Segregated Feedback with Performance-based Adaptive Sampling for Interactive News Video Retrieval

Huan-Bo Luan[1,3,4], Shi-Yong Neo[2], Hai-Kiat Goh[2], Yong-Dong Zhang[1,3], Shou-Xun Lin[1,3], Tat-Seng Chua[2]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
[2]Department of Computer Science, National University of Singapore, Singapore 117543
[3]Key Laboratory of Intelligent Information Processing, ICT, CAS, Beijing 100080, China
[4]Graduate School of the Chinese Academy of Sciences, Beijing 100039, China
{hbluan, zhyd, sxlin}@ict.ac.cn, {neoshiyo, gohhaiki, chuats}@comp.nus.edu.sg

ABSTRACT

Existing video research incorporates the use of relevance feedback based on user-dependent interpretations to improve the retrieval results. In this paper, we segregate the process of relevance feedback into 2 distinct facets: (a) recall-directed feedback; and (b) precision-directed feedback. The recall-directed facet employs general features such as text and high level features (HLFs) to maximize efficiency and recall during feedback, making it very suitable for large corpuses. The precision-directed facet on the other hand uses many other multimodal features in an active learning environment for improved accuracy. Combined with a performance-based adaptive sampling strategy, this process continuously re-ranks a subset of instances as the user annotates. Experiments done using TRECVID 2006 dataset show that our approach is efficient and effective.

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

General Terms: Algorithms, Design, Experimentation

Keywords: News Video Retrieval, Relevance Feedback, Active Learning

1. INTRODUCTION

Video retrieval has become increasingly important especially with the ever-increasing amount of multimedia data. Users demand highly intelligent systems that possess the ability to retrieve and interact. For this reason, recent video research has looked into the use of feedback from the users through an interactive process to refine the search results. The common feedback process employs the user-annotated samples to explore and rank un-judged video segments. User-labeled samples are usually able to provide the additional semantics and have been shown to be especially effective in linking the text query to low-level visual descriptors [1], and thus closing the so-called *semantic gap*. Relevance feedback and more recently active learning are two standard techniques that have received much attention towards tackling this interactive learning problem.

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.

Relevance feedback usually suffers from the small sample learning problem [2], due to limited training examples. Consequently, many potential discriminating observations go unlabeled. In recent years, attention has been given to systems that employ active learning to address these challenges. The active learning paradigm guides users to label samples that are most informative (or ambiguous) to the classifiers. The labeled data can then be used to better train the system before repeating the process multiple times. This allows the weak classifiers to improve the performance after each consecutive cycle simply because they are provided with "better" training samples. However, active learning techniques usually come with the curse of dimensionality problem which is impractical for large corpus usage. As such, active learning is usually applied only to a small subset of documents which are deemed interesting. In addition, the mentality of users during a search is tuned towards looking for correct answers in the shortest possible time, rather than spending hours annotating ambiguous materials. It is therefore necessary to balance between displaying the most relevant and most ambiguous instances to the users.

In this paper, we propose to segregate the process of relevance feedback into 2 distinct facets: the recall-directed feedback and precision-directed feedback. The recall-directed facet employs general features such as text tokens from automated speech recognition (ASR) and high level features (HLFs) to maximize the efficiency during the feedback process, making it suitable for large corpuses in a real time search. These general features have been shown to be effective in high recall retrievals [3]. The precision-directed facet, on the other hand, uses other multimodal features in an active learning environment targeted at improving precision. Fused with a performance-based adaptive sampling strategy, this process continuously re-ranks a subset of instances as the user annotates. The performance-based sampling strategy will adaptively choose instances either most ambiguous or most relevant from the classification output with emphasis on maximizing precision in a minimal time. Experiments done using TRECVID 2006 dataset show that our relevance feedback approach can outperform reported systems.

2. INTERACTIVE VIDEO RETREIVAL

Video retrieval has primarily evolved from text-based search using ASR to incorporate low-level video features (such as color, motion, volume) and, more recently, HLF for specific objects or phenomenon (such as cars, fire and applause). The system analyzes the query from users and returns short precise news video segments as answers. In particular, interactive retrieval systems further interact with the users to refine retrieval using user annotated

samples. TRECVID [4], an international video retrieval evaluation forum, has been carrying out video processing benchmarking since 2001. One of the primary tasks, the interactive search, is targeted at uncovering as many relevant shots as possible, for a given free text query, based on user interaction in a stipulated time frame (15 minutes). The task definition is shown in Fig 1.

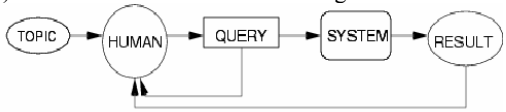

Figure 1: Interactive search task proposed by TRECVID

During the interactive search, the user can either manually refine the given query or provide relevance feedback by annotating the retrieved shots as relevant or irrelevant to the query. As the retrieval must be realizable in real-time and the search time is fixed, top performing participating teams have proposed different strategies to handle this task. Informedia [5] focused on maximizing human annotation efforts. They proposed an extensive annotation strategy based effectively on human reaction time and set an unbeatable record of annotating 5000 shots in 15 minutes. MediaMill [6] focused on building an intuitive interface with flexible display capabilities to interact with the users and made use of relevance feedback mainly based on HLFs. IBM [7] similarly proposed various feedback models with emphasis on visual features and text. The performance of the above systems show that a combination of maximizing user effort, having an intuitive interface, and providing effective relevance feedback can contribute significantly to better results. With these considerations we design our interactive news video system as shown in Fig 2.

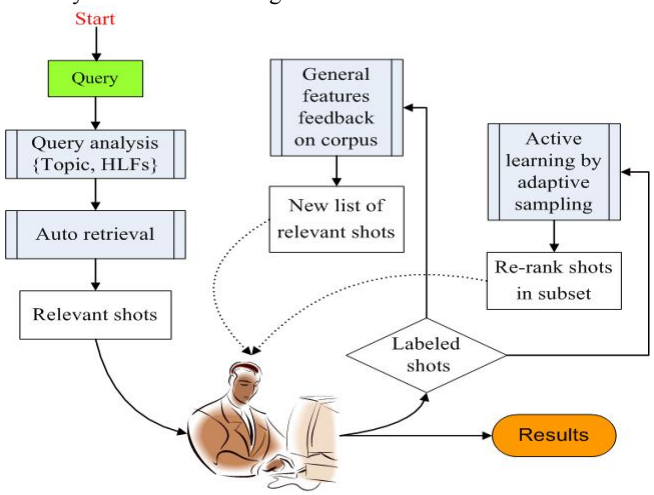

Figure 2: Overall framework for interactive video retrieval

We first employ our automated search techniques as presented in [8] to perform query analysis and initial retrieval. After that the user will annotate and iterate the relevance feedback process described in the next section. Query analysis extracts ***Q_topic*** and ***Q_HLF[]*** from the query. *Q_topic* provides the most essential key terms required for ASR retrieval while *Q_HLF[]* suggests the importance of various HLFs to the query. After using *Q_topic* for initial retrieval, *Q_HLF[]* is used to re-rank the ASR retrieved shots to give priority to shots which contains query-relevant HLFs. This automated search system is highly competitive as it is one of the top performing systems in the category of automated search task [4].

3. SEGREGATED RELEVANCE FEEDBACK

To effectively leverage users' annotations, we segregate the feedback process into 2 distinct stages. The first stage emphasizes on analyzing and applying the correlation of general features obtained from labeled positive and negative instances to provide for high recall retrieval on the entire corpus. The second stage uses active learning with a precision-emphasis sampling strategy to continuously refine the re-ranking model using a combination of multimodal features.

3.1 General features for high-recall feedback

Given a large corpus, re-ranking in the later stages might be redundant if relevant shots are not presented in the initial retrieval. Hence, in order to maximize recall performance and minimize computational complexity, we choose to employ feedback using general features such as text and HLF. Text features have been shown in many related works to provide an excellent recall measure. Results in [3] show that more than half of the positive shots in a video corpus can be found by using simple text retrieval. In particular, Informedia interactive system [5] demonstrates that the use of text feature when coupled with their extensive user interface can even outperform systems based on combination of multimodal features. Besides using ASR text in a given shot, we further enhance our recall by representing each shot with an entire ASR phrase segment which overlaps across shot boundaries on a time scale. These phrases usually contain coherent segments of text and are readily available from the Machine-Translation output provided by TRECVID for non-English news. In order to detect an entire phrase for English news, we segment the continuous English ASR by looking for substantial long silence or a change of speaker.

Due to the drawbacks of ASR text, HLFs are also leveraged as one of the general features. One serious problem faced by text-alone retrieval systems is the inability to retrieve shots which do not have any corresponding text. A particular shot without text will not be retrievable at all. It is therefore important to have other semantic features which are representative and can complement text-based retrieval. In this work, we make use of 50 HLF concepts in our experiments from [9]. This includes the 39 TRECVID concepts, such as people, car, boat, etc, as well as 11 genre and audio concepts, such as sports, commercial, applaud, etc.

Given the positively annotated shots *Np*, the feedback process at this stage relies on the text and HLF scores to iteratively adjust the retrieval function. The text from *Np* will first be analyzed and then feature selection will be performed to find highly discriminating text tokens for retrieval. We employ the 0.5 formula [10] as shown in Eqn 1 to select the top *k* terms which are most salient.

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

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*.

The HLF score for each labeled shot is used to estimate the new $Q\_HLF^{(1)}$, which is the new relevancy of HLFs to the query. This is done by averaging the detection confidence of each HLFs in the *Np* relevant shots as shown in Eqn 2.

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

where $S_i^{HLF}[]$ is the detection confidence of various HLFs in shot i.

This detection confidence of HLFs can be taken as appearance likelihood. If many positive shots have high scores for certain HLF, it will explicitly mean that the query is closely related to that HLF. After obtaining the text and HLF scores, we compute the new scores for each individual shot by using Eqn 3.

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

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 $Q\_HLF^{(1)}[]$. A high SD will signify that the query shows reliance to certain HLFs. Alternatively, if the SD is low, it means that there is low correlation between HLF and the query.

3.2 Adaptive sampling for active learning

To complement the high-recall feedback, the active learning based on multimodal features is also carried out on a subset of retrieved shots (set to 3000 in our experiments). We choose to use Support Vector Machine (SVM) for active learning since our problem can be simplified into a simple probabilistic binary classification problem (either positive or negative), and SVM is able to perform this task with high efficiency. The selected visual features consist of the 50 HLFs used in section 3.1, color moment features (1st, 2nd, 3rd) obtained at a 3x3 block, 8 directional motion features and one global motion feature. The overall dimension of the feature space is 86. We define our decision boundary function of SVM with RBF kernel setting as follows:

$$f(x) = w\cdot\Phi(x),\ where\ w = \sum_{i=1}^{n}\alpha_i\cdot\Phi(x_i) \quad (4)$$

where α_i is the coefficient of the corresponding support vector. Given a set of labeled training data set $\{(x_1,y_1)...(x_n,y_n)\}$, where x_i is the feature vector and y_i is the binary class of positive or negative, we encapsulate a Mercer kernel $K(x_i, x_j) = \Phi(x_i).\Phi(x_j)$ and project the original training data from the original data space to a higher dimensional feature space F. The classifier can then be used to output the distance of unseen instances from support vectors or classification boundaries.

Most active learning algorithms [11] sample instances close to classification boundaries for users to judge as this will enable classifiers to converge faster. However, in normal searching circumstances, searchers are more interested in finding correct shots in the shortest amount of time. This implies that shots that are deemed more relevant should be presented rather than shots that are likely to be irrelevant. This is especially true for retrieval tasks with a strict time constraint. To handle this phenomenon, we propose an adaptive sampling strategy based on the classification output quality. The system will first sample instances furthest from the boundaries (those deem to be most relevant) instead of instances near the boundaries, and if many of these instances are indeed marked positive by the user, the system will continue to sample in that particular region. On the other hand, if these samples are mostly negative, the system will then switch to sample more instances near the classification boundaries. An adaptive sampling strategy based on this rationale is developed to allow users to have quick access to highly probable relevant shots. In addition, it provides a fast convergence to deliver more accurate classification. The strategy is shown in Fig 3.

Input: L, U, α, β, k
Definitions: L: Labeled samples, U: Unlabeled samples, α: set of adaptive range, β: adaptive threshold, k: sample size
Output: m
Procedures:
- SVM training: $SVM()$
- Classifier learned by SVM on L: $f()$
- Initialization: $\alpha = \alpha_{max}$, $\beta = 0$, $positive_m$=% of positive shots in previous m_f

BEGIN
1. $f \leftarrow SVM(L)$;
2. For each $x_i \in U$
   $x_i.distance() \leftarrow f(x_i)$;
3. If ($positive_m > \beta \cap \alpha < \alpha_{max}$) $\alpha.increase()$;
4. If ($positive_m < \beta \cap \alpha > \alpha_{min}$) $\alpha.decrease()$;
5. Sampling m: $m_f + m_n$
   a) m_f: $k\alpha$ instances of $x \in U$ *furthest from boundaries*
   b) m_n: $k(1-\alpha)$ instances of $x \in U$ *nearest to boundaries*

END

Figure 3: Adaptive sampling strategy

In our implementation, α is empirically set at [0.3, 0.8] and β at 0.25.

4. INTUITIVE USER INTERFACE

Besides having satisfactory system performance, a good and intuitive user interface (UI) is required to maximize user's annotation efforts. A sample of our interactive UI is shown in Fig 4. The UI will display three images at a time in a central active row, with the previous and next rows in view. We experiment with various types of display and discover that the user reaction time is quickest when annotating three images at a time. The user will determine the images' relevance to the query and then annotate the positive ones by hitting pre-defined keys on the keyboards. The system will then capture the user's input and automatically refresh itself to display the next row of new images. In the event that no image is relevant to the query, the user can hit the "Space" key to skip a row. In addition, the "Space" can also be pressed and hold to "fast forward". Alternatively, the "Backspace" key is used to undo changes and also backtrack when the user need to perform corrections. This interface is intuitive and is demonstrated to be capable of annotating about 1000 images in 5 minutes.

Figure 4: Interactive user interface

5. EXPERIMENTS

We make use of the same evaluation methodology as in TRECVID 2006 interactive search task in our experiments. The user is given 15

minutes to interact with the system after the query is issued. The system will return a maximum of 1000 shots for each query. This TRECVID 2006 testing corpus consists of 160 hours of news video collected in late 2005 with a total of 24 queries. The performance measure used is the mean average precision (MAP). To test the various techniques mentioned in the paper, a series of runs (T2 to T5) have been designed and carried out using our interactive UI. The first run T1 consists of using only results from the automated search without any user interaction. The main objective is to establish the baseline performance, i.e. without implementing any additional techniques. The second run T2 includes user interaction without any relevance feedback. T3 employs the relevance feedback techniques in [7]. T4 is based on our 2-stage segregated feedback model with sampling based on most informative instances. T5 enhances T4 with adaptive sampling. All five runs are carried out by a user who is familiar with the capabilities of the system.

Table 1: MAP performances of runs T1-T5

Run	Average no. of shot judged	Average no. of positive shot found	Mean Average Precision (top 1000)
T1 *baseline*	Auto	63	0.077
T2 (user only)	2933	118	0.234
T3 (user, [5])	**2535**	**130**	**0.258**
T4 (user, segregated)	**2510**	**148**	**0.304**
T5 (user, segregated, adaptive)	**2488**	**161**	**0.334**

From the results, we can conjecture that runs that employ feedbacks (indicated in bold) generally perform better, thus implying the use of feedback is important. In particular, both T4 and T5 which uses the segregated feedback approach can achieve significantly better results than T3 which uses a single layer relevance feedback. The best performing run comes from T5 which employs adaptive sampling. The MAP performance of T5 is also statistically better than the best reported interactive search run in TRECVID 2006 (with a MAP of 0.308 [5]). In addition, we observe that as we improve the searching techniques from T1 to T5, the MAP increases while the number of judged samples decreases. This happens naturally as more positive shots are presented to the user for judging and users tend to take more time to confirm a positive shot. To further study the performance of various runs against interaction time, we plot the MAP performance of runs T2 to T5 using the rank list generated at fixed time interval as in Fig 5.

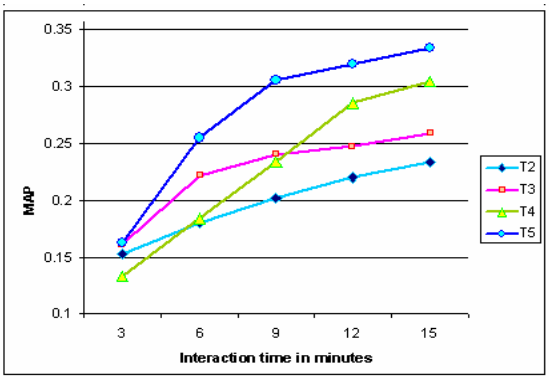

Figure 5: MAP performance against interaction time

From Fig 5, we can see that T5 is also the best run in terms of efficiency. It only requires 9 minutes of user interaction time to attain the same MAP performance as T4 which has a recorded MAP of 0.304 after 15 minutes of user interaction. We also observe that the MAP performance of T3 to be higher than T4 during the first few minutes and subsequently become almost standstill, which signifies the depletion of positive instances.

6. CONCLUSIONS

As video retrieval evolves from the traditional one-way retrieval to a more interactive process, appropriate strategies have to be deployed to ensure efficiency and accuracy. In this paper, we segregated the process of relevance feedback into 2 distinct facets, one focusing on recall and the other on precision. The recall-directed facet employs general features to maximize efficiency during the feedback process, making it suitable for large corpuses. The precision-directed facet exploits multimodal features in an active learning environment that incorporates a performance-based adaptive sampling strategy to continuously re-rank a subset of instances as the user annotates. Experiments on TRECVID 2006 dataset showed that our approach outperformed reported systems.

7. ACKNOWLEDGMENTS

This work is jointly supported by Beijing Science and Technology Planning Program of China (D0106008040291) and I^{2}R-A*STAR (Singapore) research grant number R-252-000-192-593. We also like to thank Dr Tang Sheng, Zhang Xu, Hua Xiufeng and Tian Jiuming for the help in performing the experiments.

8. REFERENCES

[1] R. Yong, T.S. Huang, M.M. Ortega, "Relevance feedback: a power tool for interactive content-based image retrieval," *IEEE Trans on Circuits and Systems for Video Technology*, 644-655, Sep 1998.

[2] X. Zhou and T.S. Huang, "Small Sample Learning during Multimedia Retrieval using BiasMap," in *IEEE Conference Computer Vision and Pattern Recognition*, 2001.

[3] T-S. Chua, S-Y. Neo, H-K. Goh, M. Zhao, Y. Xiao, G. Wang, "*TRECVID 2005 by NUS PRIS,*" *TREC Video Retrieval Evaluation Online Proceedings*, TRECVID 2005.

[4] TRECVID, TREC Video Retrieval Evaluation, http://www.-nlpir.nist.gov/projects/trecvid.

[5] A.G. Hauptmann, M.-Y. Chen, et al., "Multi-Lingual Broadcast News Retrieval," *TREC Video Retrieval Evaluation Online Proceedings*, TRECVID 2006.

[6] C.G.M. Snoek, J.C. van Gemert, et al., "The MediaMill TRECVID 2006 semantic video search engine," *TREC Video Retrieval Evaluation Online Proceedings*, TRECVID 2006.

[7] M. Campbell, S. Ebadollahi, et al., "IBM research TRECVID-2006 video retrieval system," *TREC Video Retrieval Evaluation Online Proceedings*, TRECVID 2006.

[8] T.-S. Chua, S.-Y. Neo, Y. Zheng, H.-K. Goh, Y. Xiao, and M. Zhao, "TRECVID 2006 by NUS-I2R," *TREC Video Retrieval Evaluation Online Proceedings*, TRECVID 2006.

[9] S.-Y. Neo, J., M. Kan, T.-S. Chua, "Video Retrieval Using High-level features: Exploiting Query-matching and Confidence-based Weighting," *International Conference on Image and Video Retrieval*, Arizona, USA, July 2006.

[10] 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.

[11] E.Y. Chang, S. Tong, K.-S. Goh, and C.-W. Chang, "Support Vector Machine Concept-Dependent Active Learning for Image Retrieval," *IEEE Transactions on Multimedia 2005.*