$.

We then statistically select 20 keywords for each cluster to represent the underlying concept/concepts (see Section 2.1). The semantic hierarchy is built and semantic expression is obtained for each cluster. We then use each unlabelled image as a query image and perform retrieval. During the retrieval process, our system retrieves the top 100 images from both training and unlabelled image sets. Based on the ground truth, the system simulates users' relevance feedback by selecting the top 20 most relevant images as the positive feedback examples.

The retrieval is repeated for $1,100$ times (total number of unlabelled images) and feedbacks are accumulated based on Equation 5. Finally we choose keywords with the relative large weights in the semantic expression as the annotation for a unlabelled image. We compare our annotation with the previously removed annotation for each unlabelled image. The annotation accuracy Acc is computed based on the following equation,

$$Acc = \frac{\parallel FA \bigcap MA \parallel}{\parallel FA \parallel} \tag{6}$$

where FA is the set of our final annotation and MA is the set of manual annotation (ground truth).

The average accuracy in terms of the exact keyword match is reported in Figure 3 with respect to the number of keywords in the annotation. It is obvious that our system is able to learn from users' interaction and constantly improve the annotation accuracy by 30% when compared with low-level feature based initial annotation. For example, Figure 4 shows the annotations (top 5 keywords) for a "snow mountain" image. The initial keyword list contains the keyword "waterfall" related to the category "waterfall" because the categories "snow mountain" and "waterfall" contain similar low level features. After the long-term learning through relevance feedback, "waterfall" are excluded in the final annotation. Actually our annotation accuracy is much higher when considering the semantic relations between keywords. Even though some keywords in the final annotation do not appear in the ground truth, they are actually strongly related to the meaning of the image if we look closely. For instance, the keyword "nature" is not in the ground truth of the "snow mountain" image in Figure 4, but it is actually a correct annotation.

We also evaluate our annotation model in terms of retrieval precision (the ratio of relevant images in the retrieved results). Image retrieval is performed on the data set before and after annotation refinement. The average retrieval precision over $1,100$ queries (all unlabelled images) is dramatically improved from 24% to 48%. This result demonstrates that the proposed long-term learning is quite effective and efficient.

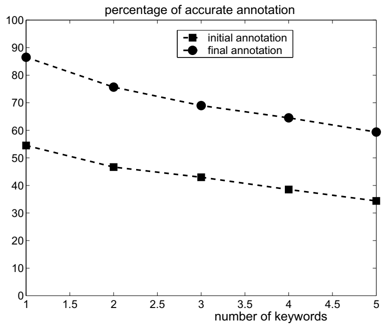

Figure 3: Annotation accuracy vs the number of keywords in the annotation.

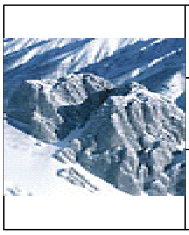	Initial annotation: rock, stone, snow, wilderness, waterfall
	Final annotation: snow ,winter, ice, mountain, nature
	Ground truth: calm, cold, mountain, cliff, view, geology, gray, grey, ice, snow, sunlight, wilderness, sunshine, winter

Figure 4: Annotations of a "snow mountain" image.

5 Conclusion

In this paper, we propose an interactive image annotation system. The major contributions of this work include,

- The connection between image contents and keywords is initiated by statistical modelling and refined by users' relevance feedback. Thus our method is able to provide accurate annotation for images.
- Our system uses semantic expression to reflect the meaning of an image instead of separated individual keywords. The semantic expression is generated based on the natural language processing model in the semantic space and could capture the meaning of an image more precisely.

References

[1] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth, "Object recognition as machine translation :learning a lexicon for a fixed image vocabulary," in *ECCV 02*, 2002, vol. 4, pp. 97–112.

[2] K. Barnad, P.Duygulu, N.de Fretias, D. Forsyth, D.Blei, and M.I.Jordan., "Matching words and pictures," *Journal of Machine Learning Research*, 2001.

[3] R. Jain, ," in *Proc. of US NSF Workshop Visual Information Management Systems*, 1992.

[4] S. Santini, A. Gupta, and F. Jain, "Emergent semantics through interactuion in image databases," *IEEE Trans. on KDE*, vol. 13, no. 3, pp. 337–351, 2001.

[5] Y. Lu, H. Zhang, L. Wenyin, and C. Hu, "Joint semantics and feature based image retrieval using relevance feedback," *IEEE Trans. on Multimedia*, vol. 5, no. 3, pp. 339–347, 2003.

[6] C. Yang, M. Dong, and F. Fotouhi, "Learning the semantics in image retrieval - a statistical natural language processing approach," *MDDE 2004*, 2004.

[7] C. Fellbaum, *Wordnet: An electronic lexical database*, MIT Press, 1998.