Exploring the Synergy of Humans and Machines in Extreme Video Retrieval

Alexander G. Hauptmann, Wei-Hao Lin, Rong Yan, Jun Yang, Robert V. Baron,
Ming-Yu Chen, Sean Gilroy, and Michael D. Gordon

School of Computer Science, Carnegie Mellon University,
Pittsburgh, PA 15213, USA
{alex+, whlin77, rongy, juny, rvb, mychen,
sgilroy, michael.gordon}@cmu.edu
http://www.informedia.cs.cmu.edu

Abstract. We introduce an interface for efficient video search that exploits the human ability to quickly scan visual content, after an automatic system has done its best to arrange the images in order of relevance. While extreme video retrieval is taxing to the human, it has also been shown to be *extremely* effective. Two variants of extreme retrieval are demonstrated, 1) RSVP which automatically pages through images with the user controlling the page speed, while a user marks the relevant ones and 2) MBRP where the user manually controls the paging and can also adjust the number of images per page, depending on the density of relevant shots found. Pros and cons of each variant are discussed.

1 Interactive vs. Automatic Video Search

When comparing results of fully automated video retrieval to interactive video retrieval [5], one finds a big gap in performance. The fully automated search (no user in the loop) succeeds with good recall for many topics, but relevant shots tend to be distributed throughout the top 3000 to 5000 slots in the ordered shot list, causing the standard metric of average precision for automated search to lag well behind most interactive runs. From this insight, we developed an interface that relies on superior human visual perception to compensate for low precision in automatic search of the visual contents of video [1]. The human user can filter the best automatically generated results and produce a better set that retains the relevant shots, resulting in much greater precision. We named this approach extreme video retrieval (XVR), as it combines the best machine performance with maximal use of human perception skills. Our interface explores two types of approaches to human filtering: rapid serial visual presentation and manually controlled browsing with resizing of pages.

The success of XVR relies heavily on the ability of automatic retrieval systems to recall more relevant at as lower depth as possible. To study the machine extremes of our automatic retrieval system we take a one automatic run [3,4] with query classes and plot MAP over 24 TRECVID 2005 search topics at the depth k of shots, as shown in Figure 1. The automatic run demonstrates respectable performance, achieving MAP of around 0.1 at the depth of 1000 shots commonly chosen in TRECVID.

After depth of 1000 shots MAP reaches the plateau, mainly due to the severe penalty for ranking relevant shots low in the calculation of average precisions. However, with the optimal ranking function, the optimal curve becomes the recall at the depth k, and clearly our automatic retrieval systems have decent recall. For easy comparison we plot the best performance of all search submission in TRECVID 2005. The results show that anyone who can browse through the top 2000 shots (merely 2.56% of TRECVID 2005 test set) for each topic, she could have achieved the best search performance in TRECVID 2005, and even better performance if she can look deeper/faster!

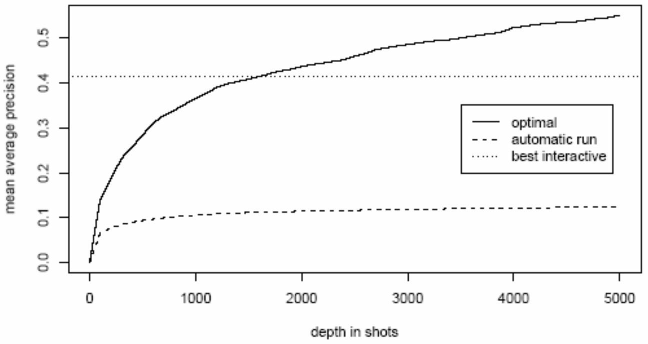

Fig. 1. The MAPs over 24 TRECVID 2005 search topics of one CMU automatic runs, best interactive run in TRECVID 2005, and a hypothetical run with an optimal re-ranking function.

2. Human Extremes – RSVP

Rapid Serial Visual Presentation (RSVP) is a technique of rapidly presenting a serial of images, and has been widely used in visualization and psychophysics experiments [2]. The basic version of RSVP, known as the keyhole mode [2], presents a sequence of images in the same position of the screen, where the following image replace the previous one every n milliseconds, n is thus the interval between two images. Users can vary the presentation speed (adding or subtracting 100ms from n) with two keys *A* (advance) and *S* (slow). When a relevant image is shown on the screen, users press the *J* key to mark the current image, plus the previous image because of the human reaction time delay between the presentation of the relevant image and the human motor response. Since two images are marked for each relevant key press, a second, *correction* phase is needed to carefully page through all marked images and validate the judgments.

In the TRECVID 2005, we submitted one complete run using this variable speed keyhole RSVP interface, where it ranked 4th among all TRECVID 2005 interactive runs [4]. The 24 topics were completed over three consecutive days, with 4 topics in the morning session, and 4 topics in the afternoon session. Before each session one topic from 2004 was used as practice to “warm up” the participant. We found that subjects can correct around 100 images per minute in the second, *correction*, phase, and thus the length of the correction phase was dynamically determined by the program based on the number of relevant shots already marked in the first phase and the total available time.

While no other existing video retrieval system uses RSVP, several reason argue tfor it: 1) RSVP is an interface specifically designed to present images rapidly, which match the human capability to quickly react to visual stimuli. 2) Keyhole mode requires no eye movements and therefore optimizes the time a user looks at an image. More complex displays such as grids or collages demand eye movements and extra time for eye fixation on every image. 3) RSVP automatically updates the next image in the sequence without manual paging, which reduces user cognitive load of pressing extra keys for each following display. 4) Variable speed control allows users to adjust the presentation speed. If we take the first derivatives of the optimal curve in Figure 1 we note that the rate of relevant images is not constant. There are more relevant shots in the early, top ranks than later. Thus it makes sense to use slower speeds for the earlier top-ranked shots reducing the chance of missing relevant shots, and speeding up for later, lower-ranked results. Variable speed also allows users to slow down for a break when their attention is failing.

A second TRECVID 2005 RSVP submission used a 2 image simultaneous display on each page. Each key press then marked both images on the current page, as well as the two images on the previous page as relevant, requiring four images to be verified in the validation/correction phase. Since there were many images marked, subjects were frequently not able to correct all images selected during the initial RSVP phase, resulting in lower mean average precision.

Manual Browsing in XVR

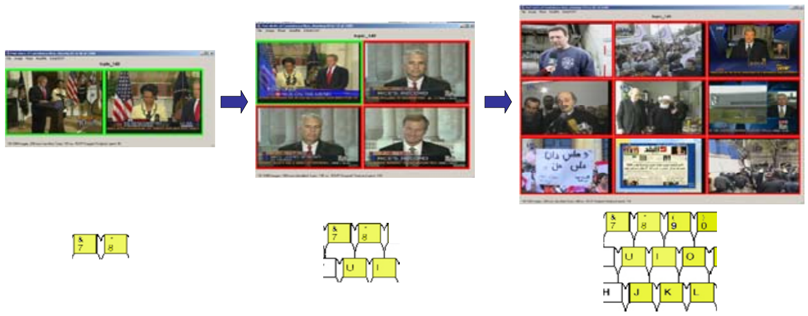

Fig. 2. Manual browsing with different page layouts: 1x2 at the beginning, 2x2 in a later stage, and 3x3 for the rest of the shots. The green bounding boxes indicate the shots labeled relevant, and the keyboard section below a page shows the keys for labeling the respective shots

Manual Browsing with Resizing Pages (MBRP) is a strategy for interactive search, which, unlike RSVP, where the same number of shots per page are used throughout the search, allows adapting the page size according to the (decreasing) percentage of relevant shots. At the beginning when relevant shots are frequent, we use a small page size since multiple relevant shots are likely on one page, which demands more attention (per image) and key presses to label them. Later when relevant shots become infrequent, large page sizes become efficient since it is unlikely that multiple relevant shots will appear even on a large page. MBRP thus reduces the overhead of page turning and the number of necessary key presses for relevant images on a page.

Since the time a user spends browsing each page depends on the page size, the visual complexity of the answer, and the number of correct shots, this time can vary dramatically with different pages. The user may occasionally need to turn back to previous pages to correct erroneous labels. Thus MBRP gains an advantage by letting users turn pages using a forward and backward keyboard key.

Unlike the RSVP, where a single key is used to mark all shots in a page, MBRP allows up to 16 keys (in a 4x4 layout on the keyboard) for labeling 16 shots simultaneously, with one key corresponding to each presented image. Moreover, another key can be used to label all the shots on the current page and automatically turn the page.

Although a page can include any layout of images (e.g., 3x3, 2x5, etc), we use only 1x2, 2x2, and 3x3 for two reasons. First, with practice, one hand can conveniently label any shot(s) in layouts up to 3x3 shots, but not more than 9 shots per page. Second, visually inspecting more than 9 shots per page is less time-efficient.

As the user must label as many shots as possible in a fixed time, errors are inevitable due to time pressures. While missed relevant shots cannot be found during the verification phase, usually one or two minutes are used to correct false alarm errors. In addition, if the user is unsure about the relevance of a shot, it can be marked as “maybe”; where all “maybe” shots will be sorted after those ranked as “relevant”.

A TRECVID 2005 submission using MBRP averaged in looking at about 2000 shots within the 15 minutes for each topic. Typically, this number is higher for queries that are easily identifiable visually, and vice versa. For example, for the query of “tennis”, a user could browse almost 5,000 shots in the allocated 15 minutes time.

The MBRP run achieved the mean average precision (MAP) of 0.408 on the TRECVID 2005 evaluation, which ranked second among a total of 50 interactive submissions and was only marginally behind the best run (MAP = 0.414). It also outperformed the submission using the best RSVP method (MAP = 0.366).

References

1. Derthick, M., Interfaces for Palmtop Image Search. Proc. JCDL (Portland, OR, July 2002), 340-341.
2. Spence, R., Rapid, serial and visual: A presentation technique with potential. Information Visualization, 1(1):13–19, 2002.
3. Yan, R., Yang, J., Hauptmann, A., Learning Query-Class Dependent Weights in Automatic Video Retrieval, Proceedings of ACM Multimedia 2004, New York, NY, pp. 548-555, October 10-16, 2004
4. A. G. Hauptmann, M. Christel, R. Concescu, J. Gao, Q. Jin, W.-H. Lin, J.-Y. Pan, S. M. Stevens, R. Yan, J. Yang, Y. Zhang, CMU Informedia's TRECVID 2005 Skirmishes, TRECVid 2005 - Text REtrieval Conference TRECVID Workshop, Gaithersburg, MD, 14-15 Nov. 2005.
5. Over P, Kraaij W and Smeaton A.F. TRECVID 2005 - An Introduction. TRECVid 2005 - Text REtrieval Conference TRECVID Workshop, Gaithersburg, MD, 14-15 Nov. 2005.