An Approach To Automatically Generate Digital Library Image Metadata For Semantic And Content-Based Retrieval

Eugen Zaharescu

MFP - Bilkent University of Ankara
ezaharescu@ee.bilkent.edu.tr

Abstract. Metadata represents textual information attached to an image or resource to aid identification and retrieval of that resource. In this paper it is revealed an approach to automate the creation of digital library image metadata embedding semantic and content features. These features will make more precise image content indexing and will allow fast retrieval of images in digital libraries, based on similarities with a given image. The proposed system will have the capacity to automatically generate Digital Library images metadata embedding visual features extracted after image processing. This system will read features arrays, previously created by image processing algorithms and saved them in XML documents that describes the image metadata. Further, using the XSLT(eXtensible Stylesheet Language for Transformations) language, the developed system will automatically generate digital library image metadata allowing the Content-based Image Retrieval in Service Providers.

1 Introduction

Metadata is information attached to an image or resource generally in the form of keywords or free text. The information contained in a metadata is searchable and therefore aids the identification and retrieval of resources. Metadata helps users both to discover the existence of information objects and to understand the nature of what they have found. Information added to a resource will also help the user to evaluate a resource, make a judgment about a resource, compare it with another resource or assess its suitability for the intended use. Metadata is used by Digital Libraries to aid the retrieval of their resources. Digital Libraries are organizations that provide the resources including the specialized staff. Their purposes are to select, to structure, to offer intellectual access, to interpret, to distribute, and to preserve the integrity of collections of digital works. They also ensure the persistence over time of such digital collections so that they are readily and economically available for use by a defined community or set of communities [2].

2 Metadata and Digital Libraries

An important thing about Digital Libraries is the dissemination of their information. In the last years, due to the Internet progress, several Digital Libraries appeared with mainly purpose of exposing thesis and others digital materials. However, needs to establish common agreement on adoption and use of standards to facilitate the efficient dissemination of content in Internet lead to the establishment of the Open Archives Initiative (OAI) http://www.openarchives.org and to the development of the OAI-Protocol for Metadata Harvesting (OAI-PMH) http://www.openarchives.org/OAI/openarchivesprotocol.html.. The essence of the open archives approach is to enable access to Web-accessible material through interoperable repositories for metadata sharing, publishing and archiving. The OAI-PMH protocol defines a mechanism for harvesting records containing metadata from repositories. OAI have two groups of 'participants': Data Providers and Service Providers. Data Providers maintain resources in a repository and "expose" for harvesting the metadata about resources in the repository. Digital Libraries are Data Providers. Service Providers harvest and store metadata from Data Providers. They use the harvested metadata for the purpose of providing one or more services across all the data. OAIster http://oaister.umdl.umich.edu/o/oaister/ is an example of Service Provider. OAIster is a project of the University of Michigan Digital Library Production Service, which goal is to create a collection of freely available, previously difficult-to-access, academically-oriented digital resources that are easily searchable by anyone. Through OAIster, is possible to search metadata of the more than 440 Data Providers registered in OAI. OAIster provides interfaces for information retrieval based on text and keywords.

The proposed system is a Semantic and Content-Based Search System from the category of Service Provider. The professionals will have textual and visual information of images. Afterwards, they are harvested by Service Providers through OAI-PMH protocol, stored in a database and used for searches. At this moment, content-based image retrieval is possible in Service Providers. In this case, the user gives a query image to find similar images. Algorithms will process the query image and extract its array of features. Theses features will be used for measuring distances between the query image and images metadata available in Service Providers to find a set of similar images.

3 Content-based Image Retrieval

The Information Retrieval (IR) term, describes the process where a user converts a consult in a useful reference collection as defined by A. Gupta & R. Jain, Visual information retrieval, Communications of the ACM, 40, 1999, 71-79. This concept can be used also to visual information retrieval, as images. The proposed Semantic and Content-based Image Retrieval System will try to solve this task and to overcome difficulties like human perception subjectivity in case of rich images contents. The

images should be indexed by attached semantic information and their own visual content.

The Similarity Search is the main tool used in Content-based Image Retrieval Systems. If visual features are included in images metadata, a considerable initial step will be achieved to implement image content retrieval in Service Providers.

3.1 Automatic Indexing

The creation and maintenance of index to large collections of images involve cost and time. For better information retrieval the computer will extract automatically visual features that describe the images under analysis. These visual features will be used in doing automatic indexing and in comparing images during the content-based retrieval process. If the index to semantic information is created by OWL Systems like Protégé or WorldNet, the index to color images, for example, will be automatically created by computer. This method will also ensure an objective abstraction of image, given that this process won't involve human subjective. Thus, if the image metadata contain both semantic information and visual features it will be possible to index that metadata automatically and lower the delay in the subsequent information retrieval step.

4 Extracting Content Features from Large Collection of Images

A common method to deal with images similarity is extracting image content features. This extraction process is decisive to storage and content-based images retrieval. This process produces an array of n features that constitute the Image Feature Array. Basically all systems that work with images use color and grey level features, mostly in the form of a histogram (L.H. Tang e.a.[10,11]). In connection with segmentation, the shape and texture of the segments can be used as a powerful feature (H. Muller e.a.[7]). I found several works presenting the recover of images using techniques based on texture and shape as well as semantic information (P.W. Huang e.a.[12], G. Sheikholeslami e.a. [13], M. Safar e.a. [14], D. Zhang e.a. [15]). A suitable shape descriptor must be invariant to rotation, translation, and scale transforms like centroid, Euler number (i.e. morphological measure of the topology of an image defined as the total number of objects in the image minus the number of holes in those objects) and moments are. The texture descriptors used could be correlation and entropy, which are the simplest texture statistical descriptors. A proposed tool in our research activity could be CVIPtools 3.9, developed by Computer Vision and Image Processing Laboratory of Southern Illinois University http://www.ee.siue.edu/CVIPtools/. Initially, the image should be segmented and afterwards, the resulting segments can be described by shape features that commonly exist, including those with invariance with respect to shifts, rotations and scaling. These features will be included in the images metadata and will serve for content image retrieval. The images bellow are the original and waterfalls or merging jump connection segmented images, together with their separate gradient and finally superimposed images:

(a) Airplane original image

(b) Waterfalls segmented image

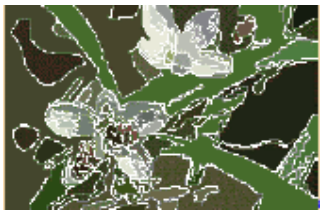
(c) Image gradient

(d) Superimposed images (b) and (c)

(a) Flowers original image

(b) Waterfalls segmented image

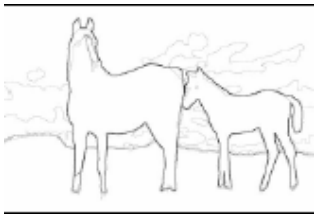
(c) Merging jump connection segmented image



(d) Superimposed gradient of (c) image

(a) Horses original image

(b) Waterfalls segmented image



(c) Merging jump connection segmented image

(d) Superimposed gradient of (c) image

Fig. 1. Two examples showing original images, segmented images, image gradients and final superimposed images after image processing.

Following is shown the features array generated for the first airplane image by image processing:

289 258 -110 0.167794 0.000721 0.000031 0.000001 0.000000 0.000000 0.000000 0.970453 0.014869 6.530753 0.258148 0.139354 0.831247 0.521371

This array has 17 elements respectively: Centroid (row and column), Euler number, Moments (7 rst-invariants moments), Correlation (average, range), Entropy (average, range), Morphological indexes (hue, luminance, saturation).

5 Automating the Image Metadata Generation

To read the feature array resulted by image processing and insert then in the image metadata (XML document), we should develop a stylesheet (i.e. a document written in XSLT [8], which is, mainly, a language to transform a XML document in another XML document).

5.1 XML Document Processing Using XSLT

XSLT was designed for use as part of XSL, which is a stylesheet language for XML, but was, also, designed to be used independently of XSL.

A transformation expressed in XSLT describes rules for transforming a source XML tree into a result tree expressed as a well-formed XML document. The transformation is achieved by associating patterns with templates. The result tree is separate from the source tree. The structure of the result tree can be completely different from the structure of the source tree. In constructing the result tree, elements from the source tree can be filtered and reordered, and arbitrary structure can be added.

A transformation expressed in XSLT is called a stylesheet. This is because, in the case when XSLT is transforming into the XSL formatting vocabulary, the transformation functions as a stylesheet.

A stylesheet contains a set of template rules. A template rule has two parts:

1. a pattern which is matched against nodes in the source tree and
2. a template which can be instantiated to form part of the result tree.

This allows a stylesheet to be applicable to a wide class of documents that have similar source tree structures.

Template description. A template is instantiated for a particular source element to create part of the result tree. A template can contain:

- elements that specify literal result element structure
- elements from the XSLT namespace that are instructions for creating result tree fragments.

When a template is instantiated, each instruction is executed and replaced by the result tree fragment that it creates. Instructions can select and process descendant source elements. Processing a descendant element creates a result tree fragment by finding the applicable template rule and instantiating its template. Note that elements are only processed when they have been selected by the execution of an instruction. The result tree is constructed by finding the template rule for the root node and instantiating its template.

A single template by itself has considerable power because it can:

- create structures of arbitrary complexity;
- pull string values out of arbitrary locations in the source tree;
- generate structures that are repeated according to the occurrence of elements in the source tree.

For simple transformations where the structure of the result tree is independent of the structure of the source tree, a stylesheet can often consist of only a single template, which functions as a template for the complete result tree.

When a template is instantiated, it is always instantiated with respect to a current node and a current node list. The current node is always a member of the current node list. Most of XSLT operations are relative to the current node and only a few instructions change the current node list or the current node. During the instantiation of one of these instructions, the current node list changes to a new list of nodes and each member of this new list becomes the current node in turn. After the instantiation of the instruction is complete, the current node and current node list revert to what they were before the instruction was instantiated.

5.2 Generating Image Metadata XML Document

For processing XML input document by applying the XSLT stylesheet few XSLT processors are available and can be used like the following:

• ***Saxon*** (http://saxon.sourceforge.net/) open source application written in Java, based on SAX (Simple API for XML) ;
•***Xalan*** (http://xml.apache.org/) open source project developed by Apache organization. Two versions are available Xalan-Java and Xalan-C++ based on XML Xerces processor.

When the stylesheet is processed, the image metadata would be like this:

```
<?xml version="1.0" encoding="UTF-8"?>
<metadata> <dc:title>Airplane Image</dc:title>
<dc:format>JPEG</dc:format>
<centroid>
<row>289</row>
<column>258</column>
</centroid>
<eulernumber>-1 10</eulernumber>
<moment>
<rst1>0. 167794</rst1>
<rst2>0.000721</rst2>
<rst3>0.00003 1</rst3>
<rst4>0.000001 </rst4>
<rst5>0.000000</rst5>
<rst6>0.000000</rst6>
<rst7>-0.000000</rst7>
</moment>
<correlation>
<average>0.970453</average>
<range>0.014869</range>
</correlation>
<entropy>
<average>6.530753</average>
<range>0.258148</range>
</entropy>
<morphoindex>
<hue>0.139354</hue >
<luminance>0.831247</luminance >
<saturation>0.521371</saturation >
</morphoindex >
</metadata>
```

6 Conclusions

In this work I proposed a method that will help the creation of image metadata. It is understandable that generating images metadata is a difficult process, and intro-

duces some difficulties. Therefore I have presented a new method for doing this: adding to the image metadata visual features extracted automatically by color image processing algorithms. These features complement the traditional metadata qualitative descriptors and will serve as more precise image content indexes. Furthermore, systems that promote search for a digital library image metadata, as Services Providers, will be able to provide searches based on similarities. Service Providers must implement systems to process an image given by a user, extract its features and calculate the similarities measures. This is a not easy task, and much work will be necessary. This method is a considerable initial step to achieve such systems. To illustrate the method, shape, texture and morphological descriptors are used. The shape, texture or morphological features were used only as example (the choice of the images content features will depends on particularities of each Content-based Image Retrieval System).

At the end, the idea of image metadata automatic generation and content search proposed in this work is valid to other types of resources in Digital Libraries, like audio and video. Algorithms also can process theses resources and extract their features, which could be automatically inserted in the metadata.

References

1. Baldonado, M., Chang, C.-C.K., Gravano, L., Paepcke, A.: The Stanford Digital Library Metadata Architecture. Int. J. Digit. Libr. 1 (1997) 108–121
2. Huang, Y. Rui, Chang, S. F. : Image retrieval, past, present and future, Journal of Visual Communication and Image Representation, 10, 1999, 1-23.
3. Huang, P. W., Dai, S. K.: Image retrieval by texture similarity. Pattern Recognition, 36(3), 2003, 665-679.
4. Gupta, A., Jain, R.: Visual information retrieval, Communications of the ACM, 40, 1999, 71-79.
5. Muller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content-based image retrieval systems in medical applications - clinical benefits and future directions, Journals of Medical Informatics - Elsevier, 2003.
6. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). http://www.openarchives.org/OAI/openarchivesprotocol.html
7. Open Archives Initiative (OAI). http://www.openarchives.org
8. OAIster. http://www.oaister.org
9. CVIPtools. http://www.ee.siue.edu/CVIPtools/
10. Tang, L.H., Hanka, R., Lan, R.: Automatic semantic labelling of medical images for content-based etrieval. Proceedings of the International Conference on Artificial Intelligence, Expert Systems and Applications, Virginia Beach, USA, 1998, 77-82.
11. Tang, L.H., Hanka, R., Ip, H.H.S., Lam, R.:Extraction of semantic features of histological images for ontent-based retrieval of images. Proceedings of the IEEE Symposium on Computer-Based Medical Systems, Houston, USA, 2000.
13. Sheikholeslami, G., Chang, W., Zhang, A.: Semquery: Semantic clustering and querying on heterogeneous features for visual data. IEEE Transactions on Knowledge and Data Engineering, 14(5), 2002, 988-1002.
14. Safar, M., Shahabi, C., Sun, X.: Image retrieval by shape: A comparative study, New York, 1999.

15. XSL Transformations (XSLT). In http://www.w3.org/TR/xslt
16. Waters, D. J.: What are digital libraries? Digital Library Information Resources in Berkeley Digital Library SunSite, CLIR Issues, (4), 1995.
15. Zhang, D., Lu, G.:Content-based shape retrieval using different shape descriptors: A comparative study, Japan, 2001, 317-320