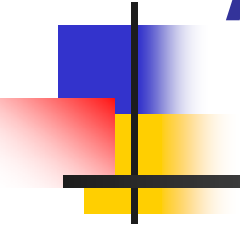


# Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach





# Outline

---

- Objective
- Approach
- Experiment
- Conclusion and Future work



# Objective

---

- Automatically establish linguistic indexing of pictures for a CBIR system
- CBIR: Content-based image retrieval is aimed at efficient retrieval of relevant images from large image databases based on automatically derived imagery features.



# Objective

---

- Categories of CBIR technology
- High level semantics description: 1). Image knowledge base e.g. MINDSWAP, The Helsinki University Museum; 2). Drawbacks
- Low level feature-based classification: 1). PicHunter, PicSeek etc. 2). Drawbacks e.g “Give me one image of this dog when it was a puppy”
- Link high level semantic description to low level feature information by automatically assigning comprehensive textual description to pictures.



# Problems of automatic annotation to images

---

- Automatic mapping between low level feature and knowledge  
e.g. Who's that guy in the picture? What is he doing?
- How does one model the semantic content? Can computer use what kind of domain knowledge to describe an image and how can computer acquire and store this knowledge?
- This article solves a part of the first problem.

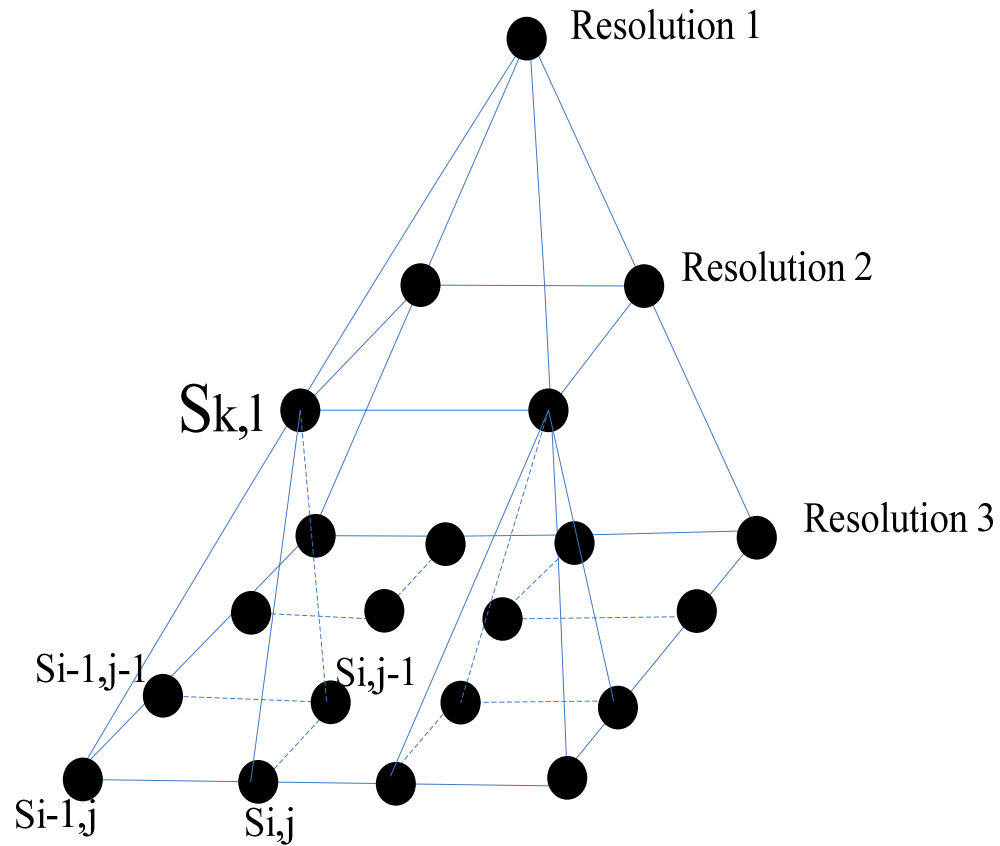


# Approach

---

- 2D MHMM (2-dimensional multiresolution Hidden Markov Model)
- Why does it choose 2D MHMM?
  - 1-D HMM is suitable for block-based image classification. For HMM, the image's category is the hidden state, and its feature vector is observation symbols for the state.
  - Compared 1-D HMM, 2-D HMM solves the problem of **overlocalization**.

# Approach



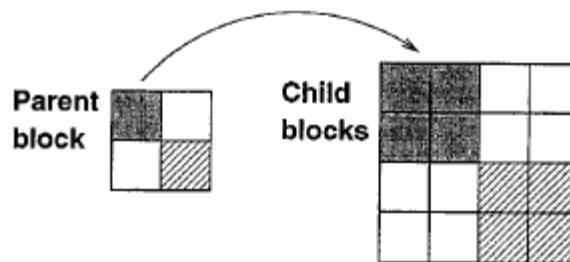
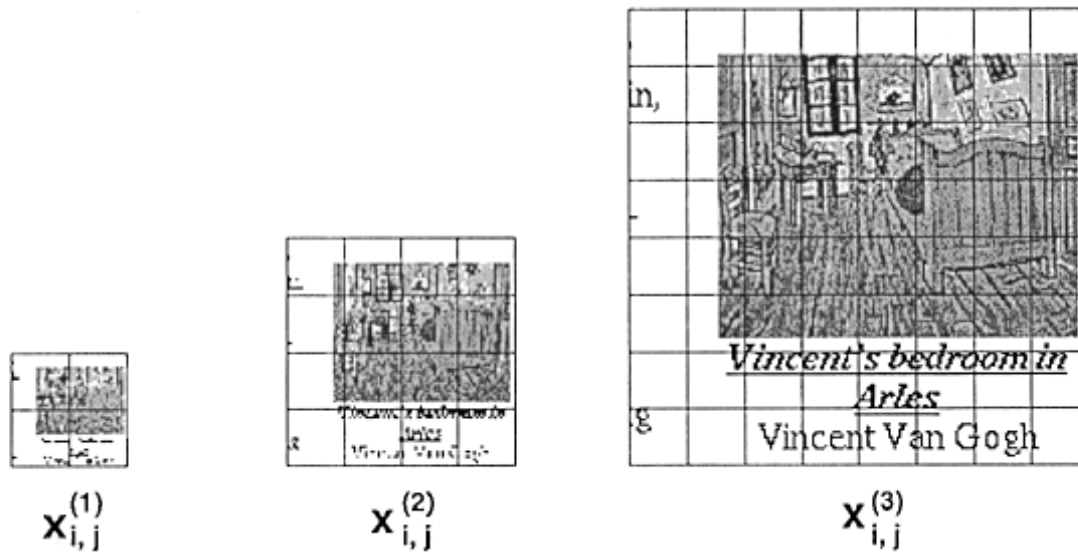


# Approach

---

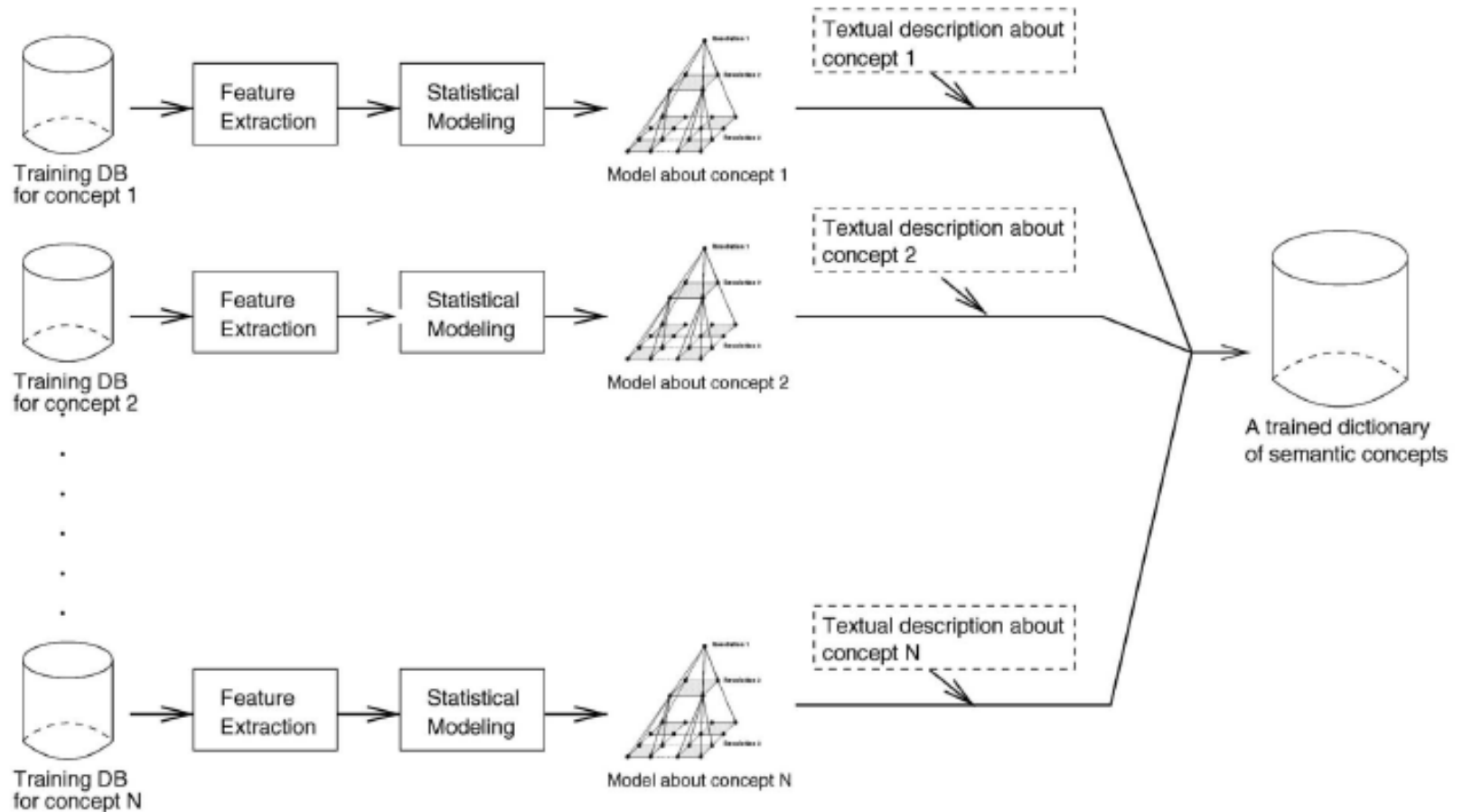
- From the view of computational complexity, it's necessary to increase the size of one block and prevent from including too many objects in 2-D HMM. For this purpose, the authors introduce multiresolution.

# Approach



Here one block is represented by a feature vector.

# Application architecture





# Select one category of images to train for one concept

---

- A concept corresponds to a particular category of images. (A concept doesn't just correspond to one word. A cluster of words can be considered as a concept.)



# Extract features from this category of images

---

- Every picture's size is  $384 * 256$ . An image is partitioned into  $4 * 4$  blocks. For each block, the system extracts a feature vector of six dimension using wavelet transform.



# Statistical Modeling

---

## ■ Assumptions

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

- Where  $S_{i,j}$  the state of block  $(i, j)$ ,  $U_{i,j}$  the feature vector of block  $(i, j)$ ,  $m = S_{i-1,j}$   $n = S_{i,j-1}$   $l = S_{i,j}$ ;  $(i, j)$  is the coordinate of the state.
- This formula means the  $S_{i,j}$  depends on  $S_{i',j'}$ .  $(i', j') < (i, j)$  means the block  $(i', j')$  is before the block  $(i, j)$



# Assumptions 2

---

- Given every state, the feature vectors follow a Gaussian distribution. And the parameters of the Gaussian distribution only depend on this state in its resolution.



# Assumptions 3

---

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

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

Where  $w$  is the maximal X coordinate of the state in the finest resolution,  $z$  is the maximal Y coordinate of the state in the finest resolution.



# Assumptions 4

---

- This formula means that given the states at the parent resolution, the states at the current resolution are conditionally independent of the other preceding resolutions, so that

$$P\left\{s_{i,j}^{(r)} : r \in \mathcal{R}, (i, j) \in \mathbb{N}^{(r)}\right\} = P\left\{s_{i,j}^{(1)} : (i, j) \in \mathbb{N}^{(1)}\right\}$$
$$\prod_{r=2}^R P\left\{s_{i,j}^{(r)} : (i, j) \in \mathbb{N}^{(r)} \mid s_{k,l}^{(r-1)} : (k, l) \in \mathbb{N}^{(r-1)}\right\}.$$



# Assumptions 5

---

- The feature vector is conditionally independent of information on other blocks once the state of a block of the feature vector is known.



# Assumptions 6

---

- According to the above assumption, we can get the joint probability of a particular set of states and the feature vector:

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



# Assumptions 7

---

- Child blocks descended from different parent blocks are conditionally independent. The state transition probabilities depend on the state of their parent block. So compute the transition probabilities in this formula:

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

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



# Statistical Modeling

- The joint probability of states and feature vectors at all the resolutions is then derived as

$$P\left\{s_{i,j}^{(r)}, u_{i,j}^{(r)} : r \in \mathcal{R}, (i,j) \in \mathbb{N}^{(r)}\right\} =$$
$$P\left\{s_{i,j}^{(1)}, u_{i,j}^{(1)} : (i,j) \in \mathbb{N}^{(1)}\right\} \times \prod_{r=2}^R \prod_{(k,l) \in \mathbb{N}^{(r-1)}} \left( P\left\{s_{i,j}^{(r)} : (i,j) \in \mathbb{D}(k,l) \mid s_{k,l}^{(r-1)}\right\} \prod_{(i,j) \in \mathbb{D}(k,l)} P\left\{u_{i,j}^{(r)} \mid s_{i,j}^{(r)}\right\} \right).$$

- To summarize, a 2D MHMM captures both the inter-scale and intra-scale statistical dependence.
- This model is trained using EM algorithm.



# Automatic Linguistic Indexing of Pictures

---

- Use the models of every concept to compute the log probabilities of generating  $u_{i,j}^{(r)}$ , that is

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

- Sort the log value to find K top ranked categories.

# Automatic Linguistic Indexing of Pictures

- After getting K candidate concepts, the author doesn't use these concepts to annotate the image.

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

$j, k$ : The word appear  $j$  times in  $k$  categories.  $p$  is the percentage of image categories in the database that are annotated with this word.



# Experiment

TABLE 1  
Examples of the 600 Categories and Their Descriptions

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



# Experiment

---

- 600 categories, 100 images, 40 images per each concept. 4,630 test images outside the training set are used to test classification accuracy.



# Accuracy

- Accuracy means the “match” percentage of 4,630 images. “match” means the test image annotated by this system is actually included in this category.

Comparison between the Image Categorization Performance of ALIP and that of a Random Selection Scheme

Accuracy	11.88%	17.06%	20.76%	23.24%	26.05%
Number of top-ranked categories required by ALIP	1	2	3	4	5
Number of categories required by a <i>random selection</i> scheme	72	103	125	140	151

*Accuracy is the percentage of test images whose true categories are included in top-ranked categories. ALIP requires substantially fewer categories to achieve the same accuracy.*



# Conclusion

---

- My opinion  
They link concepts and features in order to establish the concept indexing to make keyword queries a little intelligent. Because it doesn't care the second problems, it still isn't intelligent enough for content-based image retrieval.
- Proposed a 2D MHMM modeling approach to solve the problem of automatic linguistic indexing of pictures.



# Conclusion

---

- Advantage of this approach
  - Models for different concepts can be independently trained and retrained. Hence the system has good scalability;
  - Spatial relation among image pixels within and across resolutions is taken into consideration with probabilistic likelihood as a universal measure.



# Conclusion

---

## ■ Limitation

- Train the concept dictionary using only 2D images without a sense of object size.
- 40 training images are insufficient for the computer program to build a reliable model for a complex concept.



# Future work

---

- Improve the indexing speed of the system by using approximation in the likelihood computation.
- A rule-based system may be used to process the words annotated automatically to eliminate conflicting semantics.



# Reference

---

- [1] J. Li, R.M. Gray, and R.A. Olshen, “Multiresolution Image Classification by Hierarchical Modeling with Two Dimensional Hidden Markov Models,” *IEEE Trans. Information Theory*, vol. 46, no. 5, pp. 1826-41, Aug. 2000.
- [2] J. Li, A. Najmi, and R.M. Gray, “Image Classification by a Two Dimensional Hidden Markov Model,” *IEEE Trans. Signal Processing*, vol. 48, no. 2, pp. 517-33, Feb. 2000.
- [3] J.Z. Wang, J. Li, and G. Wiederhold, “SIMPLiCity: Semantics-Sensitive Integrated Matching for Picture Libraries,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 9, pp. 947-963, Sept. 2001.