Assessing the Filtering and Browsing Utility of Automatic Semantic Concepts for Multimedia Retrieval

Michael G. Christel
Carnegie Mellon University
Pittsburgh, PA 15213
christel@cs.cmu.edu

Milind R. Naphade, Apostol (Paul) Natsev, Jelena Tesic
IBM Thomas J. Watson Research Center
19 Skyline Drive, Hawthorne, NY 10532
{naphade, natsev, jtesic}@us.ibm.com

Abstract

The contributions of automatic semantic concept classifiers for interactive filtering (classifiers in conjunction with query rankings) and browsing (classifiers in lieu of query rankings) are tested against three test corpora: an amateur photo collection, documentary video, and news video. Results show that current classifiers offer browsing utility twice as good as having no classifier at all, and that continuous improvements in the classifiers produce comparable improvements in the browsing utility. For filtering a well-ordered set of results (e.g., a set retrieved from text search), concept classifiers need greater accuracy: current classifiers showed worse performance than not filtering at all, even when the classifiers' accuracy is nearly doubled. Results are consistent for all test corpora. Hence, automatic semantic concepts can offer significant utility for browsing at current levels of accuracy, but the requirement is much higher for filtering a well-ordered set of results, where extreme accuracy is necessary before benefits are seen.

1. Introduction

Many researchers have been developing classifiers like face, people, outdoors, and buildings to improve the representation and retrieval of information from vast multimedia collections. This paper tests the utility of such classifiers under control by a user searching for relevant image or video material.

The most familiar image search interface today is that used by Web image search engines, in which users enter keyword terms, and images are shown in a table ordered by some measure of relevance. These systems can be effective for searching for very specific items, but do not support browsing tasks well [5]. They also rely on associated text that may not document the aspect of the image of interest to the user [11]. The text will also likely be incomplete, as different people have been shown to label the same images with different words [3]. Finally, the text may not be present, especially for personal digital photo collections, as users are reluctant to invest in the time to label images with text descriptors, even when the annotation can be done through a speech interface [10]. Other systems like QBIC retrieve images based on attributes like color and texture [11], but studies have questioned the utility of image searching according to such low-level properties [5]. Digital photo collections are often navigated by the date the picture was taken [10], but for date ranges many images may need to be browsed to find the subset of interest.

Because of these issues, the nature of image retrieval today is that it is imprecise, often returning a large candidate set full of irrelevant detail. The problem is intensified in video retrieval, where an hour-long video might be decomposed into thousands of shots. These shots can each be represented by a "keyframe" image extracted from the video, and the numerous keyframes can then be subjected to image retrieval strategies.

Digital imagery retrieval can be considered a transaction sequence in which the user initiates a query or browsing action that generates a candidate set. User interaction is critical to better express the information need and generate a new, more precise candidate set. The user could filter a candidate set into a subset that drops out irrelevant images and focuses in on relevant ones. One common way to filter down imagery is through pre-classified semantic concepts such as "indoors" and "outdoors", as defined and studied in the NIST TREC video retrieval evaluation (TRECVID) forum since 2001 [8].

Many computer vision and multimedia researchers classify imagery with semantic concepts like "indoors" through machine learning and other automated approaches [7], with the hopes that these classifiers can lead to improved interactive retrieval. Past TRECVID experiments have yet to strongly validate this hope [4], perhaps because the automatic classification accuracy is not yet good enough for use as an interactive filter. The following quote from a TRECVID retrieval study is typical: "Semantic concepts' contribution to search was minor but could probably be improved by developing more accurate concept detectors" [9]. The investigation reported here directly addresses this issue: how does the accuracy of a concept detector affect its utility for filtering and browsing? We present the data used for the investigation, and summarize the results of a 2003 study with TRECVID 2002 classifiers and just filtering tasks. We utilize that study's framework and conclusions regarding interactive searching behavior to conduct the newer study reported here, examining current automated classifiers and refined classifier accuracy exploration with respect to both filtering and browsing activity.

2. Experimental data

We define three sets of data, concepts, information needs (topics), and truth sets to investigate whether concept filters are useful as follow-up actions for all sorts of topics and various types of corpora. We define topics with small to very large ranked candidate sets ranging in size from 120 images to 960, for three corpora: documentaries, news, and photos. The documentary corpus is drawn from the TRECVID 2002 set, with the news corpus taken from the TRECVID 2003 set of ABC, CNN, and C-Span news. A photo corpus is assembled from the personal collections of 3 university employees. We purposely choose personal photo collections, rather than commercial image collections as are typically used, to examine the utility of filtering against such digital photos, which have different characteristics and more closely resemble the collections of end users.

In order to focus on the characteristic of candidate set size and its relationship to concept filter accuracy, we fixed two other variables for a 2003 study with 36 students and university staff [2]. The number of correct answers in a candidate set was held constant at 10%. The distribution of correct answers was exponential in that half of the correct answers were randomly distributed among the first 20% of the candidate set, half of the remaining were randomly distributed in the next 20%, etc., down to a remainder of at least 1 in the final 20%. For 12 correct answers, the distribution across quintiles was 6/3/1/1/1, for 24 12/6/3/2/1, for 48 24/12/6/3/3 and for 96 48/24/12/6/6. We chose a 10% precision rate and this distribution based on the composition of ranked shot sets, i.e., storyboards, following reasonably good queries in TRECVID interactive query evaluations over the years. The storyboard is reasonably ranked so that more answers are found at the top, but the imprecise nature of imagery queries means that there are many irrelevant shots and images, and that correct answers may appear at the end of the storyboard ranked list. Because storyboards typically have a secondary ordering where adjacent images are related by date or coming from the same source, we likewise kept related imagery together in runs as we generated the candidate sets.

For each corpus, 4 topics are defined based on the TRECVID topics created over the years to reflect the sorts of queries real users pose [8] as shown in Table 1. These 12 topic sets contain 1405 photographs, 2451 documentary shots, and 1834 news shots for the 3 corpora, which are never mixed (e.g., news topics only draw from the 1834 news shots), limited in size compared to the complete TRECVID corpus but allowing for absolute truth as opposed to pooled truth for topics and concepts.

The independent variable for anticipated experiments is the concept classifier accuracy, which requires significant preparatory work. The 1405 photos and 4285 keyframe images for the shots were looked at individually by two independent human graders, who made binary decisions on each of 6 visual concepts, taken from TRECVID 2002 and described in Table 2: each image either has the concept or does not. The interrater reliabilities for the concepts are listed in the table. Note that perfect 100% truth is unrealistic, e.g., personal interpretation comes into play when deciding whether a face is too much in shadow, too small, or too turned to see both eyes, nose, and mouth. Given the overall high interrater reliability of 91.2%, we make use of the first rater's data for subsequent steps and reports of "truth."

Table 1. 12 topics defined with candidate sets (N = set size).

	Topic	**N**
Photo	Waterfalls with no people	120
	Road traffic	240
	Snow-capped mountains	480
	This particular adult female (shown)	960
Documentary	Leisure time at the beach or pool	120
	One or more people in the kitchen	240
	Automobiles	480
	Two-story or taller buildings	960
News	Missiles in the air or being launched	120
	This person, Pope John Paul II (shown)	240
	Airplanes on the ground or in the air	480
	People walking in an urban environment	960

The 2003 study [2] showed that for candidate set sizes of 240 and smaller, users consider the task of using storyboards to find the answers as easy and satisfying, and perform well regardless of the concept filter accuracy. Concept filtering does not come into play for these small set sizes: small candidate sets of 240 or less can indeed be navigated via the storyboard mechanism without the need for the filtering interface. Large candidate sets of 480 and greater are viewed as more difficult to navigate successfully, and for these sets concept filtering is used.

Furthermore, the accuracy of the filter directly affects the efficiency and effectiveness of the filter use. Classifiers with accuracy in the range of automatically produced classifiers from circa 2002 are not effective or well received for large candidate sets. We conducted a follow-up study for the following reasons:

- Assess whether current (2005) classifiers show improvement, and determine with greater precision the performance effects of concept classifier accuracy on *filtering*.
- Determine the contributions of concept classifiers to *browsing* activity.

For many corpora, such as digital photo collections without any date or text metadata, or a foreign broadcast news corpus with no searchable text or closed captions, there may not be a well-defined query mechanism with which to produce ranked candidate sets like those of Table 1. For example, instead of 120 candidates for waterfalls with no people, there is the whole corpus of 1405 photographs. The study reported here looks at both filtering and browsing utility for automatically derived concept classifiers.

Table 2. 6 image concepts defined for experiment data, along with interrater reliability (R), and percentage of the photo (P), documentary (D), and news (N) image sets marked by the first rater as having that particular concept.

R	Concept	Description	P	D	N
90%	Indoors	indoor location	9%	18%	35%
91%	Outdoors	outdoor location	86%	39%	38%
93%	Face	human face with 2 eyes, nose and mouth clearly visible	30%	12%	34%
91%	People	2 or more humans large enough to identify as (portions of) 2+ people	38%	19%	35%
90%	Cityscape	city/urban/suburban location	16%	13%	10%
93%	Text	at least 1 clearly readable alpha-numeric character: overlaid or in scene	10%	21%	34%

3. Automated concept classification

We pose the problem of detecting visual semantic concepts as a statistical machine learning problem. We represent images with a set of low-level visual features, such as colors, textures, and shapes. In the training phase, we then learn feature representations corresponding to the binary hypotheses for each concept (presence/absence) using generic supervised machine learning algorithms like Gaussian Mixture Models, Hidden Markov Models, and Support Vector Machines (SVMs) [7]. In the detection phase we use the existing models to score target images for the presence/absence of the concept. This approach has been evaluated on multiple NIST TRECVID benchmark data sets and has typically been one of the best performing approaches throughout this evaluation. It presents us with the opportunity of scaling the detection to a large number of concepts that can then be used to enrich the video content semantically without having to invent individual learning strategies and algorithms that are concept specific. This approach thus represents a reasonable performance level that can be practically attained using the state of the art in multimedia signal processing and machine learning.

For the experiments reported here, we use key-frames for feature extraction, modeling and detection. There is a need to tune the parameters of the learning algorithm as well as perform feature selection from the large set of features that we extract. For this we need a validation set from which we can perform parameter and feature selection. We partition the TRECVID 2003 common annotation data set into several partitions, including a training partition of 28055 keyframes and three validation sets of which one validation set with 4420 keyframes is used for the parameter and feature selection reported in this paper. Figure 1 shows the automated approach employing the training set and the validation set to derive the optimal parameter and feature selection based on average precision using support vector machine classifiers.

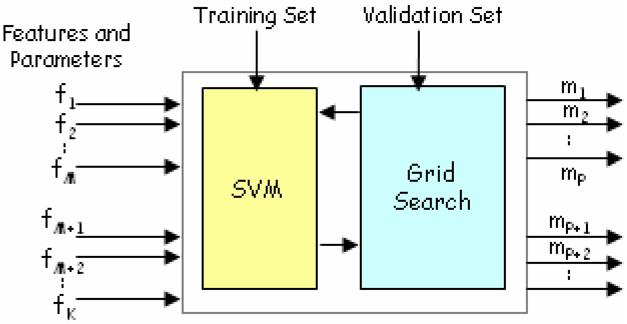

Figure 1. The process of parameter and feature selection using a training set and a validation set: at left are a variety of features and learning parameters being presented to the learning system. On the right are the selected features and parameters that turn out to be optimal with respect to the average precision performance on the validation set.

The performance evaluation metric we used is that of non-interpolated average precision over the ranking of all target images with respect to given semantic concept. This measure approximates the area under the precision recall curve and is measured by averaging the precision at all depths where there is a positive hit until reaching a predefined depth and then dividing this by the total number of relevant items in the dataset, or the predefined depth, whichever is smaller, consistent with the "average precision" metric defined for use with TRECVID 2005.

For the experiments reported in this paper, we experimented with the following features for model building:

- Color Correlogram (166): Single-banded auto-correlogram coefficients extracted for 8 radii depths in a 166-bin HSV color space.
- Edge Histogram (64): Using a Sobel filtered image and quantized to 8 edge orientations and 8 edge magnitudes.
- Co-occurrence Texture (96): Based on entropy, energy, contrast, and homogeneity features extracted from gray-level co-occurrence matrices at 24 orientations.
- Moment Invariants (6): Based on Dudani's moment invariants for shape description.

For each concept, we train a set of configurations of binary SVM classifiers using the features extracted from the keyframes. We use SVMs with Radial Basis Function (RBF) kernels and optimize RBF parameter settings using validation as detailed in [6], selecting the parameter configuration leading to best average precision performance on the validation set for the final statistical model of the given concept. The concept confidences are then normalized to fall in the range of [0, 1] as discussed earlier.

For the experiments in this paper, we thus trained models for the 6 semantic concepts Indoors, Outdoors, Face, People, Cityscape and Text. Figure 2 shows the average precision at 1000 for the three corpora using the produced 2005 classifiers.

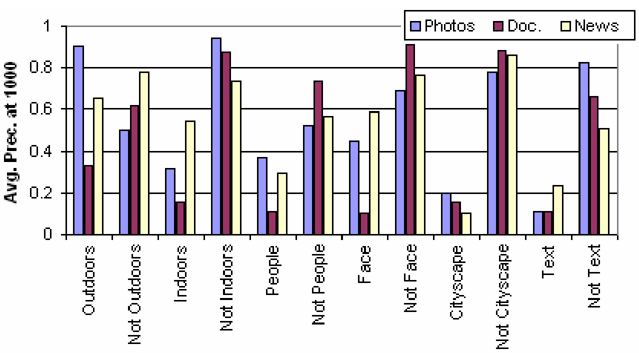

Figure 2. Average precision for 2005 classifiers for the presence/absence of each of six concepts.

All these models are based on globally extracted features, which is certainly a disadvantage for regional concepts such as *face* and *people*. Thus, the results presented here represent performance based on global models, which can be improved using regional models (e.g., see [7]). Once the models are trained selecting optimal parameters and features, we then extract identical features from the three test sets: the "news" set from TRECVID 2003, the "documentaries" set from TRECVID 2002, and the amateur "photographs" set as presented earlier. It is important to note that the TRECVID 2003 training and test data sets are identically distributed, but this is not at all the case with respect to the documentaries or photographs sets. The documentaries set is dominated by keyframes from videos captured in the 1940s with very poor color characteristics almost on the verge of tinted monochromatic images. The photographs collection has been captured using consumer-grade digital cameras. As can be noticed, these two data sets are totally different in visual characteristics from the produced broadcast news video content from 1998 which makes up the TRECVID 2003 data set. While we could have built models for each data set using training samples which are identical in distributions to that particular set, we try to push the envelop in evaluating the generalization capability of the models across very different data sets.

4. Browsing and filtering experiment

The baseline concept confidence is given by the models discussed in the prior section. For each video shot or image classified with given confidence G (in range [0, 1]) and manually identified as truth T, $T = 1$ if judged to have the concept or $T = 0$ otherwise, we compute the following weighted average of the base classifier with the ground truth:

$$C_i = G + \delta_i * (T - G) \qquad \delta_i = i * 0.01; i = \{0, 5, 10, \ldots, 100\}$$

C_0 is the baseline confidence, C_{100} is truth, with 19 steps interpolating improvements to the baseline from the given up to the truth. Rather than repeat the 2003 experiment by running users through each of the C_i settings, we take advantage of results from that experiment and confirming evidence from additional TRECVID interactive video search sessions over the years (see [4]) to note that users will accurately select images relevant to a topic from a candidate set shown in a storyboard as long as the set is not too large. While empirical evidence suggests novice users will exhaustively inspect storyboards of 240 or fewer thumbnail images, we conservatively set the limit to 200 in our analysis. If users are given a set of 200 images, the number of relevant images in that set of 200 is a good predictor for the number of relevant images the user will return for the topic, i.e., the user's recall performance will be at or near the recall value when considering only the set of 200 images, with the user's precision expected to be at 0.9 or higher. Hence, our metric in evaluating the relative merits of the C_i settings, i.e., the concept classifier accuracy, will be recall from a set of at most 200 images.

For *filtering*, the same set of topics and candidate sets as described in Table 1 are used, but in accordance with earlier findings, only the 6 topics returning 480 or 960 images in the candidate set are considered. The smaller candidate sets do not warrant post-filtering: the presumed query returning the ranked set of 120 or 240 is precise enough to avoid the need for a post-filtering step. For *browsing*, we test whether the semantic concepts could limit the full corpus to derive a good candidate set of 200. The goodness of the set must be compared against a random pull of 200 images from the data set. These numbers are given in Table 3. For example, there are 144 shots of people walking in an urban environment in the test corpus of 1834, so a random pull of 200 shots would hold 200 * (144/1834) = 15.7 shots.

Table 3. Topics, relevant count in whole corpus (R_{total}), and expected number of relevant images in set of 200 (R_{200}).

Topic	R_{total}	R_{200}
Waterfalls with no people	12	1.7
Road traffic	24	3.4
Snow-capped mountains	48	6.8
This particular adult female (shown)	184	26.2
Leisure time at the beach or pool	13	1.1
One or more people in the kitchen	26	2.1
Automobiles	118	9.6
Two-story or taller buildings	151	12.3
Missiles in the air or being launched	13	1.4
Person, Pope John Paul II (shown)	24	2.6
Airplanes on the ground or in the air	59	6.4
People walking in urban environment	144	15.7

We bypass the question of whether users can identify the optimal manner in which to apply concepts for filtering and browsing these topics. While not a trivial task, the focus on user selection of concepts to topics is more directly addressed in a separate study [1]. Here, we focus on the investigation that if users can judiciously select the right concepts to apply to topics, then how accurate must those concept classifications be before the users see any benefit? As an example of use, consider the topic "people walking in an urban environment." With C_0 accuracy concept classifiers, the user could browse the "best cityscape-people" storyboard shown in Figure 3. The thumbnails relevant to the topic are bordered in yellow to help the reader in this scaled view, which of course would not happen for the user. Users would also likely view these 200 shots at one-quarter horizontal and vertical resolution thumbnails for MPEG-1 video (i.e., 88 by 60 pixels per thumbnail), and in two pages. A storyboard of 10 x 10 thumbnails requires 1000 by 700 pixels accounting for scrollbars and borders, with the 2003 study and years of additional TRECVID studies by the authors and others [4] indicating that users are willing to page the storyboard once to see the 100+100=200 thumbnails. More patient users might investigate further, but we leave our metric conservative at a limit of 200.

Armed with better concept classification accuracy, the storyboard of highest accuracy cityscape-people shots would contain more shots that are both cityscape and people, the best a priori concepts to apply toward the "people walking in an urban environment" query. We would expect to therefore find more shots relevant to the topic when browsing with concepts of higher accuracy. For example, at C_{25} accuracy the best 200 cityscape-people shots includes 76 relevant to this topic versus only 30 in the 200 shots shown in part in Figure 3 using C_0 accuracy. Our experiment systematically explores the relationship between concept accuracy and filtering and browsing utility, using recall at 200 as the evaluation metric.

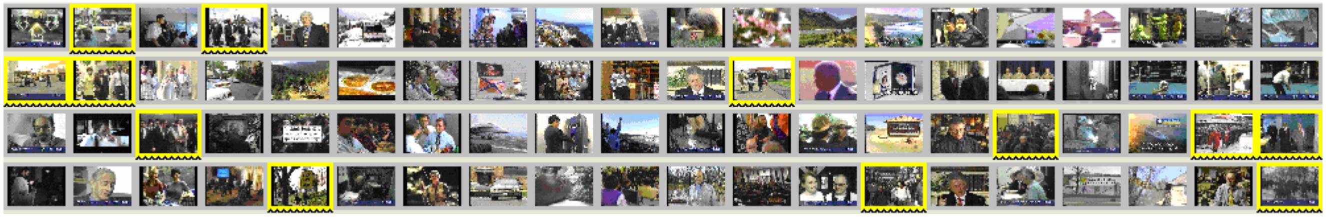

Figure 3. Best cityscape-people news shots, using C_0 concept accuracy; 30 shots relevant to "people walking" topic are in top 200, 12 in the top 80 shown here framed in yellow and underlined with wavy line marker.

5. Filtering results

With the concept classifiers described in Section 3 and a good candidate set generator returning half of the correct answers among the first 20% of the candidate set, half of the remaining in the next 20%, etc. (see Section 2), there is no benefit to filtering by concepts at given C_0 accuracy, and in fact recall is worse than if the user simply inspected the first 200 images from the candidate set. Figure 4 shows the recall at 200 averaged across the 6 test topics, where the ideal recall was returning all 48 relevant shots for the candidate sets of size 480 and all 96 for the sets of 960. This result of C_0 being worse than no filter is in agreement with experiences over the years for TRECVID search tasks, where concept filters have not contributed significant additional benefit to a text search shot-ranking mechanism [4]. Results from text search against closed captions for news and documentaries have the characteristics of the good candidate sets discussed in Section 2. Once the classifier accuracy improves to C_{10}, there is no longer a disadvantage to employing the filters, with a marked improvement in the recall considering only the top 200 images as classifier accuracy improves from C_{15} to C_{30}. C_{30} produces three-fourths of the benefits possible with true concept classification (C_{100}).

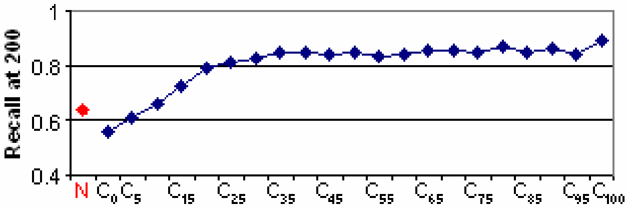

Figure 4. Recall (considering only first 200 images) when no filters are applied (N) and when filters of varying accuracies used in range C_0 (baseline), to C_{100} (truth).

Note that recall is not at 100% even with truth classification at C_{100}. Truth data holds ambiguity, too, as interrater reliability was not 100% in Table 2. Also, concepts do not fit topics exactly, e.g., most but not all two-story buildings are cityscapes, as a farmhouse could be tall, too. Also, six concepts are too few to filter all topic candidate sets down neatly, e.g., the "this adult female" topic was addressed by only the face concept, but there are 280 faces in the candidate set of 960 for this topic. In contrast, the six topics might fit the topic very well, as is the case with the airplane query fitting the filter "no face, no indoors, no people, outdoors" so well that 100% recall is achieved at C_5 through C_{100} with C_0 producing 96% recall as well, the lone case where C_0 was better than not using a filter. We return to the issue of concept fit to topics in the conclusions.

6. Browsing results

Table 4 indicates which concepts were used to generate the browsing sets evaluated for each topic, based on human inspection of the topic and concept truth data to select the most appropriate concept-topic mappings, e.g., for the topic "people walking in an urban environment" the most appropriate concepts were people and cityscape. Table 4 shows that often the negation of a concept (~people indicates "not people", i.e., people concept confidence is at or near 0) is utilized, and that for our test set of 12 topics, the text concept was rarely employed.

Again, our focus is on evaluating concept accuracy's effects on recall performance, and so we hand-tuned the settings of the confidence thresholds in generating the browsing candidate sets as shown partially in Figure 3 to maximize the number of relevant shots included in the top 200, for each of the 20 C_i accuracy settings, for each of the 12 topics. For example, Figure 3 for C_0 accuracy is generated with the setting "people >= 0.76 and cityscape >= 0.51" sorted by descending cityscape, while for C_{25} accuracy and the same topic the setting used is "people >= 0.5 and cityscape >= 0.5" sorted by descending cityscape. We wanted to completely bypass the question of whether users could make judicious use of concept classifiers and look just at the question of concept accuracy. If optimal concept choices are made in accordance with Table 4 (sort order given by first concept listed in concept choices), what are the effects on recall as concept accuracy changes?

Table 4. Topics (same order as prior tables but abbreviated), and best choice of concepts to apply to generate browsing sets, along with the number of images relevant to the topic and the total images included when using concept truth C_{100}.

Topic	Concept Choice	T_{rel}/C_{total}
Waterfalls no people	~people, ~face, outdoors, ~city	12/510
Road traffic	outdoors, city	24/225
Snowy mountains	outdoors, ~city	48/978
Adult female	face	173/422
Beach/pool	people, ~face, ~city, ~text	11/257
Person in kitchen	indoors, face, ~city	24/138
Automobiles	outdoors, ~face	109/897
2+ story buildings	city, ~indoors, ~face	126/302
Missiles	outdoors, ~city, ~face, ~people	13/318
Pope John Paul II	people, ~city	20/556
Airplanes	outdoors, ~indoors, ~face	58/542
Urban people	city, people	86/189

Figure 5 shows the effects of concept accuracy on recall at 200. As the accuracy improves, recall improves dramatically. Even the baseline concept classifier that we currently can automate today, C_0, provides double the recall performance versus the expected number of relevant images in a random draw of 200 shots, the R_{200} control metric shown in Table 3. As concept classifier accuracy improves a bit to C_{25} through C_{35}, a sweet spot is reached in which most of the benefit from concept accuracy toward retrieval performance is achieved.

The plot for truth, C_{100}, is separated out in Figure 5 because as with R_{200} we did not actually count the number of topic-relevant shots in a *ranked* set of the best 200, since there is no ranking with concept truth where all shot concept confidences are 0 or 1. Unlike the filter experiment where candidate sets came with their own rankings, the generated browse set is ranked by the confidences of the chosen concepts (Table 4), but in the case of truth, the resulting set might be greater than our limit of consideration. The right-most column of Table 4 gives the number of topic-relevant images over the number of total images returned with the listed concept choice for the data sets, when we have concept truth of C_{100}. For example, with the C_{100} truth concepts outdoors and no people and no face and no cityscape used for the waterfalls topic, 510 of the 1405 photographs remain, of which all 12 waterfalls topic-relevant shots are included. To generate the "recall at 200" plot for C_{100} when $C_{total} > 200$ in Table 4, we assume the same percentage of topic-relevant images will appear in the first 200 under consideration, i.e., for the waterfalls topic 200 * (12/510) = 4.7, for a recall at 200 value of 4.7/12 = 0.39.

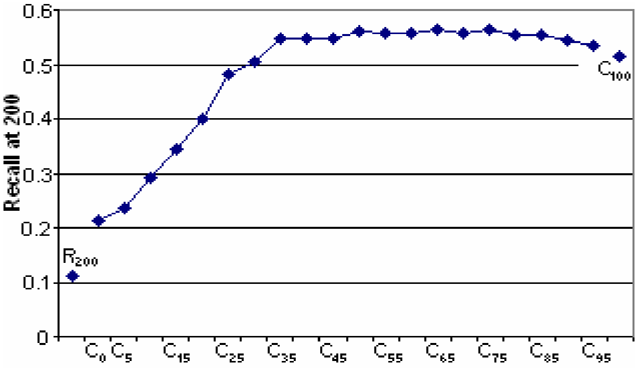

Figure 5. Recall at 200 when no concepts are applied and when concepts of varying accuracies are used to generate a "best-of" browsing set to explore in relationship to a topic.

Figure 6 shows that the sweet spot holds for all three tested corpora, with vast improvements in recall provided by early gains in concept classification accuracy from C_0 to C_{25} and then recall improvements levelling off with continued accuracy improvements on up to truth.

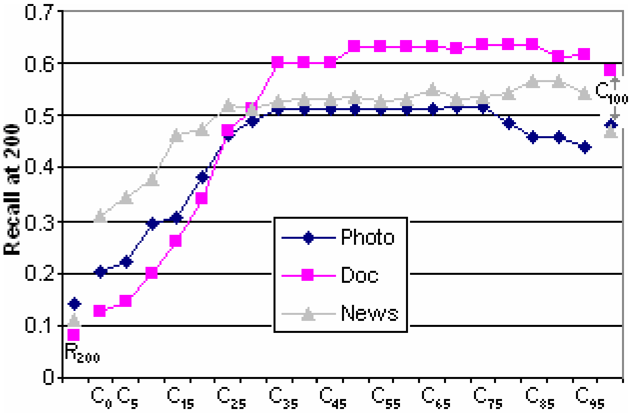

Figure 6. Recall when no concepts are applied (R_{200}) and when concepts of varying accuracies are used to generate "best-of" browsing sets, by corpus type.

The recall at 200 results for the highest quality confidences of C_{75} through C_{95} dip down, a non-intuitive result but caused by the loss of ranking between images as all confidences are pushed toward either 0 or 1, the fact that we store C_i values down only to the hundredths precision, and the use of only six concepts to generate browse sets which sometimes still leaves a large value much greater than 200 for C_{total} (see Table 4). As an example, consider the adult female topic where only the face concept was appropriate. Table 5 shows the changes in recall at 200 from C_{75} through C_{100} due to the loss in ranking between photographs with respect to a concept. In the table, N_T and N_K are image counts after tight and kept settings of "face >= X", with "first" 200 images from N_K using a consistent but default ordering of the corpus that carries no meaning (order by image ID). The * notes that for C_{100} and generated browsing sets, we assume the same percentage of topic-relevant images will appear in the first 200 under consideration, as discussed with respect to Table 4.

Table 5. Examples of recall-at-200 variability due to loss of ranking as all confidences move to 0 and 1.

	Tight setting: Face >=	N_T	Kept setting: Face >=	N_K	T_{rel} in "First" 200	R(200)
C_{75}	0.92	171	0.91	217	89	0.48
C_{80}	0.94	156	0.93	214	92	0.5
C_{85}	0.95	139	0.96	202	90	0.49
C_{90}	0.96	182	0.95	280	85	0.46
C_{95}	0.99	139	0.98	335	76	0.41
C_{100}			1	422	82*	0.45

Looking at the best ranked faces where enough ranked images still exist to draw a "best" set of near 200 (C_{80} and C_{85}) produces a better R(200) score than looking at a representative set of 200 from a larger unranked set of 300 plus (C_{95} and C_{100}). We conclude that automated semantic concept classifiers that offer confidence ranges and hence rank video shots and imagery relative to one another offer advantages for browsing compared to pure "yes/no" binary classifiers, as binary classifiers don't offer any help in reducing a large candidate set C_{total} (see Table 4) down to a set that interactive searchers are willing and able to inspect.

7. Conclusions

Results show that current automated classifiers offer browsing utility twice as good as having no classifier at all, and that improving the classifiers to C_{35} accuracy results in dramatically increasing performance improvements. To better interpret these results dealing with C_i, we return to the complete truth for the test corpora and compute the mean average precision (MAP) at 1000 across the test data using the C_i classifiers. MAP at 1000, shown in Figure 7, is computed by combining (i.e., averaging) the average precision at depth 1000 across all 12 concept classifications (see Figure 2), across each of the three test sets to create the non-interpolated mean average precision (MAP) for the test data. Behavior was similar for each of the three corpora so they are averaged together, but the MAP was significantly better for the negation/absence of the six concepts than for the positive occurrence/presence, so these two curves are plotted separately. After C_{55} the MAP(1000) values are at or near 1. Hence, results of performance improvements flattening for C_i, $i >= 40$, are thus very much expected: from C_{40} on up, across all twelve concept classifications, the top 1000 results, or top N for concepts with N relevant instances and N < 1000, are correct.

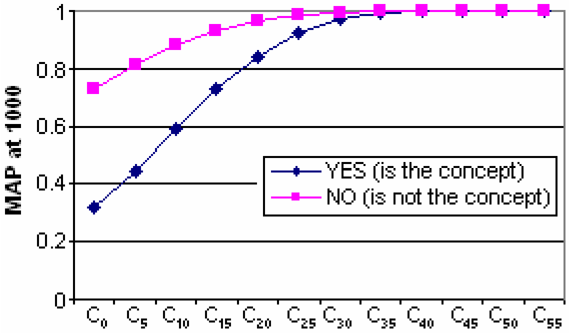

Figure 7. MAP at 1000 for 6 yes/no classifiers at C_i distributions, MAP(1000) for C_i at or near 1, $i > 55$.

The results of filtering and browsing improvements (Figures 4-6) are very much correlated with the MAP(1000) improvements shown in Figure 7. The study provides us with quantitative numbers from which to draw conclusions regarding the state of the practice and these filtering and browsing tasks. The threshold for success is much higher for filtering than for browsing. For filtering a well-ordered set of results (e.g., a set retrieved from text search), concept classifiers need greater accuracy than C_0 with its MAP of 0.32 on the presence of concepts and 0.73 on their absence. Indeed, a near doubling of accuracy to C_{10} (MAP of 0.59 and 0.88 for yes and no, respectively), still shows nearly the same recall at 100 or 200 performance as not using any filtering at all and relying on the ranking from the query service. Only with extreme accuracy of MAP better than 0.7 (C_{15} and greater) do concept-based filtering offer improvements when used in conjunction with query-ranking mechanisms.

In contrast, today's classifiers represented by C_0 show immediate utility for browsing image and video collections. As the performance of these classifiers improve, the browsing utility improves as well, right up to MAP at 1000 of 1.0, i.e., even late-stage improvements of MAP from 0.92 to 0.99 (C_{25} to C_{35}) result in recall improvements at the top range of imagery that will be investigated by human users. Results are consistent for all 3 test corpora: an amateur photo collection, documentaries, and news. Hence, automatic semantic concepts can offer significant utility for browsing at current levels of accuracy, but the requirement is much higher for filtering a well-ordered set of results, where extreme accuracy is necessary before benefits are seen.

Acknowledgments

This material is based on work supported by the Advanced Research and Development Activity (ARDA) under contract number H98230-04-C-0406 and NBCHC040037, and supported by the National Science Foundation under Grant No. IIS-0535056.

References

[1] M. Christel and A. Hauptmann. The Use and Utility of High-Level Semantic Features. *LNCS 3568 (CIVR 2005)*, 134-144.

[2] M. Christel, N. Moraveji, and C. Huang. Evaluating Content-Based Filters for Image and Video Retrieval. *Proc. ACM SIGIR '04* (Sheffield, UK, July 2004), 590-591.

[3] P.G.B. Enser. Pictorial information retrieval. *Journal of Documentation*, 51(2): 126-170, 1995.

[4] A. Hauptmann and M. Christel. Successful Approaches in the TREC Video Retrieval Evaluations. *Proc. ACM Multimedia '04*, ACM Press (2004), 668-675.

[5] M. Markkula and E. Sormunen. End-user searching challenges indexing practices in the digital newspaper photo archive. *Information Retrieval*, 1:259-285, 2000.

[6] M.R. Naphade, C.-Y. Lin, A. Natsev, B. Tseng, and J.R. Smith. Framework for Moderate Vocabulary Semantic Visual Concept Detection. *Proc. ICME 2003*, 437-440.

[7] M.R. Naphade and J.R. Smith. On the Detection of Semantic Concepts at TRECVID. *Proc. ACM Multimedia '04*, ACM Press (2004), 660-667.

[8] NIST TREC Video Retrieval Evaluation (TRECVID), 2001-current, http://www-nlpir.nist.gov/projects/trecvid/.

[9] M. Rautiainen, T. Ojala, and T. Seppänen. Analysing the Performance of Visual, Concept and Text Features in Content-Based Video Retrieval. *Proc. MIR '04*, ACM Press (2004), 197-204.

[10] K. Rodden and K.R. Wood. How Do People Manage Their Digital Photographs? *Proc. ACM CHI 2003*, 409-416.

[11] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content based image retrieval at the end of the early years. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1349-1380, 2000.