M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 358
Also to appear: Springer Verlag Workshops in Computing, MIRO 95, Invited Paper, Glasgow, Sep. 95

Toward a Visual Thesaurus

Rosalind W. Picard
MIT Media Laboratory, 20 Ames St., Cambridge, MA 02139
picard@media.mit.edu, http://www.media.mit.edu/~picard/

Abstract

A thesaurus is a book containing synonyms in a given language; it provides similarity links when trying to retrieve articles or stories about a particular topic. A "visual thesaurus" works with pictures, not words. It aids in recognizing visually similar events, "visual synonyms," including both spatial and motion similarity. This paper describes a method for building such a tool, and recent research results in the MIT Media Lab which contribute toward this goal. The heart of the method is a learning system which gathers information by interacting with a user of a database. The learning system is also capable of incorporating audio and other perceptual information, ultimately constructing a representation of common *sense* knowledge.

1 Introduction

Collections of digital imagery are growing at a rapid pace. The contexts are broad, including areas such as entertainment (e.g. searching for a funny movie scene), education (e.g. hunting down illustrations for a book report), science (e.g. analyzing satellite imagery), medicine (e.g. retrieving images with similar abnormal tissue), law (e.g. researching similar trademarks), business (e.g. finding footage for a promotional video), and design (e.g. shopping for fabric patterns). In all these applications and more, providing easy access to image and video content is a significant service, one that should expand people's access to imagery, while saving them time and effort.

One of the biggest problems with providing services for image and video databases is that, unlike text or numerical data, pictures cannot be easily indexed – there is no alphabetical or numerical order for most images. This lack of an order greatly complicates the problem of organizing visual information. Our approach to this problem is to design vision and pattern recognition tools that learn descriptions of image and video contents from users of the content. These tools help group similar regions under user-provided labels, group similar shots and scenes together, and identify prototypical shots. The tools also learn cumulatively from one or more users, becoming smarter along the way.

Although tools which can "see" and "understand" the content of digital images and video are still in their infancy, they are now at the point where they can provide substantial assistance to users in digital library tasks such as browsing, retrieval and annotation. However, this is just the beginning, and significant research hurdles remain before the systems will approximate human abilities to understand and describe scene content.

1.1 Visual languages

One of the key problems with automating the description of pictures is that there is no general "visual language" for describing an image. When computers listen to speech there is an a priori language, with associated limited vocabulary and syntax. If a picture were really "worth a thousand words," i.e able to be uniquely described by those words, then image retrieval would be relatively easy for computers: form an index by compressing the thousand words (which occupy far fewer bytes than most pictures), and apply existing text-based query methods, including an online text thesaurus such as Wordnet [1]. But, this presumes a solution to the problem of generating the best set of a thousand words; which words uniquely describe the picture, and who will decide what they should be for all pictures in the world?

Although progress is being made with computer vision tools to assist in annotation [2], [3], [4], the choice of the right words for a picture is still up to an individual. The words are domain-dependent, knowledge-dependent, and may also depend on subjective influences or visual associations. One picture might be validly annotated as, "a group of skydivers are forming a star pattern in the sky," and as "a hundred people wearing helmets and brightly-colored suits are holding each other's arms and legs in a giant formation in the sky." Visual patterns and textures often lack a vocabulary. Supplying one long text annotation is not a complete solution to the visual retrieval problem.

Very few imagery domains have an associated visual language, but where it exists at all, it should be used to simplify the retrieval problem. In sports such as football, there is a language of players, their positions and their plays. This high-level language, coupled with computer vision techniques, can be used for example to simplify the retrieval of similar plays from digital video, an important aid in analyzing successful games for improving winning strategies [5]. Another domain where there is syntax is photography. Romer [6] has described a number of useful syntactical components which occur repeatedly in photographs, such as horizontal structure (e.g. sunset photos), or aerial view (e.g. looking down from high buildings or from airplanes). Different applications also define visual structures that are important – e.g. location of smooth open spaces determines where an advertiser can overlay text on the photo. When an application relies on a special vocabulary, it is smart to exploit this in indexing and retrieval tools.

The problem addressed in this paper goes beyond extraction of domain specific descriptions, and therefore cannot rely upon the simplification of a domain-specific language. This paper addresses the construction of a new domain-independent tool, the "visual thesaurus," which helps group visual similarities much like a text thesaurus helps group semantically similar words. In the next section, I'll describe the idea of a visual thesaurus, and follow that in Section 3 with a description of our latest results toward this goal. Section 4 follows with some closing remarks and extensions of these ideas to perceptual thesauri, and combinations thereof.

2 A visual thesaurus

A thesaurus traditionally helps with retrieval when a query uses a different vocabulary than a stored item; for example, the text query "find a house with a big lawn" can get matched via a text thesaurus to a picture annotated as "a house, lots of grass." A text thesaurus is an important part of a retrieval system for annotated multimedia data. The goal of a visual thesaurus is not to replace it, but to augment it in important ways specific to vision, where language fails. Additionally, the visual thesaurus aids in annotation by helping find visually similar regions that should have the same labels. Ideally, a text thesaurus, visual thesaurus, and other tools (mentioned below) would work together in intelligent multimedia information retrieval.

2.1 Three types of relationships

In a traditional text thesaurus, each term represents a concept which is related to the other concepts usually in one of three ways: equivalence, (synonymous or nearly so, e.g. award and accolade), hierarchical (broader or narrower, e.g. an Oscar is an award by the Academy of Motion Picture Arts and Sciences), and associative (similar conceptually, but not hierarchical or synonymous, e.g. celebration tends to co-occur with award). In specific thesauri, these kinds of relationships are often detailed further. The most flexible systems allow user-defined relationships; these permit arbitrary associations which may encode subjective information such as "look humorous juxtaposed."

What do these three kinds of relationships mean visually? This is a new question, one which is only beginning to be understood. I will propose some answers here, not claiming these to be an exhaustive list.

2.1.1 Visual synonyms

Consider a video taken by a skydiver falling out of an airplane from 10,000 feet (camera mounted on her head.) Until she reaches about 2000 feet and opens her parachute, the video frames look essentially the same– there is some change around the boundaries, but there is little difference from frame to frame; the slight zoom is only perceivable between frames spaced hundreds apart. Each frame, although all its pixels may have changed, can be said to be "visually synonymous" with the frame that came before.

A gradual zoom or pan of the camera results in frames that are visual synonyms, where each frame looks essentially the same as its neighboring frame. People are surprisingly good at not discriminating such small viewpoint changes, even when there are strong perspective effects. For example, if asked to sketch a tall building, most people will draw its (usually rectangular) shape with vertical lines and equally spaced windows going up the building. However, in practice, a building is usually viewed with perspective from the ground looking up; hence the vertical lines will tilt toward each other at the top, and the windows will "chirp," moving closer together toward the top, as can be seen in Figure 1. (The word "chirp" comes from the audio equivalent, where the frequency increases with respect to time.) Although the latter is the way people usually see a tall building, they have to be taught to draw it this way; the human visual system appears to effortlessly "undo perspective," seeing images of different perspectives as the same image, as if they were interchangeable – perhaps visual synonym replacement.

This type of perspective visual synonym can currently be identified by a computer under two precise conditions: 1) the camera[1] is at a fixed center of projection, and allowed to pan, tilt, zoom, and rotate about that center; 2) the camera is allowed to do all of the above and move its center of projection, but the image can only contain a planar patch (or nearly so, e.g. an aerial image). In these two cases, all the photos taken by the camera essentially lie in the same "orbit of the projective group" so that they are related by a simple coordinate transformation [7].

Visual synonyms can also occur with patterns, colors, shapes, and textures, including motion patterns or temporal textures [8], [9]. An arrangement of chairs at an outdoor wedding viewed from above may have the same pattern as rows of hedges and flowerbeds in a formal garden. A crowd of people pouring out of a stadium exhibits motion flow similar to candies flowing down a chute in a candy factory. These are examples of events that are similar *visually*, not necessarily semantically. They can be grouped by visual features such as color, shape, and spatiotemporal texture.

2.1.2 Visual hyponyms, hypernymns, meronyms, and holonyms

Hierarchical relations are also found in digital thesauri; for example, a hypernym of "book" is "publication;" a hyponym is "tradebook." A book also has part "binding" which is a meronym, and the book itself is a type of "textual matter" which is a holonym. These four relations are very useful semantically, but it is an open question what their visual counterparts would be, or if they exist *visually* in all cases. Of course objects having these relationships can be seen in images – but their existence is usually semantic, not visual. To the extent the hierarchical relations are visual and not semantic, they are by nature difficult to describe semantically.

One case where there is a useful visual counterpart to the hierarchical text relations is in looking at an item over different scales, especially where one does not have a vocabulary to define what is seen at each scale. Consider a picture of a tree taken from several distances. At one scale (distant) the bark looks smooth and brown. A little closer and the bark might look like a flow-pattern, perhaps reminiscent of the ripple patterns made in sand as the ocean washes over it. Up closer still, the texture is rough and pitted, and may remind you of similarly rough surfaces. Similarly, the pattern on a brick and the periodic structure

[1] Assumed for these two cases to be an ideal pinhole camera pointed at a static scene.

Figure 1: Left: Building photo taken by camera. Right: Same photo digitally "de-chirped" to appear as a human would tend to draw it.

of a brick wall are different visually, and their hierarchical co-occurrence characterizes a brick wall over several scales.

The "type of" (holonym) and other hierarchical relations also may pertain when an image reminds you of part of another image, and you don't have words to describe either one, but your visual system still recognizes the relation. For example, showing a tiny piece of a famous painter's painting to an art student might evoke recognition of the painter, even if the piece is unidentifiable semantically, and even if the student has never seen that painting. If the visual style is captured in the piece, it is recognized as belonging to that painter, a type of their work. This phenomenon also occurs in audition when people hear a snippet of music that they've never heard before, but recognize its composer or performer. A possible explanation is the presence of a perceptual signature in the signal (for example, some textural characteristics), which you can associate with the author's other works, but for which you have no vocabulary.

2.1.3 Other visual associations

The word "book" has semantically associated words, such as printing, author, and publisher. Visual associations can also occur with great variety. As mentioned, patterns, textures, and artistic style can lead to visual associations, even when there is no semantic association. Sometimes these associations can be very specific – for example, an advertiser browsing a database might request images with strong directional lines – perhaps long dark shadows on a snowy hill, rows of telephone lines, or entirely different contents, as long as they evoke the right visual effect. One might also want to link pictures taken under different illuminating conditions – outdoors as the sun moves, indoors as different lights are switched on, or as camera shutter speed and film speed are varied. Links could be set for any user-defined visual effect that relates pictures, despite their otherwise unrelated contents.

Filmmakers know that visual relations between images can be exploited to arouse their audience; for example, the MTV style of fast-scene changes with lots of brightness variations is an attention-arousal mechanism. Associations such as "similar brightness levels" would help narrow down choices for smooth scene transitions.

Motion patterns, scene-change rhythms, and other visual effects can also lead to unusual associations. The Doublemint gum TV commercials exploit a visual relationship which could be named "double-ness" by showing twins synchronously swinging, kicking, running, and engaged in visual motions that have no association with gum-chewing, but which nonetheless are associated by their visual double-ness.

3 Constructing a visual thesaurus

A visual thesaurus is basically a collection of groupings of spatiotemporal data. Each grouping implies a kind of visual equivalence, and there may be additional visual relations connecting the groupings. I have described some of the possible groupings and visual relations in the previous section. In this section, I describe how the thesaurus can be assembled using a learning system, and the progress made to date with this fundamental problem.

3.1 FourEyes

The problem of learning groupings is a difficult one, no easier than the basic learning problem known to the AI and pattern recognition communities. Our latest efforts in this area appear in [4], a paper describing the new FourEyes system which learns groupings in interactive-time based on positive and negative visual examples provided by a user. Although the reader should turn to that paper for details, I will highlight the key features of FourEyes here, and new

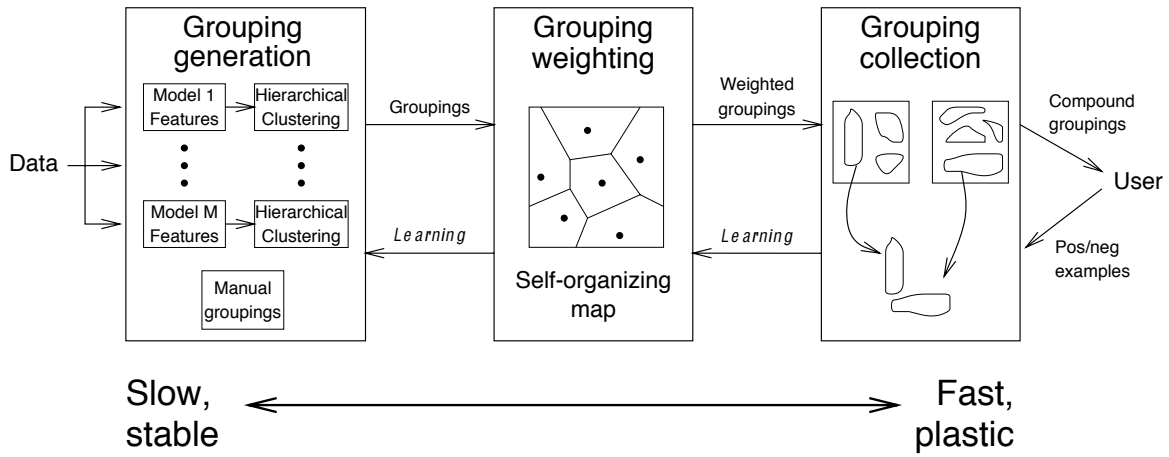

Figure 2: Three stage learning system of "FourEyes." The arrow at the bottom describes the rate at which the three stages learn.

to this paper, discuss how FourEyes contributes toward building a visual thesaurus.

3.1.1 Society of models

The structure of FourEyes is shown in Figure 2. Stage 1 at the left generates groupings of image information based on a society of models. For example, it might have shape, texture, color, motion, and position models which compute features for every patch in every image (this is done off-line). The features may also be non-parametric, perhaps provided by a user. In Stage 1 the features are used to group similar images patches together. For example, a shape model might group a cloud patch with car and cigar patches; a texture model might group clouds with cotton or smoke. Even a text thesaurus can be plugged in as a model if patches have been annotated; its features (associative links) could group a labeled cloud patch with labeled smoke or water patches. Groupings can be contained in others, expressing hierarchical meronymy or holonymy.

The groupings allowed in FourEyes are more general than just visual groupings. In FourEyes, the groupings can be anything the user wants them to be – the user could provide arbitrary groupings directly to Stage 1. In this sense, FourEyes goes beyond what is needed for a visual thesaurus. I will touch on the importance of this again later, but for the scope of this paper, I will continue with examples focusing on the visual groupings.

3.1.2 Learning in FourEyes

Unlike most database systems where the user has to specify how "much" color, texture, shape and other features to use, FourEyes automatically learns which features are most relevant, based on the user's examples. Most of this learning occurs in interactive time, in Stages 2 and 3 of Figure 2.

Consider applications where the user might be trying to annotate data, or might be trying to find other images (or regions of images) that have particular contents or qualities. In either type of application, the interaction between the user and FourEyes is basically the same. The user clicks on regions of images to indicate a set of positive and negative examples, and may (if annotating) provide labels for the positive examples. The FourEyes system adaptively computes weights for all the groupings generated by all the models, and then combines the ones which best match the user's positive examples, without including negative examples. (Criteria for "best" and other details are given in [4]). In this way, FourEyes implicitly chooses the most relevant features, combining features from multiple models if that gives the best result.

The system does not just adapt one set of weights (as is the case in most neural net learning systems) but allows for multiple weighting schemes (Stage 2) which are currently clustered by a self-organizing map (SOM). Each point in the SOM is a vector of weightings on all the groups. Different units in the SOM correspond to significantly different weighting schemes. After the learner in Stage 3 collects combinations of weighted groupings and learns which the user likes best, it enhances the winning weights in Stage 2. This feedback is similar to that[2] which inspired Werbos in creating back-propagation [10], although the mathematical update rule here is different. Thus, Stages 2-3 are necessary for the multiple models to form a society, interacting to give more powerful and efficient descriptions than any one model can provide, and learning as they interact.

What the user sees during all of this is the regions they've selected and possibly labeled, and the highlighted image data retrieved by the system. All the groupings, weightings, and learning processes are otherwise transparent to the user. It runs in interactive time.

3.2 From FourEyes to a visual thesaurus

FourEyes is close to being able to construct a visual thesaurus. Stage 1 does more than necessary since it allows visual as well as other groupings. Visual synonyms may be grouped either in Stage 1, if there is a model (such as the video orbits [7]) that can group them, or in the latter units, if a combination of models is required. Stage 2 weights combinations of groupings that serve together in

[2] Werbos was actually inspired by Freud's idea of cathexis, a feedback of emotional energy.

useful ways discovered by Stage 3. All that is needed for FourEyes to create a visual thesaurus is a restriction of Stage 1 to only visual groupings, and an addition to Stage 2 of describing specific relationships among the groupings.

Currently, Stage 2 provides part of this functionality. One unit of the SOM in Stage 2 might weight highest those groupings which correspond to "unusually bright regions." Another unit might weight highest those groupings of "high-contrast regular patterns." In so doing, a unit is combining different groupings of Stage 1 under a new label.

Although the current system does not explicitly label the associations for each unit in the SOM, this is a minor addition if the labels are provided by the user. In some cases, where model features have semantic associations, e.g, the features in the Wold model correspond to the adjectives of periodicity, directionality, and randomness [11], then the labels might be inferred from the model features directly.

The only addition remaining to make a visual thesaurus using FourEyes is to allow directed associations for the hierarchical relations mentioned above (Section 2.1.2). So far, the hierarchical visual associations appear to be the least important of the three types of relations employed by thesauri. The best way to add them to FourEyes has yet to be determined; it may require allowing directional links between the groupings (currently the links are undirected) or it may be able to be handled via other methods such as hierarchical structures on the groupings, some of which are already in place.

An advantage of building a visual thesaurus with FourEyes is it gives many important features automatically. For example, weights on links arise in Stage 2, allowing more useful links to receive higher values and consequently be found faster. This importance of weighted links has been argued by others; for example, Gao *et al.* have developed a "fuzzy" text thesaurus for help retrieving trademark images [12]. The FourEyes-based visual thesaurus automatically provides this multiple-membership advantage. FourEyes also automatically updates the weightings and groupings as users interact with it; it thus accumulates new knowledge.

The FourEyes advantage also extends to the third category of associative relations, which includes user-defined relations and subjective associations. A designer, for example, might want to annotate "attractive combination" associations among fabric patterns. A comedian might want to annotate unattractive associations for amusement. Arbitrary associations are possible with FourEyes due to its abilities to incorporate user-defined groupings and to learn.

3.2.1 A clarification: combining text, visual, and other thesauri

A text thesaurus is a relatively new part of image and video retrieval systems, but is only useful with annotated data. Systems such as FourEyes save the user time in making annotations, and can work with a text thesaurus for combining synonymous annotations.

Additionally, FourEyes goes beyond this traditional use of consulting a text thesaurus. As mentioned above, word similarities can be directly included in the cluster-generation of Stage 1. An image patch can be clustered with other patches based on semantic content, and with still others based on color, shape, texture, or other visual features. Audio and other sensory features can be similarly combined, especially in the SOM of Stage 2, where one might find a unit that favors visual and auditory groupings, e.g. "dark scenes with rumbling sounds."

4 Concluding remarks

This paper has described the idea of a visual thesaurus, a tool for recognizing visually similar events, "visual synonyms" using color, texture, pattern, motion, and other user-defined features. Such a tool would augment existing text thesaurus tools, allowing for more powerful image and video retrieval systems. Construction of the thesaurus is described, based on our existing FourEyes system, which learns similarity groupings through interacting with people as they are using image and video databases. The combination of text, visual, audio, and other perceptual thesauri is also discussed; the proposed system could easily combine these, facilitating cross-modal associations in multimedia databases.

4.1 Perceptual thesauri

It is important to add that a visual thesaurus should not operate solo, but should be combined with other tools such as a text thesaurus and an audio thesaurus. The combination is important, as people do not naturally separate associations during retrieval. Consider one of many possible paths after hearing a train whistle: you might associate it with a train, then with train tracks, then with a railroad bridge, then with interesting bridge designs, steel lattices, garden trellises and perhaps even trellis algorithms. The associations in this particular example are by audio (train whistle), by text (train tracks), by text (railroad bridge), by vision (bridge designs with steel lattices), by vision (garden trellises), and then by text again (trellis algorithms). The human shifts between different perceptual modes when connecting concepts.

Thesauri could exist not just for vision but also for all the other sensory domains – haptics, taste, olfaction, and audio. For instance, there is a huge industry associated with olfactory/taste vocabularies, e.g. perfumes, wines, cleaning products, food products, even the "new-car" smell. As computers become equipped with "artificial noses" [13], they can construct an olfactory thesaurus that allows one to compare similar odors, and retrieve products with those odors. The most progress on perceptual similarities appears to have been done with retrieval of similar audio patterns (e.g. [14], [15]), which provide an important aid to musicians and sound effects artists.

Note that the same models may be used for different senses. After all, one human brain processes all the perceptions, so that re-use of a descriptive model for vision and audition would suggest some efficiency in the brain. The work of [15] demonstrates this cross-over by successfully using a model for audio patterns that was previously used successfully for visual patterns [16].

The society of models approach used in FourEyes allows the same or different models to be combined into one set of groupings. This effortless mingling of cross-sensory groupings is a feature shared by the human brain. Thus, the perceptual thesauri do not need to be separate tools, but can co-exist in the single structure of Figure 2 as one giant representation of perceptual knowledge.

4.2 A brief note on common *sense*

Common sense, and how it can be learned by computers, is a perplexing and difficult topic. Most effort to construct common sense learning systems has been in the AI community, using language and rules, not using vision, audition, and the other "senses" common to humans. Although a set of perceptual thesauri is not equivalent to a common sense system, their acquisition bears a strong resemblance. Although the details will be left for a future publication, let me suggest here that the construction process described in this paper for perceptual thesauri parallels the construction of a common sense database, with emphasis on common *sensory* learning. At the root of common sense is expectations that certain things occur together (e.g., sky or tall buildings behind treetops) and that common associations are shared by most people. As the thesaurus gathers these associations and their relations, it builds up a store of knowledge that people largely take for granted. This "common perceptual knowledge" forms a sensory complement to the rule-based common sense systems being developed today.

Acknowledgements

I first heard the name "visual thesaurus" from Donna Romer, although she indicates the name has been suggested by others with different proposed incarnations. Figure 1 was made by Steve Mann who also provided helpful remarks on a draft of this paper. The work of FourEyes is the thesis research of Tom Minka, who has inspired several of my ideas in this paper through our numerous discussions. Tom also made Figure 2 and contributed helpful comments on this paper. I would like to thank HP Labs, BT PLC, and NEC, for their sponsorship of this research.

References

[1] "Wordnet," 1995. http://www.cogsci.princeton.edu/~wn/.

[2] R. W. Picard and T. P. Minka, "Vision texture for annotation," *Journal of Multimedia Systems*, vol. 3, pp. 3–14, 1995.

[3] E. Saber, A. M. Tekalp, R. Eschbach, and K. Knox, "Annotation of natural scenes using adaptive color segmentation," *IS&T/SPIE Electronic Imaging*, Feb. 1995. San Jose, CA.

[4] T. P. Minka and R. W. Picard, "Interactive learning using a 'society of models'," *Submitted for Publication*, 1995. Also appears as MIT Media Lab Perceptual Computing TR#349.

[5] S. Intille and A. Bobick, "Exploiting contextual information for tracking by using closed-worlds," in *Proceedings of the Workshop on Context-based Vision*, (Cambridge, MA), pp. 87–98, June 1995.

[6] D. Romer, "The Kodak picture exchange," April 1995. seminar at MIT Media Lab.

[7] S. Mann and R. Picard, "Video orbits of the projective group: A new perspective on image mosaicing," *Submitted for Publication*, 1995. Also appears as MIT Media Lab Perceptual Computing TR#338.

[8] R. Polana and R. C. Nelson, "Recognition of motion from temporal texture," in *Proceedings CVPR '92* (C. Harris, ed.), (Champaign, IL), pp. 129–134, Computer Vision and Pattern Recognition, IEEE Computer Society Press, June 1992.

[9] M. Szummer, "Temporal texture modeling," Master's thesis, MIT, Cambridge, MA, May 1995.

[10] P. Werbos, "The brain as a neurocontroller: New hypotheses and new experimental possibilities," in *Origins: Brain and Self-Organization* (K. H. Pribram, ed.), Erlbaum, 1994.

[11] F. Liu and R. W. Picard, "Periodicity, directionality, and randomness: Wold features for perceptual pattern recognition," in *Proc. Int. Conf. Pat. Rec.*, vol. II, (Jerusalem, Israel), pp. 184–185, Oct. 1994.

[12] Y. J. Gao, J. J. Lim, and A. D. Narasimhalu, "Fuzzy multilinkage thesaurus builder in multimedia information systems," 1995. Institute of Systems, Science, National University of Singapore.

[13] I. Hunter, 1995. Personal Communication.

[14] F. Matsumoto, "Using simple controls to manipulate complex objects: Application to the drum-boy interactive percussion system," Master's thesis, MIT, Cambridge, MA, Sept. 1993.

[15] N. Saint-Arnaud, "Classification of sound textures," Master's thesis, MIT, Cambridge, MA, September 1995.

[16] K. Popat and R. W. Picard, "Novel cluster-based probability models for texture synthesis, classification, and compression," in *Proc. SPIE Visual Communication and Image Proc.*, vol. 2094, (Boston), pp. 756–768, Nov. 1993.