A SEMANTIC REPRESENTATION FOR IMAGE RETRIEVAL

Lei Wang and B.S. Manjunath

Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106-9560
{lwang, manj @ece.ucsb.edu }

ABSTRACT

Robust semantic labeling of image regions is a basic problem in representing and retrieving image/video content. We propose an SVM-MRF framework to model features and their spatial distributions, leading towards a "semantic" representation. Eigenfeatures of Gabor wavelet features and Gaussian mixture model are used for feature clustering. Since similar feature vectors in one cluster can come from several different semantic classes, SVM is applied to represent conditioned feature vector distributions within each cluster, and a Markov random field is used to model the spatial distributions of the semantic labels. A semantic layout representation is proposed to describe the semantics of the images. Experiments show that this method can improve semantic labeling and is useful in similarity search.

1. INTRODUCTION

Describing image/video content with semantic labels is a challenging problem. Semantic representation plays an important role in the analysis and retrieval of image/video content. It is difficult for people to retrieve raw data or annotate images/videos manually. Automated semantic analysis is necessary for efficient image/video retrieval and mining [10]. The challenge in semantic labeling is that there might be a wide variation in appearance within the same class.

Support vector machine (SVM) [4] has been used widely for modeling complex feature distributions in high-dimensional spaces. Likewise, the Markov random field (MRF) model has proven its usefulness in image analysis [3,6,8,12]. However, despite their effectiveness, a simple SVM or MRF model is inadequate to represent the wide variation of visual features within a semantic class.

In this paper, a novel combination of MRF and SVM is proposed for the semantic analysis and representation of aerial image/video content. The major contribution of the paper is the SVM-MRF framework for clustering and classification. A casual greedy algorithm is proposed for optimization. The sites for the MRF are blocks of pixels, each of which is described by a visual feature descriptor. The spatial distribution of semantics labels of each image can be considered as a Markov random field (MRF). Therefore, the analysis phase involves the automatic identification of the semantic classes in a given image, and is a three-step procedure. The first step is to cluster the features of the image blocks using the Gaussian mixture model (GMM) [2]. The GMM is used to model the principal components of the original feature vectors. Each Gaussian represents a cluster in this model. This can improve the clustering performance if the number of clusters is not large. Since similar feature vectors in one cluster can come from several different semantic classes, the second step is the application of the "one-against-others" SVMs to classify the image blocks into candidate semantic classes within each of the clusters. The third step uses a combination of the SVMs and the MRF model to refine the classification. Semantic layouts based on these labels are then generated for similarity search.

The paper is organized as follows. Section 2 describes the use of GMM and SVM for modeling feature distributions. Section 3 focuses on the SVM-MRF model for the identification of semantic classes in a given image. Section 4 discusses the application of this approach to similarity retrieval. Experimental results are presented in Section 5 and we conclude with a discussion in Section 6.

2. FROM FEATURES TO SEMANTICS

In order to recognize the semantics of image blocks, it is necessary to describe the blocks using a visual feature and learn the statistical properties of these features considering their semantic origins. The feature used in our experiments is derived from the homogeneous texture descriptor of MPEG-7. The 2-D frequency plane is partitioned into 24 channels with Gabor filters (6 directions on 4 scales)[9]. The mean of the texture energy in each channel is computed, resulting in a 24-dimensional feature vector. In order to simplify the computation, we propose to classify the features into several clusters before the semantic classification. Due to the non-linearity, each cluster can

contain feature vectors coming from several semantic classes. In order to obtain a good clustering result, we try to reduce the dimensionality of the input feature vectors (24-dimensions) via principal component analysis (PCA) [5]. In the reduced feature space, it is easier to obtain good clustering results using a Gaussian mixture model [2]. SVMs are then applied to the original feature vectors to model the feature distribution within each Gaussian cluster. This choice is motivated by the following observations: (a) visually similar textures tend to form clusters in a sparse feature space, (b) there is a wide variation in visual appearance within each semantic class, and (c) similar feature vectors can come from several different semantic classes. Selecting a suitable kernel, the features in each cluster can be classified with SVM effectively.

3. MODELING LABEL DISTRIBUTIONS WITH SVM-MRF

SVMs produce uncalibrated values that are not probabilities. Platt [11] proposed to train the parameters of an additional sigmoid function to map the SVM outputs into probabilities. Since SVMs are binary classifiers, in order to apply SVM in the multi-class classification, a set of "one-against-others" SVMs [7] are constructed. This works by applying SVMs that first separate one class from all the other classes, and then arbitrating between several SVMs. Using the set of SVMs, an image block with feature vector z can be classified into a semantic class m using a margin-based distribution $P(z \mid x) = \frac{1}{1+\exp(\eta f_x(z))}$ and the label x of the feature vector z can be given as:

$$\begin{aligned} x &= \arg\max_m P(m \mid z) \\ &= \arg\max_m P(z \mid m) P(m) \\ &= \arg\max_m \frac{1}{1+\exp(\eta f_m(z))} p(m) \end{aligned} \quad (1)$$

where $\mathfrak{m} \in \mathfrak{M} = \{1, 2, ..., M\}$ is a semantic label, η is a constant, $f_m(z) = \sum_i y_{m,i} \alpha_{m,i} k(z_i, z) + b$ is the output of the m-th SVM, where K(.,.) is a kernel function and

$$y_{m,i} = \begin{cases} -1, & \text{if the label of } z_i \text{ is } m \\ 1, & \text{otherwise} \end{cases}.$$

Since the spatial relationships between neighboring blocks are not considered, many inconsistent labels can be generated. For example, it is possible that a "street" block is shown to be surrounded by "sky" blocks. An MRF applied to the spatial layout of class labels addresses this problem.

The class label of a block at site s can be modeled as a discrete-valued random variable X_s, taking values from the semantic label set $\mathfrak{M} = \{1, 2, ..., M\}$, and the set of random variables $X = \{X_s, s \in S\}$ constitutes a random field where S is the lattice of image blocks. This process can be analyzed as an MRF with two basic assumptions: (a) the label of a block depends only on the labels of the neighboring blocks, and (b) the class-conditioned feature distributions at all sites are independent of each other. This combination of SVM and MRF is referred to as *SVM-MRF* framework. The MRF X is modeled with a Gibbs distribution $p(x) = \frac{1}{Z}\exp(-U(x))$, where x is a realization of X. The Gibbs energy function $U(x)$ can be expressed as the sum of clique potential functions,

$$U(x) = \sum_{c \in Q} V_c(x) \quad (2)$$

where Q is the set of all cliques in a neighborhood. We reinforce the MRF model by incorporating the class-conditioned feature likelihoods into the energy function, as follows:

$$U(x) = \sum_{s \in S} \left(\sum_{s' \in N_s} -\gamma LP_{s-s'} - \kappa LP_s \right) \quad (3)$$

where $LP_{s-s'} = \log(p_{s-s'}(x_s, x_{s'}))$ and $LP_s = \log(p(z_s \mid x_s))$. N_s is the set of neighbors of the site s, γ and κ are the weights of LP_s and $LP_{s-s'}$ respectively. $LP_{s-s'}$ represents the spatial relationship between neighboring sites s and s' where s-s' indicates the direction of neighborhood. LP_s takes into account the conditional probability density of feature vector z_s given the label x_s. $p_{s-s'}(x_s, x_{s'})$ is the joint probability of x_s and $x_{s'}$ along the direction s-s' and it can be approximated with a co-occurrence matrix from the labeled training set. For each type of clique s-s', a co-occurrence matrix is constructed from the joint probabilities $P_r(i, j)$ between pairs of semantic labels i and j in a given direction r.

In order to simplify the model, a second order pair-sites neighboring system is applied. Each site thus has eight neighbors. Four types of cliques are considered, wherein s-s' makes angles of 0, 45, 90, and 135 degrees with respect to the x-axis. In any neighborhood, cliques along the same direction are considered equivalent. Four co-occurrence matrices are constructed along these four directions of label distribution. Given an image realization with block feature set Z, and M class labels, the goal now is to find the optimal labeling X* by maximizing the posterior distribution, i.e.

$$X^* = \arg\max_X P(X \mid Z) = \arg\max_X P(Z \mid X) P(X). \quad (4)$$

In practice, the above problem is tackled by locally optimizing over each site sequentially. Besag [1] proposed

a deterministic *Iterated Conditional Modes* (ICM) algorithm that maximizes local conditional probabilities at each site. As is well known, the result obtained by ICM depends much on the initial labeling. This is a serious drawback because, as discussed before, the initial labeling may have semantic inconsistencies. While the ICM algorithm tries to avoid these inconsistencies, it could converge to a poor local minimum. In order to improve the classification performance, we propose a causal greedy algorithm (CGA) to estimate the labels of the image blocks.

As shown in Fig. 1, $N_s = \{s_1, ..., s_8\}$ are the neighbors of site s. The sites $P_s = \{s_1, s_2, s_3, s_4\}$ are termed the *predecessors* and $S_s = \{s_5, s_6, s_7, s_8\}$ are termed the *successors* of site s. In order to obtain a consistent classification, the label of site s is optimized over all possible labels of the successors S_s. Given the labels of P_s, the label of site s can be determined as follows. For all labels l and $s\,s' \in S_s$, calculate

$$x_{s',l} = \max_j p_{s-s'}(x_s = l, x_{s'} = j) P(z_{s'} \mid j) \;\; \forall s' \in S_s, l \in \mathfrak{M}. \tag{5}$$

Then calculate the label of site s as follows:

$$x_s = \max_l \left(\prod_{s' \in P_s} p_{s-s'}(x_s = l, x_{s'}) \prod_{s' \in S_s} p_{s-s'}(x_s = l, x_{s'} = x_{s',l}) \right) P(z_s \mid l) \tag{6}$$

The algorithm needs only one pass over all sites, and it maximizes the local conditional probability at each site, and at each step, only the successors are revised. This ensures that the algorithm is both causal and stable, and results in a consistent labeling.

4. A SEMANTIC REPRESENTATION FOR RETRIEVAL

After the labeling process, the labels of the blocks of a given image (or video key frame) are used for interpreting its semantic content. The arrangement of the semantic labels analyzed from an image is called a semantic layout. Let the semantic layout of the query image be X^q and that of the stored image be X^I. A soft classification scheme is adopted for better retrieval result. For a given image block, the labels with the three largest local conditional probabilities are selected to represent this block. The semantic layout similarity between the query image and each stored image is given by

$$\begin{aligned} S = & \sum_{s \in S} \left(\sum_{j=1}^{3} a_1 a_j \delta(x^q_{s,j}, x^I_{s,1}) \right) \\ & + \sum_{s \in S} \left(\sum_{j=1}^{3} a_2 a_j \delta(x^q_{s,j}, x^I_{s,2}) \right) \left(1 - \sum_{i=1}^{3} \delta(x^q_{s,i}, x^I_{s,1}) \right) \\ & + \sum_{s \in S} \left(\sum_{j=1}^{3} a_3 a_j \delta(x^q_{s,j}, x^I_{s,3}) \right) \left(1 - \sum_{i=1}^{3} \delta(x^q_{s,i}, x^I_{s,1}) \right) \left(1 - \sum_{i=1}^{3} \delta(x^q_{s,i}, x^I_{s,2}) \right) \end{aligned} \tag{7}$$

where $a_i = \dfrac{1}{2^{i-1}}, i = 1, 2, 3$ are the weights for different label similarities, and $x^q_{s,j}$ is the j-th candidate label of the query image block at site s.

In Equation (7), the similarity measure is computed by comparing each candidate label of a query image block with all the candidate labels of the corresponding stored image block.

5. EXPERIMENTAL RESULTS

The dataset used in our experiments consists of infrared aerial video from urban and natural scenes. We selected 2484 key frames of 256×256 pixels from these videos. Each image block is 64×64 pixels and neighboring blocks are not overlapping. Each block is labeled manually. The dataset is divided into two sets of 1242 key frames each. One set is used for training the model parameters and the other used for testing. In the feature clustering stage, 5 eigenfeatures of the original feature vector extracted with the eigenvectors with 5 largest eigenvalues are selected as features. GMM is applied to cluster the feature vectors into 47 clusters. A set of one-against-others SVMs are constructed for each cluster. Table 1 shows the overall results for the 1242 testing images. The method "SVM" refers to the initial classification using the SVM approach alone. A polynomial kernel is applied in the SVMs with $\eta = 6$ in Equation (1). The method "ICM" refers to the classification results of ICM. The method "CGA" refers to the classification results of CGA. "N=1" is the classification accuracy that the first candidate semantic label is the desired label and "N=3" is the classification accuracy that one of top 3 candidate semantic labels is the desired label. Table 2 shows the "N=1" results for the individual labels. M is the number of feature vectors within each class. It is observed that CGA can improve the classification results of SVM and it is much better than ICM. The result of ICM is even worse than initial result of SVM. The reason is that ICM converges to a local minimum that is quite sensitive to the initial condition. The CGA attempts to overcome this problem by considering domain knowledge from the predecessor labels in a causal manner. Figure 2 shows a semantic retrieval result with semantic layout. With this method, some of the retrieved images appear visually different from the query image, while their semantic layouts are similar. Note that this kind

of semantic similarity cannot result from a low-level visual feature query.

6. DISCUSSION

SVM-MRF framework is proposed to model the distribution of semantic classes in images. GMM with eigenfeatures is applied to cluster the features. A set of SVMs models the texture feature distributions in each cluster. An MRF is used to model the spatial relationship between semantic classes. A causal greedy algorithm is presented to estimate the labels of the image blocks. The paper also demonstrates the application of semantic layout for similarity retrieval. Our experiments show that the proposed approach can be used to obtain retrievals with similar semantics but with varying visual appearances.

Acknowledgments: This work was supported by the following grant: ONR#N00014-01-0391.

7. REFERENCES

[1] J. Besag, "On the Statistical Analysis of Dirty Pictures," *Journal of the Royal Statistical Society*, Vol. B48, 1986, pp. 259-279.

[2] S. Bhagavathy, S. Newsam, and B.S. Manjunath, Modeling Object Classes in Aerial Images Using Texture Motifs, *International Conference on Pattern Recognition (ICPR)*, 2002

[3] R. Chellappa and A. Jain, "Markov Random Fields: Theory and Applications," *Academic Press*, 1993.

[4] C. Cortes, V. Vapnik, "Suppor- vector networks", *Machine Learning*, Vol.20, No. 3, pp.273-297, 1995

[5] R.Duda, P. Hart, D. Stork, Pattern Classification, *John Wiley & Sons, USA* (2001).

[6] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.6, No. 6, pp. 721-741, 1984.

[7] K. Kim, K. Jung, S. Park, H. Kim, Support vector machines for texture classification**,** *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No.11, pp.1542- 1550, 2002

[8] S. Li, "Markov Random Field Modeling in Computer Vision," *Springer-Verlag,* 1995.

[9] B. S. Manjunath, P. Salembier, and T. Sikora, "Introduction to MPEG-7: Multimedia Content Description Interface," *Wiley*, 2002.

[10] M. R. Naphade and T. S. Huang, "A Probabilistic Framework for Semantic Indexing and Retrieval in Video," *Proc. International Conference on Multimedia and Expo* (I), 2000, pp. 475-478.

[11] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods", *Advances in Large Margin Classifiers*, pp. 61-74, MIT Press, 1999

[12] L. Wang, J. Liu and S. Li, MRF Parameter Estimation by MCMC Method, *Pattern Recognition*, Vol.33, No.11, pp. 1919-1925, 2000

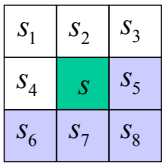
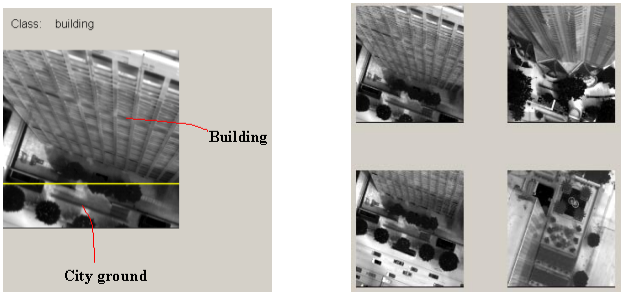


Fig. 1. The second order neighborhood



(a) Query image and its semantic layout

(b) Retrieval results

Fig. 2. Semantic retrieval example

Method	N=1	N=3
SVM	0.6345	0.7736
ICM	0.6239	0.7340
CGA	0.7151	0.7971

Table 1. Comparison of labeling accuracy of SVM and SVM-MRF on a test data set of 1242 images.

ID	Labels	M	Labeling Accuracy		
			SVM	ICM	CGA
0	street	1416	0.4640	0.4204	0.6243
1	roof	1438	0.4228	0.3935	0.4951
2	parking_lot	697	0.4763	0.4334	0.6514
3	building	167	0.5449	0.5339	0.7545
4	tree	844	0.2903	0.2889	0.3495
5	freeway	1521	0.5003	0.4643	0.6042
6	streetside	257	0.2763	0.2298	0.3658
7	house	1878	0.7572	0.7416	0.8259
8	yard	326	0.1933	0.2386	0.4018
9	sand	1242	0.6248	0.6849	0.7158
10	pipeline	553	0.4828	0.4397	0.5497
11	road	549	0.1858	0.1453	0.2477
12	waterpond	43	0.9302	0.8974	0.7674
13	farm	2980	0.8104	0.7842	0.9161
14	dirt	3422	0.6639	0.6209	0.7858
15	grass	2539	0.6187	0.5798	0.8295

Table 2 Semantic labels and corresponding labeling accuracies of GMM, ICM, and CGA on a test data set of 1242 images