Adaptive Multiple Feedback Strategies for Interactive Video Search

Huanbo Luan[1, 2], Yantao Zheng[3], Shi-Yong Neo[3],
Yongdong Zhang[1], Shouxun Lin[1], and Tat-Seng Chua[3]
[1]Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
[2]Graduate School of the Chinese Academy of Sciences, Beijing 100039, China
[3]School of Computing, National University of Singapore, Singapore 117590
{hbluan, zhyd, sxlin}@ict.ac.cn, {yantaozheng, neoshiyo, chuats}@comp.nus.edu.sg

ABSTRACT
In this paper, we propose adaptive multiple feedback strategies for interactive video retrieval. We first segregate interactive feedback into 3 distinct types (recall-driven relevance feedback, precision-driven active learning and locality-driven relevance feedback) so that a generic interaction mechanism with more flexibility can be performed to cover different search queries and different video corpuses. Our system facilitates expert searchers to flexibly decide on the types of feedback they want to employ under different situations. To cater to the large number of novice users (non-expert users), an adaptive option is built-in to learn the expert user behavior so as to provide recommendations on the next feedback strategy, leading to a more precise and personalized search for the novice users. Experimental results on TRECVID news video corpus demonstrate that our proposed adaptive multiple feedback strategies are effective.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Relevance feedback*

General Terms
Algorithms, Design, Experimentation

Keywords
Interactive Video Search, Relevance Feedback, Adaptive Model

1. INTRODUCTION
The amount of multimedia information especially video data is skyrocketing in recent years. Such massive amount of video data requires intelligent systems that are capable of retrieving relevant video that the users are looking for accurately. Many fully automated video search systems have been developed in the last two decades to cater for precise retrieval. However, with the limited semantic features and the inability to understand user's intention, most of these automated video search systems are still far from satisfactory. To enhance searching in large video corpuses, the multimedia community has recently explored the use of interactive retrieval systems which require user's effort for improving search performance [1, 2]. Interactive video retrieval can be viewed as a two-way process since the user can play the role of leading the retrieval system toward better retrieval by: a) enabling the user to annotate based on the result rank list; and b) selecting the type of feedback based on user's judgment for refinement.

Many algorithms and systems for interactive search have been proposed. Among which the relevance feedback (RF) [3, 4] techniques are the most effective to improve the performance of content-based information retrieval [5]. Early RF approaches attempt to find better query points as close as possible to the assumed "ideal query point" and adjust the weights of various features [6]. However, they are not all-purpose methods due to the presence of certain limitations or strong assumptions. Currently RF in Content-based Image Retrieval (CBIR) is an online learning problem since RF can be generally viewed as a particular type of pattern classification that regards positive and negative samples as two different groups. Among the various learning algorithms [7, 8, 9], the RF based on Support Vector Machines (SVM) [8, 10] is the most popular because of its inherent advantages such as fast learning, multi-choices of kernel, and reliance on only support vectors. However, SVM-based RF usually faces big challenges including small-sized labeled training set, imbalance between the positive and the negative samples, and high-dimension features. To tackle these problems, a new asymmetric bagging and random subspace mechanism is designed in [11]. Besides, semi-supervised learning [12, 13] and active learning [14, 15] are also successively introduced. In spite of its promising performance from previous studies, semi-supervised learning still suffers from high computation cost.

In addition to refinement using various feedback methods, many researchers find that well-designed user interfaces (UI) with good visualization are extremely helpful to improving the search performance. Hence, many efficient and novel interfaces are designed for communication between the users and the system. In fact, interactive systems designed by CMU and MediaMill [16, 17] for TRECVID evaluations have demonstrated that efficient UI is crucial for interactions which jointly maximize the performance of both human and computer.

While interactive search systems have been shown to demonstrate good performance in contrast to fully automatic search, there are also many identifiable problems. One of the problems is the limited feedback strategies as most interactive video search systems usually offer only one kind of feedback strategy (for

CIVR'08, July 7-9, 2008, Niagara Falls, Ontario, Canada.

example: "query point movement" [6] relevance feedback). Due to the complexity and variety of multi-modal features in videos, it is usually insufficient to apply a single feedback method. In reality, it is not possible for a generic feedback strategy to be suitable for different search queries and across different video corpuses. The use of single strategy also often results in a stalemate for the users as they have no other choice if the given feedback strategy does not work well for a certain query. Even with the availability of multiple feedback strategies like in [18], novice users might not be able to make the correct selection as they lack technical backgrounds to enable them to operate the system effectively.

To overcome the above problem, adaptive multiple feedback strategies are proposed for interactive video retrieval. We follow the two-fold approach of: a) Segregating interactive feedback into three distinct types, namely recall-driven, precision-driven and temporal locality-driven feedbacks to provide users with multiple feedback functions, each of which aims at leveraging different aspects of users' feedback data to allow for more flexibility. b) Providing non-expert users a recommendation mechanism, which automatically recommend a suitable strategy to perform feedback based on the current situation. The mechanism aims to help novice users approaching a level of search performance achievable only by the experts. Experiments carried out on the large-scale TRECVID dataset demonstrated the effectiveness of the proposed adaptive multiple feedback strategies as novice users can achieve search performance close to expert users.

The rest of the paper is organized as follows: Section 2 presents an overview on our interactive video search system. Section 3 introduces three multiple feedback techniques, and Section 4 presents detailed techniques about adaptive recommendation mechanism. Experiments are presented in Section 5, and Section 6 concludes the paper.

2. INTERACTIVE SEARCH FRAMEWORK

The overall framework of our interactive video retrieval approach is shown in Figure 1. The retrieval starts with the user query and the auto search will process the query and perform initial retrieval to return a rank-list of initial results to users. The user will then browse the returned results and judge whether they are indeed relevant or not. After that, the labeled positive and negative shots are sent back to the model of Adaptive Multiple Feedback Strategies to improve the subsequent search. At this stage, the users can interact with the system to choose an appropriate feedback strategy to use. Alternatively, users can let the Recommendation Mechanism adaptively suggests a strategy to use at a given time point. The recommendation is extremely useful for non-expert users or when users are unsure about their choices. Subsequently, the chosen feedback strategy is used to improve the search engine based on labeled shots and return a new rank list to the user.

Prior to the interactive stage, we employ our previous automatic search techniques [19] to return an initial rank list, which functions as a good basis for next feedback step. Our approach first performs multimedia query analysis on the original user's query in order to understand user's intention and expand the query to include the necessary context. It then performs a two-leveled retrieval (pseudo story retrieval and shot level re-ranking) to produce a high-quality initial rank list. The set of video features used for feedback also corresponds to the features used in [19]. In brief, the following features are used:

Text Feature. ASR (Automatic Speech Recognition) text is an important feature for informational video data and video retrieval usually begins with text query.

High-level features (HLF). HLFs denote a set of predefined concepts such as cars, buildings, sky, vegetation, desert, face, and people walking. HLFs are extremely useful as they provide additional semantics which are not available from text. There are 39 predefined concepts in TRECVID 2006 evaluation.

Low-level features. These low level features, such as color, edge, motion and SIFT points, present hidden semantics of video content. Due to the curse of dimensionality of visual features, we restrict the feature size to a 116-feature vector for each keyframe including color moment feature, local edge histogram text feature, and motion feature.

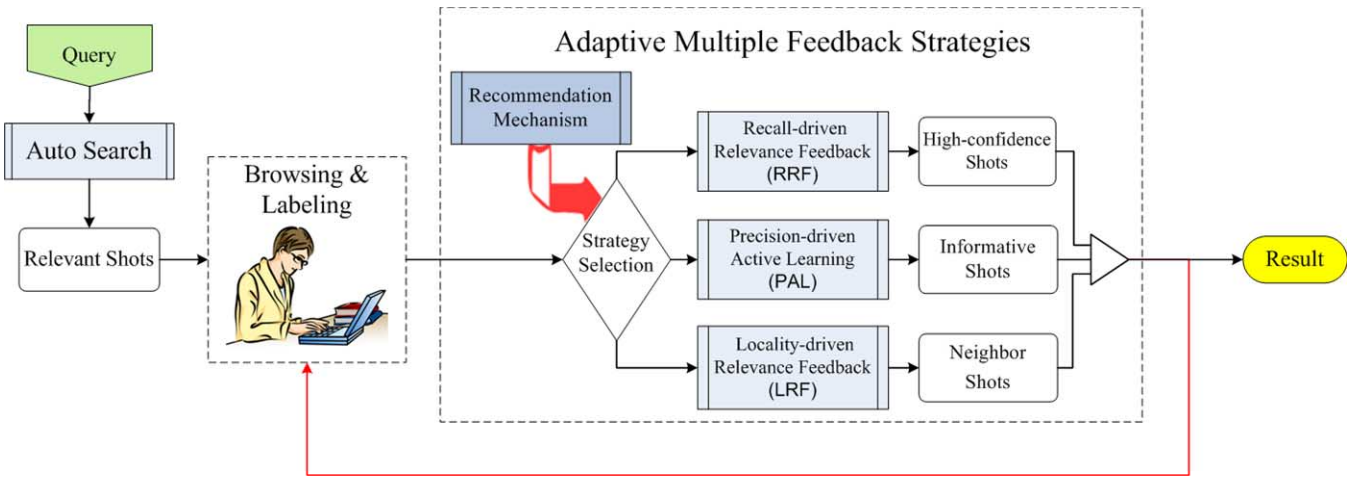

Figure 1. Overall framework for interactive video search

3. MULTIPLE FEEDBACK STRATEGIES

Interactive feedback has been found in previous studies to improve the accuracy of video retrieval. Hence, it is an essential part of any interactive video retrieval system. Although many effective feedback techniques have been proposed to refine the search results, most systems usually rely only on a specific type of relevance feedback technique. In general, it is not enough to utilize only one single feedback technique to tackle different types of queries and for different video domains. This is because the multi-modal nature of video makes video content understanding a fairly complicated process and the wide range of video domains makes user queries quite diverse. Usually, one feedback method works well only for specific classes of queries or on certain domains. A generic interaction mechanism with more flexibility is therefore demanded to cover different search queries and different video genre or corpuses. Thus, we propose to segregate the interactive feedback into three distinct types: a) **Recall-driven Relevance Feedback (RRF)**; b) **Precision-driven Active Learning (PAL)** and c) **Locality-driven Relevance Feedback (LRF)**. Each strategy aims at leveraging different aspects of user feedback data. The first emphasizes on analyzing and applying the correlation of general features obtained from labeled positive and negative instances to provide high recall retrieval on the entire corpus. The second uses active learning with a precision-driven sampling strategy to continuously refine the re-ranking model using a combination of multimodal features. The final strategy exploits high temporal coherence among the neighboring shots within the same story.

Each strategy works well for different situations. For example, RRF works well in news video corpus because ASR text is accurate; whereas LRF is the most effective in documentary video as neighboring shots have high temporal semantic coherence. By leveraging three different feedback strategies, the user can achieve good search performance in a limited time.

3.1 Recall-driven Relevance Feedback (RRF)

At the beginning of interactive search, it is necessary to provide enough labeled positive samples as early as possible, as they are the basis for later refinement. Moreover, we need to achieve high mean average precision (MAP), which requires many relevant shots to be located at high ranking positions. Thus, the recall-driven relevance feedback strategy is designed to maximize recall performance. In this recall-directed facet, we choose to employ general features such as the ASR text and HLFs so as to minimize computational complexity [20]. This process comprises of the following 3 steps:

Step1: Analyze ASR text from labeled positive shots in order to find common relevant text tokens that are highly related to user query using "0.5" formula [21] given in Eqn(1).

$$FS(term_k) = \log \frac{(r+0.5)(N-n-R+r+0.5)}{(n-r+0.5)(R-r+0.5)} \quad (1)$$

where N is the number of shots in the collection, R is the number of shots found to be relevant to the query, n is the number of shots containing term k and r is number of relevant shots containing term k.

Step2: Select distinguishable HLFs from labeled positive shots that are highly relevant to search target at the semantic level as shown in Eqn(2).

$$Q\_HLF^{(1)}[1..50] = \frac{1}{Np}\sum_{i=1}^{Np} S_i^{HLF}[1..50] \quad (2)$$

where $Q\_HLF^{(1)}$ is the new relevancy of HLFs to the query, Np is the number of positively annotated shots and $S_i^{HLF}[]$ is the detection confidence of various HLFs in shot i.

Step3: Re-compute relevance similarity score for each shot based on the above-updated relevant text tokens and HLFs which are used to perform another search to return new results by using Eqn(3).

$$Score(S_i) = \lambda \cdot \frac{1}{k}\sum_{k=1}^{k}(FS(term_k) \mid term_k \in S_i) + (1-\lambda)\cdot\frac{1}{50}\sum_{j=1}^{50}(Q\_HLF^{(1)}[j]\cdot S_i^{HLF}[j]) \quad (3)$$

where $\lambda=[0..1]$ is set according to the importance of text or HLF for a particular query. In our experiments, λ is empirically set at 0.7 and adjusted accordingly by calculating the standard deviation SD of $Q\_HLF^{(1)}[]$.

This option has been found to be the most effective in finding many new relevant shots in the initial stage. More details can be found in our previous work [18].

3.2 Precision-driven Active Learning (PAL)

To complement the high-recall feedback, an SVM-based active learning is carried out on a subset of retrieved shots using multimodal features including visual, motion and HLF, targeting at improving precision. Here, the search problem is simplified into a simple probabilistic binary classification problem (either positive or negative), and SVM is able to perform this task with high efficiency. It uses active learning to provide long term improvements to classifiers. In each iteration, two steps are carried out as follows:

Step1: Sampling & Annotation

Utilize a performance-based adaptive sampling strategy to choose a certain number of instances from unlabeled shots for annotation.

Step2: Learning

Train a new SVM classifier based on above labeled shots.

The performance-based sampling strategy will adaptively choose instances that are either most ambiguous or most relevant from the classification output with emphasis on maximizing precision in a minimal time. Ambiguous samples near classifier boundary are used to refine the classifier quickly and to get a quick convergence in order to maximize precision; while high-confidence shots far away from classifier boundary are used to select other types of relevant shots in order to achieve a high MAP performance. The resulting SVM classifier is used to generate a new result rank list; which is combined with Eqn(3) to obtain the final rank list. More details of the precision-driven active learning are presented in [18].

Figure 2. Examples of temporal semantic coherence for neighboring shots for different topics

3.3 Locality-driven Relevance Feedback (LRF)

Taking shots as search units, we can regard video as a series of temporally related shots. An intuitive method which has been demonstrated to be rather efficient and effective for video retrieval is the use of temporally related shots [22]. The main idea is that video stories usually span across several shots and there is high temporal semantic coherence among the neighboring shots. In other words, there is a high probability that neighboring shots will be relevant if a given shot is deemed relevant. Temporal coherence differs much among different domain corpus. For example, the TRECVID 2006 dataset (news video) has low temporal coherence as every news story is very short, whereas TRECVID 2007 dataset (documentary video) is of high temporal coherence as many temporally continuous shots are related to the same topic. Three examples from TRECVID dataset are shown in Figure 2. The keyframes with white rectangles are the user labeled relevant shots for various retrieval topics. From top to bottom, the topics are "George Bush", "Snow" and "Street Market" respectively. The first topic which belongs to news video has less neighboring coherence as compared to the other two. In general, the length of shot or story is directly proportional to the temporal coherence of a given topic.

The temporal locality-based relevance feedback (LRF) aims at automatically learning which nearby shots to return when a shot is judged to be relevant. Neighboring shots are selected according to the story boundaries for each video and length of the storyboard. A longer storyboard would likely mean higher temporal coherence, and thus more shots to be returned. As shown in Figure 3, the story constraint forbids the return of any shot, such as *Shot* L_2 which is outside the story boundary of the video that contains the marked relevant shot.

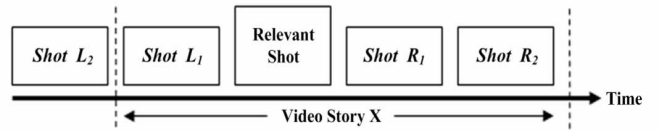

Figure 3. Temporal nearby shots based on storyboard

Given the story boundaries and restrictions, for each marked relevant shot, we attempt to select L shots left and R shots right on the temporal time scale. Different from existing systems where a fixed number of shots from the left and right are returned [22], we determine L and R for normal users based on the mean average precision (MAP) on 24 questions in the TRECVID 2006 corpus. Three expert users are asked to use our system based only on locality-driven relevance feedback for the Interactive Search Task as in TRECVID 2006, using various values of L and R. When evaluating the effectiveness for various values of L, we set R to be zero and vice-versa. From experiments, we found that L and R work best when they are set to 2 and 3 respectively. Using these values of L and R, our system returns all the neighboring shots of the k marked relevant shots in each iteration. Note that these values of L and R are applicable to the news video corpus only.

4. DESIGN OF ADAPTIVE STRATEGY

The common goal of Interactive Retrieval is to leverage as much of user intervention and interaction as possible, so as to maximize the search performance. In practice, most users desire to retrieve their required videos in the shortest possible time, which means time is a critical factor in any Interactive Video Retrieval system. Most existing video retrieval systems, however, neither use the correct techniques at appropriate times, nor fully utilize the characteristics of the videos. As we discussed earlier that different queries have different characteristics, there must be a mechanism to provide recommendations to users on the type of feedback technique they should apply in order to maximize the efficiency. Thus, we propose to resolve this problem by using an adaptive learning strategy which chooses an appropriate feedback technique under different conditions. We strongly believe that the use of a correct feedback technique at appropriate time can significantly improve the performance. Experiments in Section 5 confirm our hypothesis above, especially for non-expert users.

4.1 Considerations from Experts

Choosing RRF, PAL or LRF can greatly affect the system performance and efficiency. Expert and experienced users who have a good idea on the intuition behind each strategy usually can avoid wasting precious search time. However, general users or novice users will not be able to tap onto the segregated techniques effectively. To adapt the system and formulate recommendations to novice users, we therefore explicitly consider the following factors that are recommended by the experts as part of their selection criteria when choosing the respective techniques.

Quality of Initial search results from auto retrieval A_{init}: The quality of "good initial results" varies with the difficultly of the query as well as the size of corpus. From prior observations made from past queries in the same TRECVID corpus, we classify the quality of initial search results A_{init} into a 5-leveled scale basis using the density D_c of correct shots in the top 500 automated search results. The 5 quality levels are: (a) Excellent (5): $D_c >$ 0.5; such queries are either easy or have abundant amount of correct shots in the entire corpus. (b) Good (4): $0.5 \geq D_c > 0.25$. (c) Normal (3): $0.25 \geq D_c > 0.1$; this is the range where most past queries tend to fit in. (d) Bad (2): $0.1 \geq D_c > 0.05$; such queries are either difficult or have few correct shots. (e) Random (1): $D_c \leq$ 0.05; range of most difficult queries, or having limited amount of positive shots in the corpus, or when auto search technique fails.

Quality of current display results A_{curr}: The quality of the current set of results which the user is looking through. Since the shots are displayed in such a way that the more probable shots are shown first, it makes sense for us to gauge the quality A_{curr} based on the speed of locating the positive shots. A_{curr} is thus dynamically computed by counting the number of positive shots found in the last T seconds (where T=30). As the user screens through the list of relevant shots, we are expecting A_{curr} to drop considerably to signify the need for a change of strategy. Alternatively, if there is a sudden increase in $A_{curr,}$ it makes sense for the user to continue with the current set.

Proportion of current positive over negative A_{pro}: The A_{pro} is a good indicator of the difficulty level of the query. A constant high A_{pro} can signify that the query may be easy or simply means that there are many positive instances in the corpus. Usually, the number of negative shots found can greatly overwhelm the positive ones, but there are also special cases such as the query like "find shots containing a face" where the A_{pro} remains high throughout.

Besides the main factors A_{init}, A_{curr}, and A_{pro}, other possible factors that may affect the decision made by the experts include: the total number of positive shots A_{pos}; total number of negative shots A_{neg}; time lapse into a retrieval A_{time}; and query-class A_c. It is noted that the above factors can only model the situation accurately given that annotation is done in a steadfast manner. That is, the users will label the retrieved images as positive or negative as soon as they see the images. There are no other special circumstances like: a situation when the user pauses indefinitely (e.g. leaving the computer) or periodically (e.g. being distracted by performing other concurrent tasks).

4.2 Adaptive Feedback Selection Model

To train the selection model, we collect the performance statistics of 5 expert users on the TRECVID 2005 dataset. We utilize all the 24 queries defined in TRECVID 2005 and adopt the standard guideline of letting the users perform 15 minutes of interaction with the system for each query. We gather the search log $\Phi(A, S)$ of the 5 expert users consisting of A: $\{A_{init}, A_{curr}, A_{pro}, A_{pos}, A_{neg}, A_{time}, A_c\}$ and the expert's selection S: $\{(RRF \mid PAL \mid LRF)\}$.

With the collected $\Phi(A, S)$, we reduce the selection problem to a multi-class classification problem. The overall framework for adaptive strategy selection is shown in Figure 4. The observations denoted by $A\{\}$ at a particular time instance can be used as the determinants for class S which is the type of feedback techniques. In particular the adaptive feedback selection model follows a multi-class SVM trained accordingly to the collected set $\Phi(A, S)$. To obtain a 3-class classifier, we first construct a set of binary classifiers $\{f_1, f_2, f_3\}$, with each trained to separate one class from the rest using Eqn(4). We then combine the outputs of $\{f_1, f_2, f_3\}$ by performing the multi-class classification according to the maximal output before applying the *sgn* function [23].

$$\arg\max_{j=1..3} g^j(x), \text{ where } g^j(x) = \sum_{i=1}^{3} (y_i \alpha_i^j k(x, x_i) + b^j)$$
$$f^j(x) = sgn(g^j(x)) \quad (4)$$

where $k(x,x_i)$ is the kernel function. All the other parameters such as threshold b are found during training by solving a quadratic programming problem. x_i is a subset of the training set (the Support Vectors) and $y_i\alpha_i$ is computed from the Lagrange multipliers.

The final output $f^j(x)$ determines the choice of feedback strategy to recommend. Besides, the values of $g^j(x)$ can also be used for reject decisions [23]. To see this, we consider the difference between the two largest $g^j(x)$ as a measure of confidence in the classification of x. If that measure falls short of a threshold, the classifier rejects the pattern and does not assign it to a class. In this case, no recommendation will be made to the user. .

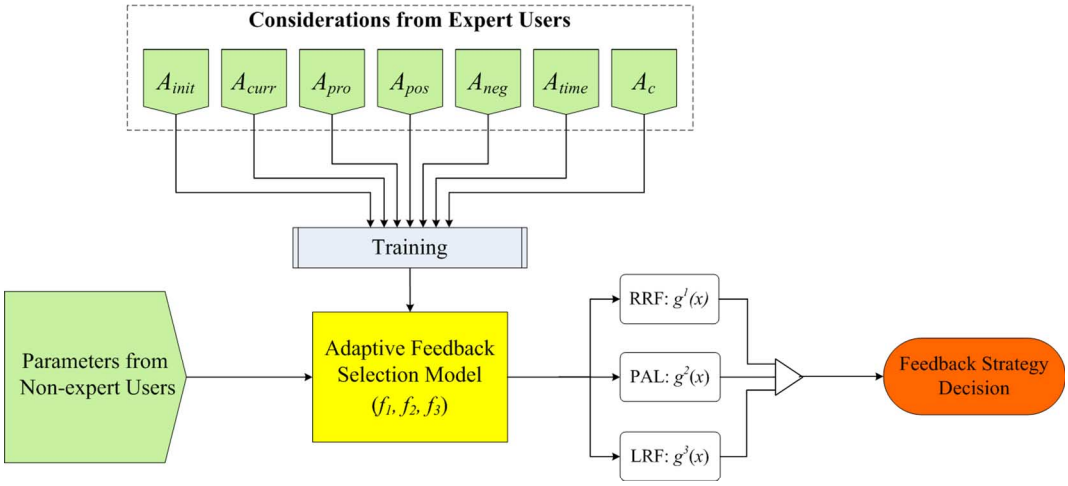

Figure 4. Framework for adaptive strategy selection

One issue with such classification is that the binary classifiers used are obtained by training on different binary classification problems, and thus it is unclear whether their real-valued outputs are on comparable scales. This is a difficult problem, which often arises when several binary classifiers attempt to assign the pattern to their respective class (or where none does) [23]. We will explore this issue in our future work. But one class must be chosen by comparing the real-valued outputs in our current implementation

5. EXPERIMENTS

We adopt the TRECVID 2006 dataset as the testing set to perform the evaluation. This dataset consists of 160 hours of English, Chinese and Arabic news videos recorded in late 2005 with a total of 24 queries. It is a widely used benchmark dataset for testing video retrieval performance. We adopt the same evaluation methodology as in TRECVID interactive video search task with MAP (Mean Average Precision) criterion, which is suitable for evaluation in information retrieval over large corpuses where recall rate is hard to determine. The user is given 15 minutes to submit a maximum of 1,000 shots for each query. The result is then automatically evaluated using the ground truth provided.

In order to test the effectiveness of our proposed approaches mentioned above, we divide the experiment into three parts. The first part is designed to test the various feedback techniques and the next two parts are designed to test the adaptive selection strategy for non-expert users. In particular, the second part tests the retrieval performance of 10 selected users using the adaptive selection strategy, while the third part conducts further experiments with the adaptive selection model to enhance the accuracy.

5.1 Multiple Feedback Strategies

To understand the improvements brought about by the use of multiple feedback strategies (RRF, PAL and LRF), we design 6 runs as detailed in Table 1 to evaluate the performance of each strategy as well as their combinations.

Table 1. Descriptions of runs

RUN NAME	DESCRIPTION
S1	Auto search (baseline)
S2	S1 + user's annotation only
S3	S1 + Precision-driven active learning (PAL)
S4	S1 + Recall-driven relevance feedback (RRF)
S5	S1 + Locality-driven relevance feedback (LRF)
S6	S1 + Multiple feedback strategies
R1, R2, R3	Best reported interactive runs from TRECVID 2006, MAP of 0.303, 0.267, 0.226 respectively

S1 is generated by automatically collating the top 1000 returned shot and will establish the baseline performance for automated retrieval. S2 focuses on leveraging the efforts of users, who will attempt to label as many shots as possible without performing any form of feedback technique. In short, S2 can be viewed as a baseline run for interactive retrieval, in which the labeled relevant shots are simply re-ranked to the top of the result list for MAP computation. S3, S4 and S5 allow the user to use different forms of feedback strategies, whereas S6 allows the user to flexibly choose among the three strategies. For comparison purposes, we also include the top 3 best reported runs (R1 to R3) from TRECVID 2006.

The experimental results are shown in Figure 5. From the Figure, we can see that the distinction in results especially between S1 and the rest of interactive runs. The runs with a human annotator perform 3 times better than S1 because of the precision jump caused by returning the positive shots nearer to the top of the rank list. Due to the lack of feedback techniques, S2 performs worse than S3, S4 and S5. Also, S3 is slightly better than S2 whereas S4 and S5 yield significant improvement. These imply that the use of feedback strategies can be important in the overall retrieval process. When using all the three feedbacks in S6, we achieve significantly better results than all previous runs and outperform the best reported interactive search run in TRECVID 2006 [24].

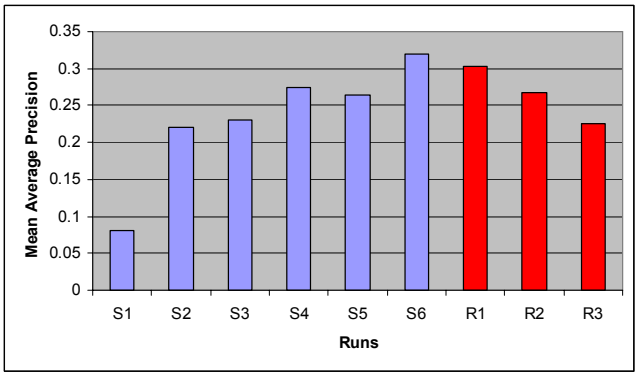

Figure 5. Performance of runs based on different feedback strategies

5.2 Adaptive Selection Model

To further access the effectiveness of the selection model, we carried out additional testing of our system with the help from 10 users. We selected 5 novice users and 5 expert users to carry out the retrieval with the help of the adaptive selection model. As the expert users are familiar with the capabilities and limitations of the retrieval system, they are able to capitalize on their familiarity and maximize the throughput. However, such domain knowledge may not be available to the novice users or first-time users. We would also like to see how the searchers perform with and without the help of the selection model. Thus, we design 2 set of runs: T6 without the use of adaptive selection model and T7 with the use of adaptive selection model. The users are first given a short briefing to the various functions on the user interface and various feedback techniques, RRF, PAL and LRF. The 24 queries are then split into 12 queries for T6 and 12 queries for T7 for each user in order to ensure that the users do not have to repeat queries so as to avoid the bias towards the experiment. In addition, we did a query rotation to ensure that each set of the 12 queries are different. The performances (average MAP) of the 5 novice users and the 5 expert users are shown in Figure 6.

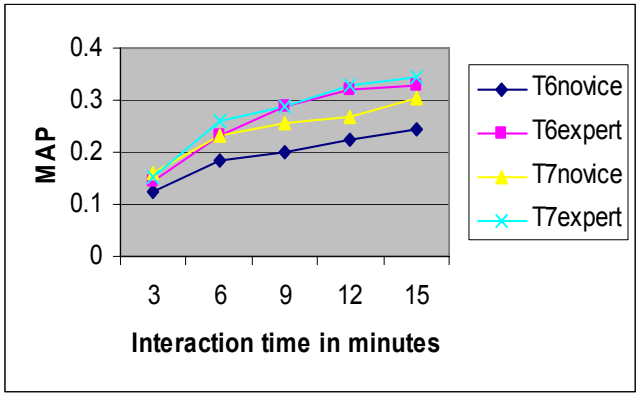

Figure 6. Performance against interaction time

From Figure 6, we can derive the following observations. First, the use of adaptive selection strategy is very effective for novice users. It results in significantly better performance for *T7noice* as compared to *T6novice*. Moreover, the best MAP in *T6novice* (0.25) can be reached by *T7novice* by using only half the interaction time needed. Second, the average MAP obtained by the 5 novice users is not comparable to that of the 5 expert users. The main reason is that the novice users are not trained to annotate the images in a competitive manner. This is further confirmed by noting that the total number of keyframes judged by the novice users is only 60% that of the expert users. Third, the *T7expert* run outperforms the *T6expert* as it is able to achieve higher MAP in a shorter amount of time.

Overall, the adaptive strategy hides the complexity of having the users to decide which feedback techniques to use, and helps to enhance the search performance of both the novice and expert user groups.

5.3 Enhanced Adaptive Selection Model

From Section 5.2, we conjecture that the use of recommendations provided by the selection model is crucial to helping novice users and even saving precious search time for the expert users. However, the current training set $\Phi(A, S)$ is solely based on human experiences. This selection S is thus not an optimal representation as it is based on human-level intuition rather than actual performance for precision. To further understand which strategy really works best under different circumstances, we design another experiment which is based on actual ground truth to automatically generate an S' during selection. Here, S' is defined to have the "greedy" property which maximizes the number of positive shots in the new top n returned list $(n=100)$. The process first generates 3 lists of returned results using the 3 feedback strategies, given the same conditions $A\{\}$. Each list is subsequently checked against the ground truth. The feedback technique S which results in the list containing the most number of positive shots ranked within the top 100 positions will be selected. Table 2 shows the statistics of user agreement between S and S' for the 5 human experts. Here the positive aggregated agreement is the total number of choices that tally with the S', and vice versa for negative aggregated agreement. For comparison purposes, we also added S_n which is the selection decision made by the 5 novice users as in Section 5.2.

Table 2. Positive aggregated agreement

	S (decision from experts)	S_n (decision from novices)
Positive Aggregate Agreement	801 (89.4%) Range: (87.3%~94%)	545 (72%) Range: (56.2%~80.5%)
Negative Aggregate Agreement	95 (10.6%) Range: (7.7%~12.7%)	212 (28%) Range: (20.3%~34.9%)

Considering that there are only 3 choices, the deviation by random chance is 66%. From the statistics, we can see that expert users deviate only about 10% of the times from S', whereas novice users range from 20% to 35%. It is therefore justifiable to say that expert users generally make better choice than novice users. This also further explains the significant improvement in Section 5.2 when novice users tap onto the recommendations generated by the adaptive selection model.

Using S', we re-train and obtain the enhanced adaptive selection model. To test the effectiveness of this enhanced model against the old model, the same users are asked to perform a final round of search. In this experiment setup, each user is asked to perform the search using the 24 search queries. From a total of 10 searchers, this amounts to a total of 240 queries, from which we pre-assign 80 queries to use the recommendations based on the enhanced model; 80 queries based on the old model; and 80 queries based on an algorithm which randomly selects one of the 3 techniques (dummy model). The user performing the search will not be told which model is currently applied on a per query basis so as to keep the experiment fair. The results are tabulated in Table 3. The results clearly show that the enhanced model is significantly better than the old model.

Table 3. Performance for various selection models

	Enhanced Model	**Old Model**	**Dummy Model**
MAP	0.367	0.345	0.289

5.4 Analysis of Results

Our analysis of results shows that by using the adaptive selection model learned from the search behaviors of expert users, the novice users are able to ignore the complicated process of feedback selection and make the best choice most of the time. For example, when querying the topic "George Bush walking" in TRECVID 2006, we found in our experiments that many novice users would select the PAL strategy to help improve the search performance. In fact, the RRF strategy is a much better choice because the ASR text is basically reliable and rich in content for news video corpus. The adaptive model would often recommend the RRF strategy to novice users by using the extracted relevant text tokens such as "white house", "president" and "US", which are really useful to find more relevant shots relating to "George Bush". Another example is the query "Street market" in TRECVID 2007, where novice users hardly know which strategy to select. But they were often recommended by the adaptive model to use the LRF strategy because the documentary video has high temporal coherence. This has resulted in improved search performance for this query. Therefore, our adaptive selection model can help novice users to cross different queries and different domains.

6. CONCLUSIONS

In this paper, we propose multiple feedback strategies, namely recall-driven, precision-driven and temporal locality-driven feedback, to provide the searcher with more options to handle different search situations. The user can choose the most suitable feedback techniques based on his/her intuition or experience to maximize the performance. In order to cater to the large number of non-expert users, we further present a recommendation mechanism for adaptive feedback selection, which can be done by learning from the expert users' search behaviors. The recommendation model suggests a suitable feedback strategy to the user at a proper time so as to enable the novice users to achieve a similar level of search performance as the expert searchers. Experiments show that our proposed approaches are effective and useful.

For future work, we will look into studying different contributions of three feedback functions for different query classes and different video domains, leading to more precise and robust searches for the users.

7. ACKNOWLEDGMENTS

This research work is performed while Huanbo Luan is an intern student at the National University of Singapore. The research is jointly supported by I2R-A*STAR (Singapore, R-252-000-192-593), the National Basic Research Program of China (973 Program, 2007CB311100), the National Nature Science Foundation of China (60773056), the Beijing New Star Project on Science & Technology (2007B071), and the National High Technology and Research Development Program of China (863 Program, 2007AA01Z416). The authors would like to thank Hai-Kiat Goh, Weifeng Zhuo, Gabriel W.Z. Leong and Xiufeng Hua for their helpful supports and discussions.

8. REFERENCES

[1] Christel M., Huang C., Moraveji N., Papernick N.: Exploiting multiple modalities for interactive video retrieval. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 1032-1035, Montreal, Canada, 2004.

[2] Snoek C. G. M., Worring M., Koelma D. C., Smeulders A. W. M.: A learned lexicon-driven paradigm for interactive video retrieval. IEEE Trans. on Multimedia, 9(2):280-292, 2007.

[3] Rui Y., Huang T. S., Mehrotra S., Ortega M.: A relevance feedback architecture in content-based multimedia information retrieval systems. In: Proc. IEEE Workshop on Content-based Access of Image and Video Libraries, pp.82-89, Puerto Rico, 1997.

[4] Rui Y., Huang T. S., Mehrotra S.: Content-based image retrieval with relevance feedback in MARS. In: Proc. IEEE Int. Conf. on Image Processing, pp.815-818, USA, 1997.

[5] Wei C. H., Li C. T.: Content-based multimedia retrieval. In: Proc. Encyclopedia of Multimedia Technology and Networking, pp.116-122, Hershey, PA, USA, 2005.

[6] Crucianu M., Ferecatu M., Boujemaa N.: Relevance feedback for image retrieval: A short survey. In: State of the Art in Audiovisual Content-Based Retrieval, Information Universal Access and Interaction, 2004.

[7] Guo G., Jain A.K., Ma W., Zhang H.: Learning similarity measure for natural image retrieval with relevance feedback. IEEE Trans. Neural Networks, 12(4):811-820, 2002.

[8] Hong P., Tian Q., Huang T. S..: Incorporate support vector machines to content-based image retrieval with relevant feedback. In: Proc. IEEE Int. Conf. Image Processing, pp.750-753, Vancouver, Canada, 2000.

[9] Tao D., Tang X.: Random sampling based SVM for relevance feedback image retrieval. In: Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition, pp.647-652, Washington, DC, USA, 2004.

[10] Zhang L., Lin F., Zhang B.: Support vector machine learning for image retrieval. In: Proc. IEEE Int. Conf. Image Processing, pp.721-724, Thessaloniki, Greece, 2001.

[11] Tao D., Tang X., Li X., Wu X.: Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. IEEE Trans. on Pattern Analysis and Machine Intelligence, 28(7), 2006.

[12] Hoi S. C. H., Lyu M. R.: A semi-supervised active learning framework for image retrieval. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.302-309, San Diego, CA, USA, 2005.

[13] Szummer M., Jaakkola T.: Partially labeled classification with markov random walks. In: Proc. Advances in Neural Information Processing Systems, pp.945-952, Vancouver, Canada, 2001.

[14] Hoi S. C. H., Jin R., Lyu M. R.: Large-scale text categorization by batch mode active learning. In: Proc. Int. World Wide Web conference, pp.633-642, Edinburgh, England, UK, 2006.

[15] Tong, S., Chang, E.: Support vector machine active learning for image retrieval. In: Proc. ACM Int. Conf. on Multimedia, pp.107-118, New York, NY, USA, 2001.

[16] Hauptmann A. G., Lin W. H., Yan R., Yang J., Chen M. Y.: Extreme video retrieval: Joint maximization of human and computer performance. In: Proc. ACM Int. Conf. on Multimedia, pp.385-393, Santa Barbara, CA, USA, 2006.

[17] Rooij O., Snoek C. G. M., Worring M.: Query on demand video browsing. In: Proc. ACM Int. Conf. on Multimedia, pp.811-814, Augsburg, Germany, 2007.

[18] Luan H. B., Neo S. Y., Goh H. K., Zhang Y. T., Lin S. X., Chua T. S.: Segregated feedback with performance-based adaptive sampling for interactive news video retrieval. In: Proc. ACM Int. Conf. on Multimedia, pp.293-296, Augsburg, Germany, 2007.

[19] Chua T. S., Neo S. Y., Zheng Y. T., Goh H. K., Xiao Y., Zhao M.: TRECVID 2006 by NUS-I2R. In: Proc. TRECVID, Gaithersburg, USA, 2006.

[20] Chua T. S., Neo S. Y., Goh H. K., Zhao M., Xiao Y., Wang G.: TRECVID 2005 by NUS PRIS. In: Proc. TRECVID, Gaithersburg, USA, 2005.

[21] Robertson S.E. and Sparck Jones K, "Relevance Weighting of Search Terms," Journal of the American Society for Information Science, 27(3), 129-146, 1976.

[22] Christel M., Yan R.: Merging storyboard strategies and automatic retrieval for improving interactive video search. In: Proc. Int. Conf. on Image and Video Retrieval, pp.486-493, Amsterdam, The Netherlands, 2007.

[23] Bernhard Schölkopf and Alex Smola: Learning with Kernels. MIT Press, Cambridge, MA, USA, 2002.

[24] http://www-nlpir.nist.gov/projects/t01v/