Feature Extraction for Image Mining

Patricia G. Foschi
Romberg Tiburon Center for Environmental Studies
San Francisco State University

Deepak Kolippakkam[*], Huan Liu and Amit Mandvikar
Department of Computer Science & Engineering
Arizona State University, Tempe, Arizona, USA
Email: {n.kolippakkam, hliu, amit.mandvikar}@asu.edu

Abstract:

Due to the digitization of data and advances in technology, it has become extremely easy to obtain and store large quantities of data, particularly Multimedia data. Fields ranging from Commercial to Military need to analyze these data in an efficient and fast manner. Presently, tools for mining images are few and require human intervention. Feature selection and extraction is the pre-processing step of Image Mining. Obviously this is a critical step in the entire scenario of Image Mining. Our approach to mine from Images – to extract patterns and derive knowledge from large collections of images, deals mainly with identification and extraction of unique features for a particular domain. Though there are various features available, the aim is to identify the best features and thereby extract relevant information from the images. We have tried various methods for extraction; the features extracted and the techniques used are evaluated for their contribution to solving the problem. Experimental results show that the features used are sufficient to identify the patterns from the Images. The extracted features were evaluated for goodness and tested on test images. An interactive system was developed which allows the user to define new features and to resolve uncertain regions.

1. Introduction

The Computer Industry has seen a large growth in technology – access, storage and processing fields. This combined with the fact that there are a lot of data to be processed has paved the way for analyzing and mining data to derive potentially useful information. Various fields ranging from Commercial to Military want to analyze data in an efficient and fast manner. Particularly in the area of Multimedia data, images have the stronghold. However there is a general agreement that sufficient tools are not available for analysis of images [ZHL01]. One of the issues is the effective identification of features in the images and the other one is extracting them. One of the difficult tasks is knowing the image domain and obtaining a priori knowledge of what information is required from the image. This is one of the reasons the Image Mining process cannot be completely automated.

[*] Contact Author

Current techniques in image retrieval and classification (two of the dominant tasks in Image Mining) concentrates on content-based techniques [RHC99]. Various systems like the QBIC [NB94], RetrievalWare [D93] and PhotoBook [PPS96] etc have a variety of features, but are still used in particular domains. Jain et al [JV96] use color features combined with shape for classification. Ma et al [MDM97] use color and texture for retrieval. Smith and Chang [SC96] use color and the spatial arrangements of these color regions. Since perception is subjective, there is no single feature which is sufficient [RHC99, ZHL001]; and, moreover, a single representation of a feature is also not sufficient. Hence multiple representations and a combination of features are necessary.

This paper is based on the partial results of the on-going research project - Egeria Mining [RTC02, EDM02]. *Egeria densa*, an aquatic weed, has grown uncontrolled in the Sacramento-San Joaquin Delta and is presently causing reservoir-pumping and navigational problems. To monitor the areal extent of Egeria, aerial photographs have been collected, scan-digitized and visually interpreted. Until the present time, the detection process has been carried out manually, which is laborious and time consuming. In addition, the images have been taken during adverse environmental conditions – wind, sun glint, reflections from water, high/low tides, etc. All this makes the detection process very difficult. Our task is to implement an effective methodology to automatically identify these infested regions by means of features and thus use a general technique to process all images. This can be done either at the pixel level [ZHL01, RHC99, CS95] or at a block level, which we use. The idea is to exploit one image (called the training image) and then use the technique on other images. It is impossible to get the same results from the other images as that of the training image. However, the goodness of the features defined initially and the techniques used to extract them, would aid us in the detection process. The output would be a systematic method of feature extraction and an interactive system to support user-defined features that can be employed in Image Mining applications.

The goals of this paper are to discuss the methods used to quickly extract/derive features and to evaluate the efficiency of these features. The paper is organized as follows: Section 2 deals with the approaches used in selecting and extracting the features along with the experimental results. Section 3 comprises the Evaluation of these features. Finally in section 4 we conclude with our findings.

2. Our Approach

As mentioned earlier, the manual detection task is being automated (semi-automated, as there will still be human intervention). The images are 300X300, RGB, TIF (Tag Based Image File) format. They are of high quality in general, but the aerial photographs were taken during varying environmental conditions, causing the images to be of varying quality. This was one of the considerations in selecting a training image. The image used to train the system initially must ideally have all the “occurring” conditions of Egeria. In this way, it is assured that we are in fact “covering” the various cases. Domain knowledge and feedback from image experts also were valuable input for selecting the training image. We have a set of 30 images (including the training image). The rest of the 29 images are testing images. Some images in the set were dark when compared to the others.

In general, images have the following features – color, texture, shape, edge, shadows, temporal details etc. The features that were most promising were color, texture and edge. The reasons are as follows:

1. **Color:** Egeria occurs in 2 colors – pink (rusty rose) and black. Hence the picture elements can be compared to these spectra.
2. **Texture:** Texture is defined as a neighborhood feature [RHC99] – as a region or a block. The variation of each pixel with respect to its neighboring pixels defines texture. In our case, Egeria occurs in open water or in water at the shoreline. Hence the textural details of similar regions can be compared with a texture template.
3. **Edge:** Edge is simply a large change in frequency. This is particularly important here, as the distinction between the dark Egeria and the lighter water bodies or land can be considered as an edge.

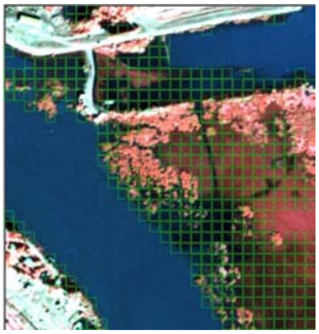
The training image was 1000_2m_lvi2.tif. After the features have been identified, a step-wise procedure extracts the features and combines them using rules that would detect the maximum coverage of Egeria in the image. Two sets of images were provided to us. The first set is the set of 30 images. The second set is the corresponding coverage of Egeria, which was manually detected and was provided for experimental verification. In the individual image each pixel corresponds to a particular intensity value. In order to correctly identify Egeria in the images, we divide the image into blocks of size nXn. We define block as a "block" of pixels – say 10X10 or 8X8 (both were used in the experiments). These sizes were considered after experimenting with various other sizes. If the size is too small, then texture features cannot be described. If it is too large, small patches of Egeria cannot be detected. Block size of 10X10 or 8X8 was an appropriate size both in terms of processing time and accuracy. The training image is divided into blocks and the domain expert makes the parameter adjustments and sets the thresholds only once.

The general procedure, which involves all the automatic feature extraction tasks, is called IClass. For texture features we have templates from the training image with representative properties for that feature. The following are the methods that were tried on this training image.

2.1 Color Feature Extraction:

Some of the techniques tried were – Average color in Gray scale, Average color in RGB format [GW92] and Average color in YCBCR (Y is the luminance and CB, CR are the chrominance components) [GW92]. We evaluated the various methods using Precision and Recall (introduced in the next section which compares the Precision and Recall values of the methods), and found that YCBCR performs better than the other two. Hence we used it as the basis of color extraction as shown in the image below (*1000_2m_lvi2.tif)*.

$$\text{Average color} = \frac{\sum(\text{intensity of all pixels in the current block})}{(\text{total pixels in the block})}$$



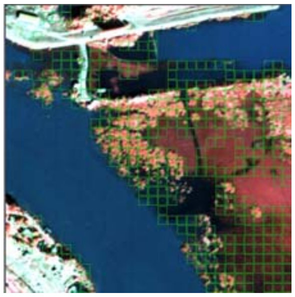
The output of this procedure would be a region matrix, of *30X30* (for 10X10 block or 37X37 for 8X8) size, with '1' in the areas corresponding to the presence of color match and '0' in the areas without color match.

2.2 Texture Features Extraction:

For texture extraction, we chose textures with pure Egeria, Egeria with land and Egeria with land and water. We tried Histograms without bins, with bins [JV96], Normalized Histogram with bins

[JV96] and Discrete Cosine Transform [GW92]. In the Histogram methods, the template is compared with each image block in terms of its histograms. The difference between the individual peaks is taken and the mean squared difference is determined. Each block with relatively smaller difference matches the template and hence can be extracted as part of that texture. If the difference between the template and the current block is smaller than a particular threshold, then that feature is marked YES for that particular block. As in the case of color, we also evaluated the performance of these methods and found that for the training image the Histogram with bins method was the most accurate.

Similarity measure $= \sqrt{\sum \left| \text{(Means of bins for template) - (Means of bins for block)} \right|}$

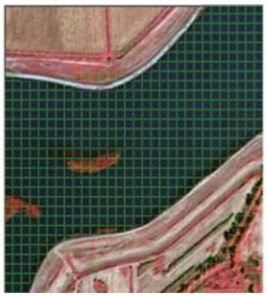


In the end, all three textures were combined (ORed) to obtain the final texture region matrix as shown in the left image (*1000_2m_lvi2.tif)*. The histogram method was time-consuming due to the extensive calculations. Our intent is to develop an interactive system that allows the user to define his/her own features and extract them. The DCT method [CS95, CB02] is a simple and fast method that does not have the hassles of peak values (histograms) and laborious calculations. The DC coefficient (first value in the DCT matrix) is representative of the block. Hence it can be compared with other DC values to provide similarity measure. The goodness of the DCT method is also given in the next section.

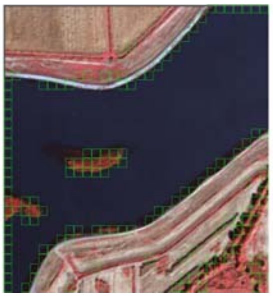
2.3 Edge Feature extraction:

Edge features are particularly important for some of the darker images. Fortunately, the training image was of normal quality and hence we did not use the edge feature. However, we do use it for some of the darker images in the set for testing. The Canny edge detection [MAT02] method with default threshold (0) was used. Edge feature alone has very little efficiency; hence we need to combine it with a stronger feature, like color. It is combined with the color feature to describe the boundaries and inner regions of Egeria.

The following images are the results obtained when a dark image from the set was tested.

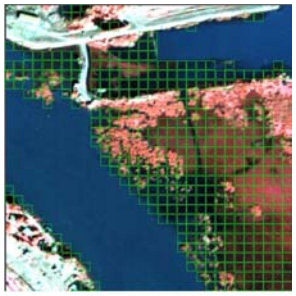


Color extracted image – water body is also covered



Edge extracted image – Egeria alone is extracted

2.4 Combining Features:



All the extracted features are combined to get the final extracted image to the left. Every block has a similarity measure for each of the features. Hence after the feature extraction process each instance (block) is a sequence of 1s (YES) and 0s (NO) of length equal to the number of features extracted. Combining these extracted features is synonymous to forming rules. One rule that combines the three features is *color & edge | textures*, which means color AND edge OR texture.

Color	Textures	Edge	Class
1	0	1	1
0	0	0	0
1	1	0	1
1	0	0	2 (Uncertain)

Depending on the features used and the domain, the rules vary. If a particular feature is very accurate for a domain, then the rule will assign the class label as YES (1) (1 in the table on the left). For those instances when IClass is not certain the class label is 2. This denotes uncertain regions that may or may not be Egeria. The same rule used during the training phase is also used in the testing phase. If there are 3 features, for example, the following table shows a part of a set of rules that could be used. The first and third rules say that color along with texture or edge conclusively determines that Egeria is present in that block. The second rule says that when none of the features is 1 then Egeria is absent for sure. Fourth rule states that color on its own is uncertain in determining the presence of Egeria. An interactive system is used for resolving the uncertain, misclassified and missed regions by showing these to the experts and then recording the actual classes. For the training image we have the “actual” coverage so we can generate the class labels for all the instances from the cover. After the Egeria regions in the training image have been covered comprehensively, a dataset is formed with the feature class label as the columns and instances or blocks as rows.

3. Evaluation of features in Image Mining

All these features were extracted for this particular domain. We used the features extracted from the training image, on the testing images. Hence in essence, we have developed a system, which is trained once and then applies the same technique to other images. The goodness of these features can be judged by certain evaluation criteria. The second set of images (with the manually interpreted Egeria) was used for comparison and validation. The validation images have the same size and hence can be compared with the extracted images in terms of blocks.

For a two-class problem, there can be 4 possible outcomes of a prediction [WF00]. The outcomes are True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN), where TP are those extracted regions that are correct, TN are the regions that are incorrect and are not retrieved, FP are regions that are actually incorrect, but have been extracted (these correspond to false alarms), and FN are regions which were supposed to be extracted but were missed.

Method	Precision	Recall
Color:		
Gray scale	0.7530	0.6349
RGB	0.7668	0.8600
YCbCr	**0.7306**	**0.8927**
Texture:		
Histo bins	**0.6880**	**0.6732**
Norm Histo - bins	0.5851	0.6547
DCT	**0.7465**	**0.5006**
Edge:	0.3220	0.4675
Edge and Color:	**0.5724**	**0.4707**

- *Precision:* Defined as the fraction of the retrieved information, which is relevant.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- *Recall:* Defined as the fraction of the relevant retrieved information versus all relevant information.

$$\text{Recall} = \frac{TP}{TP + FN}$$

In the above table, the methods used (in bold face) are better than their counterparts. Precision and Recall are used below for performance evaluation.

Image Name	Calculation			
	Certain	Uncertain	Precision	Recall
lvi2	406	0	0.8188	0.9010
bb1	338	191	0.9595	0.6614
di1	301	133	0.7181	0.7461
hr1	248	342	0.2956	0.9309
ft1	847	29	0.3778	0.9795
lps1	227	426	0.2032	0.9056
ls1	281	306	0.3180	0.6807
lvi1	150	505	0.2863	0.8909
orh1	209	466	0.1901	0.7571
qi1	235	217	0.6099	0.8471
vc1	213	76	0.8931	0.6383
wi1	305	539	0.5393	0.8186
ds1_7-02	236	107	0.1847	0.7562
7ms_7-02	467	94	0.1727	0.9785
ft1_7-02	471	164	0.5834	0.8381
ft2_7-02	494	82	0.4994	0.7247
ft3_7-02	231	115	0.2707	0.4820
lvi1_7-02	233	168	0.3600	0.4755
ri1_7-02	140	71	0.4494	0.6555
vc1_7-02	191	391	0.1111	0.8192
wdc1_7-02	524	279	0.1297	0.9944
wi1_7-02	436	112	0.5487	0.6450
bb1-7-02	459	78	0.1637	0.4214
bb2-7-02	397	151	0.2699	0.3736
ls1-7-02	344	141	0.2789	0.8558
ps1-7-02	506	48	0.0893	0.8609
sl1-7-02	737	192	0.2967	0.7161
sl2-7-02	218	87	0.5106	0.6695

4. Conclusion

We have presented a tool that copes with the research issues discussed in the previous sections of this paper. The prototype system tries to improve the detection process and also tries to reduce human intervention. High recall values obtained are necessary so that the instances retrieved are comparable to the actual (relevant) instances. We also tried to use this feature extraction process combined with Data Mining algorithms in which the training data are provided to the algorithm and the test data are given without the class labels. The algorithm labels instances from the knowledge obtained from the training data. We found that the tested images required less human interaction for resolving the uncertain regions.

In our future work, we plan to introduce more features and to work with more accurate features like those derived from wavelet transforms. The next prototype will be more automatic by requiring less expert interaction and also is expected to be better in terms of accuracy rates.

References

[CB02] Sharlee Climer, Sanjiv K. Bhatia. *Image Database indexing using JPEG coefficients*. The journal of the Pattern Recognition Society, Pattern Recognition 35 (2002) 2479-2488

[CS95] S. F. Chang and J.R. Smith. *Extracting Multi-Dimensional Signal Features for Content-Based Visual Query*. SPIE Symposium on Visual Communications and Signal Processing, May 1995

[D93] James Dowe. *Content based retrieval in multimedia imaging*. In Proc. SPIE storage and Retrieval for Image and Video Databases, 1993

[EDM02] The Egeria Densa Mining project: www.public.asu.edu/~nkolipp/egeria.html

[GW92] R. Gonzalez and R. Woods. *Digital Image Processing*, Addison-Wesley publications Co, March 1992.

[JV96] A. Jain, A. Vailaya. *Image Retrieval using Color and Shape*. Pattern Recognition, 29(8): 1233-1244, August 1996.

[MAT02] MATLAB Image Processing Toolbox User's guide, Version 3, 1993-2001 by The Mathworks Inc., www.mathworks.com

[MDM97] W. Ma, Y. Deng and B. S. Manjunath. *Tools for texture/color based search of images.* SPIE International conference - Human Vision and Electronic Imaging: 496-507, February 1997.

[NB94] W. Niblack, R. Barber, and et al. *The QBIC project: Querying images by content using color, texture and shape*. In Proc. SPIE Storage and Retrieval for Image and Video Databases, Feb 1994

[RHC99] Y. Rui, T. Huang and S. Chang. *Image retrieval: current techniques, promising directions and open issues.* Journal of Visual Communication and Image Representation, 10(4): 39-62, April 1999.

[RTC02] Romberg Tiburon Center for Environmental Studies. http://romberg.sfsu.edu/~egeria.

[PPS96] A. Pentland, R.W. Picard, and S. Slaroff. Photobook: *Content based manipulation of databases*. Int. J. Comput. Vis., 18(3): 233-254, 1996

[SC96] J. Smith and S. Chang. *VisualSEEK: A fully automated content-based image query system.* ACM Multimedia: 87-98, November 1996.

[WF00] I. Witten and E. Frank. *Data Mining: Practical Machine Learning tools and techniques with Java Implementations*, Morgan Kaufmann publications, 2000.

[ZHL01] Ji Zhang, Wynne Hsu, Mong Li Lee. *An Information-driven Framework for Image Mining*, in Proceedings of 12th International Conference on Database and Expert Systems Applications (DEXA), Munich, Germany, 2001.

[ZHL001] Ji Zhang, Wynne Hsu, Mong Li Lee. *Image Mining: Issues, Frameworks and Technique*s, in Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), San Francisco, CA, USA, 2001.