

# Learning-based Linguistic Indexing of Pictures with 2-D MHMMs

James Z. Wang<sup>\*</sup>

School of Information Sciences and Technology  
The Pennsylvania State University  
University Park, PA 16802  
jwang@ist.psu.edu

Jia Li

Department of Statistics  
The Pennsylvania State University  
University Park, PA 16802  
jjali@stat.psu.edu.

## ABSTRACT

Automatic linguistic indexing of pictures is an important but highly challenging problem for researchers in computer vision and content-based image retrieval. In this paper, we introduce a statistical modeling approach to this problem. Categorized images are used to train a dictionary of hundreds of concepts automatically based on statistical modeling. Images of any given concept category are regarded as instances of a stochastic process that characterizes the category. To measure the extent of association between an image and the textual description of a category of images, the likelihood of the occurrence of the image based on the stochastic process derived from the category is computed. A high likelihood indicates a strong association. In our experimental implementation, the ALIP (Automatic Linguistic Indexing of Pictures) system, we focus on a particular group of stochastic processes for describing images, that is, the two-dimensional multiresolution hidden Markov models (2-D MHMMs). We implemented and tested the system on a photographic image database of 600 different semantic categories, each with about 40 training images. Tested using 3,000 images outside the training database, the system has demonstrated good accuracy and high potential in linguistic indexing of these test images.

**Index Terms** – Content-based image retrieval, image classification, hidden Markov model, computer vision, machine learning, image segmentation, region matching, wavelets.

## 1. INTRODUCTION

A picture is worth a thousand words. As human beings, we are able to tell a story from a picture based on what we have seen and what we have been taught. 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 col-

---

<sup>\*</sup>James Z. Wang is also with the Department of Computer Science and Engineering.

lection 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 this work.

*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, and Web searching. Decades of research has 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.

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 [12]. 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

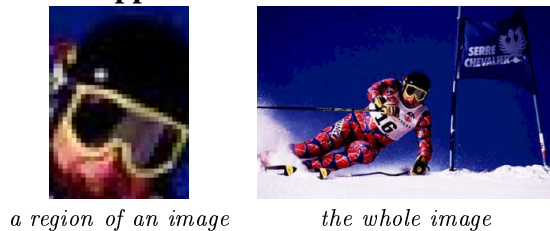
Many content-based image retrieval (CBIR) systems have been developed since the early 1990s [11, 7, 20, 22, 18, 4, 25, 19, 6]. Most of the above mentioned projects aimed at general-purpose image indexing and retrieval systems focusing on searching images visually similar to the query image or a query 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.

A growing trend in the field of image retrieval is to linguistically index images using computer programs relying on statistical classification methods. The Stanford SIMPLcity system [24] 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 Barnard and Forsyth 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. The work has achieved some success for certain image types. But as pointed out by the authors in [1], one major limitation is that the algorithm relies on semantically meaningful segmentation, which is in general unavailable to image databases. Automatic segmentation is still an open problem in computer vision [27, 21, 26]. Moreover, some concepts may not be learned from a single region of an image or a region within an over-segmented image. For example, when a tiger object of a test image is segmented into the head and the body segments, the computer program assigns the keyword “building” to the head portion of the tiger due to the similarity between the features of the head and those of buildings.

## 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 not just one region of an image. 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 only a region 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 cloth 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 model the entire images statistically. In our experimental implementation, we use a 2-D multiresolution hidden Markov model (MHMM) [15]. 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. Readers are referred to Li and Gray [16] for details on 2-D MHMM. For clarity of this paper, we present basics regarding the models.

Statistical image modeling is a research topic extensively studied in various fields including image processing and com-

puter vision. Detailed review on some models used in image segmentation is provided in [15]. Theories and methodologies related to Markov random fields (MRFs) [10, 13, 14, 5] have played important roles in the construction of many statistical image models. For a thorough introduction to MRFs and their applications, see Kindermann and Snell [14] and Chellappa and Jain [5].

The 2-D multiresolution hidden Markov model has been successfully applied to image segmentation and compression. This model explores statistical dependence among image pixels or blocks across multiple resolution as well as within a single resolution. It possesses a flexible structure to allow different proportions of emphasis on inter-resolution and intra-resolution dependency. As a result, users have ample freedom in choosing a particular form of the model according to application targeted. Analytic formula for estimating the model by the maximum likelihood criterion and for computing likelihood of an instance based on the model are available. A 2-D MHMM estimated from training images summarizes two types of information: clusters of feature vectors at multiple resolutions and the spatial relation between those clusters. The clusters of feature vectors normally reflect color and texture. Given its modeling efficiency and computational convenience, we consider 2-D MHMM an appropriate starting point for exploring the statistical modeling approach to linguistic indexing.

## 1.3 Outline of the paper

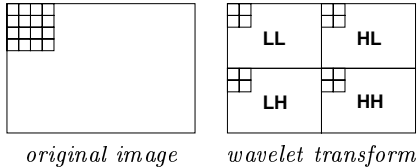
The remainder of the paper is organized as follows: an overview of our ALIP (Automatic Linguistic Indexing of Pictures) system is introduced in Section 2. The model learning algorithm is described in Section 3. Linguistic indexing methods are presented in Section 4. In Section 5, experiments and results are described. We conclude and suggest future research in Section 6.

## 2. SYSTEM OVERVIEW

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 the basics about these individual components and their relationships.

### 2.1 Feature extraction

The ALIP 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  $4 \times 4$  to 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 [9]. 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. Unser [23] has shown that moments of wavelet coefficients in various frequency bands can be used to effectively discern local texture. 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.

## 2.2 Multiresolution statistical modeling

We first manually enlist a series of concepts to be trained for inclusion in the *dictionary* of concepts. For each concept in this dictionary, we prepare a training image database with a set of images capturing the concept. 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 descriptions about each image in the training database.

Block-based features are extracted from each of these training images at several different 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.

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.

## 2.3 Statistical linguistic indexing

The ALIP system automatically indexes images with linguistic terms based on statistical model comparison. 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.

## 2.4 Major advantages

Our ALIP system has several major advantages:

1. If new images are added to a given concept training database, only the particular statistical model of this given concept needs to be retrained. We need not change the trained models about other concepts in the same dictionary. This property is very different from conventional training-based classification approaches based on neural networks [2], classification and regression trees (CART) [3], and support vector machines (SVM) [8].
2. Because models of different concepts are independent of each others in our system, we can train a relatively large number of concepts at once. A statistical model, established for a category of images, serves as a pictorial description for the entire category and enables efficient association of textual annotations with image pixel representations. In our experiments, we have trained the system to automatically create a dictionary of 600 concepts.
3. In our initial 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. We can avoid segmenting images or defining a similarity distance for any particular set of features. Likelihood can be used as a universal measure of similarity. With this statistical likelihood approach, images used to train the same semantic concept do not have to be all visually similar.

## 3. THE MODEL-BASED LEARNING OF CONCEPTS

In this section, we provide details about our statistical image modeling process which learns a dictionary of a large number of concepts automatically. We present here assumptions of the 2-D MHMM which is modified from the model originally developed for the purpose of image segmentation [15]. The model studies the collection of training images within a concept category in their entirety.

### 3.1 Image Modeling

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 [9] 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.

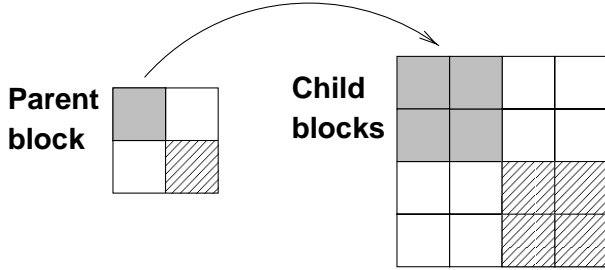


Figure 3: The image hierarchy across resolutions

In the 2-D MHMM, features are regarded as elements in a vector. They can be selected flexibly by users and are treated integrately 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. As is indicated by Figure 3, 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.

We first review the basic assumptions of the single resolution 2-D HMM as presented in [17]. In the 2-D HMM, feature vectors are generated by a Markov model that may change state once every block. Suppose there are  $M$  states, the state of block  $(i, j)$  being denoted by  $s_{i,j}$ . The feature vector of block  $(i, j)$  is  $u_{i,j}$ . We use  $P(\cdot)$  to represent the probability of an event. We denote  $(i', j') < (i, j)$  if  $i' < i$  or  $i' = i, j' < j$ , in which case we say that block  $(i', j')$  is before block  $(i, j)$ . The first assumption is that

$$\begin{aligned} P(s_{i,j} \mid \text{context}) &= a_{m,n,l}, \\ \text{context} &= \{s_{i',j'}, u_{i',j'} : (i', j') < (i, j)\}, \end{aligned}$$

where  $m = s_{i-1,j}$ ,  $n = s_{i,j-1}$ , and  $l = s_{i,j}$ . The second assumption is that for every state, the feature vectors follow a Gaussian distribution. Once the state of a block is known, the feature vector is conditionally independent of information in other blocks. The covariance matrix  $\Sigma_s$  and the mean vector  $\mu_s$  of the Gaussian distribution vary with state  $s$ .

Given an image, only feature vectors are observable, the reason that the model is named as a hidden Markov model. The state of a feature vector is conceptually parallel to the cluster identity of a vector in unsupervised clustering. As with clustering, the state of vector is not provided directly by the training data and hence needs to be estimated. In clustering, feature vectors are considered as independent samples from a given distribution. In the 2-D HMM, feature vectors are statistically dependent through the underlying states characterized by Markovian properties.

For the MHMM, denote the collection of resolutions by  $\mathcal{R} = \{1, \dots, R\}$ , with  $r = R$  being the finest resolution. Let the collection of block indices at resolution  $r$  be

$$\mathbb{N}^{(r)} = \{(i, j) : 0 \leq i < w/2^{R-r}, 0 \leq j < z/2^{R-r}\}.$$

Images are described by feature vectors at all the resolutions, denoted by  $u_{i,j}^{(r)}$ ,  $r \in \mathcal{R}$ . The underlying state of a feature vector is  $s_{i,j}^{(r)}$ . At each resolution  $r$ , the set of states is  $\{1^{(r)}, 2^{(r)}, \dots, M_r^{(r)}\}$ . Note that as states vary across resolutions, different resolutions do not share states.

To structure statistical dependence among resolutions, a Markov chain with resolution playing a time-like role is assumed in the 2-D MHMM. Given the states and the features at the parent resolution, the states and the features at the current resolution are conditionally independent of the other previous resolutions, so that

$$\begin{aligned} P &\{s_{i,j}^{(r)}, u_{i,j}^{(r)} : r \in \mathcal{R}, (i, j) \in \mathbb{N}^{(r)}\} \\ &= P\{s_{i,j}^{(1)}, u_{i,j}^{(1)} : (i, j) \in \mathbb{N}^{(1)}\} \times \\ &\quad P\{s_{i,j}^{(2)}, u_{i,j}^{(2)} : (i, j) \in \mathbb{N}^{(2)} \mid s_{k,l}^{(1)} : (k, l) \in \mathbb{N}^{(1)}\} \times \dots \times \\ &\quad P\{s_{i,j}^{(R)}, u_{i,j}^{(R)} : (i, j) \in \mathbb{N}^{(R)} \mid s_{k,l}^{(R-1)} : (k, l) \in \mathbb{N}^{(R-1)}\}. \end{aligned} \quad (1)$$

At the coarsest resolution,  $r = 1$ , feature vectors are assumed to be generated by a single resolution 2-D HMM. At a higher resolution, the conditional distribution of a feature vector given its state is also assumed to be Gaussian. The parameters of the Gaussian distribution depend upon the state at the particular resolution.

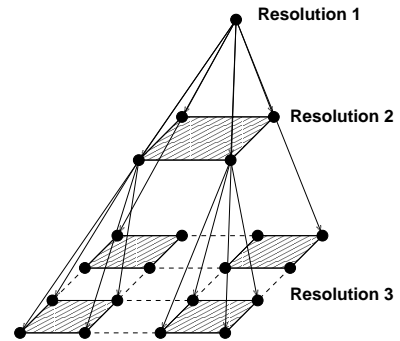


Figure 4: The hierarchical statistical dependence across resolutions

Given the states at resolution  $r - 1$ , statistical dependence among blocks at the finer resolution  $r$  is constrained to sibling blocks (child blocks descended from the same parent block). Specifically, child blocks descended from different

parent blocks are conditionally independent. In addition, given the state of a parent block, the states of its child blocks are independent of the states of their “uncle” blocks (non-parent blocks at the parent resolution). State transitions among sibling blocks are governed by the same Markovian property assumed for a single resolution 2-D HMM. The state transition probabilities, however, depend on the state of their parent block. To formulate these assumptions, denote the child blocks at resolution  $r$  of block  $(k, l)$  at resolution  $r - 1$  by

$$\mathbb{D}(k, l) = \{(2k, 2l), (2k + 1, 2l), (2k, 2l + 1), (2k + 1, 2l + 1)\}.$$

According to the assumptions,

$$\begin{aligned} & P\{s_{i,j}^{(r)} : (i, j) \in \mathbb{N}^{(r)} \mid s_{k,l}^{(r-1)} : (k, l) \in \mathbb{N}^{(r-1)}\} \\ &= \prod_{(k,l) \in \mathbb{N}^{(r-1)}} P\{s_{i,j}^{(r)} : (i, j) \in \mathbb{D}(k, l) \mid s_{k,l}^{(r-1)}\}, \end{aligned}$$

where  $P\{s_{i,j}^{(r)} : (i, j) \in \mathbb{D}(k, l) \mid s_{k,l}^{(r-1)}\}$  can be evaluated by transition probabilities conditioned on  $s_{k,l}^{(r-1)}$ , denoted by  $a_{m,n,l}(s_{k,l}^{(r-1)})$ . We thus have a different set of transition probabilities  $a_{m,n,l}$  for every possible state in the parent resolution. The influence of previous resolutions is exerted hierarchically through the probability of the states, which can be visualized in Figure 4. The joint probability of states and feature vectors at all the resolutions in (2) is then derived.

To summarize, 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. Figure 5 illustrates the inter-scale and intra-scale dependencies modeled. At all the higher resolutions, feature vectors 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 referred to [15].

## 4. THE AUTOMATIC LINGUISTIC INDEXING OF PICTURES

In this section, we describe the component of the ALIP system that automatically indexes pictures with linguistic terms. For a given image, the system compares the image statistically with the trained models in the concept dictionary and extract the most statistically significant index terms to describe the image.

For any given image, a collection of feature vectors at multiple resolution  $\{u_{i,j}^{(r)}, r \in \mathcal{R}, (i, j) \in \mathbb{N}^{(r)}\}$  are computed as described in Section 3. We regard  $\{u_{i,j}^{(r)}, r \in \mathcal{R}, (i, j) \in \mathbb{N}^{(r)}\}$  as an instance 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 [15]. 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, much 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 selected categories. This probability is given by

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

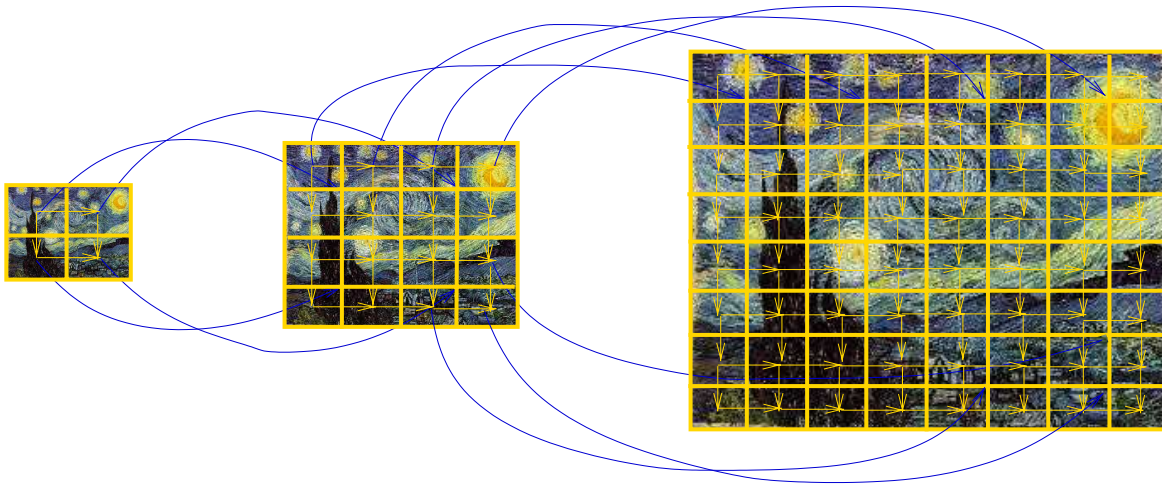
where  $I(\cdot)$  is the indicator function that equals 1 when the argument is true and 0 otherwise,  $n$  is the total number of image categories in the database, and  $m$  is the number of image categories that are annotated with the given word. The probability  $P(j, k)$  can be approximated as follows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

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

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.

## 5. 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.



**Figure 5:** In the statistical modeling process, spatial relations among image pixels and across image resolutions are both taken into consideration. Arrows, not all drawn, indicate the transition probabilities captured in the statistical model.

### 5.1 Training concepts

We conducted experiments on learning-based linguistic indexing with a large number of concepts. The ALIP 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 represent 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. Table 1 shows a sample of these annotations.

### 5.2 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 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. 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.

Figure 8 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 clothes. 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.

### 5.3 Systematic evaluation

To provide numerical results on the performance, we evaluated the ALIP system based on a subset of the COREL database, formed by 10 image categories (Africa people and villages, beach, buildings, buses, dinosaurs, elephants, flowers, horses, mountains and glaciers, food), each containing 100 pictures. Within this database, it is known whether any two images are of the same category. We trained each concept using 40 images and test the models using 500 images outside the training database. Instead of annotating

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
110	flower
120	decoration, man-made
130	Alaska, landscape, house, snow, mountain, lake
140	Berlin, historical building, European, landscape
150	Canada, game, sport, people, snow, ice
160	castle, historical building, sky
170	cuisine, food, indoor
180	England, landscape, mountain, lake, tree
190	fitness, sport, indoor, people, cloth
200	fractal, man-made, texture
210	holiday, poster, drawing, man-made, indoor
220	Japan, historical building, garden, tree
230	man, male, people, cloth, face
240	wild, landscape, north, lake, mountain, sky
250	old, poster, man-made, indoor
260	plant, art, flower, indoor
270	recreation, sport, water, ocean, people
280	ruin, historical building, landmark
290	sculpture, man-made

Table 1: Examples of the 600 categories and their descriptions. Every category has 40 training images.

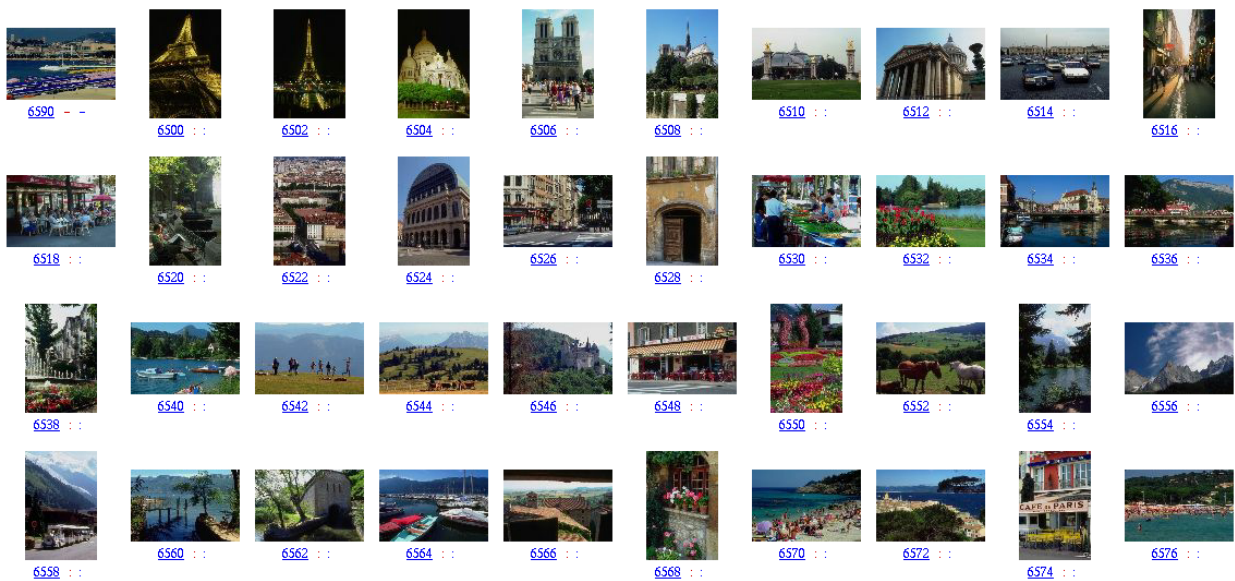


Figure 6: Training images used to learn a given concept are not necessarily all visually similar. For example, these 40 images were used to train the concept of *Paris* with the category description: “Paris, European, historical building, beach, landscape, water”.



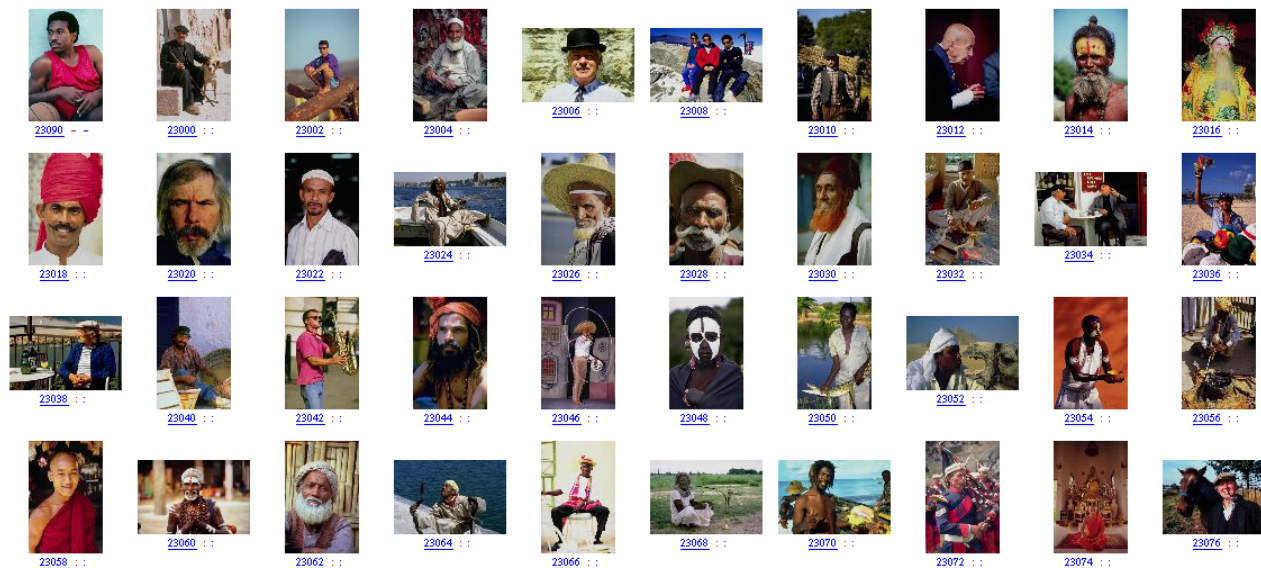


Figure 7: Training images used to learn the concept of *male* with the category description: “man, male, people, cloth, face”.

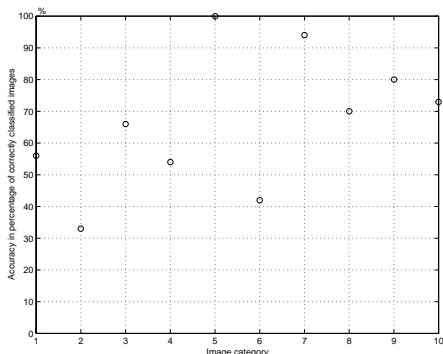


Figure 10: Percentages of images classified to the same category as the manual classification. The experiment is conducted on a test image database of 10 image categories. (also see Figure 9)

the images, the program was used to select the category with the highest likelihood for each test image. That is, we use the classification power of the system as an indication of the accuracy. An image is considered to be classified correctly if the computer predicts the original category the image belongs to. This assumption is reasonable since the 10 categories were chosen so that each depicts a substantially distinct semantic topic.

Figure 9 shows the automatic classification result as compared with the original image classification. Figure 10 plots the accuracy in each category as the percentage of images correctly classified. As indicated in the plots, some of the concepts can be related. For example, both the “beach” and the “mountains and glaciers” categories contain images with rocks, sky, and trees. Moreover, the system is designed for the purpose of automatically annotating images where the

major challenge is to model hundreds of different concepts rather than to classify images into a small set of classes with high accuracy. As a result, the evaluation method we use here can be used to assess the lower bounds of the annotation accuracy of the system for the given concept categories.

## 6. CONCLUSIONS AND FUTURE WORK

In this paper, we demonstrated our statistical modeling approach to the problem of automatic linguistic indexing of pictures for the purpose of image retrieval. In our ALIP system, we used categorized images to train a dictionary of hundreds of concepts automatically. Wavelet-based features are used to describe local color and texture in the images. After analyzing all training images for a training concept, a two-dimensional multiresolution hidden Markov model (2-D MHMM) is created and stored in a concept dictionary. Images in one category is regarded as instances of a stochastic process that characterizes the category. To measure the extent of association between an image and the textual description of a category of images, we compute the likelihood of the occurrence of the image based on the stochastic process derived from the category. We have demonstrated 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 our approach are (1) models for different concepts can be independently trained and re-trained; (2) a relatively large number of concepts can be trained and stored; (3) spatial relation among image pixels and across image resolutions is taken into consideration with probabilistic likelihood as a universal measure.

The current ALIP system has several limitations.

- We are training the concept dictionary with only 2-D images without a sense of object size. It is believed



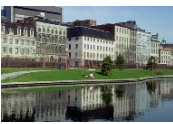

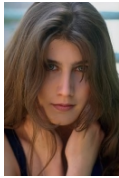




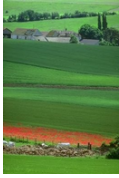






Image	Computer pre- dictions	Image	Computer pre- dictions	Image	Computer pre- dictions
	building,sky,lake, landscape, Euro- pean,tree		snow,animal, wildlife,sky, cloth,ice,people		people,European, female
	food,indoor, cui- sine,dessert		people, European, man-made, water		lake,Portugal, glacier,mountain, water
	skyline, sky, New York, landmark		plant,flower, gar- den		modern,parade, people
	pattern,flower, red,dining		ocean,paradise, San Diego, Thai- land, beach,fish		flower,flora, plant,fruit, natu- ral,texture
	ancestor, drawing, fitness, history, indoor		hair style, occupation,face, female,cloth		night,cyber, fash- ion,female

Figure 8: 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.

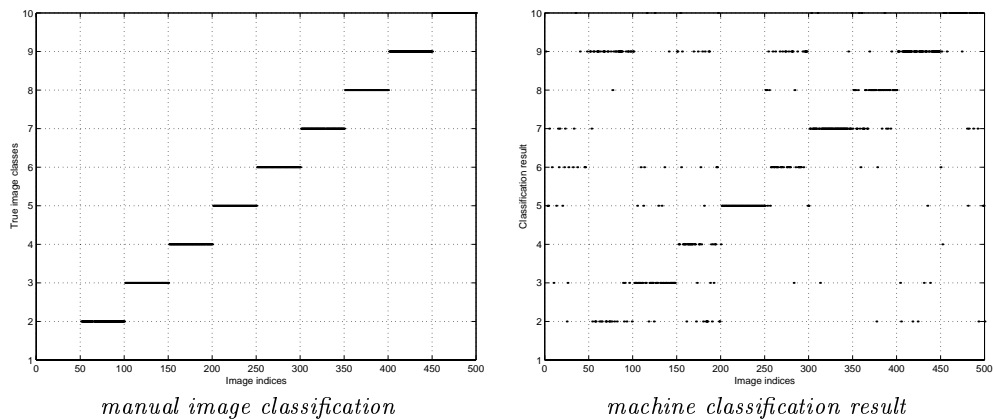


Figure 9: Results for the application of ALIP to the automatic image categorization problem. Left: the true class identities of the test images indexed from 1 to 500. Every 50 images belong to the same class. For images with indices  $50(i - 1) + 1$  to  $50i$ , the correct class identities are  $i$ . Right: the class identities predicted by our system for the 500 images. The experiment is conducted on a test image database of 10 image categories.

that the object recognizer of human beings are usually trained using 3-D stereo with motion and a sense of object sizes. Training with 2-D still images potentially limits the ability of accurate learning of concepts. We are currently attempting to work on training with 3-D images.

- Related concepts can sometimes be confusing if the models are trained with only positive examples. The current statistical model captures the association between a set of positive training example and their semantic concept. However, when a part of an association is not desired, the model should include a mechanism to make necessary corrections. We are exploring the possibility of adjusting the model based on negative examples provided to the system.
- For very complex concepts, i.e., when images representing the concept have very different appearances, it seems that 40 training images are not enough for the computer program to build a reliable model. The more complex the concept, the more training images are needed and the more CPU time is required. This is expected as it takes human beings more experiences and longer time to learn more complex concepts.

## 7. ACKNOWLEDGMENTS

The SIMPLicity work was supported in part by the US National Science Foundation under grant IIS-9817511 and Stanford University. This work is supported primarily by The Pennsylvania State University, the US National Science Foundation, the PNC Foundation, and SUN Microsystems under grant EDUD-7824-010456-US. We would like to acknowledge the comments and constructive suggestions from anonymous reviewers. Conversations with Michael Lesk have been very helpful.

## 8. REFERENCES

- [1] K. Barnard, D. Forsyth, "Learning the Semantics of Words and Pictures," *Proc. ICCV*, vol 2, pp. 408-415, 2001.
- [2] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.
- [3] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, *Classification and Regression Trees*, Wadsworth Int. Group, Belmont, Calif., 1984.
- [4] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, J. Malik, "Blobworld: A system for region-based image indexing and retrieval," *Proc. Int. Conf. on Visual Information Systems*, D. P. Huijsmans, A. W.M. Smeulders (eds.), Springer, Amsterdam, The Netherlands, June 2-4, 1999.
- [5] R. Chellappa and A. K. Jain, *Markov Random Fields: Theory and Applications*, Academic Press, 1993.
- [6] 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, 2002.
- [7] W. W. Chu, I.T. Leong, R.K. Taira, "A semantic modeling approach for image retrieval by content," *The VLDB Journal*, vol. 3, no. 4, pp. 445-477, 1994.
- [8] N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods*, Cambridge University Press, 2000.
- [9] I. Daubechies, *Ten Lectures on Wavelets*, Capital City Press, 1992.
- [10] R. L. Dobrushin, "The description of a random field by means of conditional probabilities and conditions of its regularity," *Theory Prob. Appl.*, vol. 13, pp. 197-224, 1968.
- [11] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, P. Yanker, "Query by image and video content: The QBIC system," *Computer*, vol. 28, no. 9, pp. 23-32, September 1995.
- [12] D. A. Forsyth, J. Ponce, *Computer Vision: A Modern Approach*, Prentice Hall, 2002.
- [13] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 6, pp. 721-741, Nov. 1984.
- [14] R. Kindermann and L. Snell, *Markov Random Fields and Their Applications*, American Mathematical Society, 1980.
- [15] 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.
- [16] J. Li, R. M. Gray, *Image Segmentation and Compression Using Hidden Markov Models*, Kluwer Academic Publishers, 2000.
- [17] J. Li, A. Najmi, R. M. Gray, "Image classification by a two dimensional hidden Markov model," *IEEE Trans. on Signal Processing*, vol. 48, no. 2, pp. 517-33, February 2000.
- [18] W. Y. Ma, B. Manjunath, "NeTra: A toolbox for navigating large image databases," *Proc. Int. Conf. on Image Processing*, Santa Barbara, pp. 568-71, 1997.
- [19] Y. C. Park, P. K. Kim, F. Golshani, S. Panchanathan, "Technique for eliminating irrelevant terms in term rewriting for annotated media retrieval," *Proc. of ACM Multimedia*, pp. 582-584, 2001.
- [20] A. Pentland, R. W. Picard, S. Sclaroff, "Photobook: Tools for content-based manipulation of image databases," *Proc. of SPIE*, vol. 2185, pp. 34-47, San Jose, February 7-8, 1994.
- [21] J. Shi, J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888-905, 2000.
- [22] J. R. Smith, S.-F. Chang, "An image and video search engine for the World-Wide Web," *Proc. of SPIE*, vol. 3022, pp. 84-95, 1997.
- [23] M. Unser, "Texture classification and segmentation using wavelet frames," *IEEE Trans. Image Processing*, vol. 4, no. 11, pp. 1549-1560, Nov. 1995.
- [24] J. Z. Wang, *Integrated Region-based Image Retrieval*, Kluwer Academic Publishers, Dordrecht, 2001.
- [25] 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.
- [26] J. Z. Wang, J. Li, R. M. Gray, G. Wiederhold, "Unsupervised multiresolution segmentation for images with low depth of field," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 1, pp. 85-91, 2000.
- [27] S. Zhu, A. L. Yuille, "Region competition: Unifying snakes, region growing, and Bayes/MDL for multi-band image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, No. 9, pp. 884-900, 1996.