

# ALIP: The Automatic Linguistic Indexing of Pictures System

Jia Li

The Pennsylvania State University  
University Park, PA 16802  
Email: [jjali@psu.edu](mailto:jjali@psu.edu)

James Z. Wang

The Pennsylvania State University  
University Park, PA 16802  
Email: [jwang@ist.psu.edu](mailto:jwang@ist.psu.edu)

## ABSTRACT

*In this demonstration, we present the Automatic Linguistic Indexing of Pictures (ALIP) system. The system annotates images with linguistic terms, chosen among hundreds of such terms. The system uses a wavelet-based approach for feature extraction, a statistical modeling process for training, and a statistical significance processor to annotate images. We implemented and tested our ALIP system on a photographic image database of 600 different concepts, each with about 40 training images. The ALIP system has been used to annotate about 60,000 photographic images. In this demonstration, we illustrate the algorithms in the system and show the annotation results. With distributed computation, the annotation of an image can be provided in real-time. The demonstration system is available online at the site <http://riemann.ist.psu.edu>.*

## THE ALIP SYSTEM

Automatic linguistic indexing of pictures is essentially important to content-based image retrieval [5] and computer object recognition. It can potentially be applied to many areas. Decades of research have shown that designing a generic computer algorithm that can learn concepts from images and automatically translate the content of images to linguistic terms is highly difficult. Since 2000, we have been developing our Automatic Linguistic Indexing of Pictures (ALIP) system [4], [2]. In our system, we trained a dictionary of 600 concepts using statistical modeling techniques. An extension of the ALIP work has been applied to the studying of ancient paintings [3].

In our work, categories of images, each corresponding to a concept, are profiled by statistical models, in particular, the 2-dimensional multi-resolution hidden Markov model (2-D MHMM) [1]. The pictorial information of each image is summarized by a collection of wavelet-based 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. In particular, a 2-D MHMM summarizes two types of information: clusters of feature vectors at multiple resolutions and the spatial relation between the clusters, both across and within resolutions. As a 2-D MHMM is estimated separately for each category, a new category of images added

to the database can be profiled without repeating computation involved with learning from the existing categories. The system naturally has good scalability without invoking any extra mechanism to address the issue. The scalability enables us to train a relatively large number of concepts at once.

Since each image category in the training set is manually annotated, a mapping between profiling 2-D MHMMs and sets of words can be established. For a test image, feature vectors on the pyramid grid are computed. 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 the image, words are selected from those in the text description of the categories yielding highest likelihoods.

## THE TRAINING CONCEPTS

We conducted experiments on learning-based linguistic indexing with a large number of concepts. The system was trained using a subset of 60,000 photographs based on 600 CD-ROMs published by COREL Corp. Typically, each COREL CD-ROM of about 100 images represents one distinct topic of interest.

We manually developed a series of concepts to be trained for inclusion in the *dictionary* of concepts. For each concept in this dictionary, we prepare a training set containing images capturing the concept. Hence at the data level, a concept corresponds to a particular category of images. These images do not have to be visually similar. We also manually prepare a short but informative description about any given concept in this dictionary. Table I shows some examples. Therefore, our approach has the potential to train a large collection of concepts because we do not need to manually create a description about each image in the training database.

## ANNOTATION RESULTS

Figure 1 shows the computer indexing results of some randomly selected images outside the training database. The method appears to be highly promising for automatic learning and linguistic indexing of images. Some of the computer predictions seem to suggest that one can control what is to be learned and what is not by adjusting the training database of individual concepts.


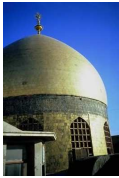







| Image   | Computer predictions                              | Image   | Computer predictions                          | Image   | Computer predictions                    |
|---|---|---|---|---|---|
|  | castle, landscape, Swiss, Europe, building, grass |  | building, silkroad, light house, architecture |  | fashion, female, mask, cloth            |
|  | subsea, fish, ocean animal, fractal               |  | landscape, desert, landmark                   |  | animal, grass, close, lion, rare animal |
|  | beach, ocean, lighthouse, travel, boat, Mexico    |  | group, city, New York, festival, life         |  | car, man-made, bus                      |

Fig. 1

ANNOTATIONS AUTOMATICALLY GENERATED BY OUR COMPUTER-BASED LINGUISTIC INDEXING ALGORITHM. THE *dictionary* WITH 600 CONCEPTS WAS CREATED AUTOMATICALLY USING STATISTICAL MODELING AND LEARNING. 36,000 TEST IMAGES WERE RANDOMLY SELECTED OUTSIDE THE TRAINING DATABASE FOR ANNOTATION.

TABLE I

EXAMPLES OF THE 600 CATEGORIES AND THEIR DESCRIPTIONS. EVERY CATEGORY HAS 40 TRAINING IMAGES.

| ID  | Category Descriptions   |
|-----|---|
| 0   | Africa, people, landscape, animal   |
| 10  | England, landscape, mountain, lake, European, people, historical building |
| 20  | Monaco, ocean, historical building, food, European, people                |
| 30  | royal guard, England, European, people                                    |
| 40  | vegetable   |
| 50  | wild life, young animal, animal, grass                                    |
| 60  | European, historical building, church                                     |
| 70  | animal, wild life, grass, snow, rock                                      |
| 80  | plant, landscape, flower, ocean   |
| 90  | European, historical building, grass, people                              |
| 100 | painting, European  |

## DISTRIBUTED ARCHITECTURE

The modeling process in ALIP allows massively parallel computation during both the training and the annotation processes. If we have  $n$  processors in a cluster computer and  $m$  concepts, the processes can be sped-up can be a factor of roughly  $m/\lceil m/n \rceil$ .

We can provide real-time annotation capability for users' images by providing a Web-based interface. The user can enter the URL of an image available on the Web in the interface. The interface handler program (master) downloads the image from the URL, extracts the features of this image, and sends the features to individual CPUs (slaves) in a cluster computer. Each CPU computes the log likelihoods of the features to a

sub-set of the trained models. The log likelihoods are then reported back to the master program. The master program analyzes the list of log likelihoods and determines a short list of keywords to annotate the image.

## ACKNOWLEDGMENTS

Jia Li is with the Department of Statistics and the Department of Computer Science and Engineering. James Z. Wang is with the School of Information Sciences and Technology and the Department of Computer Science and Engineering. This work was supported in part by the National Science Foundation. Xiaonan Lu assisted in the development of the Web-based engine wrapper and the distributed computation module. Ritendra Datta assisted in the delivery of the demonstration.

## REFERENCES

- [1] J. Li, R. M. Gray and 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, 2000.
- [2] J. Li and J. Z. Wang, "Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1075-1088, 2003.
- [3] J. Li and J. Z. Wang, "Studying Digital Imagery of Ancient Paintings by Mixtures of Stochastic Models," *IEEE Trans. on Image Processing*, vol. 13, no. 3, pp. 340-353, 2004.
- [4] J. Z. Wang and 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.
- [5] J. Z. Wang, J. Li and 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.