IEEE Workshop on Content-Based Access of Image and Video Libraries, 76-81, June, 1997.

Hidden Annotation in Content Based Image Retrieval

Ingemar J. Cox, Joumana Ghosn, Matt L. Miller
Thomas V. Papathomas, Peter N. Yianilos*

Abstract

The Bayesian relevance-feedback approach introduced with the PicHunter system [5] is extended to include hidden semantic attributes. The general approach is motivated and experimental results are presented that demonstrate significant reductions in search times (28-32%) using these annotations.

1 Introduction

Systems that retrieve images based on their *content* must in some way codify these images so that judgments and inferences may be made in a systematic fashion. The ultimate encoding would somehow capture an image's semantic content in a way that corresponds well to human interpretation. By contrast, the simplest encoding consists of the image's raw pixel values. Intermediate between these two extremes is a spectrum of possibilities, with most work in the area focusing on *low level features*, i.e. straightforward functions of the raw pixel values (see [13, 15, 3, 4, 6, 9, 8, 10] and many others [11, 16, 17, 18]). Some such features, such as color, begin to capture an image's semantics, but at best they represent a dim reflection of the image's true meaning.

The ultimate success of content based image retrieval systems will likely depend on the discovery of effective and practical approaches at a much higher level. In this paper we report conceptual and experimental progress towards this objective.

Any attempt to codify image semantics inevitably leads to design of a language with which to express them. If a human operator is required to formulate a query using this language, and interpret a database image's description in terms of the language, two serious problems arise. First, the language must not only be effective in theory, but must also serve as a natural tool with which a human can express a query. Second, inaccurate or inconsistent expression of each database image in terms of the language can lead to confusion on the part of the user, and ultimately undermine the effectiveness of, and confidence in, the system. The need for accurate and consistent expression can also limit the language's design.

For these reasons we are led to study *hidden languages* for semantic encoding, and in particular hidden boolean attributes affixed to each database image.

Our ability to follow this research direction is made possible by the general navigational paradigm introduced in [5] and used by the PicHunter image retrieval system. (see [12] for other learning-based work). With this approach a user navigates through a database by selecting similar images from the set currently displayed. No explicit query is formulated. Instead, the system chooses the next display set based on the user's earlier selections. All earlier selections influence the system's next choice – not just the most recent user response. This takes place within a simple Bayesian relevance feedback framework in which the system learns to evaluate the probability that an image is the user's target given his actions, by instead learning to predict these actions conditioned on a presumptive target – starting from a uniform prior.

Thus the focus is shifted entirely to the task of learning a predictive model to explain the users selections. The significance of this shift is that this model can rely on information beyond that which the user sees. In particular, the system's model can rely on hidden attributes affixed to each image.

As a result, we are free to consider attribute schemes that might not work well in a traditional non-hidden approach. We might, for example, use a scheme that employs 10,000 attributes, far more than a human operator could reasonably be expected to deal with. Moreover some of these attributes might correspond to complex semantic concepts that are not easily explained, or to overlapping concepts that do not fit well into the kind of hierarchies that humans frequently prefer. They might even include entirely artificial attributes that arise from a machine learning algorithm. Because the attributes are hidden, it may be that the system performs well despite considerable

*The first, third, and fifth authors are with NEC Research Institute, 4 Independence Way, Princeton, NJ 08540. The second is with the University of Montreal, Department of Computer Science. The fourth is with the Rutgers University Department of Biomedical Engineering. The fifth is also with the Princeton University Department of Computer Science. Direct Email to the fifth author at pny@research.nj.nec.com.

error in the assignment of attributes. For this reason we are free to consider attributes even if their proper identification seems very difficult.

The overall implementation of a hidden attribute approach may be divided into two components: the design of the schema of attributes, and the approach taken to assigning attribute values to each database image. Both of these contain rich opportunities for future work. PicHunter's [5] use of low-level image statistics may be viewed as a hidden attribute approach. This paper represents a first step intended to help establish the general approach's potential at a higher semantic level by focusing on a particularly simple case.

A set of approximately 125 semantic attributes was chosen and values were assigned manually to each image in our experimental database. In some sense this might be viewed as a best-case scenario since the schema is hand-designed, and the values are assigned by humans. However some existing commercial collections of images include such schemes and annotations, so beyond providing justification for future work, our positive experimental results may be of immediate practical significance.

We remark that there are errors and inconsistencies even in attributes assigned by humans. Here, the fact that the attribute values are hidden can result in more robust performance in the presence of error. We also observe that in some settings, such as the emerging area of Internet Web publication, authors are implicitly annotating their images by their choice of text to surround them. Exploiting this textual proximity represents an immediate and interesting direction for future work. This general direction is explored in [14, 1].

It is not clear how *high* in the semantic sense our approach of hidden attributes might reach. It is certainly conceivable that a large portion of an image's semantic content might be captured by a sufficiently large and rich collection of attributes – obviating the need to produce a single succinct and coherent expression of an image's meaning.

Section 2 of this paper describes our set of attributes, the manner in which their values were assigned, and other aspects of the experimental setup. Section 3 summarizes the results. In section 4 final remarks are made regarding these experiments and broader issues as well.

2 Experimental Design

Our experiments compare the performance of PicHunter based on low-level non-verbal features only (such as color content, contrast, brightness, edge content, etc.), with a new version that incorporates a vector of verbal semantic attributes (such as "sky", "hill", "person", "city", "bird", etc.).

A system of approximately 125 keywords was identified based on knowledge of our experimental database of 1,500 images. Each image was then visually examined and all relevant keywords identified. An additional set of *category* keywords were then assigned automatically. For example, the "lion" attribute causes the category attribute "animal" to be present. Altogether there are 134 attributes. These supplement the 18 low-level features used by the basic PicHunter version, and described in [5]. The 134 semantic attributes are regarded as a boolean vector, and normalized Hamming distance combines their influence to form, in effect, a 19th PicHunter feature.

The PicHunter user interface is particularly Spartan. Nine candidate images are displayed along with three buttons used to abort the search, signal that the search is successful, or request that the system display another nine candidates. Prior to requesting additional candidates the user selects the subset of the nine visible images that he/she regards as most similar to the target image.

Our experiments implement the *target testing* model of [5] in which the user seeks to locate a given target under the user interface described above. Performance is measured by the number of display iterations required to locate the target image. That is, how many nine-image displays the system had to present before the sought after image appeared.

The primary purpose of our experiments is to compare the performance of the original version of PicHunter and an annotated version. The secondary goal is to examine whether user performance improves after the user receives an explanation of the particular features in use. For notational purposes, we refer to the original version as "Orig." The version using semantic attributes is denoted "Sem." The experimental step consisting of explaining a feature set to the user is denoted "Expl."

All experiments were conducted on 1280x1024-pixel color monitors, driven by Silicon Graphics Indigo2 workstations. The monitor screen measured 38 cm by 29 cm, and was viewed from a distance of 70 cm. Individual images were either in "portrait" (4.83 x 7.25 cm on the screen) or in "landscape" (7.25 x 4.83 cm) format. They were padded with dark pixels either horizontally or vertically to form square icons. The images in the database [2] were copied from a set of CDs by Corel Inc., each CD containing 100 images. Each image is referred to by its unique identification number, which is denoted by "ID" in this paper.

Eight users, labeled A to H, participated in this experiment. Users were tested for color blindness using Ishihara test plates and found to have normal color vision. All users also had normal or corrected-to-normal vision with regard to acuity.

There were two major phases in this experiment. Each phase involved the same 17 images that users had to converge to. In the first phase, users were told to select images that they thought were similar to the target, without being told what to base their judgment of similarity upon. There were two groups of four users in this first phase. The first group, G1=A,B,C,D, was subjected first to the original ("standard") PicHunter, and then to the semantic ("word") version, while this order was reversed for the other group, G2=E,F,G,H. Before embarking on the second phase, users were divided in two new groups of four, G3 and G4, to balance performances, based on their performances in the first phase. Toward this goal, we first constrained the new groups so that each had exactly two users from each of G1 and G2, to balance previous exposure. Second, among all the partitions that were constrained as above, we selected one that resulted in two new groups which differed as little as possible with respect to their mean group performance and with respect to the standard deviation around that mean. Thus, the new groups were G3=1,2,5,6 and G4=3,4,7,8, where 1,2,3,4 and 5,6,7,8 are permutations of A,B,C,D and E,F,G,H, respectively, that minimized differences of means and standard deviations between G3 and G4.

Subsequently, the second phase consisted of first giving each individual instructions for judging image similarity, based on the algorithm's user model, and then letting them go through the picture search process, as before. Both the original and the semantic version were also used in the second phase. The sequence of versions was selected for each observer so as to obtain an overall balanced experimental design. Table 1 below shows the sequence of experiments for each observer. As can be seen from this table, pairs of users were subjected to the same experimental conditions. Phase 1 consisted of steps 1 and 2, whereas phase 2 included the rest of the steps 3-6. Again, half of the users were first subjected to the original ("standard") PicHunter, and then to the semantic ("word") version, each preceded by an explanation, while this order was reversed for the rest of the users.

Before a session with the original version of PicHunter in the second phase, users were asked to base similarity on image appearance (color, brightness, contrast, sharpness, etc.), and ignore the image semantic contents, i.e., ignore the objects, animals, people, flowers, trees, cities, buildings, etc. They were told to look at the image as if they were a machine that cannot extract any meaning from images, that has a good camera and a computer that can estimate color content, brightness, contrast, sharpness, etc., but it cannot express in words what the image contains. They were also made aware of the priority of the features in the user model, from the most to the least important, according to [5].

Similarly, before a session with the semantic version of PicHunter in the second phase, users were told to base similarity not only on image appearance, but also on image semantic contents, as one would describe them by words. In addition, they were given the list of representative semantic labels shown in Table 2, to suggest the level of semantic "resolution".

Ob.	1	2	3	4	5	6
1,2	Orig	Sem	Expl	Orig	Expl	Sem
3,4	Orig	Sem	Expl	Sem	Expl	Orig
1,2	Sem	Orig	Expl	Orig	Expl	Sem
3,4	Sem	Orig	Expl	Sem	Expl	Orig

Table 1: Sequence of experimental steps

sky	cloud	ground
tree	one subject	two subjects
many subjects	aircraft	person
water	horse	lion
snow	sand	animal
rodent	arch	church
bicycle	field	shoe
Japan	Africa	woods
art	painting	umbrella
city	boat	night
interior	wall	autumn
mountain	close up	green grass
eagle	child	house
fish	pillar	rodent

Table 2: Representative semantic labels provided to users as explanation before they run the semantic version of PicHunter in steps 4 and 6

3 Experimental Results

Our experimental results are given in tables 3 and 4. Rows correspond to target images $(1-17)$. Columns correspond to the 8 users. Each matrix entry in position (T, u) is the number of 9-image displays it took the user corresponding to column u to converge to the target corresponding to row T. The smaller the entry the better the performance was for the corresponding row-column combination. For each matrix, we provide the row and column sums, aligned with corresponding rows and columns. The inverse of a given row's sum indicates how well observers performed collectively for that row's image; similarly, the inverse of a column sum is a measure of the corresponding user's performance across all the images. We also show the sum of all the matrix elements as a figure of merit for the collective users' performance across all images under the conditions represented by the given matrix.

Although the experiments were designed with PicHunter in mind, their results can be applied to any image retrieval system and, more generally, to any system that involves judgment of image similarity by humans.

Searching the database linearly until the desired image is located requires $0.5 \cdot 1500/9 \approx 83$ 9-image displays. It is apparent that the table entries are in almost all cases much smaller than this. Moreover, the reduction in search times with the introduction of hidden semantic attributes (32% and 28%) is immediately apparent – and significant as verified by analysis of variance.

4 Concluding Remarks

It is clear that humans pay a lot of attention to semantic content when judging image similarity – but the criteria used and the nature of the composite judgment is complex indeed. All eight users were interviewed by one of the authors following completion of the experiments. In addition, eleven other users participated in shorter PicHunter searches and related pilot studies. Without exception all reported that semantic features played a key role in their judgment. For this reason we are not surprised that performance with the annotated version of PicHunter is superior to that of the non-semantic version.

Semantically annotated images are appearing in structured environments such as medical image databases, news organization archives – and the trend seems to extend to generic electronic collections. In addition to using these annotations in a hidden fashion, mature image search systems may be hybrids that include an explicit query mechanism that corresponds to the space of available annotations. Even in query-based systems learning may play a role as illustrated by related work in the field of textual information retrieval [7].

Finally, the issue of feature relevancy must be addressed. In observing the 8 users' strategies in Experiment 4, we observed that test images were sometimes selected because of similarity with the target in terms of, say, color ("it has as much blue as the target"),

Original PicHunter									
ID	User's Initials								Sum
	BZ	CE	MS	JP	KD	RN	RA	TC	
1	19	5	7	5	7	5	7	27	82
2	7	9	9	5	8	10	5	7	60
3	7	11	7	14	12	14	6	14	85
4	11	9	2	5	9	9	8	9	62
5	3	3	10	4	3	8	3	3	37
6	12	39	4	35	37	24	19	17	187
7	31	8	27	13	64	19	7	12	181
8	6	4	4	4	6	11	11	6	52
9	15	12	30	11	11	28	11	28	146
10	26	17	13	11	28	21	4	17	137
11	58	55	70	20	13	34	51	58	359
12	38	31	64	37	32	34	27	9	272
13	36	5	5	11	39	10	16	28	150
14	30	10	12	7	11	12	11	6	99
15	8	11	26	10	11	9	8	7	90
16	23	34	41	90	40	12	21	32	293
17	12	2	2	3	2	2	2	2	27
Σ	342	265	333	285	333	262	217	282	**2319**
Semantic PicHunter									
1	5	5	5	5	6	5	11	4	46
2	8	12	6	7	8	11	7	7	66
3	11	11	15	7	10	9	7	4	74
4	2	2	17	9	2	18	2	3	55
5	3	3	3	3	3	4	3	3	25
6	20	6	17	7	17	24	11	8	110
7	6	7	16	9	7	12	6	5	68
8	5	6	4	5	7	5	4	11	47
9	7	9	6	7	15	33	24	12	113
10	10	8	10	20	23	15	12	15	113
11	82	26	15	19	46	33	28	30	279
12	16	16	13	36	14	17	18	5	135
13	20	8	14	14	5	13	16	8	98
14	20	9	38	11	34	7	9	14	142
15	7	5	19	4	5	7	9	25	81
16	15	21	15	19	7	5	8	30	120
17	2	2	2	2	2	2	2	2	16
Σ	239	156	215	184	211	220	177	186	**1588**

Table 3: Results for both PicHunter versions where subjects received no explanation of the feature sets involved. Notice that the number (1588) of overall trials using semantic attributes is 32% smaller that the number (2319) of trials using the basic system.

Original PicHunter with Subject Explanations									
ID	User's Initials								Sum
	BZ	CE	MS	JP	KD	RN	RA	TC	
1	5	10	4	5	4	7	8	4	47
2	9	10	11	8	10	6	11	7	72
3	9	9	6	20	6	14	6	5	75
4	12	10	26	7	14	10	15	6	100
5	3	3	3	3	3	3	3	3	24
6	25	16	7	15	15	36	43	8	165
7	11	15	9	12	17	46	4	14	128
8	8	5	6	4	6	17	7	9	62
9	12	18	9	18	6	9	27	10	109
10	14	12	15	10	11	13	13	19	107
11	19	18	20	22	23	24	29	15	170
12	99	23	13	14	18	30	16	18	231
13	22	8	11	12	19	8	8	12	100
14	6	17	11	9	8	10	8	13	82
15	7	11	11	7	12	7	9	16	80
16	47	18	47	42	19	19	16	28	236
17	2	2	2	2	2	2	2	2	16
Σ	310	205	211	210	193	261	225	189	**1804**
Semantic PicHunter with Subject Explanations									
1	5	5	5	4	5	5	7	6	42
2	4	4	10	7	4	6	16	4	55
3	5	5	9	5	6	10	4	7	51
4	2	2	2	2	2	2	2	2	16
5	3	3	3	3	3	3	3	3	24
6	14	8	7	7	8	11	6	5	66
7	5	9	7	15	6	6	13	4	65
8	6	4	5	9	7	4	4	13	52
9	7	6	5	6	8	18	13	10	73
10	5	24	11	10	12	7	16	13	98
11	14	27	10	26	12	11	23	10	133
12	44	12	21	3	11	9	13	13	126
13	5	8	5	7	27	25	5	7	89
14	21	32	13	10	25	14	12	16	143
15	13	8	7	13	11	6	5	9	72
16	15	10	20	43	19	16	26	19	168
17	2	2	2	2	2	3	2	2	17
Σ	170	169	142	172	168	156	170	143	**1290**

Table 4: Results for both PicHunter versions where subjects did receive explanation of the feature sets involved. Notice that the number (1290) of overall trials using semantic attributes is 28% smaller that the number (1804) of trials using the basic system. Observe also that both are somewhat lower than their no-explanation counterparts.

and other times because of similarity in, say, overall brightness. To the extent that a user relies on a small number of features during a session, it may be possible to learn which are being used, and in so doing improve performance. Hybrid systems might allow explicit identification of relevant features.

Acknowledgments

The authors thank Bob Krovetz for useful discussions regarding text database search.

References

[1]

[2] Corel stock photo library. Corel Corp., Ontario, Canada.

[3] J. Barros, J. French, W. Martin, P. Kelly, and J. White. Indexing multispectral images for content-based retrieval. In *Proceedings of the 23rd AIPR Workshop on Image and Information Systems, Washington DC, Oct.*, 1994.

[4] A. D. Bimbo, P. Pala, and S. Santini. Visual image retrieval by elastic deformation of object sketches. In *Proceedings IEEE Symposium on Visual Languages*, pages 216–23, 1994.

[5] I. J. Cox, M. L. Miller, S. M. Omohundro, and P. N. Yianilos. Pichunter: Bayesian relevance feedback for image retrieval. In *Int. Conf. on Pattern Recognition*, pages 361–369, 1996.

[6] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The QBIC system. *IEEE Computer*, 28(9):23–32, 1995.

[7] D. Haines and W. B. Croft. Relevance feedback and inference networks. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieva*, 1993.

[8] M. Hirakawa and E. Jungert. An image database system facilitating icon-driven spatial information definition and retrieval. In *Proceedings 1991 IEEE Workshop on Visual Languages*, pages 192–8, 1991.

[9] K. Hirata and T. Kato. Query by visual example; content based image retrieval. In *Advances in Database Technology– EDBT '92 Sprinter-Verlag Berlin Heidelberg in Pirotte, A., Delobel, C., and Gottlob, G. (Eds.)*, 1992.

[10] T. Kato, T. Kurita, H. Shimogaki, T. Mizutori, and K. Fujimura. Cognitive view mechanism for multimedia database system. In *IMS '91 Proceedings. First International Workshop on Interoperability in Multidatabase Systems*, pages 179–86, 1991.

[11] P. Kelly and T. Cannon. Candid: Comparison algorithm for navigating digital image databases. In *Proceedings Seventh International Working Conference on Scientific and Statistical Database Management*, pages 252–8, 1994.

[12] T. P. Minka and R. W. Picard. Interactive learning using a "society of models". Technical Report 349, MIT Media Lab, 1995.

[13] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. In *SPIE Storage and Retrieval Image and Video Databases II*, volume 2185, 1994.

[14] J. R. Smith and S.-F. Chang. Searching for images and videos on the world-wide web. Columbia University CU/CTR Technical Report 459-96-25, to appear in IEEE Multimedia Magazine, 1997.

[15] H. S. Stone and C.-S. Li. Image matching by means of intensity and texture matching in the Fourier domain. In *Proc. SPIE Conf. in Image and Video Databases*, 1996.

[16] M. Stricker and M. Swain. The capacity of color histogram indexing. In *Proceedings 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 704–8, 1994.

[17] M. Swain and D. Ballard. Indexing via color histograms. In *Proceedings Third International Conference on Computer Vision*, pages 390–3, 1990.

[18] G. Yihong, Z. Hongjiang, and C. Chuan. An image database system with fast image indexing capability based on color histograms. In *Proceedings of 1994 IEEE Region 10's Ninth Annual International Conference. Theme: Frontiers of Computer Technology*, volume 1, pages 407–11, 1994.