EVALUATION STRATEGIES FOR AUTOMATIC LINGUISTIC INDEXING OF PICTURES

James Z. Wang *Jia Li* *Sui Ching Lin*

The Pennsylvania State University, University Park, PA, USA

ABSTRACT

With the rapid technological advances in machine learning and data mining, it is now possible to train computers with hundreds of semantic concepts for the purpose of annotating images automatically using keywords and textual descriptions. We have developed a system, the Automatic Linguistic Indexing of Pictures (ALIP) system, using a 2-D multiresolution hidden Markov model. The evaluation of such approaches opens up challenges and interesting research questions. The goals of linguistic indexing are often different from those of other fields including image retrieval, image classification, and computer vision. In many application domains, computer programs that can provide semantically relevant keyword annotations are desired, even if the predicted annotations are different from those of the gold standard. In this paper, we discuss evaluation strategies for automatic linguistic indexing of pictures. We provide both objective and subjective evaluation methods. Finally, we report experimental results using our ALIP system.

1. INTRODUCTION

Automatic linguistic indexing of pictures is a critically important area of research because of its demonstrated potential to narrow the semantic gap. The significant difference between similarity in low-level features and similarity in high-level semantic meanings is known as the semantic gap. The problem is considered highly challenging by computer scientists. In recent years, we experienced rapid technological advances in both machine learning and data mining. Many advanced statistical modeling capabilities are now available for image analysis. Automatic linguistic indexing systems start to emerge.

The material was based upon work supported by the National Science Foundation under Grant No. IIS-0219272 and No. EIA-0202007, The Pennsylvania State University, the PNC Foundation, and SUN Microsystems under grant EDUD-7824-010456-US. Conversations with Michael Lesk and Sally Goldman have been very helpful. J. Z. Wang is affiliated with School of Information Sciences and Technology and Department of Computer Science and Engineering. Email: `jwang@ist.psu.edu` J. Li is affiliated with Department of Statistics and Department of Computer Science and Engineering. S. Lin is affiliated with Department of Computer Science and Engineering.

There is a rich resource of prior work. Space limitation does not allow us to present a broad survey. Instead we try to emphasize some work most related to what we propose.

Content-based image retrieval (CBIR) is a closely related field. Since the early 1990s, many CBIR systems have been developed. Some recent systems incorporates machine learning techniques. A recent article published by Smeulders et al. reviewed more than 200 references in this ever changing field [4]. Readers are referred to that article and some additional references [6, 7, 8] for more information.

Developed in late 1990's, the SIMPLIcity system [6] uses statistical classification methods to group images into rough semantic classes. A work on associating images explicitly with words is that of University of California at Berkeley [1], in which a hierarchical clustering model incorporating image features and text information is established to organize images in a database.

We have recently developed the Automatic Linguistic Indexing of Pictures (ALIP) system [5, 2]. Categories of images, each corresponding to a concept, are profiled by statistical models. In our system, we used the 2-D multiresolution hidden Markov model (2-D MHMM) [3]. The pictorial information of each image is summarized by a collection of feature vectors extracted at multiple resolutions and spatially arranged on a pyramid grid. The 2-D MHMM fitted to each image category plays the role of extracting representative information about the category. For a test image, feature vectors on the pyramid grid are computed. We consider the collection of the feature vectors as an instance of a spatial statistical model. The likelihood of this instance being generated by each profiling 2-D MHMM is computed. To annotate an image, words are selected from those in the text description of the categories yielding highest likelihoods.

This ALIP approach has two major advantages:

- *High scalability*: If images representing new concepts or new images in existed concepts are added into the training database, only the statistical models for the involved concepts need to be trained or retrained.
- *Universal similarity*: The modeling approach enables us to avoid segmenting images and defining a similarity distance for any particular set of features. Likelihood can be used as a universal measure of similarity.

While exploring automatic linguistic indexing, we no-

ticed that the evaluation process for such approaches is both challenging and interesting. In many related fields including image retrieval, image classification, and computer vision, gold standards are developed and used to evaluate the results. However, It can be very difficult to develop a reasonable gold standard for automatic linguistic indexing of pictures. In many application areas, it is desirable to have computer programs that are capable of providing semantically relevant keyword annotations, even if the annotations are different from that of the gold standard.

We will briefly illustrate our automatic linguistic indexing approach and discuss both objective and subjective evaluation strategies. We will report results of our experiments.

2. AUTOMATIC LINGUISTIC INDEXING OF PICTURES

The ALIP system we developed [5, 2] has three major modules, feature extraction, model-based learning, and linguistic indexing. We now provide a brief overview to the functionalities of these modules. The architecture of our system can be very different from other automatic linguistic indexing systems. We introduce ours because the evaluation strategies we propose are suitable for the architecture.

Nearly all image indexing and retrieval systems, content-based, semantic-sensitive, or automatic linguistic indexing capable, first characterize localized features of images. For each training image, we extract localized features using the wavelet transforms [6].

We manually prepared a training database of 600 concepts, each with about 40 photographic images from the COREL CD-ROM collection. The algorithm is not limited to 600 concepts and can also handle different number of training images per concept. The images are in JPEG format. For each of the 600 concepts, we manually created a description with a few keywords. On average, 3.6 keywords are used for each concept. The concepts range from as low-level as "flowers" to as high-level as "recreation, sport, water, ocean, people". It is not required that the training images for a concept must all be visually similar. For example, the training of the concept "flowers" may include red flowers, yellow flowers, large flowers, and small flowers. While annotating these concepts, the authors attempts to use words that properly describe nearly all images in the training set. The outliers, i.e., those images not described accurately by all keywords, are handled by the robust statistical process.

In the model-based training process, we create a statistical model for each set of 40 training images depicting a concept. A *dictionary* of 600 concepts is automatically generated, each with a computed statistical model.

After the creation of the dictionary or knowledge base, computers can then use the pre-computed models for linguistic indexing. We compute the likelihoods of an unannotated image resembling the pre-computed statistical models. For each image, the best few concepts are selected by sorting the likelihoods. To annotate the image using keywords, we compute the statistical significances of each of the keywords in the best few concepts. These keywords are then ranked according to their significances [5, 2].

3. EVALUATION STRATEGIES

We propose several evaluation strategies for automatic linguistic indexing of pictures. In this section, we introduce both objective and subjective evaluation methods and report some of our experimental results.

3.1. Objective evaluation

Objective evaluation methods can often be conducted automatically with a relatively large number of tests. The numerical results obtained are more convincing. However, as discussed earlier, it can be a difficult task when no agreeable gold standards are available. We propose two methods to estimate the lower bound of the system performance. These methods use "orthogonal" or non-overlapping concepts.

3.1.1. Automated image categorization

This is probably the simplest method to evaluate the performance of linguistic indexing. We treat a linguistic indexing system as an image classification system to estimate the lower bound of the performance. For a real world linguistic annotation task, a system does not have to classify an image into the correct category. For example, a system that categorizes a Paris scene as an European scene can still be a useful system.

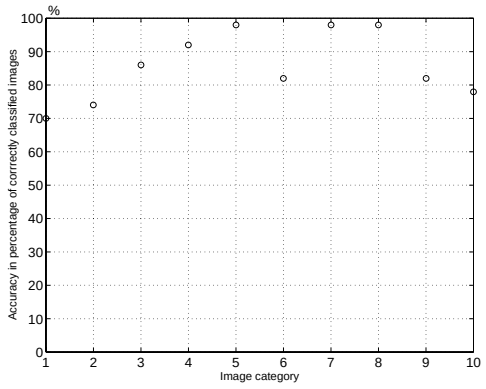

Fig. 1. Percentages of images classified to the same category as the manual classification.

In our experiment, we formed a test database using 10 image categories. The ten categories are: Africa people

and villages, beach, buildings, buses, dinosaurs, elephants, flowers, horses, mountains and glaciers, and food. We assume that images of one category are of different semantic contents from images of any other categories. We train the system using 45 images for each concept. Then we use 500 other images to test the training performance. We found that the performance varies from one concept to another. The accuracy ranges from 70% to 99% (Figure 1). This result is expected because some concepts, such as beach, are harder to learn than others. Without any prior knowledge, 40 images may not be enough even for human beings to learn the concept "beach". The results shown here is better than those published earlier [5, 2] because of some recent implementation changes and the correction of an error in the original feature extraction code.

On average, the system classifies approximately 85.8% of test images correctly. The performance of the system is clearly better than a system that classifies images by chance, when the expected accuracy is only 10%. With many more concepts in the database, the accuracy for each concept is expected to decrease using this evaluation method. Classification algorithms are more confused and less accurate with more concepts.

To demonstrate the scalability of the system, we randomly selected $5,970$ images from the entire database of 600 concepts. We used the system to put each image into one of the concepts. For each test image, the category yielding the highest likelihood is identified. If the test image is included in this category, we call it a "match". The total number of matches for the $5,970$ test images is 874. That is, an accuracy of 14.64% is achieved. In contrast, if random drawing is used to categorize the images, the accuracy is only 0.17%. If the condition of a "match" is relaxed to having the true category covered by the highest ranked *two* categories, the accuracy of ALIP increases to 20.50%, while the accuracy for the random scheme increases to 0.34%.

The absolute accuracy figures may not mean much because they depend on the number of concepts and the characteristics of these concepts. However, this evaluation strategy can reliably compare one system with another.

3.1.2. Keyword annotation

We developed a method to evaluate the keyword annotation performance of ALIP. This method can be used to provide numerical results for systems that annotate images with hundreds of possible keywords.

We randomly selected thousands of test images from our image database and processed these images by the linguistic indexing component of ALIP. For our experiments, we selected 5,970 images randomly. For each test image, the computer program selected 5 concepts in the dictionary with the highest likelihoods of generating the image. For every word in the annotation of the 5 concepts, the value indicating its significance is computed. We use the median of these values as a threshold to select annotation words from those assigned to the 5 matched concepts. A small value implies high significance. A word with a value below the threshold is selected.

We compare the system with a random annotation scheme, or the "monkey" system. The "monkey" system annotates the images randomly following the marginal distribution specified by the frequencies of words in the pool of all possible words. The "monkey" system is set to provide the same number of keywords to annotate an image as the average number of keywords per image provided by the ALIP system. For each image, we measure the percentage of manually-entered keywords that are predicted by the computer system. This method also estimates the lower bound of the ALIP performance because the system often provides meaningful keyword annotations that are not in the manual annotation of the concept.

When the system provides on average 6 (in the sense of median) keywords per image, the mean coverage percentage is 27.98% for ALIP, while that of the "monkey" scheme is only about 10%. The coverage percentage is defined by the percentage of manually annotated words that are included in the computer annotation. If all the words in the annotation of the 5 matched concepts are assigned to a query image, the median of the numbers of words assigned to the test images is 13. The mean coverage percentage for ALIP is then 56.64%, while that obtained from assigning 13 words by the random scheme is only about 18%. Clearly, ALIP has some intelligence.

Because images in the COREL collection are often not diverse enough to cover certain concepts, we examine the annotation of 250 images taken from 5 categories in the COREL database using only models trained from the other 595 categories. That is, no image in the same category as any of the 250 images is used in training. The mean coverage percentages obtained for these images by our system with on average 6 keywords and on average 13 keywords are roughly the same as the corresponding average values for the previous 5,970 test images. The mean coverage percentages achieved by randomly assigning 6 and 12 words to each image are about 11% and 18%.

3.2. Subjective evaluation

Objective evaluation methods alone may not be enough because a system with high accuracy may not produce usable keyword annotations. Subjective evaluation methods must be used to assure the developers that the system is producing useful results.

We tested ALIP using both COREL images outside the training database and images not taken by photographers of the COREL collection. Some results are available online at `http://wang.ist.psu.edu`. We closely exam

P: building, snow, sky, tree, landscape

P: flower

P: hist. bldg, (bridge), river, Italy, sky, Europe

P: animal, (squirrel), grass

Fig. 2. Test results using photos not in the COREL collection. **P:** Photographer annotation. Words appeared in the annotation of the 5 matched categories are underlined. Words in parenthesis are not included in the annotation of any of the 600 training categories. (Photos by: J. Z. Wang)

the keywords provided by the ALIP system with photographer annotations. Figure 2 shows the computer predictions of four images taken by ourselves. In 3 out of 4 cases, the ALIP system was able to predict most manually-entered keywords.

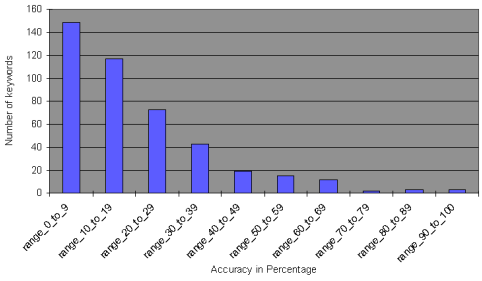

Fig. 3. A histogram of predictive accuracy over 435 words. 5970 images are tested.

The system can also be evaluated by a group of subjects. We developed a Web-based interface for human subjects to select semantically relevant computer annotations whenever the link is convincing. This evaluation method is time-consuming and difficult. Of the total of 435 possible annotation keywords used by the system, more than half of them have an over-10% predictive probability. For each of these words, over 10% of the time the human subjects clearly indicated the semantic connection between the image and the keyword. This method provides lower bound because there is no way for the human subjects to tell if an image is about Italy or not, for example. Some keywords, such as "leisure", "youth", and "green", have shown perfect accuracy. Other highly accurate words include "color", "red", "blue", "foliage", "landscape", "leaf", "orange", "sky", and "indoor".

In order to systematically consider overlapping semantics among words and concepts, we manually generated a semantic hierarchy for all the keywords. Then we used this hierarchy in the evaluation process. That is, if the word "Paris" is linked with the word "Europe" in the hierarchy, we consider a computer prediction of "Paris" to be correct if the true prediction is "Europe", or vice versa. Figure 3 shows the histogram of predictive accuracy of all the words. Because on average a category is annotated with 3.6 keywords, we considered only the top keywords selected by the computer so that the average number of keywords predicted by the computer is at the same level. For more than 65% of the words, more than 10% of the time the computer predicted words is connected with one of the manual category annotation keywords, through the semantic hierarchy.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we discussed the challenges in evaluating modern machine learning based linguistic indexing systems for pictures. We proposed several evaluation strategies for automatic linguistic indexing of pictures. Finally, we provide experimental results on our ALIP system using the proposed evaluation strategies.

A linguistic indexing system with broad applications must be capable of handling non-photographic images. We are currently applying the algorithm to other imagery types.

5. REFERENCES

[1] K. Barnard, D. Forsyth, "Learning the semantics of words and pictures," *Proc. ICCV*, vol. 2, pp. 408-415, 2001.

[2] J. Li, J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 25, no. 10, 2003.

[3] J. Li, R. M. Gray, R. A. Olshen, "Multiresolution image classification by hierarchical modeling with two dimensional hidden Markov models," *IEEE Trans. on Information Theory,* vol. 46, no. 5, pp. 1826-41, August 2000.

[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 Trans. on Pattern Analysis And Machine Intelligence,* vol. 22, no. 12, pp. 1349-1380, 2000.

[5] J. Z. Wang, J. Li, "Learning-based linguistic indexing of pictures with 2-D MHMMs," *Proc. ACM Multimedia*, pp. 436-445, Juan Les Pins, France, ACM, December 2002.

[6] J. Z. Wang, *Integrated Region-based Image Retrieval*, Kluwer Academic Publishers, Dordrecht, 2001.

[7] J. Z. Wang, J. Li, G. Wiederhold, "SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 23, no. 9, pp. 947-963, 2001.

[8] Q. Zhang, S. A. Goldman, W. Yu, and J. E. Fritts, "Content-based image retrieval using multiple-instance learning," *Proc. Int. Conf. on Machine Learning*, 2002.