

Evaluating Content-Based Filters for Image and Video Retrieval*

Michael G. Christel
CS Dept. and HCI Institute
Carnegie Mellon University
Pittsburgh, PA 15213
1-412-268-7799
christel@cs.cmu.edu

Neema Moraveji
HCI Institute
Carnegie Mellon University
Pittsburgh, PA 15213
1-412-268-7003
neema@cmu.edu

Chang Huang
CS Department
Carnegie Mellon University
Pittsburgh, PA 15213
1-412-268-5163
liz+@cs.cmu.edu

ABSTRACT

This paper investigates the level of metadata accuracy required for image filters to be valuable to users. Access to large digital image and video collections is hampered by ambiguous and incomplete metadata attributed to imagery. Though improvements are constantly made in the automatic derivation of semantic feature concepts such as indoor, outdoor, face, and cityscape, it is unclear how good these improvements should be and under what circumstances they are effective. This paper explores the relationship between metadata accuracy and effectiveness of retrieval using an amateur photo collection, documentary video, and news video. The accuracy of the feature classification is varied from performance typical of automated classifications today to ideal performance taken from manually generated truth data. Results establish an accuracy threshold at which semantic features can be useful, and empirically quantify the collection size when filtering first shows its effectiveness.

Categories and Subject Descriptors

H5.1. [Information Interfaces and Presentation]: Multimedia Information Systems – evaluation, video.

General Terms

Human Factors, Experimentation.

Keywords

Video and image browsing, visual semantic feature filters.

1. FILTER EXPERIMENT AND RESULTS

Digital imagery, used here to refer to both still images (photographs) and motion video, has proliferated in the past few years, ranging from ever-growing personal photo collections to professional news and documentary archives. Digital imagery retrieval can be considered a transaction sequence in which the user initiates a query or browsing action that generates a candidate set of photos and key frames representing video shots. User interaction is critical to better express the information need and generate a new, more precise candidate set [3]. The user can 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 features such as “indoors” and “outdoors”, as defined and studied in the NIST TREC video retrieval evaluation (TRECVID) forum since 2001 [2].

Past TRECVID experiments have yet to validate that such classifiers can lead to improved interactive retrieval [2], perhaps because the automatic classification accuracy is not yet good enough for use as an interactive filter. We investigate this issue by indexing candidate set imagery manually, and then introducing errors to the classification according to the different distributions shown in Figure 1. Figure 1a shows the distribution of confidence for our automatic indoor and outdoor classification evaluated in TRECVID 2002, with 1e showing human-generated truth data. The mean is shifted from 0.52 to 0.63, 0.74, and 0.85 for the imagery truly having a feature, and from 0.46 to 0.35, 0.25, and 0.15 for imagery truly not having a feature, to generate the OK, BETTER, and BEST distributions as shown in Figure 1.

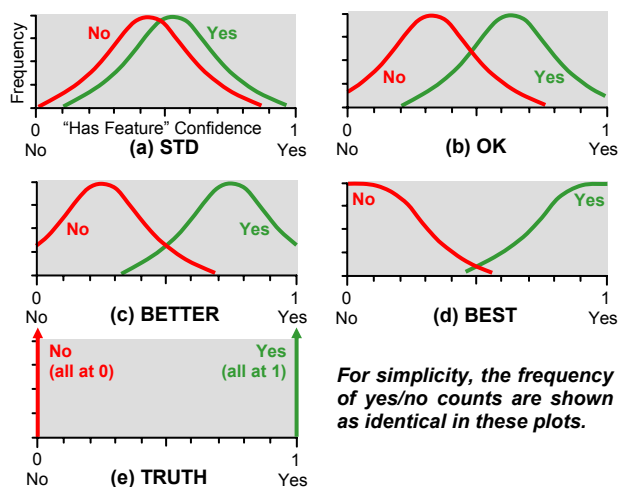


Figure 1. Plots of feature confidence distributions in study.

We look at how filters are used against candidate set sizes of 120, 240, 480, and 960, for three corpora: documentaries from TRECVID 2002, news from TRECVID 2003, and photos. The experiment employed a between-subjects design. For each corpus, 4 topics were chosen based on the TRECVID topics created over the years to reflect the sorts of queries real users pose [2], queries such as snow-capped mountains or a particular shown person. This study was not concerned with whether participants could generate good queries for these topics. Rather, the participants were given fixed candidate sets for each topic, with the task of marking the imagery satisfying the given topic. The number of correct answers in a candidate set was held constant at 10%.

*This material is based on work supported by the National Science Foundation (NSF) under Grant No. IIS-0121641 and by the Advanced Research and Development Activity (ARDA) under contract number H98230-04-C-0406.

Each participant saw the same candidate set per topic, was limited to 4 minutes per topic, and answered follow-up questionnaires.

36 students and staff (14 female, ages 18-41 [mean age 23]) at Carnegie Mellon University responded to an electronic call for participation. The corpus order was completely randomized, and order of topics within the corpus likewise completely randomized via a Latin Square design. Participants saw the candidate sets as storyboards, initially sized within the application to show 7 rows of 9 images each, with 6 filter sliders available above the storyboard canvas as shown in part in Figure 2. The filter interface widget design is based on dynamic query sliders [1]. Participants were free to resize the storyboard, and change the resolution of each thumbnail in the storyboard, but rarely did so. Participants could visually inspect thumbnails not initially visible in the storyboard page through scrolling or by using the filter to dynamically drop out some images and have off-screen ones that satisfy the filter scroll into view. Using the truth data to separate images into yes and no sets for each feature, the confidence values are set making use of the distributions shown in Figure 1. For example, an outdoors image on average will get an outdoors confidence value of 0.52 for STD accuracy, 0.63 for OK, 0.74 for BETTER, and 0.85 for BEST. Each participant then made use of filter data from exactly one of these distributions STD, OK, BETTER, or BEST throughout the 12 topics.

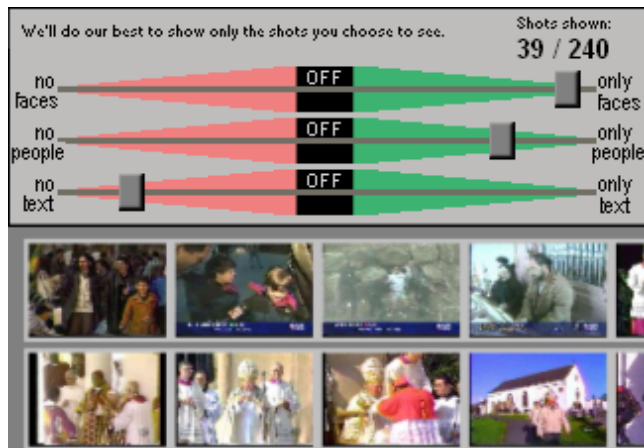


Figure 2. Storyboard showing faces, people, and text filters (indoors, outdoors, and cityscape filters cropped from view).

With respect to recall, as expected for topics with small sets of 120 or 240 images, there were no significant differences in recall for the filter accuracy. However, for the larger topic sets of 480 and 960 images, a one-way ANOVA conducted on the data yielded a significant effect on filter accuracy, $F(3,212) = 3.1, p < 0.03$; recall improved from STD to OK to BETTER and then dropped off slightly for BEST. No significant differences in precision were found across corpora, topic sizes, or filter accuracy: precision for the 4 treatment groups STD, OK, BETTER, and BEST was very high in the 0.92 to 0.96 range.

The average size of the storyboard, i.e., the number of images left in the storyboard after the users applied filtering, was computed for all topics where filtering was used, revealing a highly significant pattern across the treatment groups, shown in Figure 3. As the quality improved from STD to OK to BETTER to BEST, participants made use of the filters to drop out more imagery from the storyboards ($F(3,339) = 9.07, p < 0.0001$). The improved

feature filter accuracy directly helped with finding relevant answers deeper in the storyboard. The relevant images that initially from the fourth page onward in the storyboard were found with significantly greater counts in large data sets by users with the BETTER and BEST filters than those with the STD or OK filters ($F(3,212) = 2.7, p < 0.05$). For the STD, OK, BETTER, and BEST groups, the average number of correct answers from this far back in the storyboard was 7.8, 9.6, 11.8 and 11.4 respectively.

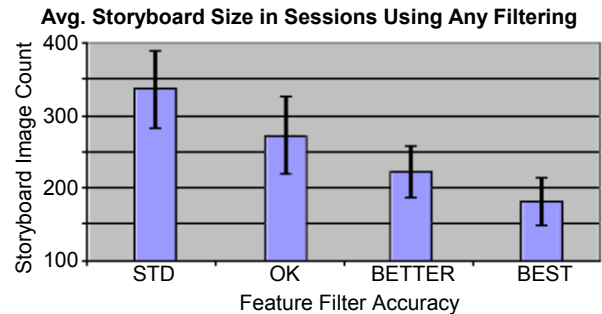


Figure 3. Average number of images (with 95% confidence intervals) in storyboard following filter use.

No differences were found in performance or satisfaction between the three corpora of personal photo collections, documentary video sets represented with key frame images for shots, and news video represented likewise, and the relative merits of the different feature filter accuracies across set sizes held for the different corpus types. For candidate set sizes of 240 and smaller, users consider the task of employing storyboards to find the answers as easy and satisfying, and perform well regardless of the feature filter accuracy. Feature 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 given the 4-minute time limit, and for these sets feature filtering is used.

The accuracy of the filter directly affects the efficiency and effectiveness of the filter use. Today's automated systems producing confidence distributions as shown in Figure 1a (STD) are not good enough to improve recall on TRECVID tasks. If the filter accuracy is at the BETTER or BEST range (Figure 1c, 1d), users recall significantly more correct images from the candidate set, especially those buried deep within the storyboard. As the accuracy improves from STD to BEST, users can more effectively apply the filter to reduce the size of the storyboard left for visual inspection. The implication for designers of digital imagery systems is that visual feature filters with an accuracy of at least the BETTER range should be provided for interactive browsing.

REFERENCES

- [1] Ahlberg, C. and Shneiderman, B. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. *Proc. CHI '94*, ACM Press, 313-317.
- [2] NIST TREC Video Retrieval Evaluation, 2001-current, <http://www-nlpir.nist.gov/projects/trecvid/>.
- [3] Worring, M., Smeulders, A.W.M, and Santini, S. Interaction in content-based retrieval: an evaluation of the state-of-the-art. *LNCS 1929*, Springer-Verlag (2000), 26-36.