

Image Mining and Retrieval Using Hierarchical Support Vector Machines

R. Brown¹, B. Pham¹

¹Faculty of Information Technology,
Queensland University of Technology, GPO Box 2434,
Brisbane 4001, Australia.
{b.pham,r.brown}@qut.edu.au

Abstract

For some time now, image retrieval approaches have been developed that use low-level features, such as colour histograms, edge distributions and texture measures. What has been lacking in image retrieval approaches is the development of general methods for more structured object recognition. This paper describes in detail a general hierarchical image classifier approach, and illustrates the ease with which it can be trained to find objects in a scene. To further illustrate the wide capabilities of this approach, results from its application to particle picking in biology and Vietnamese art image retrieval are listed.

Keywords: image mining, image retrieval, support vector machines.

1. Introduction

Many image retrieval systems have successfully used low-level features, such as colour, shape and texture, to aid the process of identifying images close to those chosen by the operator [1]. In addition, user feedback is often used to aid the process of refining the search process [2].

However, there is a need to elevate image mining and image retrieval above the level of simple features, to provide a general approach to identifying objects that can be found in image. While other work has taken steps in this direction, there appears to be no general approach for finding arbitrary objects in a scene, or any easy method for developing such detectors that can be used by the general public [3]. The method described in this paper is a step in this direction.

The original motivation for this work was the detection of inappropriate content in image collections for *Digital Forensic* applications. Digital forensics is the application of computer analysis techniques to determine potential legal evidence of computer crimes or misuse that are caused by unauthorised users or by unauthorised activities generated by authorised users. Much forensic evidence comes in the form of images or videos that contain objects and/or scenes that may be related to criminal behaviours.

Therefore, this application area required an approach which allowed the detection of objects trained from image patches, and also enabled arrangements of component detectors for image mining. This work has been detailed in [4].

From these experiments with Digital Forensics, it was realised that other application areas could benefit from this hierarchical approach. In particular, it was considered that particle picking in biology and art image retrieval would be good candidates, due to the nature of the imagery involved. Particle picking would benefit from the ability to train the classifier for any number of particles, and to then allow it to mine the microscopic images for matching particles. Art image retrieval would benefit due to the stylized manner of the objects in the scene, creating coherent features across many images that could be classified.

The rest of this paper is structured as follows. Firstly, the hierarchical support vector machine classifier training technique is described in detail. Results are shown for two new applications areas: particle picking and art image retrieval. Finally, the paper concludes with a description of future work for this approach.

2. Description of Technique

As with other machine learning pattern classifiers, the process of using the system is divided into two main sections: offline training and classifier execution. This strong separation is drawn from the fact that classifier development and usage are two different processes, as outlined in the operational model in [4]. This dichotomy was enforced by the original goal of integrating the produced software into a digital forensics system. However, the training process is not so difficult as to exclude the general public from forming their own detectors, or a search engine system being configured by a technician to provide set searches for users.

The training phase has three main sub processes to be performed: training a *Support Vector Machine* (SVM) to recognise a component and or sets of components, placing constraints on the organisation of the components and objects within an image, and then exportation of the classifier for an operator to use. SVMs are used as the base classifier, due to their ability to handle high dimensional classification problems, such as the image data used in this application.

The approach is based on the work of Mohan et al. [6], with our own enhancements as detailed in [4]. The next section describes the training tasks in detail and illustrates the process with an application to find faces within images.

2.1. Image Segmentation

Patches of the images which are able to be detected by a trained SVM are segmented into files for further preprocessing to extract features for the classifier. For the operator, this requires the identification of objects that will have invariant image components across a given data set. These features are pieced together under spatial constraints for classification purposes. For the application area of face detection, the skin colour and eye, nose and mouth areas form a consistently shaped region across an ensemble training set of images. Therefore, a training set is assembled from a series of segmented images having these invariants.

2.2. Feature Parameters

Feature parameters are required for the correct use of the system, with regards to the appropriate features to be processed. The non-decimated Haar basis wavelet transformation used by Mohan et al. [5] consists of a normal Haar basis decomposition [6],

without down sampling. Thus, a quadruple density (doubled in x and y) set of coefficients is derived for a particular level of the wavelet hierarchy. This provides more detailed information for the classifier than normal wavelet decompositions. This wavelet decomposition needs to be normalised and thresholded to extract consistent coefficients from the wavelet ensemble.

The normalisation process involves creating a single channel image from the three channel rgb wavelets producing the following ensemble image characteristics. Coefficients that are random in nature across the ensemble set tend to 0.5. The coefficients that are consistently high in value are normalised towards 1.0 in value (white pixels), and the consistently low coefficients are normalised towards zero (black pixels).

For this approach, the use of non-decimated wavelets provides two items of information from the image patches. First, the regions with consistently high coefficients have a high average value across the ensemble of the training set. The second piece of information is the regions with consistently low coefficient levels for the ensemble images. These have a consistent luminance component, due to the lack of edges and so contribute further information to characterise the image patch (refer to Figure 1).



Figure 1 Example ensemble wavelet decompositions are shown for (left) horizontal coefficient image, (middle) horizontal low threshold and (right) horizontal high threshold coefficients.

The selection of features is controlled by parameter settings of upper (for high coefficient) and lower (for low coefficient) thresholds. This is set by manipulation of an interface with results then being viewed interactively. The operator decides the threshold levels based upon the appearance of an identifiable pattern for the patch. In some cases, these coefficients may be ignored and only scaling functions are used. This is left to the discretion of the operator.

From Figure 1, two main features are used, the average colour (all three components) of the low thresholded regions (in this case blobs of skin), and the values of the coefficients of the high thresholded regions, which represent the consistent outline of the human face. Together, these form a feature vector of

the face, which is unique enough for the SVM system to be trained for face classification. This is performed for two levels of the wavelet decomposition.

Each feature vector is made up of two main groups of information: edge coefficients defining the outline of body parts and regions defining areas of continuous tones. The number of entries for each vector depends on the number of coefficients making up the two components of the patch detector. Hence, the feature vector varies for each patch detector generated. Each vector contains two levels of the wavelet pyramid; the block size being predicated on the quality of the ensemble images. Region and edge information is repeated for each level within the final feature vector, as shown in Figure 2.

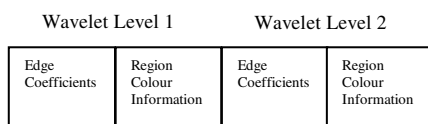


Figure 2 Diagrammatic representation of the feature vector generated for the hierarchical SVM classifier.

The SVM constructs a hyperplane from the vector data, after it has been processed with an appropriate kernel. It maximises the margin between the positive and negative examples, and this margin forms the *support vector*. In this application a *quadratic* kernel is used for the component detectors, with a *linear* kernel is used for the object detectors combined from the lower component detectors. The hyperplane is used to classify between members and non-members of a class, making SVMs useful for binary classification tasks. For further details on SVMs, the reader is referred to [7].

2.3. Detector Constraints

From the SVM and Wavelet approach described, a grammar has been constructed to enable arbitrary scene descriptions to be devised for search queries [4]. This grammar allows the search to be specified as a hierarchy of detectors, working at different structural resolutions. The grammar is made up of component detectors and object detectors and their related spatial relationships and position data.

Object Detectors and *Component Detectors* are an abstraction of the detection mechanisms used to find components. The object detectors themselves are hierarchies of other component detectors and/or other objects. An entire scene description can be used in

another object detection scene, and so on. Thus, the grammar allows the encapsulating system to store the resolution of the search at either a broad structural level (e.g. skin detection) down to fine grain informational detection (e.g. face, pelvis, torso).

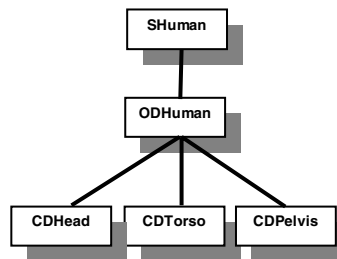


Figure 3 Illustration of tree structure for an example human detector.

Relationships encapsulate relative spatial arrangements between the object detectors within the query description. This is specified in a multi-precision manner: from rough linguistic terms like up, down, above, below; to more precise terms like north, south, east, west; and then to orientation and absolute position specifications. These terms, values and deviations facilitate the complete 2D specification of the arrangements of the detectors at levels of resolution relevant to the search task. Figure 3 illustrates the n-ary tree data structure for an example object detector finding the occurrence of human bodies in images.

As can be seen in the diagram, the image mining query is described as a scene (marked S) to be searched for using a number of object detectors (marked OD) and component detectors (marked CD). Each node contains the relevant information to fully specify the spatial attributes of the object or component detector. It should be noted that an object detector can be a collection of components and other object detectors, giving complete freedom to use predefined scenes as object detectors in other scenes.

All the object/component attributes and constraint specifications have been fully detailed as a formal grammar in another paper [8], and we refer the reader to it for details.

2.4. Training Process

The training process involves the refinement of the system until a reasonable model is developed and requires a suitable training set available to at least initiate the refinement process. This process will involve two major sub processes: the modification of

parameters with regards to features used and the constraint arrangement and training of the component detectors with training sets of positive and negative images. Object detectors, being made up of component detectors, are trained from the raw SVM scores returned by the component detectors. In other words, the raw scores from the training set are processed in a similar manner to the way components process low-level image features, and form the hyperplane used by the SVM system for object classification purposes.

The training process involves iterations of the parameter modification and retraining stages, until a satisfactory level of performance is reached. The classifier is then able to be used routinely for checking image databases for desired contents. Furthermore, any false positives can then be fed back into the previously mentioned query development process to improve the query's capabilities.

The rest of this paper now describes the implementation and results of this technique applied to various image retrieval applications.

3. Results of Further Applications

This approach has been applied to Image Mining for Computer Forensics, for the detection of inappropriate image content in unstructured image databases, and results have been reported in detail in [4]. The success of the approach suggested that other application areas may benefit from such an image classification approach. Hence, this section details the results from experiments with the implemented library in the new areas of particle picking and art image retrieval.

3.1. Particle Picking

Selection of individual particles from biological imagery is a current area of research, and forms the area of *Automatic Particle Selection* [9]. The present state of the art is for the automatic selection of particles, with a further pruning of particle images for selected for 3D reconstructions. The regions have to be chosen carefully, to weed out clumps of multiple particles, leaving good quality particle images.

However, this is not possible when the number of particles required for the reconstruction process grows to the order of one hundred thousand to a million particle images.

Previous methods have used various image registration and processing techniques, with one using

ADABOOST as a machine learning technique [9]. The newly developed SVM library was tested upon donated negatively stained images of Ferritin particles. The training involved a set of 62 positive images and 84 error images, refer to Figure 4. The ensemble images were generated with 8x8 and 16x16 pixel sized wavelets. A default quadratic SVM component detector was trained using a C penalty value of 10.

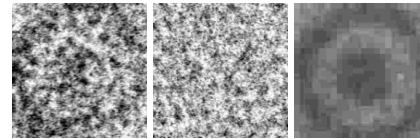


Figure 4 Examples of training images, the left image is an example positive Ferritin particle the middle is an error background image and the right is the scaling function ensemble image.

In this application, the thresholding of the ensemble images did not yield discernable shapes. Therefore, the entire set of coefficients for the scaling function were used to train the Ferritin particle detector, no coefficients from the horizontal, vertical or diagonal images were used in this detector application (refer to Figure 1). Using the scaling function equates to using a low pass filtered image of the ensemble particles. A single Ferritin component detector was trained using these parameters, and tested on three microscope images.

A detection image is thus generated by rendering a white square onto a black image of the same size as the average particle, as shown in Figure 6. The detector yielded a surface of detections across the entire image, which detected clumps of grouped particles. This result was not desirable; therefore the detector result surface was processed as an image using connected components analysis.

The post processing algorithm consisted of the following steps:

1. The image is thresholded to binary levels. Any positive score is considered a potential particle, so the surface image is thresholded at zero;
2. The image is then eroded with a structured disc element of size 50, to open up the gaps in clumps of near particles;
3. The image positive score regions are labeled using a four way connected components analysis;
4. The area and centroid of each labeled region is then calculated.

A histogram of the area of each resultant region was drawn up, showing that a detection area of less than 60,000 pixels in content identifies a single particle. Therefore, the labeled regions are processed in turn. If the region is less than 60,000 pixels in size, then a region of 188x183 pixels is cropped from the image at the centroid drawn from the labeled regions.

Figure 5 shows an example of one of the three images that were tested with the Ferritin component detector.

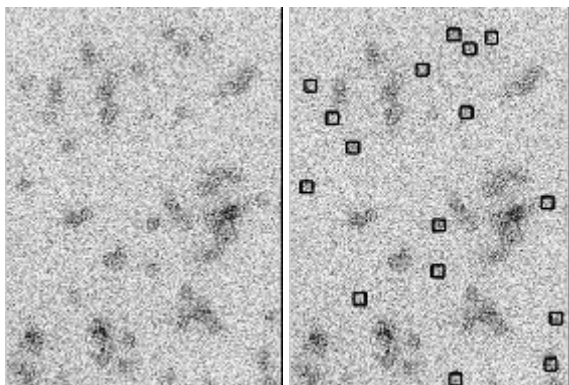


Figure 5 An example image on the left and the detected particles on the right.

There are three questions to consider when analyzing these results:

1. Does the raw detector find all particles?
2. Does the algorithm only accept single particles that are framed appropriately for further analysis?
3. Does the false negative rate run too high?

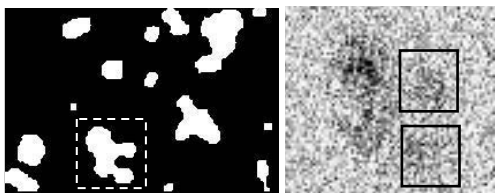


Figure 6 An example of a false negative caused by incorrect erosion results. The dashed square in the top picture is magnified in the bottom image from the original, and shows the incorrect negative classification of the two positive particles.

From the analysis of each image it can be stated that the technique is able to find every instance of a particle, clumped or not, as shown by the white regions in the eroded result image in Figure 6.

Each of the particles cropped from the image, bar very few, are single particles, with the particle being within the region size of the cropped image, and are

thus suitable for further analysis. This is indicated by a low false positive rate.

However, some false negatives are generated, as it is difficult to discern the separate particles from the surface of detection generated by the system, refer to Figure 6. The quantitative results are summarized in a table below in Table 1 for the three images.

Table 1 List of results for particle picking experiments

IMAGE	POS COUNT	NEG COUNT	FALSE POS COUNT	FALSE NEG COUNT	TOTAL COUNT
1	15	14	1	4	34
2	10	12	4	3	30
3	17	9	2	4	26
Overall Particle Count	42	35	7	11	90
Overall %	79%	83%	17%	21%	

From these experiments a number of positives are drawn. Firstly, the results are comparable to the ADABOOST results from previous work, with the median reported positive detections in the 80% range. Secondly, the SVM classifier comes ahead of the previous ADABOOST false positive rate of 23% [9]. Finally, this method is robust against noisy images, as no modeling of noise was required. These preliminary results indicate the ability of this technique in the particle picking application area.

3.2. Art Image Retrieval

In addition, some preliminary results have been generated for art image retrieval, in particular that of Vietnamese art images. This is in connection to a project being run by one of the authors to apply image retrieval techniques to Vietnamese art.

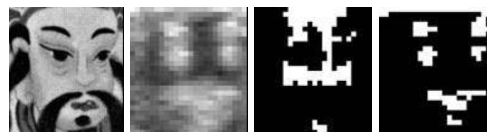


Figure 7 An example training image, horizontal coefficient ensemble and threshold images for the deity head component detector.

From analysis of the test set of images available, it was noted that images of the heads of deities are stylized in a fashion that is amenable to this approach. A test deity head detector was created using the previously described techniques. From a test set of 30 heads and 117 random error images, drawn from

cropped Vietnamese art images, the following ensemble and thresholded images were generated in Figure 7. Figure 8 shows some results from using this detector.

Whilst preliminary, the Vietnamese paintings indicate another application area for the hierarchical technique. The head component detector is a weak classifier, which will benefit from the creation of a set of related hierarchical detectors, to make it more robust. Furthermore, this will have application in a Bayesian Network-based Vietnamese art retrieval system [10]. The deity detector can be used to provide probabilistic evidence of a head being located within a particular scene, that will help further reasoning about the contents and nature of the images being analysed.



Figure 8 The first two images shows examples of correct detections with red rectangles, while the next two are of incorrect detections, including a failure with a valid deity head.

4. Conclusion

This paper has described in detail a hierarchical SVM-based image classification system, including methods of training images and the usage of the technique in classification of image contents. The paper has shown the flexibility of this approach by describing two applications of the technique to automated particle picking in biology and art image retrieval. Both of these application areas, especially the particle picking, show promise for future development.

Future work and improvements for the particle picking include: the refinement of the post processing stage and the introduction of noise cancellation techniques. Further development of the art retrieval area will investigate the use of input features other than undecimated wavelets, and the analysis of other Vietnamese image icons that can be used in an image retrieval system.

Acknowledgements

The authors would like to thank Ben Hankamer and Jasmine Banks from the University of Queensland for

the donation of Ferritin microscopic images used in the experiments.

References

- [1] W. Niblack, X. Zhu, J. Hafner, T. Breuel, D. Ponceleon, D. Petkovic, M. Flickner, E. Upfal, S. Nin, S. Sull, B. Dom, B. Yeo, S. Srinivasan, D. Zivkovic, and M. Penner, "Updates to the QBIC System.," presented at Storage and Retrieval for Image and Video Databases, San Jose, USA, 1997.
- [2] H. Muller, W. Muller, S. Marchand-Maillet, and T. Pun, "Strategies for positive and negative relevance feedback in image retrieval," presented at Proc. International Conference on Pattern Recognition ICPR2000, 2000.
- [3] R. Vetkamp and M. Tanase, "Content-Based Image Retrieval Systems: A Survey," University of Utrecht, Utrecht, Technical Report UU-CS-2000-34, Dec 2000.
- [4] R. Brown, B. Pham, and O. de Vel, "A Grammar for the Specification of Forensic Image Mining Searches," presented at Eighth Australian and New Zealand Intelligent Information Systems Conference, Sydney, Australia, 2003.
- [5] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-Based Object Detection in Images by Components," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 23, pp. 349-361, 2001.
- [6] E. Stollnitz, T. DeRose, and D. Salesin, *Wavelets for Computer Graphics*. San Francisco: Morgan Kaufman, 1996.
- [7] N. Cristianini and J. Shawe-Taylor, *AN INTRODUCTION TO SUPPORT VECTOR MACHINES (and other kernel-based learning methods)*: Cambridge University Press, 2000.
- [8] R. Brown, B. Pham, and O. de Vel, "Design of a Digital Forensics Image Mining System," presented at Submitted to the Eleventh International Multi-Media Modelling Conference, Geelong, Australia, 2004.
- [9] Y. Zhu, B. Carragher, R. Glaeser, D. Fellmann, C. Bajaj, M. Bern, F. Mouche, F. de Haas, R. Hall, D. Kriegman, S. Ludtke, S. Mallick, P. Penczek, A. Roseman, F. Sigworth, N. Volkmann, and C. Potter, "Automatic particle selection: results of a comparative study," *Journal of Structured Biology*, vol. 145, pp. 1-2, 3-14, 2004.
- [10] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. San Mateo: Morgan Kaufmann, 1988.