Using Texture to Annotate Remote Sensed Datasets

S. Newsam, L. Wang, S. Bhagavathy, and B.S. Manjunath
Electrical and Computer Engineering
University of California at Santa Barbara
{snewsam,lwang,sitaram,manj}@ece.ucsb.edu

Abstract

Texture remains largely underutilized in the analysis of remote sensed datasets compared to descriptors based on the orthogonal spectral dimension. This paper describes our recent efforts towards using texture to automate the annotation of remote sensed imagery. Two applications are described that use the homogeneous texture descriptor recently standardized by MPEG-7. In the first, higher-level access to remote sensed imagery is enabled by using the texture descriptor to model geo-spatial objects. In particular, the common textures, or texture motifs, are characterized as Gaussian mixtures in the high-dimensional feature space. In the second application, the texture descriptor is used to label regions in a large collection of aerial videography in a perceptually meaningful way. Gaussian mixtures are used to model the distribution of feature vectors for a variety of semantic classes. Frame level similarity retrieval based on semantic layout and semantic histogram is enabled by modeling the spatial arrangement of the labeled regions as a Markov random field.

1. Introduction

Remote sensed imagery is accumulating at an increased rate due to advances in sensors, storage, and other technologies. The full value of this data is not being realized, however, since there has not been similar progress in automated analysis. A fundamental challenge to managing large repositories of remote sensed imagery remains automating the annotation process. Automated techniques are urgently needed since manual annotation is prohibitively time consuming and expensive.

There has been noted success over the past few decades in using spectral information to label remote sensed imagery at the pixel level. These approaches are limited however since many land cover types appear similar through such a small aperture. Spatial context must be incorporated.

Spatial context at the finest, or pixel, scale can be considered as the low-level image primitive of texture. Texture-based analysis remains largely underutilized, however, in the analysis of remote sensed imagery. This paper describes our recent work on using texture for the automated annotation of large collections of remote sensed data. In particular two applications that use the homogeneous texture descriptor recently standardized by the MPEG-7 Multimedia Content Descriptor Interface [1] are described.

In the first application, higher-level access to remote sensed imagery is enabled by using the descriptor to model geo-spatial objects with common characteristic textures. These common textures are termed texture motifs and are represented as Gaussian mixture models (GMM) in the high-dimensional feature space. Examples of motifs include the rows of boats and water in harbors, and the grass and trees in golf courses.

In the second application, the descriptor is used to label aerial video frames in a perceptually meaningful way. Many researchers have worked on the problem of content-based image/video retrieval based on example/sketch queries. Many of these methods use low-level visual features such as color, texture, motion, and shape for retrieval [2]. Although this approach is convenient for internal representation, it is unfriendly to users who are not familiar with the systems since the semantics are not captured by low-level features. Recently, there have been attempts to bridge this gap by representing video content with probabilistic semantic models [3]. In the proposed approach, a combination of Markov random field (MRF) and Gaussian mixture modeling is used for the semantic analysis and representation of image/video content. The MRF model is reinforced by biasing the Gibbs energy function at each site with the class-conditioned feature likelihoods obtained using a GMM. The sites for the MRF are blocks of pixels, each of which is described by a texture descriptor. Similarity retrieval at the frame level is then enabled by modeling the spatial arrangement of the regions as a Markov random field.

The rest of the paper is organized as follows. Section 2 describes the homogeneous texture descriptor. Sections 3 and 4 describe the two applications. And, section 5 concludes.

2. Homogeneous texture descriptor

The homogeneous texture descriptor is based on the outputs of Gabor filters. The use of Gabor filters is motivated by several factors. The Gabor representation has been shown to be optimal in the sense of minimizing the joint two-dimensional uncertainty in space and frequency [4]. These filters can be considered as

orientation and scale tunable edge and line detectors, and the statistics of these micro-features can be used to characterize the underlying texture.

The texture feature vector consists of the first and second moments of the filter outputs. If $f_{11}(x,y),\ldots,f_{RS}(x,y)$ are outputs of a filter bank tuned to S scales and R orientations then the feature vector f is

$$f=[\mu_{11},\sigma_{11},\mu_{12},\sigma_{12},\ldots,\mu_{1S},\sigma_{1S},\ldots,\mu_{RS},\sigma_{RS}] \quad (1)$$

where μ_{ij} and σ_{ij} are the mean and standard deviation of the output of filter $f_{ij}(x,y)$ respectively.

The texture feature vector can be used to compute the visual similarity between images by defining a distance measure on the $2xRxS$ dimension space. The Euclidean distance is commonly used so the distance between images $I^{(1)}$ and $I^{(2)}$ is:

$$d(I^{(1)},I^{(2)})=\left\|f^{(1)}-f^{(2)}\right\|_2. \quad (2)$$

Figure 1 shows an example of using the texture descriptor for query-by-example similarity retrieval. The left top tile is used to retrieve similar tiles from a large collection of satellite images. The tiles are of a freeway.

Orientation invariant similarity retrieval is possible by modifying the distance measure to exploit the structure of the feature vector. Specifically, let $f_{\langle r\rangle}$ be the feature vector circularly shifted by r orientations. Orientation invariant similarity can then be computed using the following distance measure:

$$d_{RI}(I^{(1)},I^{(2)})=\min_{r\in R}\left\|f^{(1)}_{\langle r\rangle}-f^{(2)}\right\|_2. \quad (3)$$

Conceptually, this distance function computes the best match between rotated versions of the images. Figure 2 shows an example of orientation invariant similarity retrieval.

Please see [5,6] for more details on the homogeneous texture descriptor. An online demonstration of similarity retrieval in a large collection of aerial images is available at [7].

3. Geo-spatial object modeling

The texture descriptor has limited application for characterizing remote sensed imagery since it is a low-level feature. This section describes a framework for using the descriptor to enable an object-based representation [8]. In particular, geo-spatial objects consisting of multiple characteristic textures are modeled using mixtures of Gaussians. The eventual goal is to use the models to locate new object instances. The intermediate goal is to estimate the spatial extents of known object instances.

3.1. Geo-spatial objects and texture motifs

The goal is to model geo-spatial objects that consist of characteristic textures but lack 1) consistent geometry, 2) well-defined boundaries, and/or 3) well-defined spatial layouts. Such objects defy traditional modeling approaches and include harbors, golf courses, and mobile home parks. The focus of the approach is to characterize the common textures, or texture motifs, such as the rows of boats and water in harbors, or the grass and trees in golf courses.

3.2. Gaussian mixture models

Homogeneous texture feature vectors are extracted using the Gabor filters described in section 2. These vectors are the direct outputs of the filters, not the summary vectors consisting of the first two moments computed over the image. Thus, there is an RxS dimension vector c associated with each pixel. Assuming that the pixels in an object class, such as harbors, are generated by one of N possible texture motifs, modeled as Gaussians, the probability density function of c can be expressed as a mixture distribution,

$$p(c)=\sum_{j=1}^{N}P(j)\,p(c\mid j) \quad (4)$$

where $p(c|j)$ is the conditional likelihood of the feature c being generated by motif j, and $P(j)$ is the prior probability of motif j. The number of motifs N along with distribution means and covariance matrices are the parameters that completely specify the class model.

The Expectation Maximization (EM) algorithm [9] is used to estimate the parameters of the Gaussian mixture models (GMMs). Since the objects and their constituent textures occur at arbitrary orientations, the EM algorithm is extended to account for not only the unknown mixture memberships but also the unknown orientations of the feature vectors [10]. This is accomplished by exploiting the structure of the texture features in a manner similar to the orientation invariant distance measure described in section 2. The EM learning is bootstrapped using a K-means algorithm similarly modified to account for the unknown texture orientations.

The GMM can be used to label the motifs for an object instance. A maximum a posteriori (MAP) classifier assigns label i^* to a pixel with feature vector c according to

$$i^*=\arg\max_{1\le i\le N}[P(i\mid c)]. \quad (5)$$

The posterior probabilities $P(i|c)$ are computed using Bayes' rule. Figure 3 shows the motif assignments for two harbor regions.

3.3. Estimating spatial extents

As mentioned, the intermediate goal is to use the models to estimate the spatial extents of known object instances. The Alexandria Digital Library (ADL) [11] at UCSB contains an extensive gazetteer that catalogues the spatial locations of over 5 million instances of over 200 types of geo-spatial objects. However, only a single point location is available for each instance so that extending the gazetteer to include even bounding boxes has been identified by the ADL development team as essential for supporting a broader range of spatial queries [12]. This can be accomplished by modeling the spatial arrangement of the motifs for an object as a Markov random field. The model parameters are learned from a training set using a Markov Chain Monte Carlo approach [13]. The model is then used to iteratively grow a bounding box that is initialized inside a known object instance using the location information available in the gazetteer. Figure 4 shows the bounding box used to estimate the spatial extent of a mobile home park at every 75 iterations from initialization to stopping. The image region outside the manually chosen ground truth has been dimmed for clarity. Please see [14] for more details on using the models for estimating spatial extents.

4. Semantic image labeling using GMMs and MRFs

4.1. Semantic image labeling

Learning the statistical distribution of features in each semantic class is the reverse of a classification problem. Images (or key frames from videos) from a training set are partitioned into blocks, allowing the computation of localized features without getting into the details of image segmentation, which is still an outstanding problem. For each block, the texture feature vector is extracted, and a semantic label from a predetermined set is manually assigned. Since visually similar textures tend to form clusters in a sparse feature space and there is a wide variation in visual appearance within each semantic class, a GMM is used to model the feature distribution conditioned on each class label. Suppose that each semantic class can be described by K clusters in the feature space. For each class, the distribution of features $Y \in R^d$ (d=60 in our experiments) is modeled as a mixture of K Gaussians, with the following density function:

$$P(y \mid \theta_m) = \sum_{j=1}^{K} \alpha_{mj} \frac{1}{\sqrt{(2\pi)^d \mid \Sigma_{mj} \mid}} \exp\left\{-\frac{1}{2}(y-\mu_{mj})^T \Sigma_{mj}^{-1}(y-\mu_{mj})\right\} \quad (6)$$

where $\theta_m = \left\{\alpha_{mj}, \mu_{mj}, \Sigma_{mj}\right\}_{j=1}^{K}$ is the parameter set for the semantic class m. α_{mj} are the mixture coefficients, $\mu_{mj} \in R^d$ is the mean of the j^{th} Gaussian, and Σ_{mj} is the covariance matrix of the j^{th} Gaussian.

Using the set of GMMs trained as described above, an image block with feature vector y can be classified into a semantic class using a maximum likelihood classifier. However, since the spatial relationships between neighboring blocks is not considered, many inconsistent labelings occur. For example, it is possible that a "street" block is shown to be surrounded by "water" blocks. This problem is addressed by using an MRF to model the label distribution. The label of a block at site s is 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. The random field X is modeled as an MRF with a Gibbs distribution [15]:

$$p(x) = \frac{1}{Z} e^{-U(x)} \quad (8)$$

where x is a realization of X. The Gibbs energy function $U(x)$ can be expressed as the sum of clique potential functions,

$$U(x) = \sum_{c \in Q} V_c(x) \quad (9)$$

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:

$$U(x) = \sum_{s \in S} \left(\sum_{s' \in N_s} -\alpha LP_{s-s'} - \beta LP_s \right) \quad (10)$$

where $LP_{s-s'} = \log(p_{s-s'}(x_s, x_{s'}))$ and $LP_s = \log\left(p(y_s \mid \theta_{x_s})\right)$. N_s is the set of neighbors of the site s, α and β 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 represents the conditional probability density of feature vector y_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 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 used. 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.

4.2. 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. This representation of an image in terms of the semantic labels of its blocks is called its *semantic layout*. For retrieval purposes, the similarity between the query image and each stored image has to be measured from their semantic layouts. Let the semantic layout of the query image be X^q and that of the stored image be X^I. In order to improve the retrieval performance, a soft classification scheme is adopted. For a given image block, the labels with the three largest local conditional probabilities are selected to represent this block. All these candidate labels are stored along with the feature vectors for future retrieval. The modified semantic layout similarity between the query image and each stored image is given by

$$S_3 = \sum_{s\in S}\left(\sum_{j=1}^{3} a_1 a_j \delta(x_{s,j}^q, x_{s,1}^I)\right) + \sum_{s\in S}\left(\sum_{j=1}^{3} a_2 a_j \delta(x_{s,j}^q, x_{s,2}^I)\right)\left(1-\sum_{i=1}^{3}\delta(x_{s,i}^q, x_{s,1}^I)\right) + \sum_{s\in S}\left(\sum_{j=1}^{3} a_3 a_j \delta(x_{s,j}^q, x_{s,3}^I)\right)\left(1-\sum_{i=1}^{3}\delta(x_{s,i}^q, x_{s,1}^I)\right)\left(1-\sum_{i=1}^{3}\delta(x_{s,i}^q, x_{s,2}^I)\right) \quad (11)$$

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

In Eq. (12), 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. This approach for image/video similarity retrieval is expected perform better than existing methods that do not consider the underlying semantics.

The similarity measure of Eq. (12) is effective only when all images are of the same size. In order to compare the semantics of images of any size, a semantic histogram can be computed for each image. This is similar to the image histogram where the inputs are semantic labels instead of image intensities. As long as the semantic label set is the same, two images can be compared using their semantic histograms.

Figures 5 and 6 show the top four retrieval results using two different queries. Figures 5(a) and 6(a) show the query images and the corresponding semantic layouts. Observe that some of the retrieved images appear visually different from the query image, while their semantic layouts are similar. For example, in figure 5, the parking lot in the first retrieved image appears dissimilar to that in the query image. Note that this cannot result from a low-level feature query. The combination of GMM and MRF thus captures the semantic layout of a scene effectively.

5. Discussion

This paper describes two applications of using texture to annotate remote sensed datasets. In the first application, geo-spatial objects are modeled by characterizing the distribution of texture motifs using Gaussian mixtures. These models are then used to estimate the spatial extents of known object instances. In the second application, a novel combination of GMMs and MRFs is used to model the distribution of semantic classes in aerial image/video data. Frame level similarity retrieval based on semantic layout and semantic histogram is then enabled.

Acknowledgements

This work is supported by the following grants: NASA California Space Grant, USDOT Research and Special Programs Administration Contract DTRS-00-T-0002, ONR# N00014-01-1-0391, NSF Instrumentation #EIA-9986057, and NSF #IIS-9817432.

References

[1] B.S.Manjunath, P. Salembier, and T. Sikora, Eds., *Introduction to MPEG7: Multimedia Content Description Interface*, John Wiley & Sons, first edition, 2002.

[2] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik, "Blobworld: A System for Region-based Image Indexing and Retrieval," in *Proceedings of the Third International Conference on Visual Information Systems*, 1999, pp. 509-516.

[3] 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.

[4] J.G. Daugman, "Complete discrete 2D Gabor transforms by neural networks for image analysis and compression", *IEEE Trans. ASSP* 36 (July 1988) 1,169-1,179.

[5] B.S. Manjunath and W.Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.18, no.8, pp.837-42, Aug 1996.

[6] P. Wu, B. S. Manjunath, S. Newsam and H.D. Shin, "A texture descriptor for browsing and similarity retrieval," *Journal of Signal Processing: Image Communication*, Volume 16, Issue 1-2, page 33-43, September 2000.

[7] Online demo:http://vision/texture/mpeg7/instructions.html

[8] S. Bhagavathy, S. Newsam, and B. S. Manjunath, "Modeling Object Classes in Aerial Images Using Texture Motifs," *International Conference on Pattern Recognition*, Quebec, Canada, August 2002.

[9] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm," *Royal Stat. Society*, Series B, 39, 1977.

[10] S. Newsam and B.S.Manjunath, "Normalized texture motifs and their application to statistical object modeling," *ICCV*, Beijing, October 2003 (submitted).

[11] Alexandra Digital Library Project homepage: http://www.alexandria.ucsb.edu.

[12] L.L. Hill, J. Frew, and Q. Zheng, "Geographic names: The implementation of a gazetteer in a georeferenced digital library," *D-Lib*, January 1999.

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

[14] S. Newsam, S. Bhagavathy, and B.S. Manjunath, "Object localization using texture motifs and Markov random fields," *Int. Conf. on Image Processing*, September 2003.

[15] S.Z. Li, "Markov Random Field Modeling in Image Analysis," *Springer-Verlag*, 2001.

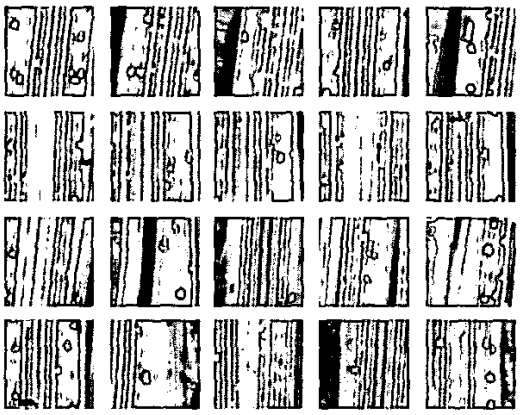


Figure 1. Similarity retrieval in a large collection of satellite images. The top left tile is the query and the other tiles are the most similar with respect to the texture descriptor.

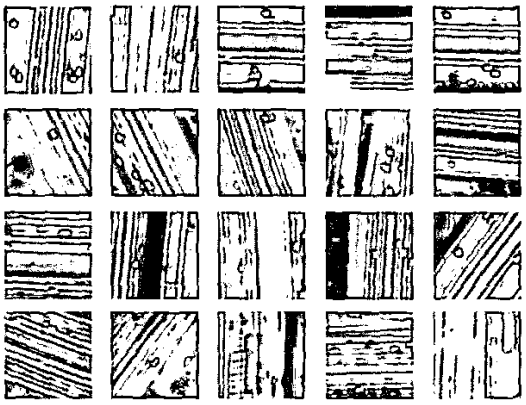


Figure 2. Orientation invariant similarity retrieval.

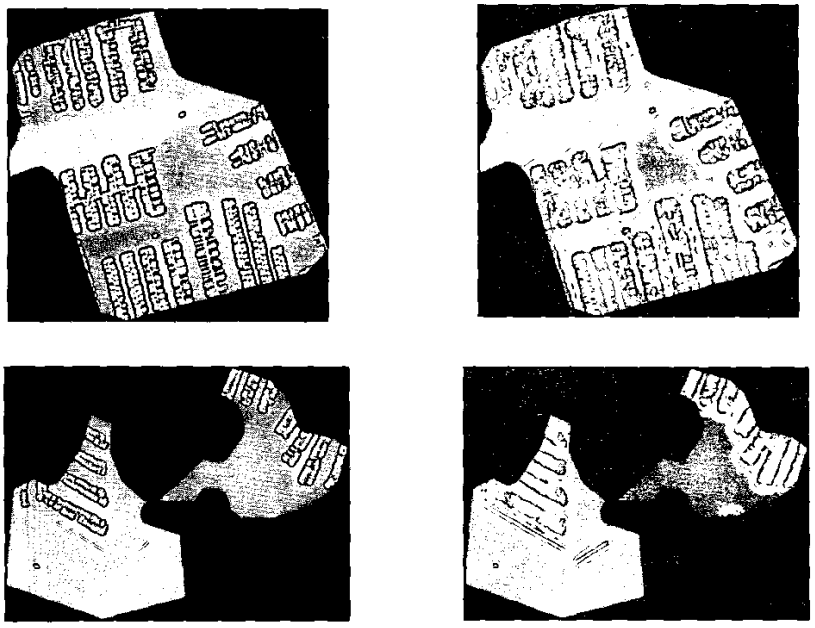


Figure 3. Two harbor regions and corresponding texture motif assignments.

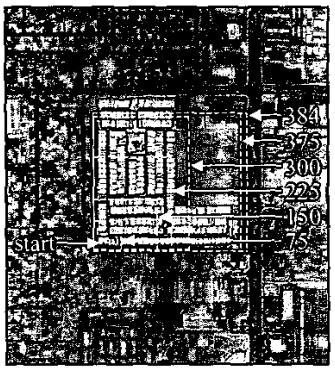
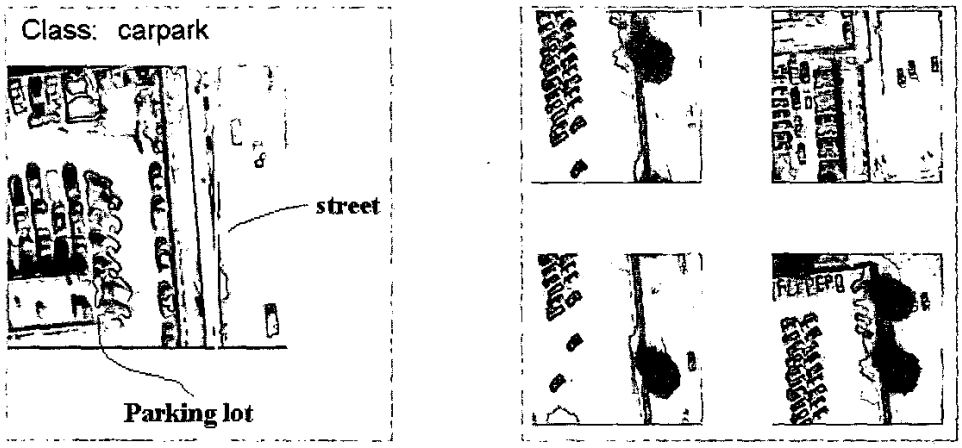
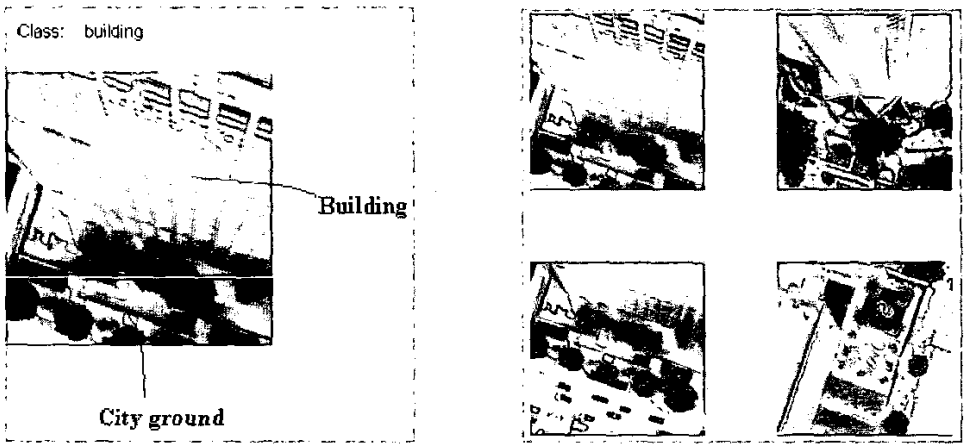


Figure 4. Bounding box for estimating spatial extent of mobile home park every 75 iterations until stopping.



(a) Query image and its semantic layout. (b) Retrieval results.

Figure 5. Semantic retrieval example #1: The query image contains a parking lot and a street. Observe that the retrieved images have the same semantic configurations, although they are visually quite different.



(a) Query image and its semantic layout. (b) Retrieval results.

Figure 6. Semantic retrieval example #2: The query image contains building and city ground.