



Content-Based Image Retrieval in Astronomy

A. CSILLAGHY

Space Sciences Laboratory, University of California, Berkeley CA 94720-7450

csillag@ssl.berkeley.edu

H. HINTERBERGER

Institute of Scientific Computing, ETH-Zentrum, CH-8092 Zurich, Switzerland

hinterberger@inf.ethz.ch

A.O. BENZ

Institute of Astronomy, ETH-Zentrum, CH-8092 Zurich, Switzerland

benz@astro.phys.ethz.ch

Received August 26, 1999; Revised August 26, 1999; Accepted June 12, 2000

Abstract. Content-based image retrieval in astronomy needs methods that can deal with an image content made of noisy and diffuse structures. This motivates investigations on how information should be summarized and indexed for this specific kind of images. The method we present first summarizes the image information content by partitioning the image in regions with same texture. We call this process texture summarization. Second, indexing features are generated by examining the distribution of parameters describing image regions. Indexing features can be associated with global or local image characteristics. Both kinds of indexing features are evaluated on the retrieval system of the Zurich archive of solar radio spectrograms. The evaluation shows that generating local indexing features using self-organizing maps yields the best effectiveness of all tested methods.

Keywords: astronomy, image archives, self-organizing maps, image feature indexing

1. Overview

Mining the contents of large image archives is challenging. The number of archived images constantly increases. Therefore, for a given application, finding relevant images gets increasingly problematic. Single relevant images are “buried” in a desert of irrelevant images. For example, in astronomical image archives, hundreds of thousands of images may be available, but searching for a given class of astronomical objects (e.g. find all images containing elliptical galaxies) remains a non-trivial task.

Search engines used by retrieval systems to find quickly relevant information must be efficient and effective. Ideally, an efficient retrieval system will respond quickly to any query. An effective retrieval system, on the other hand, will respond by returning complete information relevant only to that query. Unfortunately, efficiency competes with effectiveness. In an efficient system, relatively few computations are performed to maximize speed, hence only few tests can be made to determine the information relevance. By contrast, in an effective system, many tests are performed at the expense of speed. An inherently application-specific balance between speed and effectiveness must be found.

The optimum retrieval system is reached by summarizing the image information, and by performing search operations on this summarized information. Methods to summarize—

and eventually index—the image information have been first designed for conventional photographic images, such as press photographs, museum catalogs etc. Methods have used text association (Murtagh 1994), color histograms (Flickner et al. 1995), low-level image properties (Gupta and Jain 1997) or texture description (Carson et al. 1997). Astronomy, however, requires different indexing methods: astronomical images often contain diffuse and noisy features that cannot be handled by the same procedures (Csillaghy 1997c).

We present a method to summarize information from images containing noisy and diffuse structures. This method, called texture summarization, first partitions images into regions of similar texture. Second, it creates a list of relevant image regions. Third, it derives indexing information, i.e. indexing features, from this list. Summarizing an image into a list of relevant regions is attractive, because this list can be used to both construct a quicklook image and to define various kinds of indexing features that may eventually be used to search for similar images.

2. Field of application

The investigations presented below have been applied to the Zurich archive of solar radio spectrograms (www.astro.phys.ethz.ch/rapp/). Solar radio spectrograms show the radio spectrum of the sun (figure 1). Even though the information they contain is not spatial, they are visualized as images, showing the intensity of the radio emission as a function of time and frequency. The x -axis of the image is associated with the time of observation, and the y -axis is associated with the frequency. Frequencies increase downward because they are inversely proportional to the solar altitude, thus high solar altitudes are represented at the top of the image.

The Zurich archive contains spectrograms recorded by three solar radio spectrometers called *Ikarus* (Perrenoud 1981), *Phoenix* (Benz et al. 1991) and *Phoenix-2* (Messmer 1999). These instruments observed the radio spectrum between 0.1 and 4 GHz during two decades. The archive contains about 50,000 images. Once calibrated, images are written in a

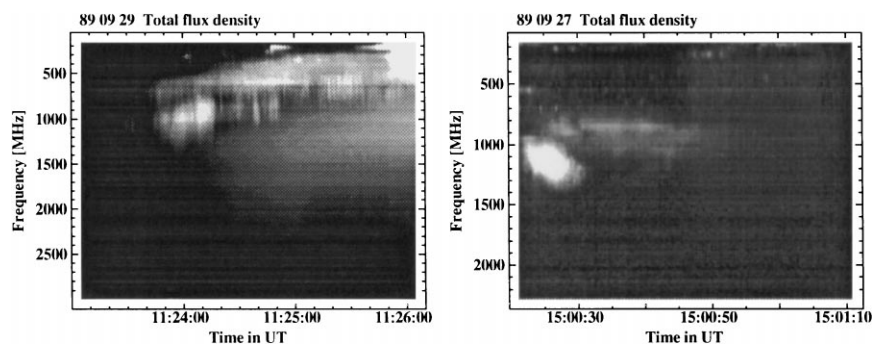


Figure 1. Two spectrograms of solar radio bursts. Left: a broadband type IV event, drifting to lower frequencies, recorded on September 29, 1989. Right: two long-duration diffuse continua (patches), slowly drifting to higher frequencies, recorded on September 27, 1989.

spectrogram-specific FITS files (Wells et al. 1981, Csillaghy 1997c). The FITS (Flexible Image Transport System) file format is widely used in astronomy (fits.gsfc.nasa.gov). A FITS file contains an ASCII header that describes the type of data it stores. Therefore, any data set can be reconstructed from the (generic) FITS definition and the specific header information.

The information content of solar radio spectrograms has the same characteristics as many other astronomical observations in image form. Structures are diffuse and noisy. Moreover, no assumptions can be made about the image information content. In addition, for solar radio spectrograms, the image size varies widely—the size depends on the observation duration. These characteristics make solar radio spectrograms well suited for a case study.

To retrieve spectrograms, queries such as “give me the list of all narrow-band type III bursts” or “which spectrograms in an archive are similar to a given spectrogram?” are typical. These types of queries involve engines searching for similar images. Multi-mission data archives, in preparation for the next solar maximum in 2001, will require such query engines to make the best use of the information they contain (see solar.physics.montana.edu/max-millennium/).

3. Summarization

3.1. *The texture summarization approach*

How do people summarize image information? For images with sharp structures, one may draw a sketch showing the edges of these structures. For images with diffuse and noisy structures, however, it may be more appropriate to distinguish roughly between regions that have different textures. Each region describing a specific texture can be described approximately by the edges that delimit the region, and by a single value characterizing the degree of “roughness” of the region. For example, image regions can be approximated by rectangles, as shown in figure 2. In this case, image regions are described by 5 parameters: the location of the center of the region (2 attributes), its “roughness” (1 attribute) and its extensions (2 attributes).

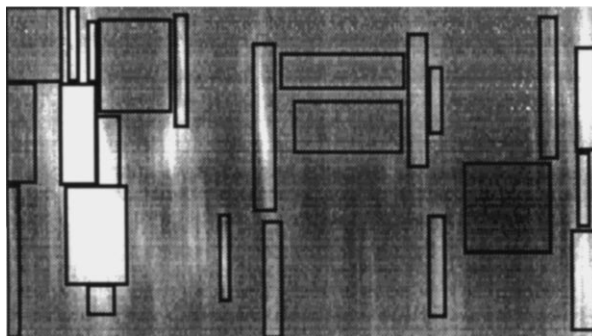


Figure 2. Texture summarization “by hand.” The rectangles delimit regions with analogous textures.

The texture summarization approach has the following characteristics:

- Large (lossy) compression. Each region is associated with a single “roughness” value, thus the noise is eliminated. Even represented as bitmap, image regions are well compressed in formats such as GIF, as they are represented by many pixels of equal values.
- Constant output size. The image summary can be described by a constant number of regions.
- Controlled information content. Irrelevant image regions can be eliminated from the list of regions if adequate assumptions on their shape can be done.
- Highlighted visualization. The information content can be enhanced for visualization purposes by enlarging specific regions in the quicklook image (this can be compared as “bold-facing” in texts).

3.2. Implementation

The texture summarization of the image can be done automatically using spatial data structures. These multidimensional storage structures are employed in many applications (Samet 1990). Many are available, including the R-Tree (Guttman 1984), and the Gridfile (Nievergelt et al. 1984). Although originally developed for fast data access, they also support the implementation of texture summarization, because the fast data access mechanisms rely on “summaries” of the original data. If adequately managed, these summaries can be made equivalent to the regions mentioned above.

The texture summarization method described here has been implemented using a Gridfile. The Gridfile is a multi-dimensional data structure that uses a *grid directory* to allow fast range queries. The grid directory is a partition of the input space. Therefore, each grid corresponds to a region of the input space. By associating the image with the input space of the gridfile, the grid directory yields the image partition wanted.

The algorithm used to implement the texture summarization with a gridfile is described by Csillaghy (1997a). In short, it involves the following steps (figure 3):

1. *Image insertion*: Image pixels are parameterized, i.e. they are represented as three-dimensional points (x, y location and color), and inserted into a three-dimensional gridfile. The grid directory yields the image partition.
2. *Region transformation*: The regions of the grid directory are transformed to eliminate regions of empty space. Grid regions are adapted to the distribution of data they really describe. The region center and extension are set to the average and standard deviation of the parameterized pixels contained in the region, respectively. The region “roughness” is translated into a (more or less arbitrary) color, by averaging the pixel values in the region.
3. *Region selection*: Given a specific application, not all regions are relevant. For instance in an astronomical image, the background may be insignificant when considering summary data. Relevant regions can be selected by specifying a number of assumptions on their shapes. Through an adequate selection, the number of boxes describing the image content can be significantly reduced.

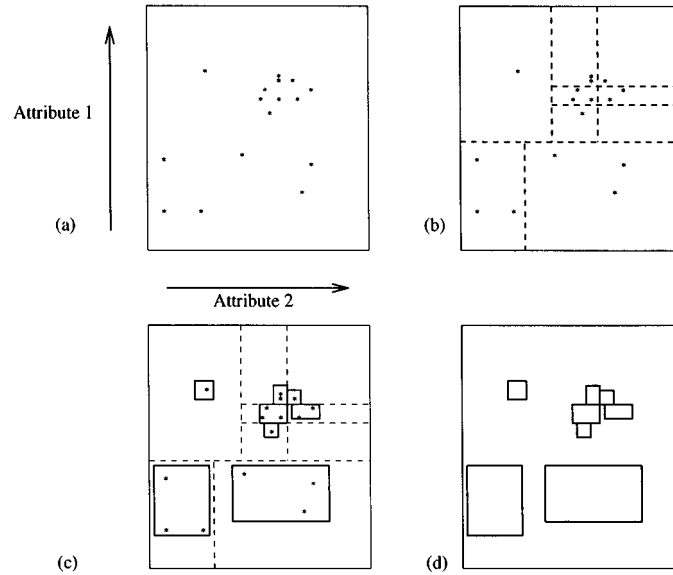


Figure 3. The texture-summarization method applied to a set of two-dimensional points. (a) Data points are considered in a multi-dimensional space. (b) The multi-dimensional space is partitioned into regions containing less than a given number of points, here 3. (c) Each region extension is adjusted to the data distribution. (d) Adjusted regions summarize the original data distribution.

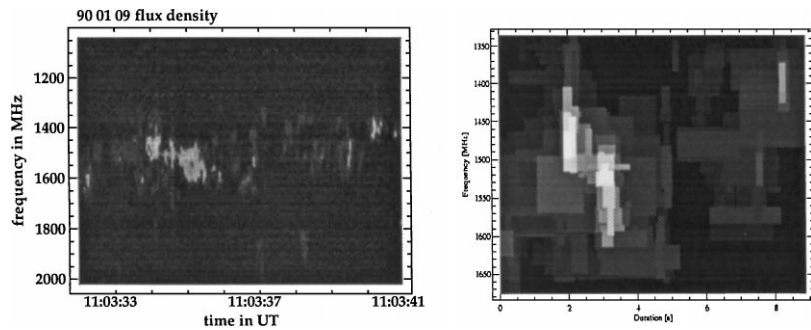


Figure 4. Comparison between an image (left) and the corresponding image icon generated by a Gridfile (right). The image icon highlights the image content and zooms into the relevant part of the original image (note the different axis scaling).

4. *Visualization:* Selected regions can be visualized by drawing their coordinates into a reference system. This visualization can be used to generate image icons stored in a compressed bitmap format such as GIF.

Figure 4 shows an image icon produced by texture summarization. The icon highlights the most relevant parts of the image. This allows getting a quick impression about the

image content—even though users may need some time to “learn” interpreting the way information is represented in the image icon.

4. Feature indexing

The goal of the texture summarization is not only to produce image icons, but also to generate indexing features. Indexing features associated with an image are represented as elements of a description vector. Description vectors represent images in a space of relatively low dimensionality (10–1000). They allow to search rapidly for similar documents among a large number of images.

Indexing features can be derived from the image summary (icon) generated by the texture summarization. An image summary contains a given number of regions. The distribution of the parameters describing these regions is related to the information contained in the original image. We observe that, for a given image, the information tends to be in regions occurring relatively rarely, while the most frequently occurring regions usually do not contain relevant information. This observation leads to an imaging analogous of the inverse document frequency used for texts. Therefore, the distribution of parameters associated with image regions can generate indexing features. This approach, can be described by “summarizing the summaries.” It is attractive for the following reasons:

- Indexing features are produced on the basis of already summarized data. Therefore, additional access to the (large size) original images is avoided.
- The information content of the image icons is represented by simple structures, i.e. a list of regions. The definition of indexing features is therefore relatively straightforward.
- The information content of the image icons is already selected for a specific application, since regions have been selected depending on their shape. Thus, the values of indexing features are not biased by irrelevant data.

Indexing features derived from icons may either describe global or local image characteristics. Global indexing features are associated with the former, and local indexing features are associated with the latter. Obviously, the production of local indexing features is slower than the production of global indexing features. But local indexing features are more accurate.

4.1. Global indexing features

Global indexing features can be derived from icons using multi-dimensional histograms describing the distribution of image-region parameters. Each bin of the histogram is considered as a single indexing feature. The number of indexing features generated can be set by choosing an adequate histogram bin size. Histograms with variable bin size may also be considered.

4.2. Neural-network generated indexing features

Histogram-based indexing features describe images only globally, because they uniformly divide the 5-dimensional “region” space, i.e. the space spanned by the regions of the

image summary. However, the region distribution in this 5-dimensional space may be fairly complex. Therefore, to describe local characteristics of the images, more sophisticated ways must be used to determine the decision (hyper-)surface dividing the region space in classes of regions with same shape. This cannot be done analytically. Thus, we consider a method where the actual distribution is “learned.” In the following, we present results on investigations using a self-organizing map, a non-linear classifier, to determine non-uniform regions of the space that correspond to a specific region class.

4.2.1. Principle. Self-organizing maps (Kohonen 1989) are a particular class of neural networks (Hertz et al. 1991). They associate a cell of a two-dimensional map with a region of the (multi-dimensional) input data space. The association between the (discrete) two-dimensional map and the input space is determined from the actual data-distribution by a learning algorithm.

The SOM can generate local indexing features from image icons. The region space is considered as the input space of the SOM. A sample of (randomly-chosen) regions, selected from all regions available (i.e. independent from which image they are associated with) is used by the map to learn the point distribution in the region space. Once trained, regions associated with a specific image are analyzed by the self-organizing map. Each region is associated with a reaction of a single cell of the map. The reactions are summed into a “total map” that shows all reactions associated with a specific image. The indexing features are then defined as the cells of the map, and their values correspond to the number of times a cell reacted.

The production of indexing features with self-organizing maps is attractive for the following reasons:

- The SOM classifies the image regions depending on their shape and color.
- The SOM uses the actual data distribution to determine the classification.
- The learning of the distribution in the region space by the SOM can rely on a large number of regions.

The SOM package used in this work has been developed by the group of Kohonen (SOM 1995). A map of a dimension of 30×30 has been used. The tuning parameters of the SOM have been determined experimentally (Csillaghy 1997a).

4.2.2. Maps for a single type of radio emission. Figure 5 compares maps associated with three images of the same burst type. These images belong to the class of “type III bursts.” The map reactions are located in specific regions (arrows). Other regions do not react at all. The maps present the following characteristics:

- There are two main domains where reactions are recorded. The first domain is around (15, 15). The second domain is at (20–25, 5–10) and is more spread.
- In some domains, no reactions at all were registered.
- In some domains of the map, a relatively small number of reactions were registered, but for each map at different locations.

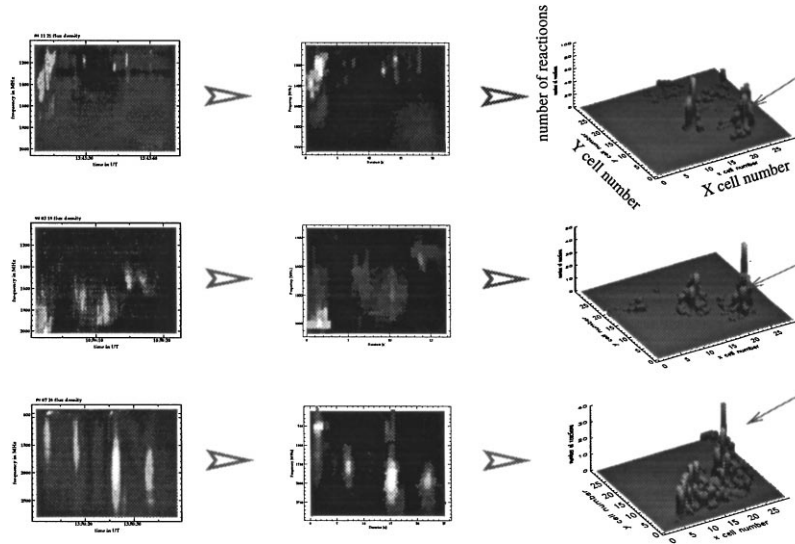


Figure 5. Left: Spectrograms of a single type of radio burst (type III). Center: Image icons associated with the images. Right: Self-organizing maps. The arrows point to domains that appear to be characteristic of the burst type.

4.2.3. Map for different types of radio emission. Figure 6 compares maps associated with three images of different burst types. The maps present the following characteristics:

- There is almost no correlation between the type-III burst map and the type-IV burst map. This corresponds to expectations: type-III and type-IV bursts correspond to signatures of different processes.
- Between the type-III burst map and the millisecond-spikes map, only a few cells have reactions in common.
- Between the type-IV burst map and the millisecond-spikes map, there is a correlation in the region (10–20, 25–30). For type IV however, a whole domain of the map has reacted with no correlation with the other maps.

5. Similarity searching

Similarity searching can be implemented by taking image description vectors (defined as an ordered set of indexing features associated with an image), and by computing the distance between all description vectors in the archive. The degree of similarity is then proportional to the inverse of the distance.

Consider the space spanned by the description vectors, called the *description vector space*. Consider also two document description vectors, \vec{d}_j and \vec{d}_k in the description vector

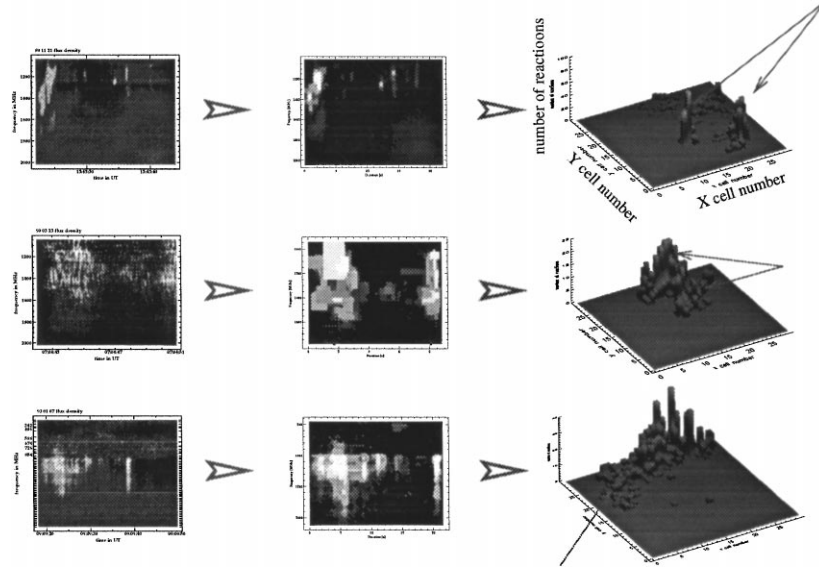


Figure 6. Spectrograms of different types of radio bursts: Upper row: Type III bursts. Middle row: Millisecond radio spikes. Lower row: Type IV bursts. The major reaction areas are in different locations of the map (arrows).

space. Their distance can be measured, for instance, using the Euclidean norm:

$$\rho_E(\vec{d}_j, \vec{d}_k) = \|\vec{d}_j - \vec{d}_k\| = \left(\sum_{i=0}^{m_q-1} (\vec{d}_j - \vec{d}_k)^2 \right)^{\frac{1}{2}}. \quad (1)$$

However, ρ_E is not necessary the best measure of the degree of similarity. It fails, for instance, if similar documents are aligned in a specific direction of the description vector space. In this case, the distance function should take into account the non-isotropic distribution of points in the description vector space.

For this purpose, another function often used (mainly in the context of textual information retrieval) is the direction cosine:

$$\cos(\theta_{j,k}) = \frac{\vec{d}_j \cdot \vec{d}_k}{\|\vec{d}_j\| \|\vec{d}_k\|}. \quad (2)$$

$\cos(\theta_{j,k})$ assumes a linear dependence between similar document description vectors. The case $\cos(\theta_{j,k}) = 1$ represents the highest correlation between \vec{d}_j and \vec{d}_k , and therefore corresponds to the highest similarity.

5.1. Retrieval effectiveness

A search engine using the indexing features described above is evaluated by considering its capacity to effectively retrieve information relevant to a user. We compute the recall and precision (van Rijsbergen 1979) for the search engine implemented on the ASPECT system, the retrieval system of the Zurich archive (Csillaghy and Benz 1999). We then plot the recall and precision values in a recall and precision graph (Frei et al. 1991, Raghavan et al. 1989, Schäuble 1997).

The recall and precision graph for ASPECT is computed as follows. 32 reference ('query') images are selected from a catalogue of solar radio bursts with 437 images by Isliker and Benz (1996). Solar radio bursts are divided into several *main types* and *subtypes* (see Table 1). The classification processed by the retrieval system is compared with a classification that have been done by hand for the catalogue. In this way, it is possible to decide whether the image selected is relevant.

Two reference images are selected for each subtype. For these images, a search for similarity is started. The SOM-based response of the system is compared with a global indexing feature-based response, where description vectors are made of 5 indexing features, as described by Csillaghy and Benz (1999). The similarity between the reference and the other images is computed using the cosine function given in Eq. 2. The resulting recall and precision graph is shown in figure 7.

Table 1. The classification of radio bursts in types and subtypes used for the evaluation of the ASPECT system. A test collection \mathcal{D} of 437 images is used. The number of images per main type and sub-type classes are given by $|\mathcal{D}_{\text{main}}^{\text{rel}}|$ and by $|\mathcal{D}_{\text{sub}}^{\text{rel}}|$, respectively.

Main type	$ \mathcal{D}_{\text{main}}^{\text{rel}} $	Subtype	$ \mathcal{D}_{\text{sub}}^{\text{rel}} $
III	265	narrowband	114
		broadband	73
		large group (>5 bursts)	26
		small group (<5 bursts)	21
		reverse drift	31
IV	56	modulated	15
		fibers	33
		zebras	8
Blips	32	III-like	11
		patchy	21
Pulsations	58		58
Patches	39	cloudy	14
		cigar	7
		caterpillars	5
		large spots	13
Spikes	50		50

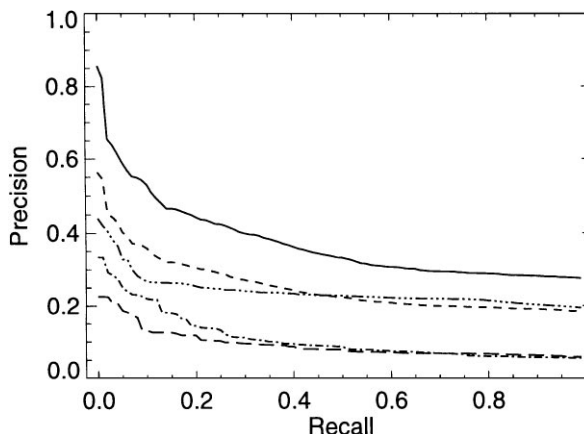


Figure 7. The recall and precision graph for the search engine tested, using the cosine function. Solid line: retrieval of type III and type IV bursts only. Dashed line: retrieval of all main types. -.-: retrieval of all subtypes. -...-...: retrieval of main types with global indexing features. — — —: retrieval of subtypes with global indexing features.

The graph shows that the precision is better for SOM-generated indexing features than for global indexing features. For low recalls, the precision is high: when considering the classes of type III bursts and type IV bursts, a precision above 50% for recalls lower than 10% is reached. Unfortunately, the precision breaks down if classes with only few elements are considered. Furthermore, sub-class retrieval precision is also much lower. Nevertheless, the SOM-generated indexing features lead to a better precision than global indexing features.

6. Conclusion

We discussed methods to summarize and index features from noisy and diffuses image structures. The texture summarization approach partitions an image into regions that contain a single kind of texture. Each region is fully described by 5 parameters, namely, the region extensions, location and a texture “roughness” value. Indexing features are produced by quantization of such regions. Regions can also be used as quicklook data. For the implementation of the image partition, a Gridfile has been used. This was successful for some situations. Nevertheless, other approaches to texture summarization should also be considered in future studies. For instance, wavelet coefficients could describe regions more precisely than ‘grid’ regions

The generation of local indexing features is needed to achieve acceptable retrieval effectiveness. This requires more complex operations than the generation of global indexing features—which just uses a histogram function to quantize the shape of the regions. A method based on a self-organizing map has been tested here. This method has shown that it can help finding relevant images in archives. The results show that self-organizing

maps allow a more effective retrieval than global indexing features. They require more computing time, however, and thus currently their use is not sufficiently efficient.

The approach to indexing features using two phases—first texture summarization and then indexing feature generation—allows to work with already summarized data for the production of indexing features. This accelerates the indexing process significantly, because summarized data is already available. A two-phase indexing thus generally has attractive aspects important for Terabyte-sized archives.

The method presented is neither limited to solar radio spectrograms, nor to astronomical images. Our scheme for retrieving images can be applied as well to any kind of images that have noisy and diffuse structures.

Acknowledgments

The authors acknowledge helpful discussions with P. Schäuble.

References

- ASPECT (1996) The Ikarus/Phoenix Image Retrieval System ASPECT, ETH Zurich. www.astro.phys.ethz.ch/rapp/
- Benz AO, Güdel M, Isliker H, Miskowicz S and Stehling W (1991) Solar Physics, 133:385. www.astro.phys.ethz.ch/papers/benz/benz_p_nf.html
- Carson C, Belongie S, Greenspan H and Malik J (1997) Proc. Workshop on Content-Based Access of Image and Video Libraries, IEEE, 4:42.
- Csillaghy A (1996) Vistas in Astronomy, 40:503. www.astro.phys.ethz.ch/papers/csillaghy/nice.ps
- Csillaghy A (1997a) In: Proc. of the 5th Int. Workshop on Data Analysis in Astronomy. www.astro.phys.ethz.ch/papers/csillaghy/erice96.ps.gz
- Csillaghy A (1997b) PhD Thesis, Swiss Federal Institute of Technology, Zurich. www.shaker.de/Online-Gesamtkatalog/Details.idc?ISBN=3-8265-3131-0
- Csillaghy A (1997c) In: Proc. of the Final CCMA Conference. www.astro.phys.ethz.ch/papers/csillaghy/munich/paper.html
- Csillaghy A and Benz AO (1999) Sol. Phys., 188:203. hessi.ssl.berkeley.edu/csillag/papers/sp99.ps
- Flickner M, Sawhney H, Niblack W, Ashley J, Huang Q, Dom B, Gorkani M, Hafner J, Lee D, Patrovic D, Steele D and Yanker P (1995) Computer, 28(9):2.
- Frei H, Meienberg S and Schäuble P (1991) In: Fuhr N Ed., Workshop on information retrieval. Informatik-Fachberichte. Springer, Berlin. Vol. 289, pp. 1–10.
- Gupta A and Jain R (1997) Comm. ACM, 40(5):71.
- Guttman S (1984) In: Proceedings of the SIGMOD Conference, ACM, pp. 47–57.
- Hertz J, Palmer RG and Krogh AS (1991) In: Introduction to the Theory of Neural Computation. Addison-Wesley, Reading MA, Lecture Notes.
- Isliker H and Benz AO (1994) A&AS, 104:145.
- Kohonen T (1989) In: Self-Organization and Associative Memory, Springer Series in Information Sciences.
- Messmer P, Benz AO and Monstein C (1999) Sol. Phys., 187:33.
- Murtagh F (1994) In: Proceedings of the 14th Int. CODATA Conference, Chambéry. <http://hq.eso.org/~fmurtagh/papers/image-retrieval-codata94.ps>
- Nievergelt J, Hinterberger H and Sevcik K (1984) ACM Transactions on Database Systems, 9(1):35.
- Perrenoud M (1981) PhD Thesis, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland. In German.
- Raghavan V, Jung G and Bollman P (1989) ACM Transactions on Information Systems, 7(3):205.

- Samet H (1989) *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, Reading, MA.
- Schäuble P (1997) *Multi-Media Information Retrieval, Content-Based Information Retrieval from Large Text and Audio Databases*. Kluwer, Dordrecht.
- SOM (1997) *SOM_PAK: The Self-Organizing Map Program Package*, SOM Programming Team, Helsinki University of Technology. `nucleus.hut.fi/nnrc/som_pak`
- Tanimoto T (1958) *Undocumented Technical Report*, IBM Corp.
- van Rijsbergen C (1979) *Information Retrieval*. Butterworth, London.
- Wells DC, Greisen EW and Harten RH (1981) *Astronomy & Astrophysics Supplement Series*, 44:363.