Indexing the Content of Multimedia Documents

Stephen W. Smoliar and Lynn D. Wilcox
FX Palo Alto Laboratory
3400 Hillview Avenue, Bldg. 4
Palo Alto, CA 94304 USA
Telephone: +1 415 813-{6703,7574}
Fax: +1 415 813-7081
{smoliar,wilcox}@pal.xerox.com

Abstract

As the concept of what constitutes a "content-based" search grows more mature, it becomes valuable for us to establish a clear sense of just what is meant by "content." Recent multimedia document retrieval systems have dealt with this problem by indexing across multiple indexes; but it is important to identify how such multiple indexes are dealing with multiple dimensions of a description space, rather than simply providing the user with more descriptors. In this paper we consider a description space for multimedia documents based on three "dimensions" of a document, namely *context*, *form*, and *content*. We analyze the nature of this space with respect to three challenging examples of multimedia search tasks, and we address the nature of the index structures that would facilitate how these tasks may be achieved. These examples then lead us to some general conclusions on the nature of multimedia indexing and the intuitions we have inherited from the tradition of books and libraries.

Introduction

With the increase in functional power of multimedia document retrieval systems, the idea of what constitutes a "content-based" search is becoming more mature. The earliest forms of such systems were based only on means of detecting and indexing *features* of different media objects, such as keywords for text and distributions of colors and textures for images [4]. However, experience with these systems quickly revealed that there was a significant distinction between *media objects* and *objects of content*.

These two concepts may best be explained through the terminology of semiology [1], which places particular emphasis on the proposition that every symbolic object consists of two interacting components, one of which is situated on the plane of *expression* while the other is situated on the plane of *content*. This distinction was introduced by semiology because any analysis of how symbolic objects communicate has to recognize the difference between what constitutes the objects themselves (media objects) and what is actually being communicated (the objects of content). From this point of view, the plane of content serves to represent the objects of content; and the plane of expression is concerned with the media objects, i.e. the *"bit-level" objects*, such as text, images, audio clips, and video excerpts, that are generated in order for communication to be achieved. Thus, given a large collection of multimedia documents, we need to address how that collection can be indexed with respect to both the plane of expression and the plane of content.

For example a 485x408 GIF image exhibiting predominantly red color in a central circular region is a media object; but there is nothing about the properties of that media object that inform us as to whether the object of content corresponding to that region is a tomato, a traffic light, a stop sign, or a sunset under particular atmospheric conditions. More recent systems have tried to support indexing with respect to the objects of content associated with the media objects. One approach to achieving this capability has been to extend feature analysis with clustering methods that allow for either automated or user-assisted classification of the objects of content [13]. These classification techniques have, in turn, led to taxonomic hierarchies that facilitate browsing and navigating the search space of media objects on the basis of the objects of content being represented [9].

Our own approach to the indexing and retrieval of multimedia documents extends the dual nature of the document defined by the planes of context and expression. This extension is based on a recent suggestion by John Seely Brown [2] that a document consists of a "weaving together" of "threads" of *context*, *form*, and *content*. We see Brown's suggestion as an approach to answering the most fundamental question behind all attempts at indexing: **How do we describe documents?** Traditional information retrieval has tried to develop description techniques based solely on content; but Brown has argued that documents require a

description *space*, whose three dimensions are content, context, and form. The best way to index a document is to account for how it is situated with respect to all three dimensions of this space, rather than trying to reduce all description to a single dimension.

What does it actually *mean* to describe and index a document with respect to content, form, and context? Indexing with respect to content may remain concentrated on accounting for features on the plane of expression and may thus continue to take advantage of existing technology. Indexing with respect to form serves to describe the internal structure of a document, such as the chapters of a book or the scenes in a video. As more and more texts are prepared with some markup language, such as SGML, the automatic generation of form indexes for text becomes more feasible. Similarly, segmentation techniques for images, audio, and video provide the basis for form indexes for these media. The more challenging problem is dealing with context. Context indexing must account for our ability to classify documents with respect to an ontology for their objects of content, but such classification is not necessarily sufficient. Once a category has been indentified, it may entail additional descriptors which must also be taken into account in a context index.

We shall now address this potential complexity of context-based indexes, as well of the interactions among context, content, and form, by examining three retrieval tasks. The first deals with retrieving an image, the second concerns an excerpt of a musical performance, and the third involves a similar excerpt from a movie. We shall see that what constitutes context, as well as the relationship between context and form and content, differs across these three examples. We shall then interpret these results in terms of their implications for indexing. We conclude with some remarks on the general nature of multimedia indexing.

The Role of Context: Three Examples

Images

It has long been recognized that having databases of images can be just as important as having databases of texts. In the past, limitations of technology tended to force us to manage such databases on the basis of text descriptions of the images, rather than the images themselves. It has only been in this decade that potentially viable systems have emerged in which

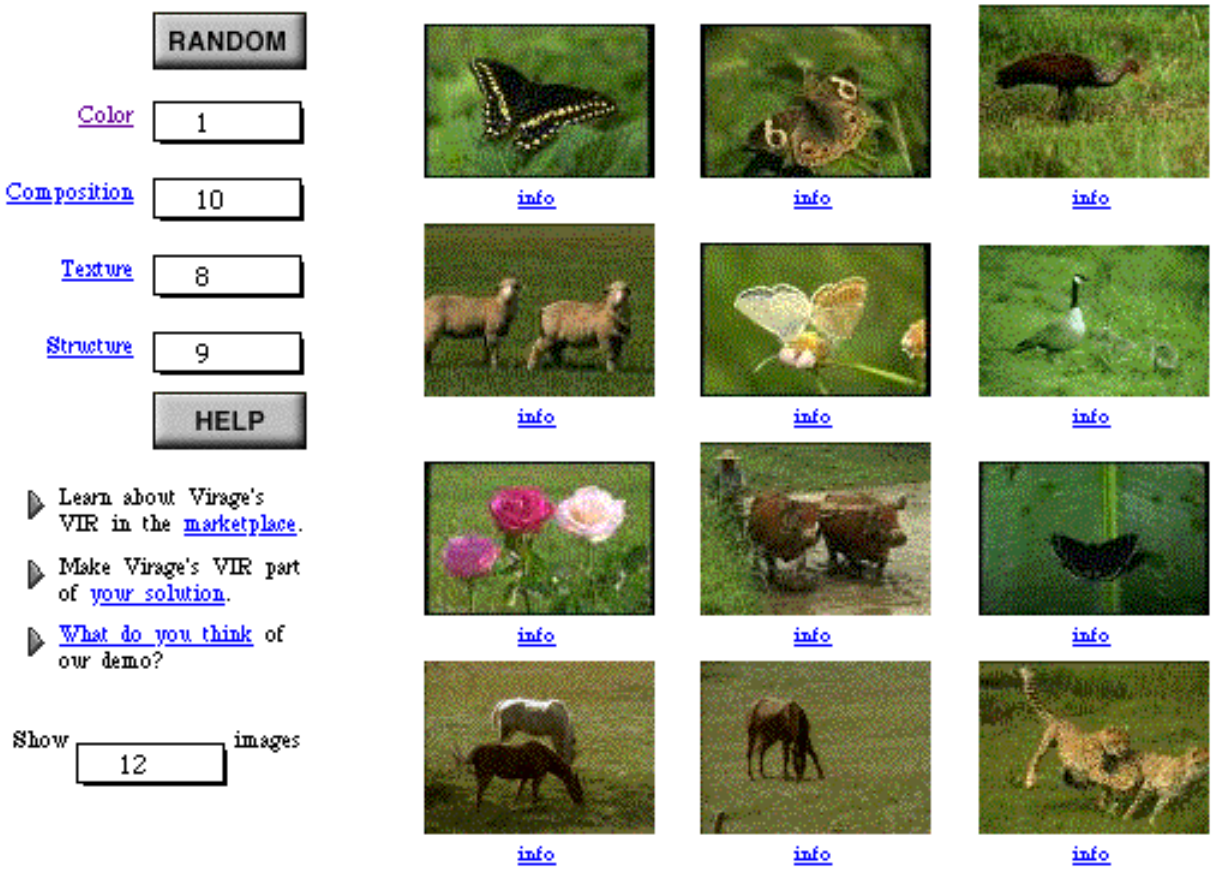

Figure 1: The Virage approach to image retrieval.

operations of indexing and search are based on those images, rather than on text descriptions [4]. However, in order to make such systems work, it is necessary to represent images in terms of quantitative models of specific features, such as color, texture, shape, and layout.

A representative example of this approach is illustrated in Figure 1, a demonstration taken from the Web site for Virage, Inc. (http://www.virage.com/). The user can select an image and then submit a query to find similar images. The query is expressed in terms of features for similarity. The Virage VIR Image Engine uses the following features:

- **Color:** The variation of hue, saturation, and intensity are used to determine color. Regional as well as global color are evaluated, considering both dominant color and color variation.
- **Composition:** The spatial relationships among elements in the image determine composition. For example, the locations of colors in the image are compared to the same information in the query image.
- **Texture:** Texture is determined by pattern variations within narrow sample regions. Patterns include granularity, roughness, and repetitiveness. Sand, for example, has a strong textural element.
- **Structure:** Structure is computed by evaluating the boundary characteristics of distinct shapes. Information from both organic (photographic) and vector sources are used, and partially obscured shapes can be extrapolated. Polka dots, for example, have a strong structural element.

The effects of these criteria can be seen, to some extent, in Figure 1. Consider the problem of trying to collect images of butterflies. The image in the upper-left corner can serve as the query image, but the actual control over the query will depend on how the criteria for similarity are weighted. For retrieval of butterfly images, the highest weights have been assigned to the composition, structure, and texture features. A low weight is assigned to the color feature under the assumption that there are many different colors of butterflies. Results of the search are shown in the other 11 images in Figure 1. We see that three butterfly images were found with variations in both orientation and color. However, we also retrieved images of flowers, birds, and other animals.

Figure 2: The WebSEEk approach to image retrieval.

One way to overcome this difficulty is to provide support for indexing based on subject classification, in addition to the representation of visual features. The WebSEEk system (http://www.ctr.columbia.edu/webseek/) provides a powerful semi-automated approach to classifying images and videos according to a general subject taxonomy based on text associated with those images and videos [10]. Thus, a user who wants the image of a whole red tomato can find the "food" category in this taxonomy, proceed to the "food/vegetables" category, and eventually arrive at the "food/vegetables/tomatoes" category. Once this category has been identified, thirteen images may be displayed, as shown in Figure 2.

Subject classification is a step in the right direction towards context-based indexing; but we would argue that it is only a first step. There still remains the question of whether or not the user can get the "right" tomato through this process. Suppose, for example, that the user is interested in cooking with tomatoes and wants to explore different ways in which tomatoes may be prepared or perhaps even the necessary preparation for a particular recipe. In this case indexing with respect to context, form, and content involves an approach to context that goes beyond the current subject classification hierarchy of WebSEEk:

1. **Context:** If the user is interested in cooking, it would be most useful to be able to restrict the search to documents about cooking. Thus, while the taxonomy of objects that forms the basis for classification in WebSEEk is valuable, what may be just as valuable, if not more so, would be the sort of subject classification that is applied to library books. It is entirely possible that, as the collection grows, the WebSEEk taxonomy will evolve in this direction; but, if it does so, it will be through the fortuitous appearance of text that accompanies the images. Library subject indexes are based, instead, on categories posited *a priori* on the basis of some model of how the world of documents should be described (an activity which is ocasionally referred to as "ontological engineering"); and classification with respect to these categories may have to be assigned manually. What is important is that such *a priori* categories are often as valuable as the categories generated by WebSEEk's automatic techniques, if not more so. As a result, a context-based index should probably take advantage of both approaches to categorization and provide a useful approach to reconciling these two points of view.
2. **Form:** Documents about cooking frequently include collections of recipes. Any recipe, in turn, is usually structured to consist of a description of ingredients and a description of the preparation process. In other words context based on subject matter provides additional information as to how a document may be segmented and which segments are most likely to be of interest. This particular search can be restricted to those segments concerned with ingredients.
3. **Content:** Once the search space has been narrowed with respect to context and form, we can revert to the sorts of feature-based techniques available through Virage, since, within the restricted set of images, a red circle is more likely to be a tomato than a stop sign, a sunset, or a traffic light.

Figure 3: Retrieval of a red tomato.

In Figure 3 we see an example of a result from a search that makes more specific use of index information based on context, form, and content. A document has been found with an illustrated ingredients list. That list includes an image of two tomatoes in a frying pan.

Audio

A feature-based approach to the indexing and retrieval of sounds is being pursued by the researchers at Muscle

Fish [12]; and, like Virage, they have made a demonstration of their technique available at their Web site (http://www.musclefish.com/). This demonstration is illustrated in Figure 4. Through this interface, the user can apply the Virage-style paradigm of trying to retrieve sounds that are similar to a selected sound template. In this case the criteria for similarity are drawn from the following acoustic features [12]:

- **Pitch:** Pitch is computed from peaks in the short-time Fourier spectrum.
- **Amplitude:** Amplitude is defined as the signal's root-mean-square (RMS) level in decibels.
- **Brightness:** Brightness is computed as the centroid of the short-time Fourier spectrum.
- **Bandwidth:** Bandwidth is a magnitude-weighted average of the differences between the spectral components and the centroid.

The pitch, brightness, and bandwidth features are computed for a series of short-time Fourier spectra over the duration of the sound. Three measurements are then stored for each feature: the mean of the feature over the sound, its variance, and the autocorrelation for a small lag. The duration of the sound is also used as a feature.

To illustrate the Muscle Fish audio retrieval system, we consider the problem of finding oboe music in a collection of audio files. To this end, we selected as a template a digitized recording of an oboe, and performed a similarity search based on the above features. Figure 4 shows the results of this search. The search correctly brings up other oboe files, but in addition retrieves files corresponding to other instruments and "heavy rainfall."

As might be expected, the problems with this technique are similar to those encountered with the Virage approach. Also, as we know from the literature concerned with the perception and cognition of music [3], the above features are not necessarily related in any useful way to how the sounds are actually perceived. Furthermore, even to the extent that such features *may* be relevant to perception, representing them by statistics computed over the duration of the entire sample is unlikely to have much perceptual significance. Further, as in the Virage example, there is no attempt to classify the object of content. Each sample is assumed to be a single, isolated sound; so for example it is not possible to find an oboe solo in a recorded symphony.

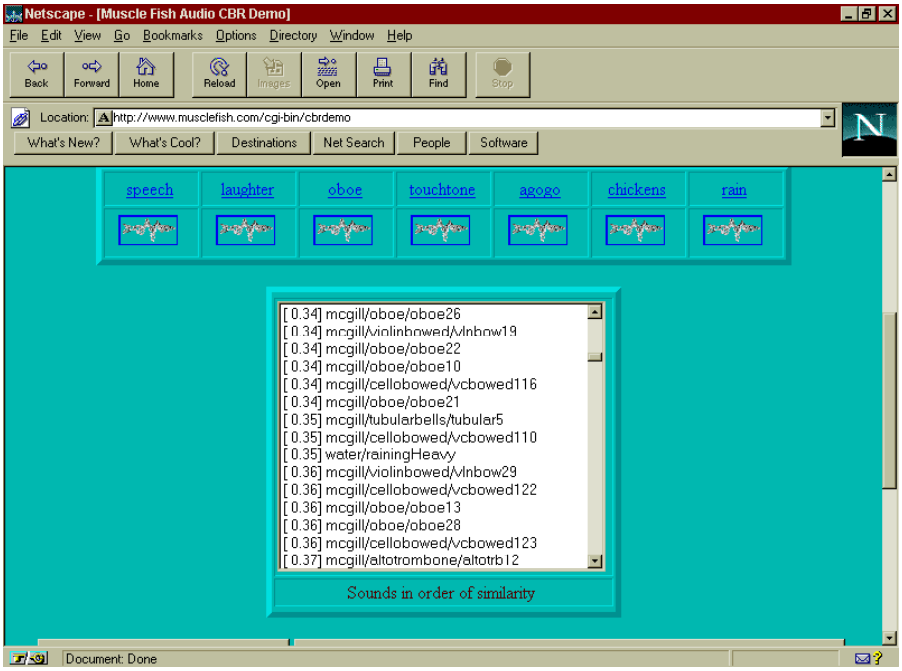

Figure 4: The Muscle Fish approach to audio retrieval.

As an alternative let us consider an example in which there is a context that can be exploited as part of the search process. Suppose that our search space is based on *music*, rather than isolated sounds; and, being even more specific, let us assume that we are restricting ourselves to performances by jazz combos. Now suppose that we want to find a performance with a "dialogue" between an alto saxophone and a trumpet. In this case we may approach indexing based on context, form, and content as follows:

1. **Context:** Almost all jazz recordings maintain catalog information for each song of who the specific performers were and what instruments they were playing. This information can be used as a major element for a context-based index. Once that information is available, the search may be restricted to performances that include an alto saxophone and a trumpet.
2. **Form:** Such performances consist of improvisations on material performed by either soloists or small groups in combination. Each such group or solo has its own characteristic sound. Therefore, the entire composition may be segmented on the basis of these sounds.
3. **Content:** Once all segments have been classified according to their respective characteristic sounds, a search can be conducted for segments of the alto saxophone sound alternating with those of the trumpet sound.

Figure 5: Results of a jazz segmentation.

Segmentation by characteristic sound is achieved using techniques similar to those employed in segmenting an audio recording of a meeting according to speaker [11]. We begin by constructing a *jazz palette*. This consists of a model for each of the various jazz sounds, where the jazz sounds consist of individual instruments or small groups of instruments. Models are created by collecting samples of jazz sounds from actual recorded performances, rather than from the more sterile conditions of the instruments playing under isolated circumstances. These samples are labeled by hand, and a model is generated by training a Gaussian mixture for each particular characteristic sound. A recording is segmented by first applying the contextual knowledge of what instruments are being played to select sounds from the palette likely to be contained in the recording. A hidden Markov model is then constructed from the Gaussian mixture models for the selected sounds. The Viterbi algorithm [7] is used to find the most likely sequence of models, thus providing a segmentation of the recording by characteristic sound.

The result is a sequence of music segments labeled by the sound types in the palette, as illustrated in Figure 5. In this example, the characteristic sounds are trumpet, alto saxophone, tenor saxophone, and duet, which is the alto saxophone and tenor saxophone in combination. Each line in the display represents a sound, indicated by a label and color. The colored bands on the lines represent time intervals during which the corresponding sounds were present.

To find the desired dialogue between an alto saxophone and trumpet, the user could visually scan a display of the type shown in Figure 5. In this case, it is apparent that the trumpet and alto saxophone conversation occurs in the first two thirds of the display. Alternately, the results of the segmentation could be stored in a data file, with the sequence of sounds and the times when these sounds occurred. These data could then be searched for the appropriate pattern.

Video

For all intents and purposes, all problems concerned with searching video are subsumed by problems concerned with searching both images and audio. Thus, WebSEEk can apply the same approach to indexing videos that it applies to images, classifying all instances with respect to its subject ontology on the basis of associated text data. Unfortunately, this approach precludes any use of the sound-track, which is often the more valuable source of cues. Consider, for example, a search problem similar to the one we just discussed, in which the object of the search is a woman singing a song. We may account for context, form, and content in such a search as follows:

1. **Context:** If we know the vintage of the song, we may restrict our search to films that were made after the song was written. In addition, Lawrence Rowe *et. al.* [8] have developed an indexing system that accounts for the genre of a film. Such an index can then restrict the search to films in which songs are sung.
2. **Form:** As was the case with recordings of jazz combos, each film may be segmented according to the different sounds on the audio track.
3. **Content:** These film segments can now be searched for those matching a "female song" type.

Form indexing is performed using unsupervised hierarchical agglomerative clustering [6]. This technique iteratively merges segments of audio with similar sound type. The result is clusters of segments which have been classified as similar in sound type. Sound segments are not labeled at this point, as they were in the jazz example, because we cannot predefine all the sound types likely to occur in a video.

Content indexing can then be performed in part by classifiying each cluster of segments according a known

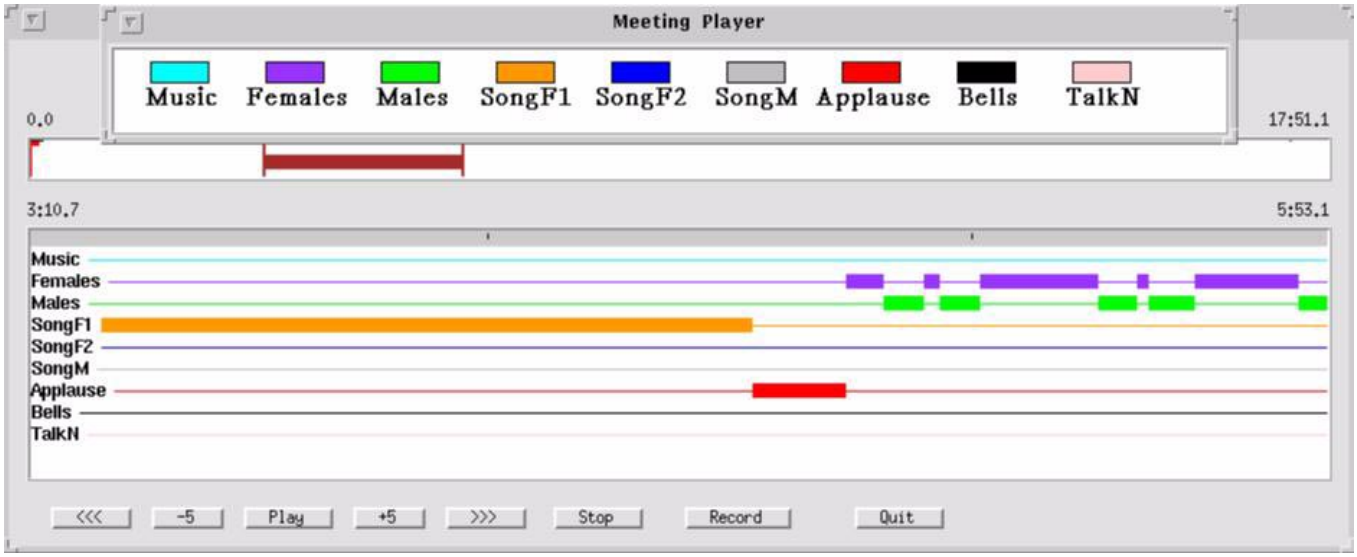

Figure 6: Segmenting a film with an audio palette.

palette of sound types for the particular film genre. For musical films, the palette would consist of male and female speech, male and female song, and music. Classification is performed by first training a Gaussian mixture model for each sound using data that had been collected for that sound. Then, the likelihood that a cluster of segments is a particular sound type can be computed based on the mixture model for that sound. If the likelihood exceeds a given threshold, the cluster can be classified as that type. Not all sound types present in the video are necessarily contained in the palette, so it may not be possible to classify all clusters of sounds. However, all that is necessary is to identify clusters for particular search objectives.

Figure 6 shows a portion of the video "Old Lace" that has been segmented according to sound type using hierarchical agglomerative clustering. As in the jazz example, the colored bands represent intervals during which each sound type was present. The clusters of sounds in this example were hand-labeled as male and female speech (Males, Females), male song (SongM), songs by two different female singers (SongF1, SongF2), Music, Applause, Bells, and talk in noise (TalkN). It is interesting that the clustering algorithm created two distinct clusters for female song corresponding to the different singers in the film. The segments labeled male and female speech, music, and male and female song are used to create the genre palette discussed above. The two female song clusters would be combined in the palette.

Note that while these techniques allow us to identify segments where a woman is singing, they do not allow us to identify the particular song. This problem can be solved by either listening to the female song segments or by using a pitch-tracking technique [5] to identify the melody. One could then index on the basis of melodic fragments, using, for example, the intervals between successive notes as a search key.

Implications for Indexing

It is important to note that indexing with respect to context, form, and content is not entirely novel. An approach involving all three types of indexes had been taken by Rowe and his colleagues in dealing with the problem of indexing video databases [8]. In their system, context is managed by a *bibliographic* index that accounts for descriptors such as title, a text abstract, genre, director, and cast. Form is handled by a *structural* index that accounts for the segmentation of a video into shots and scenes. Content is provided by an *object* index, which includes the names of actors and actresses, and a *keyword* index, for text associated with the film.

What we feel is a valuable lesson from our exercise is that the indexes for context, form, and content each provide a different way of *filtering* a large search space of documents. The context index filters the space on the basis of one or more conceptual classifications of the documents. The form index partitions any document into some structure of components, so that search may be restricted to some subset of those components. Finally, within any such component, features carry more semantic import than they would for an arbitrary collection of data objects; so content indexing may still be based on features but used in a far more efficient manner.

The distinction between these three types of indexes may best be appreciated in terms of how we use conventional libraries. If we have a particular search objective, we

usually associate it with one or more subject areas. We then search the library either by going to a room housing a specialized collection or by consulting the subject index of the master card catalog. Thus, the structure of the library itself, as well as a subject index, are examples of what we mean by *context* indexes. This process leads us to a potentially useful subset of books from the entire collection, and we now have to start examining those books individually. For each book, we tend to begin by consulting the table of contents and/or the index. These are what we mean by *form* indexes. They allow us to narrow down our search based on the structure of the book: a specific section indicated by the table of contents or some set of pages designated by the index. Having narrowed that search we can then concentrate on specific features, such as the presence of keywords in text or the color or texture of an image; and these are the items managed by what we call a *content* index.

Conclusions

What our experiments reveal, then, is that books and libraries seem to have had it right all along, long before information technology became immersed in problems of indexing and search keys. It is tempting to think of documents as digital objects, ask what features of those objects are most readily computable, and structure indexes about those features. One can clearly get a lot of mileage out of such an approach, as can be seen by the prodigious number of objects handled by existing systems. Nevertheless, we, as individual library users, have acquired our own filtering techniques on the basis of how we use both the library and the books we cull from its shelves. Those habits have served us well for a very long time. They can continue to serve us if we can use them in a manner consistent with our technology. Developing indexes for context, form, and content that are appropriate to the material being indexed and the ways in which readers are likely to use that material will be a step in the right direction.

Acknowledgments

The images in Figure 3 were taken from a televised cooking program. We wish to thank Arna Vodenos Productions for permission to use this material. We also thank Maxime Fleckner Ducey at University of Wisconson Center for Film Research for supplying the video "Old Lace." Finally, we wish to thank our colleagues, John Boreczky and Bernard Mont-Reynaud, for their valuable inputs while we have been working on this paper.

References

1. R. Barthes. *Elements of Semiology*. A. Lavers and C. Smith, translators, Hill and Wang, New York, NY, 1973.
2. J. S. Brown. Documents & the evolving workscape. *Stanford Professional Publishing Course* (Stanford, CA, July 1996).
3. D. Butler. *The Musician's Guide to Perception and Cognition*. Schirmer, New York, NY, 1992.
4. B. Furht, S. W. Smoliar, and H.-J. Zhang. *Video and Image Processing in Multimedia Systems*. Kluwer Academic Publishers, Boston, MA, 1995.
5. A. Ghias, J. Logan, D. Chamberlin, and B. Smith. Query by humming: Musical information retrieval in an audio database. *Proceedings: ACM Multimedia Conference* (San Fransisco, CA, November 1995), ACM, 231-236.
6. D. Kimber and L. Wilcox. Acoustic segmentation for audio browsers. *Proceedings: Computing Science and Statistics: 28th Symposium on the Interface* (Sydney, AUSTRALIA, July 1996).
7. L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*, 2 (February 1989), 257–285.
8. L. A. Rowe, Jr., J. S. Boreczky, and C. A. Eads. Indexes for user access to large video databases. *Proceedings: Storage and Retrieval for Image and Video Databases II* (San Jose, CA, February 1994), IS&T/SPIE, 150–161.
9. J. R. Smith and S.-F. Chang. Tools and techniques for color information retrieval. *Proceedings: Storage and Retrieval for Still Image and Video Databases IV* (San Jose, CA, February 1996), IS&T/SPIE, 426–437.
10. J. R. Smith and S.-F. Chang. Searching for images and videos on the World-Wide Web. Center for Telecommunications Research Technical Report #459-96-25, Columbia University (1996) (available at http://www.ctr.columbia.edu/webseek/paper/).
11. L. Wilcox, D. Kimber, and F. Chen. Audio indexing user speaker identification. *Proceedings: Automatic Systems for the Identification and Inspection of Humans* (San Diego, CA, July 1994), SPIE, 149–157.
12. E. Wold *et al.* Content-based classification, search, and retrieval of audio. *IEEE MultiMedia, 3*, 3 (Fall 1996), 27–36.
13. D. Zhong, H.-J. Zhang, and S.-F. Chang. Clustering methods for video browsing and annotation. *Proceedings: Storage and Retrieval for Still Image and Video Databases IV* (San Jose, CA, February 1996), IS&T/SPIE, 239–246.