MCCV BASED IMAGE RETRIEVAL FOR ASTRONOMICAL IMAGES

S. Santhosh Baboo
Department of Computer Science, D.G Vaishnav College, Arumbakkam, Chennai 106. India

P. Subashini
Department of Computer Science, Avinashilingam Deemed University, Coimbatore, India

K. S. Easwarakumar
Department of Computer Science & Engineering, Anna University, Chennai 25, India

Keywords: Astronomical Images, Colour coherence vector, Histogram, MCCV, Feature vectors.

Abstract: Content based image retrieval in astronomy, a technique that uses visual contents to search astronomical images from a large scale image databases according to the users interests, has been an active and fast advancing research area. Early techniques were not generally based on visual features but the textual annotation of images. In other words, images were first annotated with text and then searched using a text-based approach from traditional database management systems. Comprehensive surveys of early text-based image retrieval methods can be found. Text-based image retrieval uses traditional database techniques to manage images. Through text descriptions, images can be organized by topical or semantic hierarchies to facilitate easy navigation and browsing based on standard Boolean queries. The aim of this paper is to review the current state of the art in content-based image retrieval in astronomical images, a technique for retrieving astronomical images on the basis of color distributions. The paper highlights the various retrieval methods like color histogram, color coherence vector and multi scale color coherence vector. The findings are based on both review of the relevant literature and discussions with researchers and practitioners in this field.

1 INTRODUCTION

A database is a repository of data that traditionally contains text, numeric values, Boolean values and dates, known as 'printable' objects (Golshani F and Dimitrova N, 1998). A multimedia database additionally contains graphical images; video clips and sound files, known as 'presentable' objects. Users may retrieve information from a database without having any knowledge of how or where that data is stored.

The information stored within a database must be structured in such a way that the information required can be readily retrieved. A complex multimedia data object, such as a video, was treated as a single data item (Paul Lewis et al., 2002). In terms of data management this object would be treated in exactly the same way as other data items. Queries were usually based on identifying an object from its associated attributes. The deficiencies of this approach for multimedia data objects quickly became apparent and researchers are now developing ways of retrieving multimedia data objects based on their content. In order to handle the enormous volume of data, not only of the big number of images but also the size of each image file, it is necessary to summarize the information. From this perspective, we propose an approach to perform image retrieval that initially generates a representation that is independent of scale and orientation, and then generates a more compact representation, amenable to exhaustive search, using principal component analysis.

Santhosh Baboo S., Subashini P. and S. Easwarakumar K. (2006).
MCCV BASED IMAGE RETRIEVAL FOR ASTRONOMICAL IMAGES.
In *Proceedings of WEBIST 2006 - Second International Conference on Web Information Systems and Technologies - Internet Technology / Web Interface and Applications*, pages 477-480
DOI: 10.5220/0001248504770480

2 DESCRIPTION OF THE APPROACH

The method developed for image retrieval of galaxies is divided in three modules : the Vision module, the feature vector analysis module, and the Retrieval module (engine). The method works as follows: it takes as input the galaxy images, then the Vision module rotates, centers, and crops them; the FVA module finds the feature vectors, and the projection of the images onto the principal components will be the input parameters for the Retrieval module. At the end, the user can supply a query image and the Retrieval module, after processing in the same way the image, compares it against those in the collection, producing as a response the images found that are similar to the query image.

3 MULTISCALE COLOUR COHERENCE VECTOR

The MCCV algorithm allows retrieval of similar images based on the general colour distribution of the image and sub-images, with discrimination between colors in images, which are homogenous to some sizable area (Mohammed F.A. et al., 2002).

The detail finder finds sub-images (Stephen Chan et al., 2001) by dividing the query and the database image into a number of tiles over a number of resolutions, for each tile a CCV is created and stored, so that the final feature vector is a set of CCV feature vectors, one for each tile.

The multiscale ccv allows sub-image finding based on the general colour distribution of tiles within the images. This is a very basic match, which allows detection of a query image within a database image at any scale.

The highest resolution image is converted into 64x64 tiles, overlapping by 32 pixels in each dimension. The image resolution is halved and, again, divided into 64x64 tiles (of which there will be 4 times less). The lowest resolution is of 64x64 pixels and 1 single tile. For each tile a RGB histogram is created and stored, so that the final feature vector is a set of RGB histogram feature vectors, one for each tile.

Both the query image and the database image are converted into a pyramid structure, and then each of the tiles in the query image are compared against each of the features for the tiles in the database image using the RGB histogram matching algorithm (Mike Westmacott et al., 2002).

Figure 1: Multiscale pyramid structure of an image.

The query is converted to a pyramid to facilitate the database image being a sub-image (Mohammed F.A. et al., 2002) of the query image (double sub-image detection). An alternative is to assume the query is a sub-image of the database image only, and perform a RGB histogram match of the whole query image against each of the tiles in the database image (Fazly. S. Abas and Kirk Martinez, 2002). To achieve a speed up in the matching, the features for each of the tiles are compressed by an index (David Dupplaw et al., 1999)

Rather than storing every single bin in each histogram, only those bins that are populated are stored. This can reduce the storage requirements considerably. However, it introduces a matching problem, because the populated bins may vary between features, disallowing direct comparison. It would be possible to 'unpack' the compressed features and then compare those features, however, it is possible using an algorithm developed by Stephen Chan, to compare the compressed features while they are still compressed. A histogram's populated bins would be stored as {bin_number:frequency}; for example, {0:10, 3:4, 6:12} would represent a histogram with 3 populated bins. A second histogram has a compressed feature {0:1, 2:6, 8:4}. The matching algorithm steps

through these features accumulating a score as though the feature was uncompressed. These two histograms would be matched as follows:

- Compare first two values: {0:10}, {0:1}.
  - Indexes match, so score becomes sum of absolute difference: score += abs (10-1) = 9
- Compare second values: {3:4}, {2:6}
  - Increment score by value of smaller index: score += 6
- Compare next value with the same unused value: {3:4}, {8:4}
  - Increment score by value of smaller index: score += 4
- Compare next value with the same unused value: {6:12}, {8:4}
  - Increment score by value of smaller index: score += 12
- No value pairs left, so we add the left over value on:
  - Increment score by value of final index: score += 4
- Final score is: 35.

An example of the mutliscale matching is shown. The image on the left is the query image which is a sub-image of an image within the database.

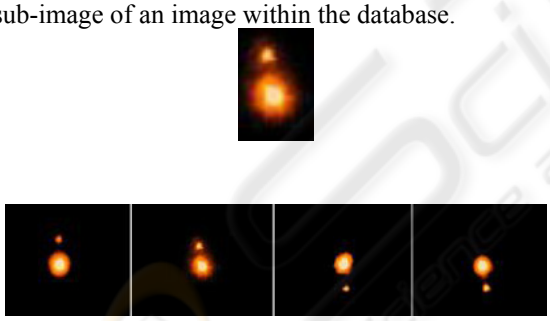


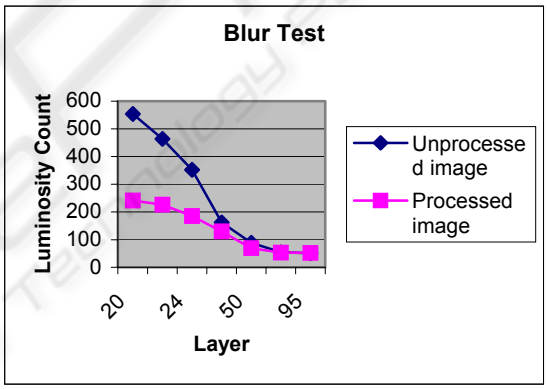
Figure 2: An example multiscalar match.

4 EXPERIMENTAL RESULTS

We tested the system using a data set that consisted images of galaxies. It was taken from the NAOA catalog on the web page of the Astronomical images (http://www.noao.edu/image_gallery). We processed the images. Figure 3 shows examples of images from the original data set and the resulting images output by the vision module.

In order to assess the effectiveness of the approach, we used leave-one-out cross validation, testing the output of the system with one image, and training with the rest, and repeating the process, until all images have been used once as test image. Using this approach, the system output as best match a galaxy belonging to the same class as the query image 89% of the time. This error rate is smaller to those reported in the literature using automated means for image-based galaxy classification (M.C. Storrie-Lombardi et al., 1992) and (A.Adams and A.Woolley, 1994). The best results were obtained when using an elliptical galaxy as query image, very likely because they present the most regular structure, while the worst results were obtained for irregular galaxies, which have very little discernible structure.

We tested the images for noise and blur also. We noticed that the luminosity of the image before and after processing is doing well. Chart 1 says the blur test and chart 2 says the noise test.



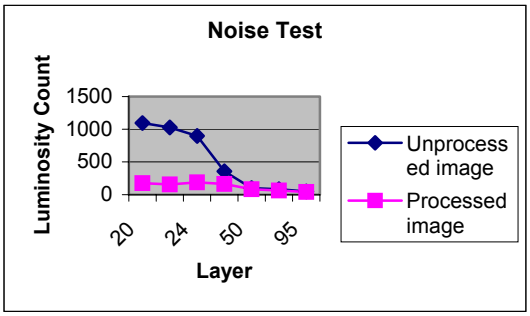


Chart one shows that the luminosity count for processed image is less than that of the unprocessed image when we applied blurring over it, which indicates that processed image is resistive to the blurring.

Chart two shows that the luminosity count for processed image is less than that of the unprocessed image when we applied noise over it, which

indicates that processed image is resistive to the noise.

We tested the efficiency of the algorithm given over the images which we applied our processing and tested. The accuracy of both the retrievals is depicted in the chart. We checked with 200 processed and tested images. When compared with the histogram, MCCV gives more accurate retrieval.

5 CONCLUSION

We have presented a system that performs content-based retrieval of astronomical images. The system executes the following steps to perform image retrieval: 1. Use computer vision techniques to find the location, orientation and size of the galaxy in the image. 2. Rotate, crop and resize the images so that in all the images are the same size, the galaxy is at the center of the image, has horizontal orientation and covers the whole image. 3. Find the feature vectors of the images and project the images. Given a query image, process it as in steps 1 and 2, project it and retrieve the n images with the smallest distance. Quantitative results show that almost 90% of the time the image deemed by the system as most similar to query belongs to the same class, and qualitative results show that the set of images retrieved by the system are visually similar to the query image. Some directions of future work include:

- Extending the experiments to a larger database of galactic images
- Efficient search methods using R-trees (Guttman S, 1984)
- Building classifiers for other types of astronomical objects, such as nebulas and clusters.
- Extending the system to deal with wide-field images, containing multiple objects. This will be done by means of a preprocessing stage to segment the objects in the images, and then processing them individually.

In summary, the global CCV is superior to the basic colour histogram in most of the tests. However, the histogram has the advantage of rapid generation and comparison. Although processing power is less of an issue the memory requirements for generating CCVs can sometimes be a limiting factor when compared to histograms.

This global CCV and histogram results translates well to the Multiscale versions of the algorithms. MCCV generally fared better than MHistogram. The monochrome histogram provides good retrieval rate in both single, and multi scale versions. However, it is rather sensitive to noise.

REFERENCES

Golshani F and Dimitrova N A Language for content based Video Retrieval, Multimedia tools and applications 6, 1998 pp 289-312.

NOAO: Image Gallery on the web page of http://www.noao.edu/image_gallery

M.C. Storrie-Lombardi, O. Lahav, L. Sodre and L.J. Storrie-Lombardi, Morphological Classification of Galaxies by Artificial Neural Networks, Monthly Notices of the Royal Astronomical Society 259, 1992.

A.Adams and A.Woolley, Hubble Classification of Galaxies Using Neural Networks, Vistas in Astronomy 38, 1994.

David Dupplaw, Paul Lewis, Mark Dobie, Spatial Colour Matching for Content Based Retrieval and Navigation, The Challenge of Image Retrieval, February 1999, Newcastle, UK.

Guttman S (1984) In: Proceedings of the SIGMOD Conference, ACM, pp. 47–57.

Stephen Chan and Kirk Martinez and Paul Lewis and C.Lahanier and J.Stevenson, Handling Sub-Image Queries in Content-Based Retrieval of High Resolution Art Images, International Cultural Heritage Informatics Meeting (ICHIM01),pp. 157-163 (2001).

Paul Lewis, David Dupplaw and Kirk Martinez, Content-Based Multimedia Information Handling: Should we Stick to Metadata?, Cultivate Interactive Issue 6, February 2002,
http://www.cultivate-int.org/issue6/retrieval/

Mike Westmacott, Paul Lewis and Kirk Martinez, Using Colour Pair Patches for Image Retrieval, Proceedings of the First European Conference on Colour in Graphics, Imaging and Vision, 245-248, June 2002.

Fazly. S. Abas, and Kirk Martinez, Craquelure Analysis for Content-Based Retrieval, Proceedings, of the 14th International Conference on Digital Signal Processing, July 2002.

Mohammed F.A. Fauzi and Paul Lewis, Query by Fax for Content-Based Image Retrieval, Proceedings of International Conference CIVR 2002, London, July 2002, pages 91-99.