Large-scale Satellite Image Browsing using Automatic Semantic Categorization and Content-based Retrieval*

Ashish Parulekar† Ritendra Datta‡ Jia Li§ James Z. Wang¶
The Pennsylvania State University, University Park, PA, USA
{ashishp,datta,jiali,jwang}@psu.edu

Abstract

We approach the problem of large-scale satellite image browsing from a content-based retrieval and semantic categorization perspective. A two-stage method for query based automatic retrieval of satellite image patches is proposed. The semantic category of query patches are determined and patches from that category are ranked based on an image similarity measure. Semantic categorization is done by a learning approach involving the two-dimensional multi-resolution hidden Markov model (2-D MHMM). Patches that do not belong to any trained category are handled using a support vector machine (SVM) based classifier. Experiments yield promising results in modeling semantic categories within satellite images using 2-D MHMM, producing accurate and convenient browsing. We also show that prior semantic categorization improves retrieval performance.

1. Introduction

Today, the need for reliable, automated, satellite image classification and browsing systems is more than ever before. Every day there is a massive amount of remotely-sensed data being collected and sent by terrestrial satellites for analysis. The use of automated tools for this analysis has become imperative due to the large volumes required to be processed on a daily basis. Applications of this classification lie in diverse fields such as Geography, Geology, Archaeology, Atmospheric Sciences, Environmental Sciences and in serving other civilian and military purposes.

Over the last three decades, there has been an overwhelming amount of work done in the field of automated classification of satellite images, one of the earliest being [7] by Haralick et al. In the more recent years, many different techniques have been proposed, some of which include texture analysis [10, 8], Markov random fields [13], genetic algorithms [14], fuzzy approaches [6, 18], bayesian classifiers [11, 5] and neural networks [2, 1]. Some others have approached satellite image classification as an application of content-based image retrieval [10, 12].

In a recent survey [17] on satellite image classification results published in the last fifteen years, it has been reported that the classification accuracy has not increased significantly over the years. The paper reports a mean overall pixel-wise classification accuracy across all these experimental results at 76.19% with a standard deviation of 15.59%. However, as suggested by the author, these results need to be taken with caution, since a number of experimental parameters have varied over the years and across different approaches towards classification. Nonetheless, we believe that this is certainly indicative of the existence of an upper bound on the classification accuracy in this type of imagery, given the current technology. As a result, this paper is more about approaches towards a fast and scalable working system while maintaining respectable classification accuracy.

Our work involves learning the semantics of satellite image patches for querying, retrieval and browsing, using a content-based image retrieval approach. The purpose is not so much to improve upon the accuracy rates but rather to show that 2-D MHMM is an effective and efficient way to jointly model the texture and spatial structure of semantic categories within satellite imagery. Moreover, we show through experiments that performing categorization prior to applying image retrieval techniques increases speed and accuracy in browsing large databases of satellite imagery. Rather than generating an automatic segmentation of the terrain at pixel level, we intend to provide a browsing platform through patch-level classification followed by a

*The material is based upon work supported by the US National Science Foundation under Grant Nos. IIS-0219272, EIA-0202007, and IIS-0347148, the PNC Foundation, and SUN Microsystems. Discussions with satellite image analyst Bradley Farster were helpful in selecting the training samples and acquiring the data set. `Riemann.ist.psu.edu` provides more information.

†A. Parulekar is affiliated with the Department of Electrical Engineering.

‡R. Datta is affiliated with the Department of Computer Science and Engineering.

§J. Li is affiliated with the Department of Statistics and the Department of Computer Science and Engineering.

¶J. Z. Wang is affiliated with the School of Information Sciences and Technology and the Department of Computer Science and Engineering.

similarity matching. The reasons it is helpful to have retrieval at the patch level are 1) within a semantic category such as urban regions or forests, it helps to find regions with comparable density, 2) it helps to track salient features such as specific patterns of deforestation or terrace farming within crop fields, and 3) from a large set of satellite images it helps to find those that contain significant coverage of a particular type such as cloud cover, so that manual effort can be concentrated only on them.

We consider publicly available[1] Landsat-7 Enhanced Thematic Mapper Plus (ETM+) images [15]. For our experiments, four categories are considered, although there is scope for seamless addition of more classes through a modular training process. The goal is to build a system with an interface to browse/retrieve small semantically homogeneous regions from large collections of remotely sensed imagery consisting of multiple (possibly dynamically growing) categories. The retrieval of relevant patches or regions can be either by example or through complex querying. Support for such querying is made easier since automatic semantic categorization is performed prior to retrieval and since we are dealing with patches rather than pixels. However, classification of patches instead of pixels leads to ambiguity, because ground-truth inter-class boundaries are at the pixel-level. How we deal with this issue is discussed in the Sec. 2. One of the reasons why semantic categorization is done prior to retrieval is that it reduces the search space by a factor roughly equal to the number of categories. This becomes more and more significant as the size of the patch database or the number of categories increases.

The rest of the paper is arranged as follows. In Sec. 2, we discuss the proposed system architecture including the user interface for browsing and retrieval. In Sec. 3, the learning based categorization and retrieval methods are discussed. The experimental setup and the results obtained are discussed in Sec. 4. We conclude in Sec. 5.

2. System Architecture

Let us first define the generic framework of the proposed system before going into specific details. There are two parts to the system. The *off-line* processing part consists of initial data acquisition, ground-truth labeling and model building, while the *on-line* part consists of querying, browsing and retrieval. A schematic diagram of the architecture is given in Fig. 1.

2.1 Off-line Processing

Consider a set of M satellite images $I_i, i = 1, ..., M$ of an arbitrary type (Landsat, ASTER, SRTM etc.). A subset of the available spectra for the given type of imagery is taken for each image. The choice of a subset is either empirical or based on past research trends. In the case of Landsat-7 ETM+ images used in our experiments, this choice is discussed in Sec. 4. We first perform histogram stretching and adjusting in order to improve the visual clarity of the images. We then divide each image into equal sized non-overlapping rectangular patches of dimensions $X_w \times Y_w$, padding the right and bottom sides with zeros appropriately, to get a total of N patches $\{p_1, ..., p_N\}$. For each patch p_i, information about the global position (z_i, w_i) of its top left corner is stored as meta-data. This can be calculated using the global position of the top left corner of the image it was a part of, and its relative position within that image.

Suppose that by some means there is a way to manually identify K non-overlapping semantic classes or categories[2] $\{S_1, ..., S_K\}$ relevant to a specific application. Note that this need not be an exhaustive set of classifications but those that can be easily identified through some reliable source. We require a small number T of patches $\{t^1_{S_k}, ..., t^T_{S_k}\}$ of each semantic category S_k for training the 2-D MHMMs. As mentioned, one issue here is the ambiguity in the semantic classification of individual patches.

Figure 2: Examples of ambiguous patches. *Left:* Urban and Residential. *Right:* Residential and Crop.

Traditionally, satellite image classification has been done at the pixel level [13, 2]. For a typical Landsat image at $30m$ resolution, a 128×128 sized image patch covers roughly $14.74Km^2$. This is too large an area to represent precise ground segmentation, but our focus is more on building a querying and browsing system than showing exact boundaries between classes. Dividing the image into rectangular patches makes it very convenient for training as well as browsing. Since users of such systems are generally more interested in getting an overview of the location, zooming and panning is allowed optionally as part of the interface. Moreover, the training process can be done at multiple scales in an identical manner, for browsing and retrieval at a scale of choice. The only consideration is that the training images need to be sampled accordingly. Nonetheless, classification at the pixel-level in itself does not provide for a good browsing strategy, even

[1] http://glcf.umiacs.umd.edu/data/landsat/

[2] The words "class" and "category" are used interchangeably throughout this paper.

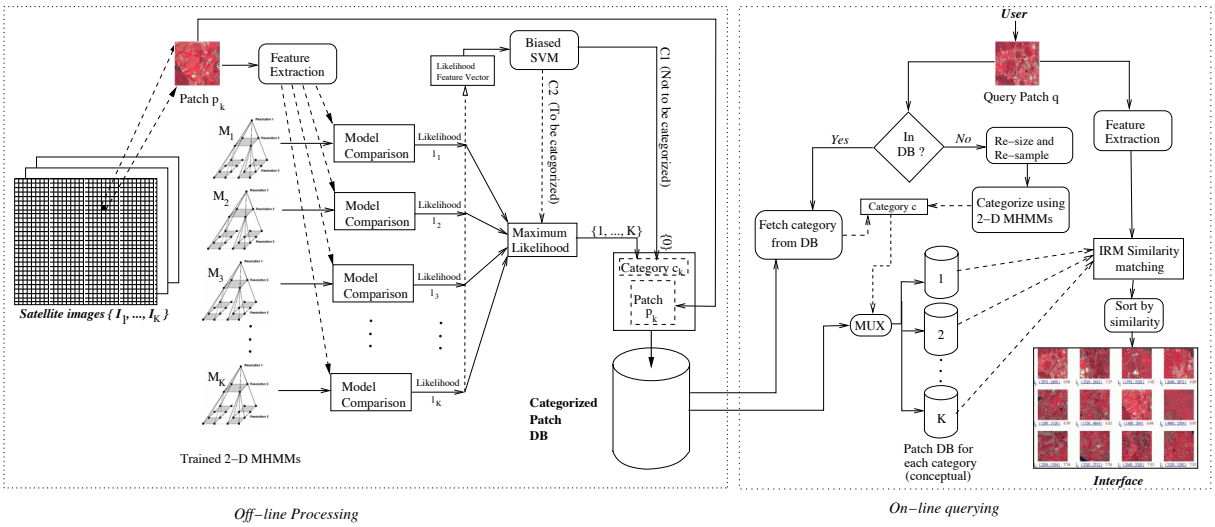

Figure 1: System architecture. *Left side:* The database building process. *Right side:* Real-time user interaction.

though it may give an overall segmented view of the land cover. Even at the pixel level there may sometimes be class ambiguity, due to which some authors have proposed *fuzzy* classification [6, 18] as opposed to *hard* classification. In our case of semantic categorization at patch level, both manual and automatic, we need a strategy to resolve this possibly greater ambiguity. For example, as shown in Fig. 2, some patches have large coverage of different categories. We do not incorporate fuzziness. Instead, our strategy is to consider a patch p_k to belong to a category S_k if p_k has roughly over 50% coverage of type S_k (dominant category). Patches which do not belong to any of the categories $\{S_1, ..., S_K\}$ or those that do not have a dominant category are given class label 0 (category unknown).

For each semantic category k, a separate 2-D MHMM is trained using visual features of the corresponding training patches $\{t^1_{S_k}, ..., t^T_{S_k}\}$, resulting in K different models $\{M_1, ..., M_K\}$. Now for each of the image patches $\{p_1, ..., p_N\}$ in the database, the likelihood l_k of belonging to class k is computed using trained model M_k. Since we are dealing with a non-exhaustive set of categories, those patches that do not belong to any of the classes are required to be labeled as class 0, as mentioned before. It does not make much sense to treat them as a separate class as far as training another 2-D MHMM is concerned, since there may not exist any spatial or textural motif among them. Instead, we perform another supervised classification using Support Vector Machines (SVM). We take two sets of randomly chosen training patches, $C1$ with manual class labels 0 (not to be categorized), and $C2$ with any of the labels $\{1, ..., K\}$ (to be categorized). The K 2-D MHMM likelihood estimations for each of the samples of the two classes are used as feature vectors for training an SVM. The sampling of $C1$ and $C2$ is done in such a way that $C2$ is predicted with high accuracy while allowing $C1$ to be predicted with moderate accuracy, producing a *biased* SVM. A new patch p_k, whose 2-D MHMM likelihood estimation vector $\{l_1, ..., l_K\}$ is classified as $C2$ by this biased SVM, is labeled $c_k = 0$, while for the rest of the patches, the class label is assigned as $c_k = \arg\max_i(l_i)$. The class label c_k is stored as meta-data for p_k. This is discussion in details in Sec. 3.

2.2 On-line Processing

Assume that there is an efficient indexing strategy for handling the database of patches and their associated meta-data. The simplest way to represent our query is the following. Given a query patch the user seeks to find patches within the same semantic category, sorted by their visual similarity. There can be two kinds of such querying:

1. *The patch is part of the database:* In this case the semantic category of the patch, say p_k is already stored as c_k. Patches in the database whose semantic categories are not c_k are eliminated from consideration.
2. *The patch is externally uploaded:* This patch is re-sized or trimmed to fit the standard dimensions $X_w \times Y_w$ and adjusted for spectral encoding (i.e., choosing the required subset from the available spectra), if needed. As mentioned before, the semantic category c_k of this patch is predicted using the 2-D MHMM likelihoods and the biased SVM. Again, all but the patches labeled c_k are eliminated from consideration.

The remaining patches are now ranked according to their visual similarity with the query. Visual similarity is computed using the Integrated Region Matching (IRM) measure, which is fast and robust and can handle large

image volumes. The top Q matched patches $\{p_{r_1}, ..., p_{r_Q}\}$ are then displayed for perusal. The choice of Q is contingent upon the specific application. Experimentation on choosing Q and how precision of retrieval varies with it are discussed in Sec. 4. Note that for the purpose of retrieval, query patches determined as $C1$ (uncategorized) are also searched from among only the $C1$ patches in the database.

Using the meta-data associated with the patches, such as precise geographic location or semantic category (either manually provided or automatically generated), more complex querying is possible. Such queries are often useful to analysts when handling large-scale image data. Some of the possible queries are "Find the closest urban area near a given crop field" and "Find satellite images that contain at least 10% residential coverage and show the associated patches". Additionally, users often require more information on the local neighborhood surrounding the retrieved patches.

Figure 3: Interface for zooming, showing urban (blue), residential (yellow) and the retrieved (green) patch. *Top:* Original $30m$ resolution. *Bottom:* Zoomed out at $60m$.

One way our system helps in this regard is by providing the interface and support for *zooming* and *panning*. These features allow the users the ability to view the position and neighborhood of the patches of interest in their parent satellite images. Haar wavelet transforms are used to achieve zooming since they preserve localization of data. These transforms decompose the images into sums and differences of neighborhood pixels. On a given query, the system only needs to retrieve the quantized coefficients of the queried region for reconstruction. Since the processing for categorization and zooming is done only once during setup, and only localized parameters are required, the response time is considerably low. The interface for zooming in our system is shown in Fig. 3.

3. Categorization and Retrieval

3.1 Categorization using 2-D MHMMs

The two-dimensional multiresolution hidden Markov model (2-D MHMM) has been used for generic image categorization. Here we present a brief overview of the model and its application to semantic categorization of satellite images. For a more detailed discussion on the topic please refer to [9]. Under 2-D MHMM, each image is

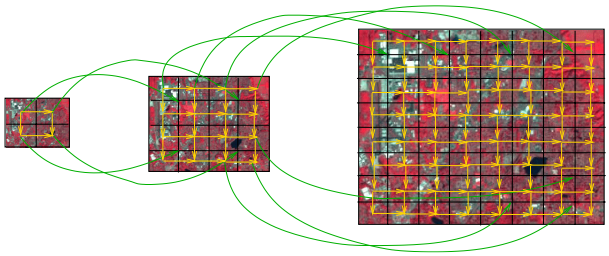

Figure 4: A conceptual diagram of the 2-D MHMM based modeling process. Arrows indicate the intra-scale and inter-scale transition probabilities among visual features.

characterized by several layers, i.e., resolutions, of feature vectors. The feature vectors within a resolution reside on a 2-D grid. The nodes in the grid correspond to local areas in the image at that resolution. A node can be a pixel or a block of pixels. The feature vector extracted at a node summarizes local characteristics around a pixel or within a block. The 2-D MHMM specifies the distribution of all the feature vectors across all the resolutions by a spatial stochastic process. Both inter-scale and intra-scale statistical dependence among the feature vectors are taken into account in this model. These dependencies are critical for judging the semantic content of satellite image patches because texture or spatial structure in these patches can be captured at a larger scale than at a block or pixel level. While the inter-scale dependence is modeled using a Markov chain over multiple resolutions, the intra-scale dependence is captured using hidden Markov models (HMM). In our experiments we use a three-level pyramidal structure in the model. A schematic diagram on this idea can be found in Fig. 4. For feature extraction, 4×4 blocks are taken and the visual features are characterized by a six dimensional feature vector. This vector consists of three moments of the wavelet coefficients in the high frequency bands (representing texture) and the three average color components in the LUV space.

3.2 Separating C1 and C2 using SVM

The training of the 2-D MHMMs is performed on a finite non-exhaustive set of categories $\{S_1, ..., S_K\}$. Generating a training set covering all possible land-cover categories is a time-consuming and expensive task, if at all it is possible. Hence it is preferred to limit the scope to only those semantic classes that are of interest. As a result, among the image patches there exist many that represent categories outside of $\{S_1, ..., S_K\}$. Also, there are many patches that are a mixture of multiple categories without any one being dominant. In both cases, these patches should ideally be assigned a category label 0 $(C1)$. As mentioned previously, all of the patches labeled $\{1, ..., K\}$ are considered as $C2$.

Using the maximum likelihood approach, we end up assigning a category label between 1 and K to every patch, regardless of whether they belong to $C1$ or $C2$. This is not a desirable outcome, and as explained in Sec. 2.1, neither can we train another 2-D MHMM to model patches in class $C1$ to solve the problem. A naive approach to solving this problem is based on the following assumption: Given a patch that does not visually resemble any of the semantic categories, the likelihood estimation from all the models should be low. Under such an assumption, if all likelihood scores are below a certain threshold, then the patch can be assigned $C1$. However, not surprisingly, it is found that for a given patch, the likelihood estimates are not independent of each other. This may be due to the fact that the 2-D MHMMs are trained on samples that have some degree of visual resemblance across categories.

Let the set of likelihood estimates for a given patch p_k be its feature vector $L_k = \{l_1, ..., l_K\}$. In our experiments, we consider 4 classes. We plot the 4-D feature vectors of 2000 patches manually labeled as $C1$ or $C2$. The plots, taken two dimensions at a time, are shown in Fig. 5. Clearly, a non-linear method can better model the class separation than thresholding or other linear methods. We experimented with Quadratic Discriminant Analysis (QDA) and Logistic Regression for classification. The accuracy rate with Logistic Regression turned out to be the best at approximately 79% with accuracy of classifying only $C2$ at about 84%. We then tried SVM on the data using the LibSVM software package [3], using the RBF Kernel $\phi(f_i, f_j) = e^{-3|f_i - f_j|^2}$. The results were still better, at around 81.7% overall accuracy and 86.4% accuracy at classifying $C2$. When a patch is classified as $C1$ it is removed from further consideration for retrieval. We do not want to leave out any patch that could be a potential target for a given query, while it is acceptable to have some $C1$ patches to be classified as $C2$, hence the focus is on higher accuracy in detecting $C2$. Hence we desire to have a *biased* classifier. This process eliminates a significant chunk of unwanted patches. One way to introduce weights into

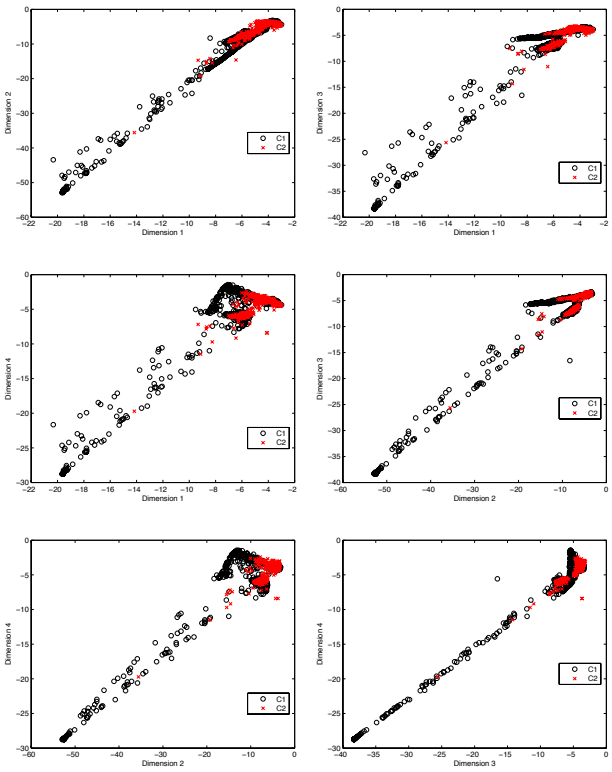

Figure 5: Plot of 4-D likelihood feature vector L for C1 (black) and C2 (red). Six pairs of dimensions are shown.

SVM learning is to sample the training classes accordingly. In our case we introduce bias by sampling $C1$ and $C2$ in the approximate ratio 13 : 25 for training the SVM, resulting in a total of about 9000 samples (with repetition). In this manner, we achieve high accuracy of classifying $C2$ (96.04%) while for $C1$ the score is moderate (53.7%) which is acceptable in our case. Hence less than 4% of the patches within categories $\{S_1, ..., S_K\}$ will be mistakenly eliminated. This may not be a problem since patches of one category in a satellite image are usually spread over a large region. It is highly unlikely that all patches in one region will be eliminated and a target patch will slip the user's attention, since our system supports panning.

3.3 Retrieval using IRM

Integrated Region Matching (IRM) [16] is a robust region-based image similarity measure. In our experiments, IRM was used to perform retrieval by ranking image patches within the query category.

The images are segmented and for each segment, a nine dimensional feature vector is composed. The feature vectors used include the same six texture and color features used in 2-D MHMM (see Sec. 3.1), and three more features characterizing the shape of the segment. The matching is

performed by a *soft* similarity measure in the following manner. For two images i_1 and i_2, suppose they have k_1 and k_2 segments respectively. The IRM distance between images i_1 and i_2 is given by

$$d(i_1, i_2) = \sum_{l=1}^{k_1} \sum_{m=1}^{k_2} s_{lm} d_{lm}$$

where d_{lm} denotes the euclidian distance between the nine dimensional feature vectors of segment l of i_1 and m of i_2 and s_{lm} is the significance credit associated with that pair of segments. The significance credit between a given pair of segments measures how much importance is to be put on the comparison of visual features between that pair. This is partly dependent on the percentage of area covered within their respective images. The significance computation is performed using the most similar highest priority (MHSP) principle [16]. Our experiments show that IRM performs well with satellite images, possibly due to the soft matching approach and the emphasis on texture features.

4. Experimental Results

For our experiments, we use $M = 3$ Landsat-7 ETM+ multi-spectral satellite images with $30m$ resolution. We choose to support $K = 4$ semantic categories in our experimental system, namely *mountain*, *crop field*, *urban area*, and *residential area*. In consultation with an expert in satellite image analysis, we choose *near-IR* (infra-red), *red* and *green* bands as the three spectral channels for classification as well as display. The reasons for this choice are as follows. Near-IR band is selected over *blue* band because of a somewhat inverse relationship between a healthy plant's reflectivity in near-IR and red, i.e., healthy vegetation reflects high in near-IR and low in red. Near-IR and red bands are key to differentiating between vegetation types and states. Blue light is very abundant in the atmosphere and is diffracted all over the place. It therefore is very noisy. Hence use of blue band is often avoided. Visible green is used because it is less noisy and provides unique information compared to Near IR and red.

The pixel dimensions of each satellite image I_i used in our experiments are 6000×6600, with geographic dimensions being approximately $180\ Km \times 198\ Km$. The choice patch size is critical. A patch should be large enough to encapsulate the visual features of a semantic category, while being small enough to include only one semantic category in most cases. We choose patch size $X_w \times Y_w$ to be 128×128 pixels. Our experiments show that 2-D MHMMs are able to capture visual features of semantic categories quite well at this size. We obtain $N = 9874$ patches from all the images in this manner. These patches are stored in a database along with the identity of their parent images and the relative location within them. Ground-truth categorization is not available readily for our patches. This is required for testing the accuracy of 2-D MHMM based categorization and retrieval. In order to build a manual categorization of the patches, an expert working on satellite image analysis in a government research lab gave 2 arbitrarily chosen subjects a tutorial on how to distinguish between the 4 semantic categories. The subjects then independently labeled each patch as either $\{1, 2, 3, 4\}$, or 0 in case it belonged to neither class or had no dominant coverage, keeping in mind the 50% coverage policy (Sec. 2). The final category labels $\{c_1, ..., c_{1984}\}$ are determined by taking the overlap of the sets as it is, and in case of conflict, randomly choosing one of the two. With the high-quality Landsat images it is not hard to visibly identify the four categories used. The overlap between these two sets is approximately 94%. This serves as our "silver standard".

For classification, we use $T = 40$ samples of each of the four categories for training the 2-D MHMMs to yield models M_1, M_2, M_3 and M_4. It is important that accuracy of these models in categorization be high, otherwise many critical patches might be wrongly eliminated from the retrieval process. In order to test the accuracy, we randomly picked up 900 patches outside of those used for training and compared the classification results with the manual "silver standard". We computed the *confusion matrix* for the 4 classes as well as for the class 0 ($C1$) patches, shown in Table 1. Note that the accuracy of classifying $C1$ patches reflects on the model accuracy of both 2-D MHMMs and the biased SVM. The overall unweighted accuracy over the 4 categories and 0 for these 900 samples is 87.22%. A measure of accuracy often used in the remote-sensing community to evaluate multi-class classification performance is Cohen's Kappa Coefficient:

$$\kappa = \frac{N \sum_{i=1}^{K} R_{ii} - \sum_{i=1}^{K} (R_{i'}\ R_{i''})}{N^2 - \sum_{i=1}^{K} (R_{i'}\ R_{i''})}$$

where K is the number of classes, N is the total number of samples R_{ij} indicates observation in row i column j, $R_{i'}$ is the total of row i and $R_{i''}$ is the total of column i. When taking only classes 1 to 4, $\kappa = 93.02\%$, while when including class 0 ($C1$) also into consideration we have $\kappa = 82.81\%$. These results are very encouraging.

Sample results obtained when querying using a *residential* patch and a *mountain* patch are shown in Fig. 6. It is worth noting that in our system, patches in untrained categories can also be effectively retrieved. For example, as shown in Fig. 7, retrieval results for a query using a *coastline* patch are rather satisfactory, albeit with less precision. The IRM similarity based ranking and display of patches should reflect relevance to the query. Of the Q patches displayed in response to each query, one measure to determine retrieval effectiveness is the percentage of relevant patches in them. We measure this as follows. For each of the four classes, we use our system to retrieve from

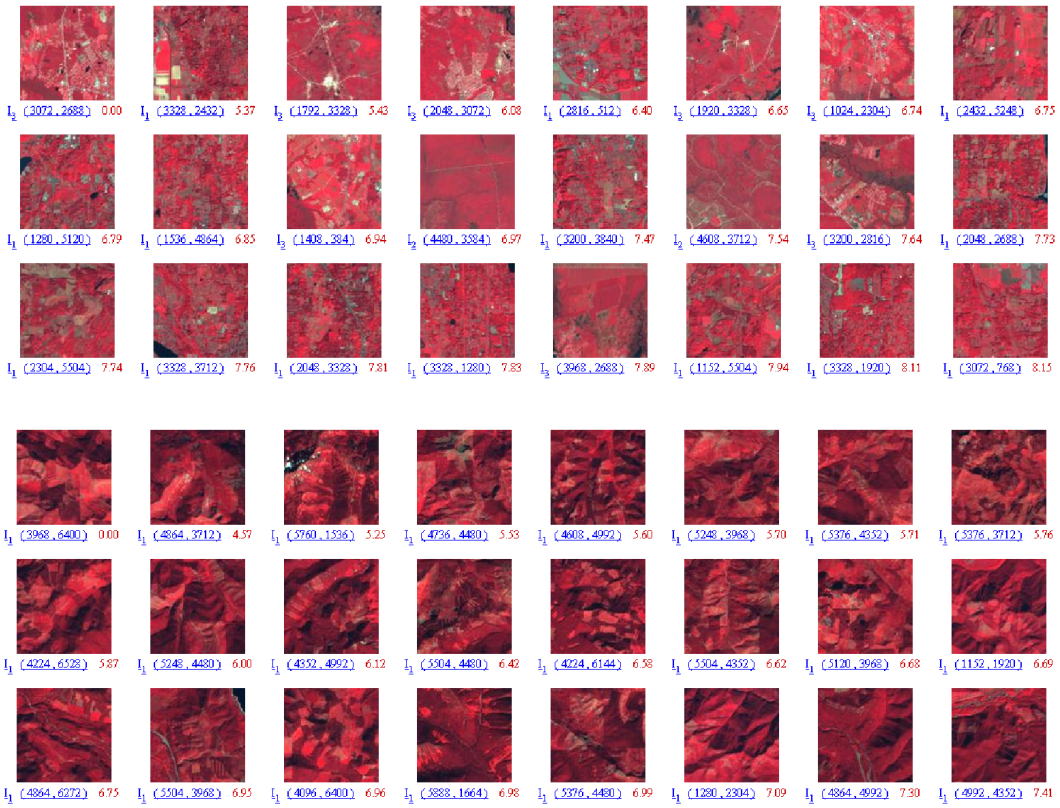

Figure 6: Ordered retrieval results on *Residential* (top) and *Mountain* (bottom) query patches. Patch labels consist of (1) Parent image, (2) Local Co-ordinates and (3) IRM distance.

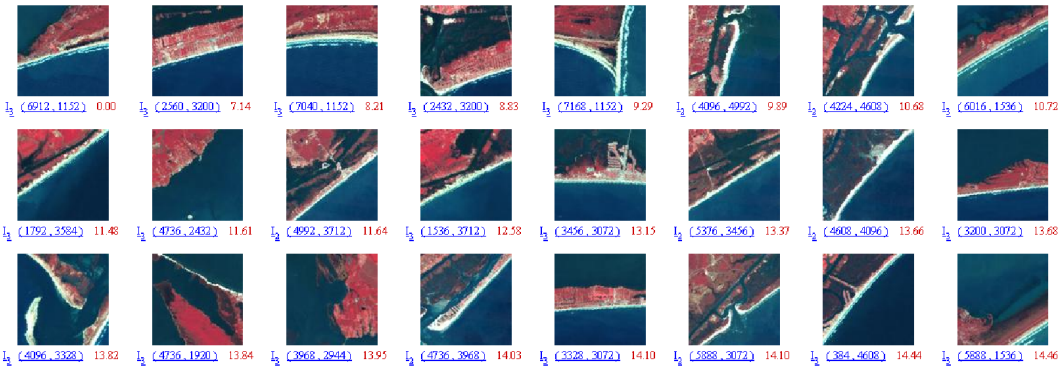

Figure 7: Demonstrating the effectiveness of retrieval within the *Other* ($C1$) category: Coast-lines, though not learnt using 2-D MHMM, are retrieved with high accuracy due the SVM classification and the IRM measure.

5 to 30 patches per query (in intervals of 5) and measure the percentage of patches retrieved that have the same manual category label as the query patch. This is repeated 5 times for each category and the average accuracy results are plotted over variation of Q, as shown in Fig. 8. The most vital observation made is that semantic categorization using 2-D MHMM results in roughly 6% to 10% improvement in retrieval accuracy. However, that accuracy drops when the number of patches retrieved increases. Yet, the values are considerably high at $Q = 30$. For specific requirements, these graphs can be used to choose suitable values of Q.

About 20 minutes are required to train each 2-D MHMM on a 1.7 GHz Intel Xeon machine, but this is not a recurring process. Subsequent indexing is done only once for each image added to the database. Our system performs retrieval in real-time. Since linear search is performed within the five-class database, the retrieval time decreases roughly five times on an average with semantic categorization.

Table 1: Classification Results using 2-D MHMM

	Mtn.	Crop	Urban	Res.	Oth.	Accuracy
Mtn. (1)	198	0	0	0	13	93.84%
Crop (2)	0	176	1	6	19	87.13%
Urban (3)	1	0	43	6	8	74.14%
Res. (4)	3	3	5	76	3	84.44%
Oth. (0)	6	4	17	20	292	86.14%

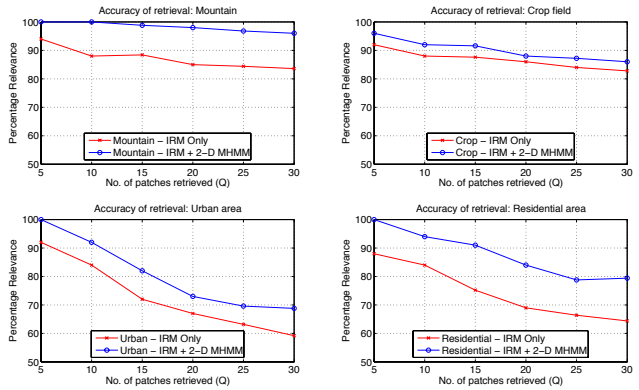

Figure 8: Average accuracy of IRM based retrieval for each category, with/without prior categorization.

5. Conclusions

We have proposed a convenient learning based approach for large-scale browsing and retrieval of satellite image patches. It has been shown that automatic semantic categorization of image patches using 2-D MHMM prior to retrieval improves speed and accuracy. Our intuition is that 2-D MHMM and IRM complement each other to boost retrieval performance. Prior categorization reduces the search space to fewer, more relevant patches, thereby also reducing search time. SVM has been effectively used to deal with patches that have not been trained for. Performing classification at patch level instead of pixel level in satellite images helps in building a more convenient interface that allows complex querying. There are still some issues which have not been tackled in our present work. Square patches are used due to the convenience in computation, but the users may desire more flexible shapes for querying. Moreover, size of the patch is a function of the user's specific needs as well as performance requirements. How the patch size affects these two factors remains to be studied. We use only three of the six available bands from the satellite images. What impact there may be to use more bands on the performance has not been tested. How performance varies with change in the number of levels of the 2-D MHMM will also be an interesting study.

References

[1] P. M. Atkinson, A. R. L. Tatnall, "Neural Networks in Remote Sensing," *Int. J. of Remote Sensing*, 18(4):699-709, 1997.

[2] H. Bischof, W. Schneider, A.J. Pinz, "Multispectral Classification of Landsat-images using Neural Networks," *IEEE Trans. Geosci. and Remote Sens.*, 30(3):482-490, 1992.

[3] C.-c. Chang, C.-j. Lin, "LIBSVM : A Library for SVM," 2001. `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[4] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, 20:37-46, 1960;

[5] H. Daschiel, M. Datcu, "Information Mining in Remote Sensing Image Archives: System Evaluation", *IEEE Trans. Geosci. and Remote Sens.*, 43(1):188-199, 2005.

[6] G.M. Foody, "Approaches for the Production and Evaluation of Fuzzy Land Cover Classifications from Remotely-sensed Data," *Int. J. of Remote Sensing*, 17(7):1317-1340, 1996.

[7] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification, *IEEE Trans. Systems, Man and Cybernetics*, 3:610-621, 1973.

[8] C.-S. Li, V. Castelli, "Deriving Texture Feature Set for Content-based Retrieval of Satellite Image Database," *IEEE ICIP*, 1997.

[9] J. Li, R.M. Gray, R.A. Olshen, "Multiresolution Image Classification by Hierarchical Modeling with Two Dimensional Hidden Markov Models," *IEEE Trans. Information Theory*, 46(5), 1826-1841, 2000.

[10] B.S. Manjunath, W.Y. Ma, "Texture Features for Browsing and Retrieval of Image Data," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(8):837-842, 1996.

[11] M. Schrder, H. Rehrauer, K. Seidel, and M. Datcu, "Interactive learning and probabilistic retrieval in remote sensing image archives, *IEEE Trans. Geosci. and Remote Sens.*, 38(5):22882298, 2000.

[12] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, "Content-Based Image Retrieval at the End of the Early Years," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1349-1380, 2000.

[13] A.H.S. Solberg, T. Taxt, A.K. Jain, "A Markov random field model for classification of multisource satellite imagery," *IEEE Trans. Geosci. and Remote Sens.*, 34(1):100-113, 1996.

[14] B.C.K. Tso, P.M. Mather, "Classification of Multisource Remote Sensing Imagery using a Genetic Algorithm and Markov Random Fields," *IEEE Trans. Geosci. and Remote Sens.*, 37(3):1255-1260, 1999.

[15] U.S. Geological Survey, "Landsat (Sensor: ETM+)," EROS Data Center, Sioux Falls, SD. Available from: `http://glcf.umiacs.umd.edu/data/landsat/`.

[16] J.Z. Wang, J. Li, G. Wiederhold, "SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture LIbraries," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(9):947-963, 2001.

[17] G.G. Wilkinson, "Results and Implications of a Study of Fifteen Years of Satellite Image Classification Experiments," *IEEE Trans. Geosci. and Remote Sens.*, 43(3):433-440, 2005.

[18] J. Zhang, G.M. Foody, "A Fuzzy Classification of Sub-urban Land Cover from Remotely Sensed Imagery," *Int. J. of Remote Sensing*, 19(14):2721-2738, 1998.