TRECVID 2007 Search Tasks by NUS-ICT

Tat-Seng Chua, Shi-Yong Neo, Yan-Tao Zheng, Hai-Kiat Goh, Xiaoming Zhang
School of Computing, National University of Singapore

Sheng Tang, Yong-Dong Zhang, Jin-Tao Li, Juan Cao, Huan-Bo Luan, Qiao-Yan He, Xu Zhang
Institute of Computing Technology, Chinese Academy of Sciences, China

ABSTRACT

This paper describes the details of our systems for our automated and interactive search in TRECVID 2007. The shift from news video to documentary video this year has prompted a series of changes in processing techniques from that developed over the past few years. For the automated search task, we employ our previous query-dependent retrieval which automatically discovers query class and query-high-level-features (query-HLF) to fuse available multimodal features. Different from previous works, our system this year gives more emphasis to visual features such as color, texture and motion in the video source. The reasons are: (a) given the low quality of ASR text and the more visual and motion oriented queries, we expect the visual features to be as discriminating as text feature; and (b) the appropriate use of motion features is highly effective for queries as they are able to model intra-frame changes. For the interactive task, we first utilize the results from the automated search results for user feedback. The user is able to make use of our intuitive retrieval interface with a variety of relevance feedback techniques to refine the search results. In addition, we introduce the motion-icons, which allow users to see a dynamic series of keyframes instead of a single keyframe during assessment. Results show that the approach can help in providing better discrimination.

1. INTRODUCTION

The overall framework of our video search and retrieval for both automated and interactive system is shown in Figure 1. There are two main stages: the auto search stage and the interactive search stage. The retrieval starts with the user query, which can simply be a free text query; or coupled with image and video (multimedia query). The auto search first processes the multimedia query and performs the retrieval. The emphasis is on understanding the query to infer the roles of HLF, motion and visual features in query processing. For the interactive search, the user will make use of the automated search results to indicate whether the results are indeed relevant or otherwise. The emphasis is on designing a high performance feedback system, from which users can make use of several auto-feedback and active learning functions to improve the retrieval performance.

The domain of corpus for this year is the Dutch documentary video. The videos are preprocessed, segmented into shots with the speech track automatically recognized using a commercial automated speech recognition (ASR) engine and translated to English text. As a result of ASR and translation, the quality of ASR text is quite low. This, coupled with a large number of visual and motion oriented queries, suggests that ASR text may not play a critical role in the retrieval process. In fact, visual and motion information will be as important as text, as we move from news video to Dutch documentary video retrievals.

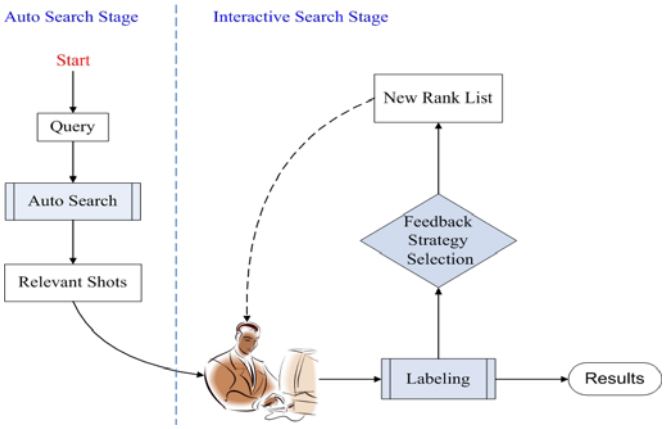

Figure 1. Framework for auto and interactive video retrieval system

2. AUTO SEARCH

As ASR text is unlikely to be very effective for Dutch documentary video, our focus this year is on effective query analysis and retrieval using HLF, motion and visual features. We induce and extract query-information like query-class and query-HLF from the text query, as well as image level features from the visual counterparts if available. This query-information is then used in shot level retrieval to rank the results. The overall framework is highlighted in Figure 2.

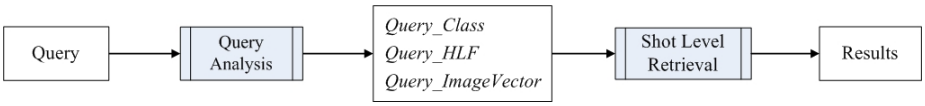

Figure 2. Framework for automatic video search

2.1 Query Analysis

Query-class has been shown in many prior works to be an important guide to fuse multi-modal features effectively. The five classes we used are {Scene, People, Object, Action, Unknown}. They are chosen because they cover most queries and can be classify using heuristic rules. In particular, 19 out of the 24 queries in the TRECVID 2007 can be classified into the first four classes. We first perform shallow linguistic parsing on the query to obtain the part-of-speech. Subsequently, we classify the query by automatically checking the lexical dictionary WordNet using the nouns and verbs as follows: scenes: {scene based nouns like vegetation, mountain, streets, etc}; people: {specific names, occupation, etc}; Object {Object nouns}; and Action {action verbs like walking, moving}. The Unknown-class is used to group queries that do not belong to any of the first four classes. This characterization is important as it is not possible to perform complex query classification for short text queries

Query-HLF on the other hand suggests possible HLFs that are important to the query in terms of visual requirements. We approach this by employing morphological analysis on the text query followed by selective query expansion using WordNet [1] on both the feature descriptions of HLFs and user's query. The stronger the match between the HLF descriptions and the query, the more important the HLF is to the query. Besides relying on lexical relationship, we also infer query-HLF from sample keyframe and shots when they are available from the multimedia query. A sample keyframe or shot containing one of the HLFs explicitly means that this particular HLF is important to the query. Additional details of our approach can be found in our prior work [2]. Finally, we combine the inferred query-HLFs from text query and video shots to obtain a better and more representative query-HLF for the query.

Query-image-feature, Q_{IMG}, is the corresponding set of video features extracted from frames of the sample video shots. We extract three visual features from all the sample keyframes, which include a 320-dimensional vector of edge histograms(***EH***) on 5 regions; a 166-dimensional color histogram(***CH***) vector in HSV space; and a set of 128-dimensional SIFT features. Then, we cluster the SIFT to 1037 visual words (***VW***) by a density-based K-means method [3], with low density visual words discarded. We regard the each dimension of ***CH*** and ***EH*** as a word, and normalize the value to [0,100]. We learn three nonparametric Latent Dirichlet Allocation (LDA) models [7] based on above three visual features (***CH, EH, VW***), and obtain the topic distribution of every shot. The optimal parameter of topic number is automatically selected through a density-based method [4].

Query-motion features: A number of query topics this year are associated with motions. For example, the queries "finding shots of train in motion" and "find shots in which a boat moves past" tend to present large horizontal translational motions in the shot, as both train and boat are highly probable to be the dominant objects in the video. Another example is the query "find shots of a road taken from a moving vehicle through the front windshield", which tends to present zoom-like global motions, as shooting through front windshield in the moving cars is similar to a zoom-in camera action. To facilitate retrieval with motion cues, we extract global motion patterns from the example shots, compute the similarity between query shot and shots in the corpus, and then fuse them into the final ranking. Specifically, we exploit eight directions of global motion: up, down, left, right, up-left, up-right, down-left and down-right, and global motion intensity. The motion cues are extracted from motion vectors stored in p-frames in compressed domain according to the method in [5]. The compressed domain motion extraction achieves efficiency by processing around 50 hours of testing videos in approximately 40 hours.

2.2 Shot Level Retrieval

The score of a shot is the fusion of its similarities to the multimedia query on multi-modal features. Query-HLF and the query-image-feature piece in nicely as they are able to suggest visual-oriented requirements and pinpoint exact relevant shots. Eqn (2) shows this shot ranking function.

$$\begin{aligned} Score(Q, Shot_j) = {} & \beta_c \cdot Text(Q, words \mid words \in Shot_j) + \\ & \gamma_c \cdot \sum_{HLF_m \in shot_j} [Conf(HLF_m) \times Sim\_Lex(Q_{HLF}, HLF_m)] + \\ & \delta_c \cdot \max_{image_n Q_{IMG}} (image\_sim(image_n, shot_j) + \\ & \chi_c \cdot \max_{image_n Q_{IMG}} (motion\_sim(image_n, shot_j) \end{aligned} \quad (2)$$

where *Text(Q, Words)* follows the vector-space model using tf.idf for document ranking. The representative words of a shot consist of the machine translated ASR text phrases that have temporal overlaps with the shot boundaries. *Conf(HLF$_m$)* denote the detection confidence of *HLF$_m$* in the shot. *Sim_Lex()* is from previous work [2] that computes the importance of various *HLF$_m$* to the Q_{HLF}, which uses WordNet to carry out lexical matching so as to infer the importance of query to HLF. The *image_Sim()* computes the image similarity in using the image features and *motion_sim* compute the motion similarity using motion features. β_c, γ_c, δ_c, and χ_c are parameters obtained heuristically through pre-run using TRECVID 2006 queries on 2007 training dataset [8].

2.3 Result Analysis

We submitted a total of 5 runs, which are as follows:
*Run1: *Required text baseline;*
*Run2: *Required visual baseline;*
Run3: Fusion without motion using only text query;
Run4: Fusion with motion using only text query;
Run5: Fusion with motion using multimedia query;

The intuition for defining these runs is to assess the effects of incorporating various features and techniques on the retrieval performance, in particular the motion. We would also like to assess the importance of text features against visual features since past results from TRECVID have found that text features are far more important than visual features. However, with the video domain changed to documentaries with poor ASR text, we speculate that the situation could be very different.

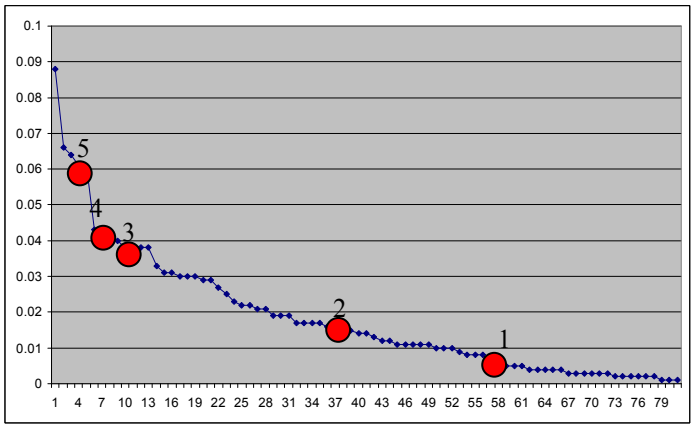

Figure 3. MAP performance of Run1 to Run5

The automated search task received 81 submissions this year. The distribution of the runs statistics are shown in Figure 3. In particular, we found that the result concurs with several of our hypotheses. Firstly, the worst performing run (Run1: MAP 0.004) comes from the text baseline. This is to be expected as speech transcripts and machine translated text are less predictive of the visual counterparts as compared to the high level feature and visual image features. In particular, the visual baseline (Run2: MAP 0.017) yields much better results than that of the text run.

The results from Run3 (MAP 0.04) and Run4 (0.043) shows that the combining HLF, text and visual features outperform both the text and visual baseline, by a large margin. In particular, Run4 that uses motion features is 8% better than Run3. The last auto run which is also our best performing run, Run5 (0.061), combines the query features from video shots with the text query for retrieval. The significant improvement indicates that query content obtained from the multimedia counterpart can be important and discriminating. This phenomenon is also observed in Run2 over Run1. Overall, our group is ranked in 2nd among all participant groups and has 3 runs in the top 10 submitted runs. We also observe that our best automated run is able to attain the best performance for at least 5 queries with many others above the median range of results.

In addition, we also observe that β_c values (text fusion parameters) used in the Eqn (2) across all query-classes are generally much lower (around 0.2) as compared to γ_c, δ_c, and χ_c values. This is very different from previous years in retrieving news video where β_c values for text are around 0.6 to 0.7. In conclusion, the overall result is indicative that the importance of text counterpart from query has dropped.

3. INTERACTIVE SEARCH

For interactive search, our emphasis is three-fold. We focus on maximizing the human annotator effort through the use of: (1) effective UI (User Interface); (2) multiple feedback strategies; and (3) motion icons.

3.1 Intuitive User Interface

To maximize user's annotation efforts, the intuitive UI is designed for fast keystroke actions with quick previews of previous and subsequent sets of shots in the ranked list. A sample of our interactive UI is shown in Figure 4. The UI will display three images at a time in a central active row, with the previous and next rows in view. We experimenting with various types of display and discovered that the user reaction time is the 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 the 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 keyframes in the rank list. For fast throughput, we designed a number of shot-cut keystrokes for quick overall actions. 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 needs to perform corrections. In our experiment, the UI enables a normal user to annotate about 3,500 shots based on motion icons or 5,000 shots based on static icons in 15 minutes.

Figure 4. Intuitive annotation user interface

3.2 Flexible Feedback Strategies

To allow for more flexibility and to provide a range of options for users to click during relevance feedback, we propose to segregate the interactive feedback into three distinct types: the **recall-driven feedback**, **precision-driven feedback** [6] and **temporal locality-driven feedback**. At any time, the users are able to select any feedback strategy to enhance the search performance when they feel that the search and feedback process is not progressing well. First, the recall-driven facet employs general features such as text tokens from automated speech recognition (ASR) and high level features (HLFs). The features from relevant shots are used to perform query expansion based on text and HLFs. This option has been found to be the most effective in finding many new relevant shots in the initial stage. Second, the precision-driven facet uses other multimodal features in an active learning environment targeted at improving precision. It uses active learning to provide long term improvements to classifiers. Third, the temporal locality-driven facet return shots from neighboring shots of the positive labeled set, as it is noted that positive shots tend to cluster near each other. Based on these multiple feedback strategies, the user will be able to choose the type of feedback that is more suitable based on his/her intuition or experience to maximize the performance. We are currently experimenting and analyzing the effectiveness of using different feedback strategies and user interface options on interactive search.

3.3 Motion icons

We observed that many queries are associated with objects in motion in the video. It is therefore necessary to provide some thing more than just static keyframes during the annotation process. Specifically, instead of displaying an icon with static keyframe for each video shot, we construct a summarized clip comprising a sequence of progressive keyframes which can show moving picture information. This method is more suitable as users are likely to waste more time if the whole shot is played. Through these motion icons, the users have a clearer idea of what information is in the shot and can identify relevant shots with better confidence. For example, the search topic 0197 is "Find shots of one or more people walking up stairs". As shown in figure 5, the image highlighted in red is the main keyframe for shot213_62. If the user were only presented with this single frame, this shot have been considered irrelevant. However, through motion icon, we can identify straight away that this is a relevant shot. The tradeoff in using motion-icon is that the display speed is slower.

Figure 5. A sequence of multiple keyframes for shot213_62

3.4 Results Analysis

Our Run6 with a MAP 0.251 is ranked as the 5th best performing run out of the 37 submitted run for interactive search. From Figure 4, we observe that 19 out of 20 queries are above median with significant margins with 2 of them obtaining the highest MAP among runs of all participants. This demonstrates that the proposed interactive search system is effective. However, we also observe that the query 210 ("Find shots of people and dogs walking") has no relevant shots found. This has contributed to our lower overall performance.

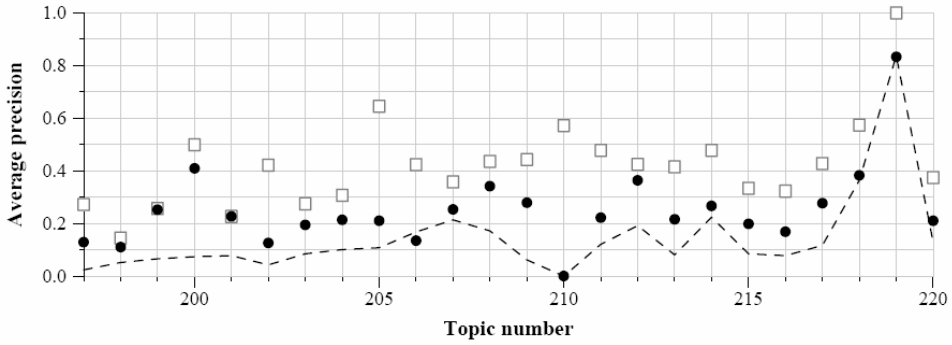

Figure 4: Average Precision Performance of Search Run6

Reference

[1] C. Leacock and M. Chodorow, "Combining Local Context and WordNet Similarity for Word Sense Identification," *WordNet: An Electronic Lexical Database,* C. Fellbaum, ed., Cambridge, Mass.: The MIT Press, pp. 265-283, 1998.

[2] S.-Y. Neo, J. Zhao, M.-Y. Kan, T.-S. Chua "Video Retrieval Using High-level features: Exploiting Query-matching and Confidence-based Weighting" In CIVR 2006, Arizona, USA, 13-15 July 2006.

[3] Y. Xie, L. Wu, S. Lao, C. Wang, Y. Hu, an efficient indexing algorithm of clustering supporting image retrieval for large image database, Mini-Microsystem, 2001

[4] J. Cao, T. Xia, S. Tang, J. Li, and Y. Zhang, A Density-Based Method for Adaptive LDA Model Selection, Neurocomputing, review.

[5] X. Zhu, A. K. Elmagarmid, X. Xue, L. Wu, and A. C. Catlin, "InsightVideo: Toward Hierarchical Video Content Organization for Efficient Browsing, Summarization and Retrieval", IEEE Transactions on Multimedia, Aug. 2005.

[6] H.-B. Luan, S.-Y. Neo, H.-K. Goh, Y.-D. Zhang, S.-X. Lin, T.-S. Chua, "Segregated Feedback with Performance-based Adaptive Sampling for Interactive News Video Retrieval" ACM MM 2007, Augsburg, Germany, 23-29 Sep 2007.

[7] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993–1022, January 2003.

[8] A. F. Smeaton, P. Over, and W. Kraaij, Evaluation campaigns and TRECVid. In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06. ACM Press, New York, NY, 321-330.