AUTOMATIC SEMANTIC ANNOTATION OF IMAGES USING SPATIAL HIDDEN MARKOV MODEL

Feiyang Yu[1] *and Horace H S Ip*[1,2]

Image Computing Group,Department of Computer Science
Center for Innovative Applications of Internet and Multimedia Technologies (AIMtech Centre)
City University of Hong Kong, HONG KONG

ABSTRACT

This paper presents a new spatial-HMM(SHMM)for automatically classifying and annotating natural images. Our model is a 2D generalization of the traditional HMM in the sense that both vertical and horizontal transitions between hidden states are taken into consideration. The three basic problems with HMM-liked model are also solved in our model. Given a sequence of visual features, our model automatically derives annotations from keywords associated with the most appropriate concept class, and with no need of a pre-defined length threshold. Our experiments showed that our model outperformed the previous 2D MHMM in recognition accuracy and also achieved a high annotation accuracy.

1. INTRODUCTION

With the growing maturity of content-based image retrieval (CBIR), researchers gradually come to a realization of its limitations. CBIR systems, which adopt visual features for similarity comparison, assume that there is an inherent mapping between low-level features and high-level semantics. It becomes clear that this assumption does not hold for many applications. How to narrow down the semantic gap still remains an open issue.

Automatic image annotation has emerged as a major approach to bridge the semantic gap. Most works in this field focus on directly deriving semantic content from low-level features. J.Li et al. [2] proposed a new 2D MHMM to classify images into categories and propagate annotations from keywords which were manually assigned to those categories. On the other hand, some researchers regard the annotation task as an unsupervised learning problem. D.Blei et al. [1] and K.Barnard et al. [3] attempted to discover the statistical relationships between keywords and image features, and infer potential annotations from the joint distribution of features and keywords. However, the prerequisite of these works is to segment images semantically, which is still an error prone process.

We regard the annotation task as a multi-classification problem. The two most crucial problems with this approach are how to build a statistical model for each concept class, and how to propagate annotations from keywords associated with some specific classes. We propose a new spatial-HMM to describe the spatial relationships of objects and investigate the semantic structures of concepts in natural scene images. At the same time, we solve the second problem by automatically deriving annotations from available keywords.

The remainder of the paper is organized as follows. Section 2 elaborates our approach for representing image features. In Section 3, we describe the mechanism of our new spatial-HMM for modeling semantic contents of natural images. Experimental results are presented in Section 4. Finally, we conclude this paper in Section 5.

2. REPRESENTATION OF IMAGES

To analyze the texture patterns of natural scene images, a bank of Gabor filters are used in our work. Gabor filter is well-known for its orientation and frequency selective properties, and its optimal joint resolution in both spatial and frequency domains. We adopt the family of two-dimensional Gabor functions in [5].

$$\begin{aligned} g_{\xi,\eta,\theta,\varphi}(x,y) &= \exp(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2})\cos(\frac{2\pi x'}{\lambda}+\varphi) \\ x' &= (x-\xi)\cos\theta-(y-\eta)\sin\theta \qquad (1) \\ y' &= (x-\xi)\sin\theta+(y-\eta)\cos\theta \end{aligned}$$

Our bank of Gabor filters are configured with eight equidistant preferred orientations (θ=0, θ=π/8, θ=2π/8 ... θ=7π/8) and two preferred spatial frequencies (λ_0=5.65 λ_1=3.77; image block size=32 × 32 pixels). For the frequency λ and orientation θ, the outputs of a pair of corresponding Gabor filters with initial phases 0 and $-\pi/2$ are denoted by $\gamma_{\lambda,\theta,0}$ and $\gamma_{\lambda,\theta,-pi/2}$respectively. The combination of these two quantities yields the so called Gabor energy [6].

$$e_{\lambda,\theta}=\sqrt{\gamma^2_{\lambda,\theta,0}(i,j)+\gamma^2_{\lambda,\theta,-\pi/2}(i,j)} \qquad (2)$$

The total local Gabor energy E for a block can then be calcu-

1-4244-0367-7/06/$20.00 ©2006 IEEE

lated as:

$$E = \sum_{i=1}^{32} \sum_{j=1}^{32} \sqrt{r^2_{\lambda,\theta,0}(i,j) + r^2_{\lambda,\theta,-\pi/2}(i,j)} \qquad (3)$$

Consequently, for each block, a 16-dimentional texture description vector can be derived. In addition to texture information, mean values of every color component of the RGB color system are also used for our representation. Appending the three color component values to feature vector described above, we finally obtained a 19-dimensional feature vector.

3. SPATIAL HIDDEN MARKOV MODEL

To capture both the visual variations across blocks and the spatial relationships of objects across a sequence of blocks, we propose a new form of HMM, which we called Spatial HMM (SHMM). The idea is that the sequence of feature vectors corresponding to all blocks in an image can be modeled as a stochastic process. We assume that the feature vectors belonging to one semantic concept follow a multivariate Gaussian distribution. Each semantic concept is mapped to a hidden state in our model,which is a strict 1:1 correspondence. Given a test image, the task is to find the best state/label sequence which best explains the corresponding feature vector sequence. Our focus is on describing the 2D spatial relationships of hidden states in a plane, and unlike 2D MHMM [2], multi-resolution information of the image across scales is not used.

Our model is a 2D generalization of the HMM[4]. The HMM was originally developed for characterizing transitions between hidden states along the 1D time axis, which is inadequate for the case of 2D images. In order to describe transitions along two orthogonal directions in a plane, we introduce the concept of vertical transition, which is defined as the probability of the state of a block reached by transition from the hidden state of its immediately upper block. Just like the HMM, the horizontal transition is defined as the probability of a state reached by transition from its immediately preceding state. Our Markov assumption can be formalized by a combination of the above two kinds of transitions.

$$\begin{aligned} P(q_{l,m}|Q_{l,m\ominus 1}) &= P(q_{l,m}|q_{l\ominus 1,m}q_{l,m\ominus 1}) \\ &= \underbrace{P(q_{l,m}|q_{l,m\ominus 1}))}_{h}\underbrace{P(q_{l,m}|q_{l\ominus 1,m})}_{v} \end{aligned} \qquad (4)$$

where the sign $\ominus$ denotes precedence relations in a raster scan, $q_{l,m}$ denotes the state for block(l,m), and $Q_{l,m\ominus 1}$ denotes the sequence of states from block$(1,1)$ to block$(l, m\ominus 1)$. The underlying idea is that the state of a block only depends on the states of two previously observed neighbor blocks in a raster scan. This assumption,which is appropriate for horizontally layered natural images, differs dramatically from that used in second-order Markov mesh models[6]. The complete specification of a spatial HMM λ requires a specification of the number of states *N*(the collection of available hidden states is denoted by S), and the four probability measures such as *H*(horizontal transition matrix), *V*(vertical transition matrix), *B* and π. For convenience, we denote λ in shorthand as $\lambda = (H, V, B, \pi)$.Given an observation sequence, the test image is regarded as belonging to the concept class which has the highest probability to generate its observation sequence. To find this generation probability, we need to extend the Forward-Backward algorithm for HMM. For an L×M image,we denote its observation sequence as: $O_{L,M} = o_{11}\ldots o_{1M}o_{21}\ldots o_{2M}\ldots o_{LM}$. we define the forward variable as

$$\alpha_{l,m}(k) = P(O_{l,m}, q_{l,m} = S_k|\lambda) \qquad (5)$$

To facilitate our calculation, we define an auxiliary variable.

$$g_{l,m}(i,j) = P(O_{l,m\ominus 1}, q_{l,m\ominus 1} = S_i, q_{l\ominus 1,m} = S_j, |\lambda) \quad (6)$$

By conditional probability calculation, we can derive the recursive relationship for g.

$$\begin{aligned} g_{l,m}(i,j) &= \sum_{t,u,w} P(O_{l,m\ominus 2}, t, u, w, i, j) b_s(o_{l,m\ominus 1}) \\ &= \sum_{t,u,w} P(O_{l,m\ominus 2}, t, u, w) h_{u,j} v_{t,j} h_{w,i} v_{u,i} b_s(o_{l,m\ominus 1}) \end{aligned} \quad (7)$$

Please refer to Fig. 1 for the definition of t, u, w. Let us denote the first term in the above equation as $P_{t,u,w}$. Then it follows:

$$\begin{aligned} P_{t,u,w} &= P(O_{l\ominus 1,m\ominus 1}, u) P(C, B, w, t|O_{l\ominus 1,m\ominus 1}, u) \\ &= g_{l,m\ominus 1}(u,t)\frac{g_{l\ominus 1,m}(w,u)}{\alpha_{l\ominus 1,m\ominus 1}(u)} \end{aligned} \qquad (8)$$

Therefore, we can solve the generation probability problem iteratively:

$$\begin{cases} \alpha_{1,1}(k) = \pi_k b_k(o_{ll}) \\ \alpha_{1,m}(k) = \sum_{i=1}^{N} \alpha_{1,m\ominus 1}(i) h_{i,k} b_k(o_{1,m}) \\ \alpha_{l,m}(k) = \sum_{i,j} g_{1,m\ominus 1}(i,j) h_{i,k} v_{j,k} b_k(o_{1,m}) \\ 2 \le l \le L \end{cases} \qquad (9)$$

$$\begin{cases} g_{2,1}(i,j) = P(O_{1,M}, q_{1,1} = i, q_{1,M} = j) \\ g_{2,m}(i,j) = \sum_{u,w} g_{2,m\ominus 1}(w,u) h_{u,j} h_{w,i} v_{u,i} b_i(o_{2,m}) \\ g_{l,1}(i,j) = P(O_{l\ominus 1,M}, q_{l\ominus 1,1} = i, q_{l\ominus 1,M} = j) \\ 2 \le l \le L \\ g_{l,m}(i,j) = \sum_{t,u,w} \frac{g_{l,m\ominus 1}(w,u) g_{l\ominus 1,m}(u,t)}{\alpha_{l\ominus 1,m\ominus 1}(u)} h_{u,j} v_{t,j} \\ \qquad h_{w,i} v_{u,i} b_i(o_{l,m\ominus 1}) \\ 2 \le l \le L \end{cases} \qquad (10)$$

where the range of parameter m runs from 1 to M. Summing all the forward variables for the last block, we can obtain the conditional probability of $P(O_{L,M}|\lambda)$.

Likewise, we extend the traditional Viterbi algorithm for SHMM accordingly.

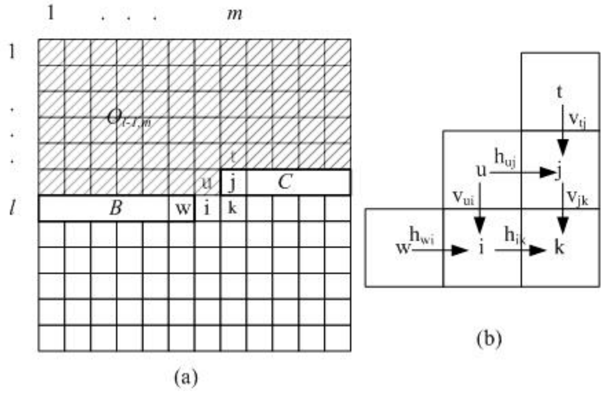


Fig. 1. Illustration of (a) observation sequence and hidden states. (b)horizontal and vertical transitions between neighbor states.

1)Initialization

$$\delta_{1,1}(i) = \pi_i b_i(o_{1,1}) \quad 1 \le i \le N \tag{11}$$

2)Recursion

$$\begin{aligned}
\delta_{1,m}(k) &= \max_{1\le i\le N} \left[\delta_{1,m\ominus 1}(i) h_{ij}\right] b_k(o_{l,m}) \\
\delta_{l,1}(k) &= \max_{1\le i\le N} \left[\delta_{l-1,M}(i) h_{ik} \delta_{l-1,m}(j^*) \nu_{j^*k}\right] b_k(o_{l,m}) \\
\delta_{l,m}(k) &= \max_{1\le i\le N} \left[\delta_{l,m\ominus 1}(i) h_{ik} \delta_{l-1,m}(j^*) \nu_{j^*k}\right] b_k(o_{l,m}) \\
\psi_{l,m}(k) &= \operatorname*{argmax}_{1\le i\le N} \left[\delta_{l,m\ominus 1}(i) h_{ik} \delta_{l-1,m}(j^*) \nu_{j^*k}\right] \\
j^* &= \underbrace{\psi_{l-1,m\oplus 1}(\ldots(\psi_{l,m\ominus 1}}_{L-1}(i))\ldots) \\
& 1 \le l \le L, 1 \le m \le M, 1 \le n \le N
\end{aligned} \tag{12}$$

3)Termination

$$\begin{aligned}
P^* &= \max_{1\le i\le N} \left[\delta_{L,M}(i)\right] \\
q^*_{L,M} &= \operatorname*{argmax}_{1\le i\le N} \left[\delta_{L,M}(i)\right]
\end{aligned} \tag{13}$$

4)Path (state sequence) backtracking

$$q^*_{L,M} = \psi_{l,m\oplus 1}(q^*_{l,m\oplus 1}) \quad 1 \le l \le L, 1 \le m \le M \tag{14}$$

A key issue for HMM-like model is to determine the model parameters to maximize the probability of the observation sequence given the model. Since in supervised training, the labels for each image have been provided by experts, the estimation problem is then reduced to a simple maximum likelihood estimate of parameters.

4. EXPERIMENTAL RESULTS

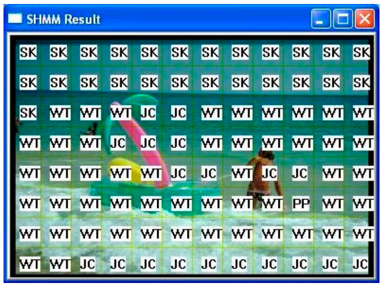
The spatial-HMM approach was implemented and tested against a number of COREL images. More specifically, four COREL CDs, i.e. Beaches, Buses, Elephants, and Mountains, was used for our experiment. Each of these CD contains one hundred high-resolution images. To compare our method with the 2D MHMM method, we down-sampled these image to the size of 384×256 and transformed them into JPEG format.

Before further processing,each image was then divided into equivalent blocks of the size of 32×32 pixels. The choice of the block size is a tradeoff between the integrity of semantic meanings and processing cost.Manual annotations are provided for all blocks in each training image.For each concept category, we arrived at a collection of semantic labels.As we mentioned in Sec.3, one Markov model is built to represent the semantic content of each concept category.In Table 1, we listed all semantic labels used in this work.

Table 1. The set of semantic labels used in our experiments

Class	Semantic Labels
Beaches	Sky(SK), Water(WT), Sand(SD), People(PP),Rock(RK),Building(BD),Boat(BT), Tree(TR),Equipment(EQ), Junction(JC)
Buses	Bus(BS), Building(BD), People(PP), Advertisement(AD), Sky(SK), Tree(TR), Ground(GD), Junction(JC)
Elephant	Sky(SK), Elephant(ET), Tree(TR), Water(WT), Plain(PL), Junction(JC)
Mountains	Sky(SK), Snow(SN), Tree(TR), Mountain(MT),Junction(JC)

Gabor filters used in this work was implemented in the spatial domain with the mask size of 9×9. Since our block size is 32×32, the available radial frequencies for configuration of Gabor filters are $\sqrt{2}, 2\sqrt{2}, 4\sqrt{2}, 8\sqrt{2}$. We adopted $4\sqrt{2}$ and $8\sqrt{2}$, because $\sqrt{2}$ and $2\sqrt{2}$ are too low to capture local texture features. For color features, we have evaluated the CIE LAB, CIE LUV and RGB color systems. Our experiments showed that the RGB system is more appropriate for this work.It is because the distance between two distinct points in the feature vector is multiplied by the inverse of the covariance matrix associated with a hidden state.



SK	SK	SK	SK	SK	SK	SK	SK	SK	SK	SK	SK
SK	SK	SK	SK	SK	SK	SK	SK	SK	SK	SK	SK
SK	WT	WT	WT	JC	JC	WT	WT	WT	WT	WT	WT
WT	WT	WT	JC	JC	JC	WT	WT	WT	WT	WT	WT
WT	WT	WT	WT	WT	JC	JC	WT	JC	JC	WT	WT
WT	WT	WT	WT	WT	WT	WT	WT	WT	PP	WT	WT
WT	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT
WT	WT	JC	JC	JC	JC	JC	JC	JC	JC	JC	JC

Fig. 2. An example of semantic labeling using SHMM.

To verify our method, we conducted the same experiment as that in [2]. That is, 40 images were selected for training each concept, and the remaining 60 images in that category were used for testing.Keywords listed in Table 1 were used to manually annotate blocks in these training images. One snapshot of SHMM annotation result is shown in Fig. 2. More example annotated images are shown in Fig. 3. To the right of these images, labels occurred in their annotation are listed.

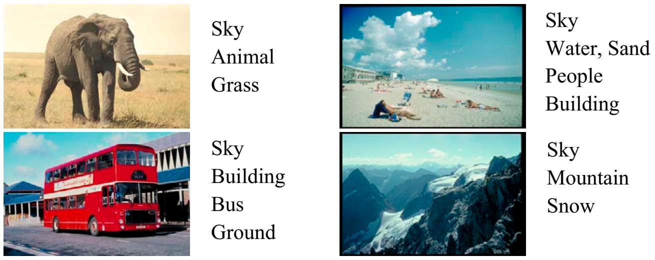


Fig. 3. Example annotations generated by our SHMM.

Our experimental results are shown in Table 2. The term ”R-rate” is defined as the ratio of correctly recognized images to the number of total test images in one concept category, while the term ”A-rate” is defined as the mean annotation accuracy, which is the ratio of correctly labeled blocks to all blocks (12×8) in a test image, over all correctly recognized images in one concept category. It can be seen that the R-rate and A-rate are both high for images of mountains by the two methods. This can be attributed to the fact that images in the mountain category have many areas dominated by brown rocky mountains which have roughly similar color and texture characteristics. Such homogeneous feature simplifies the recognition task to a certain extent. We also noticed that the two rates are both low for images of elephants due to the greater visual variation of animals and backgrounds in these images.

Table 2. Performance comparison of 2DMHMM and SHMM

Class	2D-MHMM	SHMM	
	R-rate	R-rate	A-rate
Beach	32%	76%	71.6%
Bus	46%	86%	78.5%
Elephant	40%	72%	70.4%
Mountains	84%	85%	85.9%

Our results also indicate that the SHMM consistently outperforms the 2D MHMM method for all cases. Compared with 2D MHMM, our approach improved the recognition accuracy by 1%-44%. Our method also achieved about 70%-86% annotation accuracy. Especially for categories such as bus, beach, and elephant, which have a distinctive layered structure, the performance of our method is much superior to that of the 2D MHMM. Let us use beach as an example. For images coming from that category, the upper area is occupied by sky; the middle area is occupied by ”sea water”; the lower area is occupied by ”sandy beach”. The general structure of semantic labels for beach remains constant across all images.By taking into account of both the horizontal and vertical arrangement of semantic labels,our spatial-HMM achieved as high as 76% recognition accuracy,while it is hard for 2D MHMM to classify images with such complex semantic structure.This demonstrates that spatial-HMM successfully captures the spatial relationships of semantic labels in an image which enables it to better cope with visual variations in complex images compared with 2D MHMM.

Another major advantage of our approach is the automatic selection of keywords for annotation. The most frequently used method for keyword selection is to determine a threshold for the length of keywords and to get rid of extra words which exceed the limit. However, short annotations may not provide enough information while long annotations tend to introduce redundant information. In our system, the keywords used for annotation are automatically selected by the models themselves. Generally speaking, the length of annotation keywords for the category of elephant and mountain is 4, while the lengths for bus and beach are 5 and 7 respectively.

5. CONCLUSION

A new spatial-HMM model is presented here to analyze and annotate the semantic content of Corel images. The traditional Viterbi and Forward-Backward algorithm have been extended accordingly for the spatial-HMM. In terms of recognition and annotation accuracy, the spatial-HMM achieves better performance compared with the 2D MHMM since the former captures and takes into account more contextual information. Some deficiencies associated with previous approaches of image annotation have also been overcome using this new approach.

6. ACKNOWLEDGEMENT

This work is fully supported by a City University of Hong Kong Strategic Research Grant, Project No: 7001698.

7. REFERENCES

[1] D. Blei, M. I. Jordan, ”Moldeling Annotation Data,” In Proc. SIGIR, Toronto, Aug.2003.

[2] J. Li, J.Z. Wang, ”Automatic Linguistic Indexing of Pictures by a Statistical Moldeling Approach,” *IEEE Trans on PAMI.* Vol.25(9),Sep,2003.

[3] K. Barnard, D. Forsyth, ”Learning the Semantics of Words and Pictures,” Proc.ICCV, pp. II: 408-415, 2001.

[4] L. Rabiner. ”A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” *Proc. IEEE.* Vol.77(2),257-286,1989.

[5] S.Grigorescu, N.Petkov, P.Kruzinga, ”Comparison of Texture Features Based on Gabor Filters,” *IEEE Trans. on Image Processing*. Vol.11(10),Oct,2002.

[6] Devijver, P.A., ”Probabilistic Labeling in a Hidden Second Order Markov Mesh,” Pattern recognition in Practics, II,E.S. Gelsema and L.N.Kanal eds., NorthHolland, Amsterdam, 1986.