

# Mining Digital Imagery Data for Automatic Linguistic Indexing of Pictures\*

James Z. Wang

School of Information Sciences and Technology  
The Pennsylvania State University  
University Park, PA<sup>†</sup>

Jia Li

Department of Statistics  
The Pennsylvania State University  
University Park, PA<sup>‡</sup>

## Abstract

*In this paper, we present a new research direction, automatic linguistic indexing of pictures, for data mining and machine learning researchers. Automatic linguistic indexing of pictures is an imperative but highly challenging problem. In our on-going research, we introduce a statistical modeling approach to this problem. Computer algorithms have been developed to mine numerical features automatically extracted from manually annotated categorized images. These image categories form a computer-generated dictionary of hundreds of concepts for computers to use in the linguistic annotation process. In our experimental implementation, we focus on a particular group of stochastic processes for describing images. We implemented and tested our ALIP (Automatic Linguistic Indexing of Pictures) system on a photographic image database of 600 different semantic categories, each with about 40 training images. Tested using more than 4600 images outside the training database, the system has demonstrated good accuracy and high potential in linguistic indexing of photographic images. Such a system can potentially be used in many areas such as semantic Web and counter terrorism.*

## 1 Introduction

Human beings are constantly mining visual scenes encountered and stored in our brains. Based on the models of the world we acquired during the mining process, we can tell a story by looking at a picture. Experiments have shown

\*The Website <http://wang.ist.psu.edu> provides more information and demonstrations related to this work. This work is supported primarily by the National Science Foundation under Grant No. IIS-0219272, IIS-9817511, and IIS-0112641. We have also received support from The Pennsylvania State University, the PNC Foundation, and SUN Microsystems. Conversations with Michael Lesk have been very helpful. Portions of the work will be presented at ACM Conference on Multimedia, 2002. [10]

<sup>†</sup>J. Z. Wang is also affiliated with the Department of Computer Science and Engineering. Email: [wangz@cs.stanford.edu](mailto:wangz@cs.stanford.edu).

<sup>‡</sup>Email: [jiali@stat.psu.edu](mailto:jiali@stat.psu.edu).

that a 3-year old child is capable of building models of a substantial number of concepts and recognizing them using the learned models stored in her brain. Can a computer program learn a large collection of semantic concepts from 2-D or 3-D images, build models about these concepts, and recognize them based on these models? This is the question we attempt to address in our on-going research. This can be a new direction for next generation data mining and machine learning research.

*Automatic linguistic indexing of pictures* is essentially important to content-based image retrieval and computer object recognition. It can potentially be applied to many areas including biomedicine, commerce, the military, education, digital libraries, semantic Web, and counter terrorism. One potential application of such a computerized program is that it can automatically learn possible terrorist objects, process 3-D Computed Tomography scans of luggages, and warn airport security officers if suspect objects are potentially checked in to be aboard.

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. Many people believe that it cannot be achieved by computers because of the high complexity of the concepts we learn. Much success has been achieved in recognizing a relatively small set of objects or concepts within specific domains. There is a rich resource of prior work in the fields of computer vision, pattern recognition, and their applications [4]. Space limitations do not allow us to present a broad survey. Instead we try to emphasize some of the work that is most related to what we propose. The references below are to be taken as examples of related work, not as the complete list of work in the cited areas.

### 1.1 Related work on indexing images

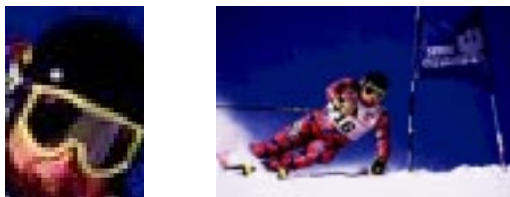
Many content-based image retrieval (CBIR) systems have been developed since the early 1990s. A recent article published by Smeulders et al. reviewed more than 200 ref-

erences in this ever changing field [9]. Readers are referred to that article and some additional references [7, 8, 13, 12, 2] for more information.

Most of the CBIR projects aimed at general-purpose image indexing and retrieval systems focusing on searching images visually similar to the query image or a query sketch. They do not have the capability to assign comprehensive textual description automatically to pictures, i.e., linguistic indexing, because of the great difficulties in recognizing a large number of objects. However, this function is essential for linking images to text and consequently broadening the possible usages of an image database.

Many researchers have attempted to use statistical data mining and machine learning techniques for image indexing and retrieval. The Stanford SIMPLicity system [11], developed by the authors of this paper, uses manually-defined statistical classification methods to classify the images into rough semantic classes, such as textured-nontextured, graph-photograph. Potentially, the categorization enhances retrieval by permitting semantically-adaptive searching methods and narrowing down the searching range in a database. The approach is limited because these classification methods are problem specific and must be manually developed and coded. A recent work in associating images explicitly with words is that of University of California at Berkeley [1]. An object name is associated with a region in an image based on previously learned region-term association probabilities.

## 1.2 Our approach



**Figure 1. It is often impossible to accurately determine the semantics of an image by looking at a single region of the image.**

Intuitively, human beings recognize many concepts from images based on the entire images. Often we need to view the image as a whole in order to determine the semantic meanings of each region and consequently tell a complete story about the image. For one example (Figure 1), if we look at a small portion of an image, i.e., the face of a person, we would not know that the image depicts the concept ‘ski’. But if we see in addition the clothing of the person, the equipment the person is holding, and the white snow in

the background, we can recognize easily the concept ‘ski’. Therefore, treating an image as an entity has the potential for modeling relatively high-level concepts as well as improving the modeling accuracy of low-level concepts.

In our work, we propose to mine and model entire images statistically. In our experimental implementation, we use a 2-D multiresolution hidden Markov model (MHMM) [5]. This statistical approach reduces the dependence on correct image segmentation because cross-pixel and cross-resolution dependencies are captured in the model itself. These models are created automatically by training on sets of images representing the same concepts. Machine-generated models of the concepts are then stored and used to automatically index images based on linguistic terms. Statistical image modeling is a research topic extensively studied in various fields including image processing and computer vision. Detailed review on some models used in image segmentation is provided in [5, 6].

## 1.3 Outline of the paper

The remainder of the paper is organized as follows: our ALIP (Automatic Linguistic Indexing of Pictures) system is introduced in Section 2. In Section 3, experiments and results are described. We conclude in Section 4.

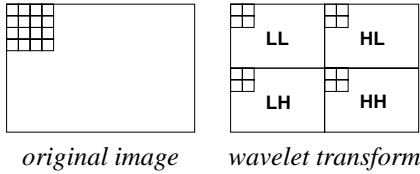
## 2 ALIP: Automatic Linguistic Indexing of Pictures

The ALIP system has three major components, the feature extraction process, the multiresolution statistical modeling process, and the statistical linguistic indexing process. In this section, we introduce these individual components and their relationships. Due to space limitation, we introduce the most fundamental ideas here. More details about the techniques are presented in [10].

### 2.1 Feature extraction

The system characterizes localized features of training images using wavelets. In this process, an image is partitioned into small pixel blocks. For our experiments, the block size is chosen to be  $4 \times 4$  as a compromise between the texture detail and the computation time. Other similar block sizes can also be used. The system extracts a feature vector of six dimensions for each block. Three of these features are the average color components in the block of pixels. The other three are texture features extracted to represent energy in high frequency bands of wavelet transforms [3]. Specifically, each of the three features is the square root of the second order moment of wavelet coefficients in one of the three high frequency bands. The feature extraction process is performed in the LUV color space, where L encodes

luminance, and U and V encode color information (chrominance). The LUV color space is chosen because of its good perception correlation properties.



**Figure 2. Decomposition of images into frequency bands by wavelet transforms.**

To extract the three texture features, we apply either the Daubechies-4 wavelet transform or the Haar transform to the L component of the image. These two wavelet transforms have better localization properties and require less computation compared to Daubechies’ wavelets with longer filters. After a one-level wavelet transform, a  $4 \times 4$  block is decomposed into four frequency bands as shown in Figure 2. Each band contains  $2 \times 2$  coefficients. Without loss of generality, suppose the coefficients in the HL band are  $\{c_{k,l}, c_{k,l+1}, c_{k+1,l}, c_{k+1,l+1}\}$ . One feature is then computed as

$$f = \frac{1}{2} \sqrt{\sum_{i=0}^1 \sum_{j=0}^1 c_{k+i,l+j}^2} .$$

The other two texture features are computed in a similar manner from the LH and HH bands, respectively.

The motivation for using these features is their reflection of local texture properties. Wavelet coefficients in different frequency bands signal variation in different directions. For example, the HL band shows activities in the horizontal direction. A local texture of vertical strips thus has high energy in the HL band of the image and low energy in the LH band. The use of this wavelet-based texture feature is a good compromise between computational complexity and effectiveness. The use of these features in the successful SIMPLIcity system [12] has demonstrated that they capture the image content.

## 2.2 Multiresolution statistical modeling

Figure 3 illustrates the flow of the statistical modeling process of the system. We first manually develop 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. Therefore, our approach has the potential to train a large collection of concepts because we do not need to manually create description about each image in the training database.

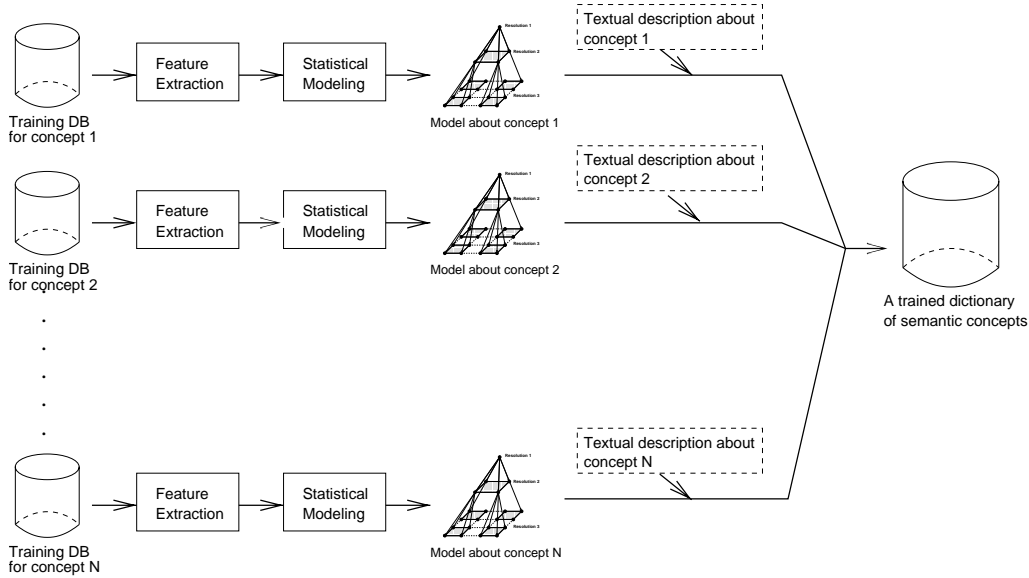
Block-based features are extracted from each training image at several resolutions. The statistical modeling process does not depend on a specific feature extraction algorithm. The same feature dimensionality is assumed for all blocks of pixels. A cross-scale statistical model about a concept is built using training images belonging to this concept, each characterized by a collection of multiresolution features. This model is then associated with the textual description of the concept and stored in the concept dictionary.

The statistical modeling process studies the multiresolution features extracted from each training image in the training database. A cross-scale statistical model about a concept is obtained after analyzing all available training images in a training database. This model is then associated with the textual description of the concept and stored in the concept dictionary.

To describe an image by a multiresolution model, multiple versions of the image at different resolutions are obtained first. The original image corresponds to the highest resolution. Lower resolutions are generated by successively filtering out high frequency information. Wavelet transforms [3] naturally provide low resolution images in the low frequency band (the LL band). Features are extracted at all the resolutions. The 2-D MHMM aims at describing statistical properties of the feature vectors and their spatial dependence.

In the 2-D MHMM, features are regarded as elements in a vector. They can be selected flexibly by users and are treated in an integrated manner in the sequel as dependent random variables by the model. Example features include color components and statistics reflecting texture. To save computation, feature vectors are often extracted from non-overlapping blocks in an image. An element in an image is therefore a block rather than a pixel. The numbers of blocks in both rows and columns reduce by half successively at each lower resolution. Obviously, a block at a lower resolution covers a spatially more global region of the image. The block at the lower resolution is referred to as a parent block, and the four blocks at the same spatial location at the higher resolution are referred to as child blocks. We will always assume such a “quad-tree” split in the sequel since the extension to other hierarchical structures is straightforward.

A 2-D MHMM reflects both the inter-scale and intra-scale statistical dependence. The inter-scale dependence is modeled by the Markov chain over resolutions. The intra-scale dependence is modeled by the HMM. At the coarsest resolution, feature vectors are assumed to be generated by a 2-D HMM. At all the higher resolutions, feature vectors



**Figure 3. The architecture of the statistical modeling process.**

of sibling blocks are also assumed to be generated by 2-D HMMs. The HMMs vary according to the states of parent blocks. Therefore, if the next coarser resolution has  $M$  states, then there are, correspondingly,  $M$  HMMs at the current resolution.

The 2-D MHMM can be estimated by the maximum likelihood criterion using the EM algorithm. Details about the estimation algorithm and the computation of the likelihood of an image given a 2-D MHMM are presented in [5].

### 2.3 Statistical linguistic indexing

The system automatically indexes images with linguistic terms based on statistical model comparison. Figure 4 shows the statistical linguistic indexing process of the system. For a given image to be indexed, we first extract multiresolution block-based features in the same manner as the feature extraction process for the training images.

This collection of feature vectors is statistically compared with the trained models stored in the concept dictionary to obtain a series of likelihoods representing the statistical similarity between the image and each of the trained concepts. These likelihoods, along with the stored textual descriptions about the concepts, are processed in the significance processor to extract a small set of statistically significant index terms about the image. These index terms are then stored with the image in the image database for future keyword-based query processing.

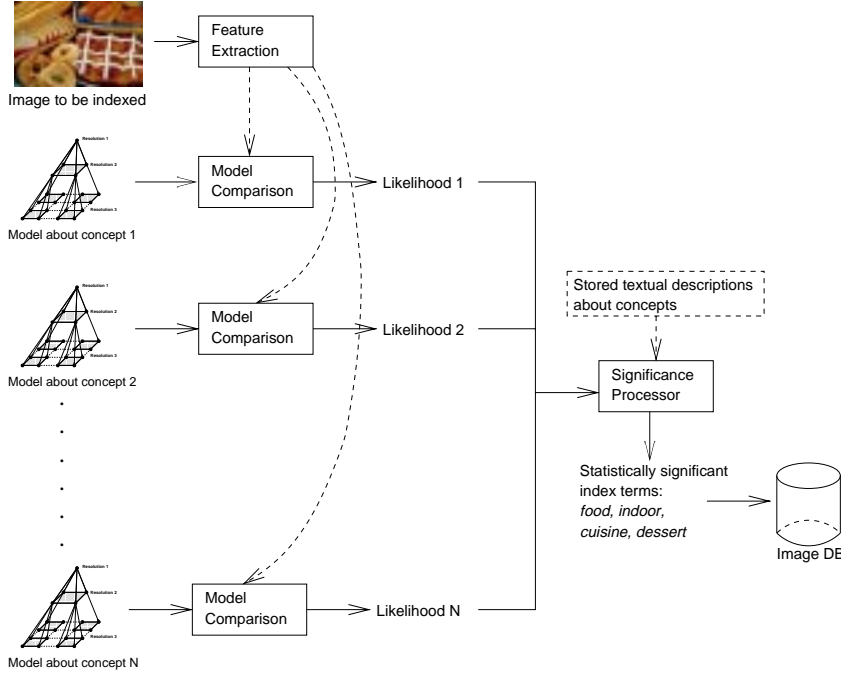
For any given image, a collection of feature vectors at multiple resolution  $\{u_{i,j}^{(r)}, r \in \mathcal{R}, (i, j) \in \mathbb{N}^{(r)}\}$  is computed. We regard  $\{u_{i,j}^{(r)}, r \in \mathcal{R}, (i, j) \in \mathbb{N}^{(r)}\}$  as an in-

stance of a stochastic process defined on a multiresolution grid. The similarity between the image and a category of images in the database is assessed by the log likelihood of this instance under the model  $\mathcal{M}$  trained from images in the category, that is,

$$\log P\{u_{i,j}^{(r)}, r \in \mathcal{R}, (i, j) \in \mathbb{N}^{(r)} \mid \mathcal{M}\}.$$

A recursive algorithm is used to compute the above log likelihood in a manner described in [5]. After determining the log likelihood of the image depicting any given concept in the dictionary, we sort the log likelihoods to find the few categories with the highest likelihoods. The short textual descriptions of these categories are loaded in the program in order to find the proper index terms for this image.

We use the most statistically significant index terms within the textual descriptions to index the image. Annotation words may have vastly different frequencies of appearing in the categories of an image database. For instance, many more categories may be described with the index term “landscape” than with the term “dessert”. Therefore, obtaining the index word “dessert” in the top ranked categories matched to an image is in a sense more surprising than obtaining “landscape” since the word “landscape” may have a good chance of being selected even by random matching. To measure the level of significance when a word appears  $j$  times in the top  $k$  matched categories, we compute the probability of obtaining the word  $j$  or more times in  $k$  randomly



**Figure 4. The architecture of the statistical linguistic indexing process.**

selected categories. This probability is given by

$$P(j, k) = \sum_{i=j}^k I(i \leq m) \frac{\binom{m}{i} \binom{n-m}{k-i}}{\binom{n}{k}} = \sum_{i=j}^k I(i \leq m) \frac{m! (n-m)! k! (n-k)!}{i! (m-i)! (k-i)! (n-m-k+i)! n!},$$

where  $I(\cdot)$  is the indicator function that equals 1 when the argument is true and 0 otherwise,  $n$  is the total number of image categories in the database, and  $m$  is the number of image categories that are annotated with the given word. The probability  $P(j, k)$  can be approximated as follows using the binomial distribution if  $n, m \gg k$ ,

$$P(j, k) = \sum_{i=j}^k \binom{k}{i} p^i (1-p)^{k-i} = \sum_{i=j}^k \frac{k!}{i! (k-i)!} p^i (1-p)^{k-i},$$

where  $p = m/n$  is the percentage of image categories in the database that are annotated with this word, or equivalently, the frequency of the word being used in annotation. A lower value of  $P(j, k)$  indicates a higher level of significance for a given index term. We rank the index terms within the short descriptions of the most likely concept categories according to their statistical significance. The terms with high significance are used to index the image.

## 2.4 Major advantages

Our system architecture has several major advantages:

1. 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. Hence, 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. This property is different from classification approaches that aim at forming decision boundaries between classes, e.g., neural networks, classification and regression trees (CART), and support vector machines (SVM). To form decision boundaries by these methods, a certain type of model needs to be established for the entire database. Therefore, every image is potentially involved in updating the decision boundaries when new images are added to the training set.
2. In our statistical model, spatial relations among image pixels and across image resolutions are both taken into consideration. This property is especially useful for images with special texture patterns. Moreover, 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.

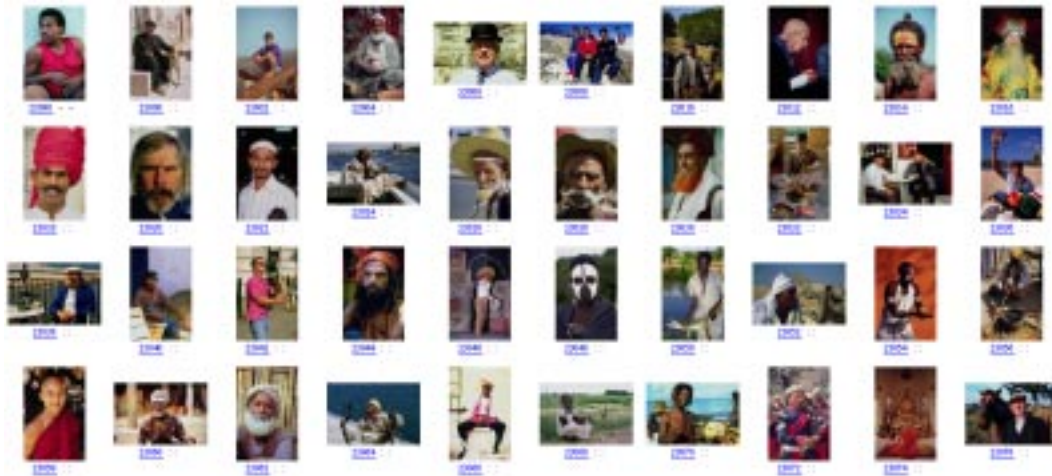


Figure 5. Training images used to learn the concept of *male* with the category description: “man, male, people, cloth, face”.

### 3 Experiments

To validate the methods we have described, we implemented the components of the ALIP system and tested with a general-purpose image database including about 60,000 photographs. These images are stored in JPEG format with size  $384 \times 256$  or  $256 \times 384$ . The system is written in the C programming language and compiled on two UNIX platforms: LINUX and Solaris. In this section, we describe the training concepts and show indexing results.

#### 3.1 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 which are based on 600 CD-ROMs published by COREL Corp. Typically, each COREL CD-ROM of about 100 images represents one distinct topic of interest. For our experiment, the dictionary of concepts contains all 600 concepts, each associated with one CD-ROM of images.

We manually assigned a set of keywords to describe each CD-ROM collection of 100 photographs. The semantic descriptions of these collections of images range from as simple or low-level as “mushrooms” and “flowers” to as complex or high-level as “England, landscape, mountain, lake, European, people, historical building” and “battle, rural, people, guard, fight, grass”. On average, 3.6 keywords are used to describe the content of each of the 600 concept categories. It took the authors approximately 10 hours to annotate these categories. For each concept category, we train the system with 40 training images (Figure 5).








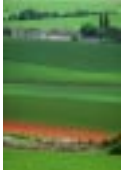













While manually annotating categories, the authors made efforts to use words that properly describe nearly all if not all images in one category. It is possible that a small number of images are not described accurately by all words assigned to their category. We view them as “outliers” introduced into training for the purpose of estimating the 2-D MHMM. In practice, outliers often exist for various reasons. There are ample statistical methods to suppress the adverse effect of them. On the other hand, keeping outliers in training will testify the robustness of a method. For the model we use, the number of parameters is small relative to the amount of training data. Hence the model estimation is not anticipated to be affected considerably by inaccurately annotated images. We therefore simply use those images as normal ones.

#### 3.2 Results

After the training, a statistical model is generated for each of the 600 collections of images. Depending on the complexity of the concept, the training process takes between 15 to 40 minutes of CPU time on an 800 MHz Pentium III PC to converge on a model. On average, 30 minutes of CPU time is spent to train a concept. The training process is conducted only once for each concept in the list.

These models are stored in a fashion similar to a dictionary or encyclopedia. Essentially, we use computers to create a dictionary of concepts that will enable computers to index images linguistically. The process is entirely parallelizable because the training of one concept is independent from the training of other concepts in the same dictionary.

We randomly selected 3,000 test images outside the training image database and processed these images by the

Image	Computer predictions	Image	Computer predictions	Image	Computer predictions
	building,sky,lake, landscape, European,tree		snow,animal, wildlife,sky, cloth,ice,people		people,European, female
	food,indoor, cuisine,dessert		people, European, man-made, water		lake,Portugal, glacier,mountain, water
	skyline, sky, New York, landmark		plant,flower, garden		modern,parade, people
	pattern,flower, red,dining		ocean,paradise, San Diego, Thailand, beach,fish		elephant,Berlin, Alaska
	San Diego, ocean side, beach,Florida, Thailand,building		relic,Belgium, Portugal,art		fitness,indoor, Christmas, cloth,holiday
	flower,flora, plant,fruit, natural,texture		travel,fountain, European, Florida, beach,building		Africa,Kenya, Zimbabwe, animal,cave
	ancestor, drawing, fitness, history, indoor		hair style, occupation,face, female,cloth		night,cyber, fashion,female

**Figure 6. Annotations automatically generated by our computer-based linguistic indexing algorithm. The dictionary with 600 concepts was created automatically using statistical modeling and learning. Test images were randomly selected outside the training database.**

linguistic indexing component of the system. For each of the 3,000 test images, the computer program selected a number of concepts in the dictionary with the highest likelihood of describing the image. Next, the most significant index terms for the image are extracted from the collection of index terms associated with the chosen concept categories.

It takes an average of two seconds CPU time on the same PC to compute the likelihood of a test image resembling one of the concepts in the dictionary. Thus, for computing the likelihoods of a test image resembling all the concepts in the dictionary of 600 concepts can take an average of 20 minutes of CPU time. The process is highly parallelizable because the computation of the likelihood to a concept is independent from the computation of the likelihood to another concept. We are planning to implement the algorithms on massively parallel computers and provide real-time online demonstrations in the future.

Figure 6 shows the computer indexing results of 21 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. As indicated in the second example, the computer predictions of a wildlife animal picture include “cloth” and “people”. It is possible that the computer learned the association between animal fur and the clothes of people from the training databases which contain images with female super-models wearing fur coats. Consequently, computer predictions are objective and without human subjective biases. Potentially, computer-based indexing of images eliminates the inconsistency problems commonly associated with manual image annotations.

## 4 Conclusions

In this paper, we presented a new direction for next generation data mining and machine learning research. We demonstrated our statistical data mining and modeling approach to the problem of automatic linguistic indexing of pictures. We have shown that the proposed methods can be used to train 600 different semantic concepts at the same time and these models can be used to index images linguistically. The major advantages with this approach are (1) models for different concepts can be independently trained and retrained so that a relatively large number of concepts can be trained and stored; (2) spatial relation among image pixels and across image resolutions is taken into consideration with probabilistic likelihood as a universal measure. The work can be potentially applied to many areas.

## References

- [1] K. Barnard, D. Forsyth, “Learning the semantics of words and pictures,” *Proc. ICCV*, vol 2, pp. 408-415, 2001.
- [2] Y. Chen, J. Z. Wang, “A region-based fuzzy feature matching approach to content-based image retrieval,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, 2002.
- [3] I. Daubechies, *Ten Lectures on Wavelets*, Capital City Press, 1992.
- [4] D. A. Forsyth, J. Ponce, *Computer Vision: A Modern Approach*, Prentice Hall, 2002.
- [5] 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.
- [6] J. Li, R. M. Gray, *Image Segmentation and Compression Using Hidden Markov Models*, Kluwer Academic Publishers, 2000.
- [7] S. Ravela, R. Manmatha, “Image retrieval by appearance,” *Proc. of SIGIR*, pp. 278-285, Philadelphia, July 1997.
- [8] G. Sheikholeslami, S. Chatterjee, A. Zhang, “WaveCluster: A multi-resolution clustering approach for very large spatial databases” *Proc. of the VLDB Conf.*, pp. 428-439, New York City, August 1998.
- [9] 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.
- [10] J. Z. Wang, J. Li, “Learning-based linguistic indexing of pictures with 2-D MHMMs,” *Proc. ACM Multimedia*, Dec. 2002.
- [11] J. Z. Wang, *Integrated Region-based Image Retrieval*, Kluwer Academic Publishers, Dordrecht, 2001.
- [12] 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.
- [13] J. Z. Wang, G. Wiederhold, O. Firschein, X. W. Sha, “Content-based image indexing and searching using Daubechies’ wavelets,” *Int. J. of Digital Libraries(IJODL)*, vol. 1, no. 4, pp. 311-328, Springer-Verlag, 1998.