Jaimes, A., Benitez, A., Joergensen, C., & Chang, S. (2000). Experiments in indexing multimedia data at multiple levels.. 11th ASIS SIG/CR Classification Research Workshop, 163-168. doi:10.7152/acro.v11i1.12783

Experiments in Indexing Multimedia Data at Multiple Levels

Alejandro Jaimes*, Ana B. Benitez*, Corinne Joergensen¥, and Shih-Fu Chang*

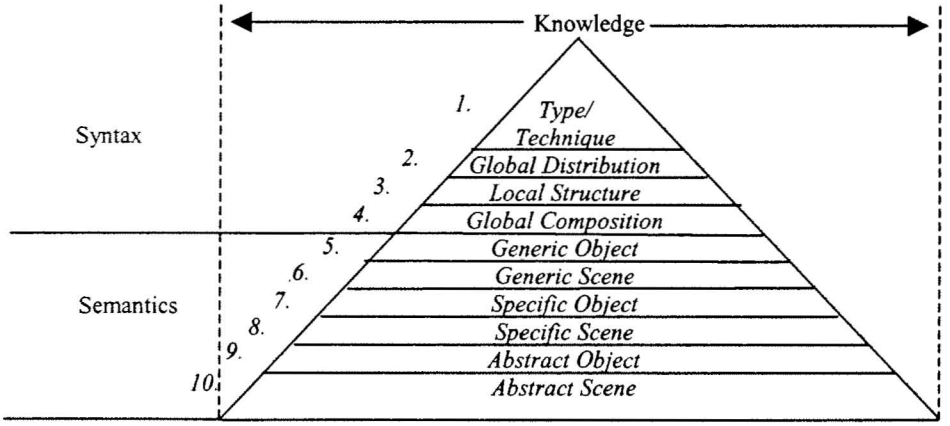
* Columbia University ¥State University of New York at Buffalo

Introduction: The importance of indexing multimedia data

The increasing availability of digital images, video, and audio has created exciting new research challenges on the organization of multimedia data for a variety of purposes. While some of these challenges relate to computational techniques (e.g., automatic extraction of visual features for automatic indexing of visual data), others are conceptual in nature (e.g., design of templates for manual indexing of visual data). The key issues are what to index from the data, how to perform the indexing of the data, and how to organize the indices obtained. The indices used to describe content as well as the organization of those indices have a tremendous impact on applications, particularly on large digital libraries where different types of media need to be stored and accessed. Relevant efforts in this direction include the emerging MPEG-7 standard [5], which aims at standardizing tools for describing multimedia data.

The 10-level indexing pyramid

In this workshop, we will present experiments we have performed for MPEG-7 [3][6] in indexing and retrieving images using structures we have developed to facilitate indexing and to organize different types of attributes. For indexing the images, we have used a template developed by one of the authors [3], which provides a framework for manually indexing visual content. The template has been mapped to a ten-level pyramid that was developed independently by the other authors [1] (see ANNEX). The ten-level pyramid (fig. 1), which draws on research in different fields such as cognitive psychology and content-based retrieval, can be used to classify attributes obtained from images, video, or audio. Although the indexing template was developed independently of the pyramid, we found that the mapping between the two structures was intuitive and worked well in practice. In particular, we found that the template is very useful in guiding the indexing process, and that the pyramid is very useful in organizing the attributes obtained using the template.



ISSN: 2324-9773

Jaimes, A., Benitez, A., Joergensen, C., & Chang, S. (2000). Experiments in indexing multimedia data at multiple levels.. 11th ASIS SIG/CR Classification Research Workshop, 163-168. doi:10.7152/acro.v11i1.12783

Figure 1. The 10-level indexing pyramid. The first four levels correspond to syntactic attributes. The remaining six correspond to semantic attributes.

The pyramid (fig. 1) distinguishes between syntactic (first four levels) and semantic attributes (next six levels). The syntactic levels hold attributes that describe the way in which the content is organized, but not its meaning. In images, for example, type could be "color image". Global distribution holds attributes that are global to the image (e.g., color histogram), whereas local structure deals with local components (e.g., lines and circles), and global composition relates to the way in which those local components are arranged in the image (e.g., symmetry). The semantic levels, on the other hand, deal with the meaning of the elements. Objects can be described at three levels: generic -every day objects (e.g., person), specific -individually named objects (e.g. Bill Clinton), and abstract -representing emotions (e.g., power). In a similar way, a scene can be described at these three levels (see ANNEX for examples). The same pyramid structure can also be applied to audio and video [3][2].

Indexing and retrieval experiments using the pyramid

We have performed experiments in which several participants manually annotated 700 images collected randomly from the World Wide Web. The images were annotated using the template or the pyramid as starting points. Once the annotations were completed, two web-based search engines were constructed, one to automatically retrieve images by specifying keywords and a pyramid level, and another one to retrieve images using keywords only (no pyramid level). In the set of experiments reported we show that organizing the attributes with the pyramid helps users search for images. In particular, we found that maintaining recall (i.e., percentage of correct images that were retrieved) at 100%, precision (i.e., percentage of returned images that are correct) drops if the pyramid structure is not used in the retrieval (table 1). We maintain 100% recall by specifying the appropriate pyramid level when we perform the corresponding query. Since the pyramid level is specified in the query, all returned images contain the query keyword at the specified level, and no errors occur. For example, a "clouds" + "generic object" query returns only images that have that annotation at that level (precision is 100%), and it returns all of them (recall is 100%). When the level is not specified, errors occur because often the same terms are used to describe images at different levels. For example, a user searching for "clouds" objects will get non-cloud images if he/she does not use the pyramid (with the generic object level specified). Using the free text query, images with the abstract annotation "clouds" could be returned (producing errors, and therefore reducing precision) since the system would not know what level of indexing the user is referring to. This demonstrates the importance of classifying the attributes at multiple levels using our structures.

Indexing Term	Pyramid Level	Free Text Matches	Pyramid Matches	Free Text Errors	Free Text Precision	Pyramid Precision
Cartoon	Type/Tech	29	19	10	65%	100%
Coarse	Glob Distrib.	6	3	3	50%	100%
Geometric	Loc Struct.	11	7	4	63%	100%
Clouds	Gen Obj.	4	3	1	75%	100%
Lake	Gen Scene	18	1	17	94%	100%

ISSN: 2324-9773

Moon	Spec Scene	5	1	4	20%	100%
Desire	Abs Obj.	7	4	3	57%	100%
Friendship	Abs Scene	26	22	4	85%	100%

Table 1. Precision results maintaining a 100% recall for some indexing terms assigned by the participants of the experiment.

Although the indexing structures we have developed have worked well in our experiments, there are still open issues and room for possible improvements. The amount of indexing required for multimedia data (i.e. number of attributes at different levels of the pyramid or a similar structure) is highly dependent on the target application and specific content being used. Abstract levels, for example, may not be needed if a database of company logos is being indexed. Structures like the ones we have developed are very useful because they allow a selective and recursive organization of attributes. The pyramid, for example, can be applied to an entire image or part of an image. The question regarding what to index, however, remains a difficult one. In that sense, these types of structures can help identify relevant content indexing dimensions. Other issues include the subjectivity of indexing and retrieval, and the integration of automatically/manually extracted features in describing visual content. In our current work for MPEG-7, we are extending the presented experiments to video and audio [2] indexing and retrieval.

References

[1] A. Jaimes and S.-F. Chang, "A Conceptual Framework for Indexing Visual Information at Multiple Levels", IS&T/SPIE Internet Imaging 2000, January 2000.

[2] A. Jaimes, A. B. Benitez, S.-F. Chang, "Multiple Level Classification of Audio Descriptors", ISO/IEC JTC1/SC29/WG11 MPEG00/M6114, Geneva, Switzerland, May/June 2000.

[3] A. Jaimes, C. Jorgensen, A. B. Benitez, and S.-F. Chang, "Experiments for Multiple Level Classification of Visual Descriptors", ISO/IEC JTC1/SC29/WG11 MPEG99/M5593, Maui, Hawaii, USA, Dec. 1999.

[4] C. Jorgensen, "Image Attributes in Describing Tasks: an Investigation", *Information Processing & Management*, 34, (2/3), pp. 161-174, 1998.

[5] MPEG-7 Web site: http://www.cselt.it/mpeg/

[6] M. Shibata, A. Tam, C. Leung, K. Hasida, A. Benitez, C. Jorgensen, A. Jaimes "Report of CE on Structured Textual Description", ISO/IEC JTC1/SC29/WG11 MPEG00/M6240, Beijing, China, July 2000.

Questions

1. Scope. The pyramid represents one view of the types of information that need to be indexed when describing multimedia content. Are the levels we propose necessary and sufficient to index visual information? What other types of visual content information need to be represented? Can they be accommodated within the existing pyramid? (please note that current work is extending the pyramid to video and audio).

ISSN: 2324-9773

Jaimes, A., Benitez, A., Joergensen, C., & Chang, S. (2000). Experiments in indexing multimedia data at multiple levels.. 11th ASIS SIG/CR Classification Research Workshop, 163-168. doi:10.7152/acro.v11i1.12783

2. Utility/Evaluation. How and where can an indexing structure such as that presented in the pyramid be useful? Is it useful as one or all of the following: -- A guide to indexing? -- A method for structuring indexing records? -- A query formulation tool? -- A retrieval mechanism? What are the major benefits/drawbacks of our approach, in comparison to existing structures, and where should we focus future research? What is the best way to evaluate these structures, and to justify the need to have multiple levels?

3. Implementation. What is the most appropriate implementation of an indexing structure such as this? Does the conceptual hierarchy need to be reflected in an actual hierarchical implementation of the levels, or could a flat or simple list of levels have the same functionalities? If a hierarchical structure is used, is there a trade off between number of levels and gain in retrieval performance? Will problems in distinguishing between levels counteract any advantages? Can automatic classification techniques work with this kind of (hierarchical) structure?

ISSN: 2324-9773

ANNEX: Image Indexing Template Mapped to Pyramid

The data fields were designed for human text input. The fields are mapped to each level of the pyramid, and examples of indexing terms are given in parentheses.

1. **Type/Technique**

   Image Type (photograph, digital image, drawing, painting, animation), Medium (oil, watercolor), Style (realism, abstract, mechanical).

2. **Global Distribution**

   Color Type (color, black and white), Global Color, Global Color Quality, Global Shape, Global Texture

3. **Local Structure**

   Local Color, Local Color Quality, Local Color Placement (center, upper right, upper half, foreground), Local Shape (square, oval, elongated, curved), Local Shape Placement, Local Texture (smooth, shiny, fuzzy), Local Texture Placement, Object Placement, Shape, Texture, Size, Number, Color, Living placement, Living size, Living number

4. **Global Composition**

   Perspective/point of view/ (bird's-eye, close-up)

5. **Generic Objects**

   Generic Objects Category (general category; what it is: tool, fruit), Generic Object Living (human, animal, plant, mythical being), Object Type (what it is - hammer, apple), Object Living Gender, Object Living Age

6. **Specific Objects**

   Specific Object Name (keyword/proper noun: ball peen hammer, Macintosh Apple, President Bill Clinton)

7. **Abstract Objects**

   Symbolism (Garden of Eden, Afterlife), Emotions/mental states (sadness, laughter), Relationships (brothers, romantic), Status (occupation, ethnicity, social status)

8. **Generic Scene**

   When (general time: Middle Ages, summer), Where (general location: city, rural, seashore, indoor, office) Genre (landscape, portrait), Category (action and adventure, drama), General Event (type of event: parade, football game), Activity (writing, camping, gambling), Pose/Action (seated, standing, lying down, running, talking, throwing)

ISSN: 2324-9773

Jaimes, A., Benitez, A., Joergensen, C., & Chang, S. (2000). Experiments in indexing multimedia data at multiple levels.. 11th ASIS SIG/CR Classification Research Workshop, 163-168. doi:10.7152/acro.v11i1.12783

9. Specific Scene

Specific When (specific time: 1108, April 15.), Specific Where (New York City, Chrysler Building), Specific Event (Rose Bowl, Super bowl)

10. **Abstract Scene**

Subject (Subject/discipline: Geography, Engineering), Symbolism (Garden of Eden,, nature, urban life, power, freedom), Emotions/mental states (awe, fear), Relationships (friendship, competition, dominance), Atmosphere (overall feeling: gloomy, mysterious, carefree)

ISSN: 2324-9773