AUTOMATIC QUERY EXPANSION FOR NEWS VIDEO RETRIEVAL

Yun Zhai, Jingen Liu and Mubarak Shah

School of Electrical Engineering & Computer Science
University of Central Florida

ABSTRACT

In this paper, we present an integrated system for news video retrieval. The proposed system incorporates both speech and visual information in the search mechanisms. The initial search is based on the automatic speech recognition (ASR) transcript of video. Based on the relevant shots selected from the initial search round, keyword histograms are automatically generated for the refinement of the search query, such that the reformulated query fits better to the target topic. We have also developed an image-based refinement module, which uses the region analysis of the video key-frames. SR-tree like indexing structure is constructed for the region features, and the image-to-image similarity is computed using the Earth Mover's Distance. By performing a series of relevance feedback processes, the set of the true relevant shots is expanded significantly. The proposed system has been applied to a large open-benchmark news video dataset, and very satisfactory improvements have been obtained by applying the proposed automatic query expansion and the region-based refinement.

1. INTRODUCTION

The problem of retrieving the desired information from the video data has attracted lots of attention in various research fields, such as multimedia processing, information retrieval, computer vision, etc. With the vast amount of video data generated every day, it is impossible for humans to manually annotate every single video before it is archived for further reference. Besides this unimaginable workload, the manual annotation is sometimes considered "incomplete" in the sense that, there is always something in the video left unlabelled, which might be meaningful and valuable to other users. This mainly is due to the subjective nature of the annotation. Thus, there is a great need for building a video search system, which can perform the search based on the computable features of the video rather than using only manual annotations.

Many methods have been developed for the content-based image/video retrieval task. These methods are based on the different design aspects. Several works involved the design and use of efficient visual features, including both low-level features [12][8][7] and high-level semantic features [1][3]. An obvious limitation of the feature-based retrieval is that one can always formulate some query, which cannot be retrieved using those particular features. Another research trend emphasizes on the refinement of the search results based on the user selected results, usually referred to as the "relevance feedback" [10][1], where the search queries are refined by finding the prominent features and the queries are expanded based on the concept ontology. Given the training data, the query that best fits to the target topic can be estimated without the relevance feedback process, such as the mixture of experts [11] and the query-dependent search [6].

In this paper, we present a content-based video retrieval system. It retrieves relevant video shots from the news programs. The proposed system does not require the pre-annotation of video shots. Rather, it uses the computable video features, including speech (using ASR) and images (key-frame regions). Firstly, the manually formulated query for the target topic is submitted to the system. A set of video shots are returned by searching the automatic speech recognition (ASR) transcript. Then, an automatic query expansion technique is applied to refine the results. In this process, a set of relevant shots are selected by the user, and a more suitable query is formulated using the keyword histograms generated from both the relevant and irrelevant shot sets. The true relevant set is further expanded by applying an image-based refinement mechanism, which utilizes the analysis of the key-frame regions. The proposed system has been tested on a large open-benchmark data set, which contains more than 43,000 video shots of news programs. With the help of the proposed automatic query expansion and image-based refinement techniques, significant improvements have been observed in the experiments of retrieving relevant shots for multiple search topics. The remainder of this paper is organized as follows: In Section 2, we describe the query expansion using keyword histograms and the image-based refinement in detail. Then, the performance evaluation is presented in Section 3. Finally, Section 4 concludes our work.

2. SYSTEM DESCRIPTION

In this section, we describe the proposed video search system in detail. Particularly, we emphasize on two contributions, automatic query expansion and image-based refinement. In our system, there are three major components, user interface, server and the feature index (as shown in Fig.1 with their corresponding functionalities). In the user interface component, the user is able to formulate a query according to the target topic and browse the search results returned from the

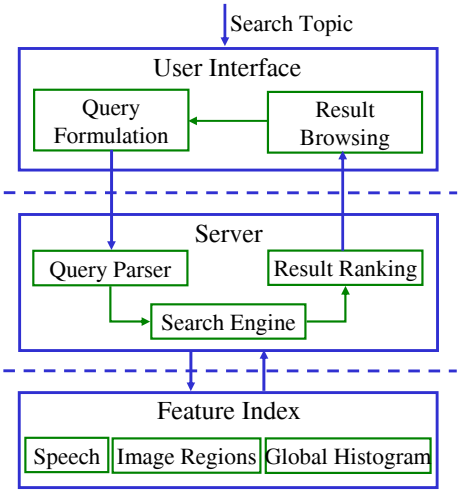

Fig. 1. System structure of the proposed video search engine. It shows the three components of the system and their functionalities.

server. The query is then submitted to the server for the result generation using the information stored in the feature index database. As described in the following sections, the search query is not necessarily in the text form. It could also be in the form of visual features. The initial results for the manual query are further refined by reformulating the query using more appropriate keywords and by performing an image-based relevance feedback process.

2.1. Query Expansion Using Keyword Histograms

Since there exist multiple ways to describe the same event or object, the initial manually formulated query often may not quite be relevant to what user really wants. Or, the query may not be complete in the sense that, it fails to relate the videos which have semantic similarities but may not have speech similarity. Assume that, the user wants to find the shots that are related to Condoleezza Rice. If the query is simply the "Condoleezza Rice", the results may miss the shots containing "secretary of state" or "national security advisor" in the ASR transcript. On the other hand, the results may find shots on the food "rice", which are irrelevant. Due to the limited knowledge of a single person, it is almost impossible to formulate a perfect query at their first attempt. To overcome this problem, there is a need for finding an automatic way to expand the query, such that the refined query better fits the target topic and covers a more complete set of relevant shots. WordNet[2], which has general-purpose and is not tuned to any particular dataset, has been widely used in the literature. However, we believe that the relations between keywords in different dataset would be different. For example, in the news data corpus, the topic "tennis" is strongly correlated with keywords "cup", "game", etc. On the other hand, in the instructional videos of tennis, "tennis" is more correlated with "forehand-stroke", "backhand-stroke" or "serve".

Fig. 2. An example for the automatic query expansion. The search topic is "soccer". The figure shows the initial query, the examples of selected "relevant" and "irrelevant" videos, and the positive and negative keywords sets.

Therefore, the relationships between keywords should be discovered based on the target data corpus rather than from a neutral source.

Here, we propose an automatic query expansion technique, which enriches the search query to cover more relevant shots. The expansion is performed using the speech (ASR) information of the videos, which is expressed in the text format. From the shots returned by the first round of search with query Q_{i-1}, the user can select a set of shots, which is considered relevant, and another set which is irrelevant. They are denoted as "positive (D^+)" and "negative (D^-)" sets, respectively. A keyword histogram $WH_{D^+} = \{(a_1^+, W_1^+), (a_2^+, W_2^+), \cdots, (a_m^+, W_m^+)\}$ is computed based on the ASR of positive set, where W_i^+ is the extracted keyword accompanied by its normalized frequency a_i^+ in the positive set. Similarly, another histogram $WH_{D^-} = \{(a_1^-, W_1^-), (a_2^-, W_2^-), \cdots, (a_m^-, W_m^-)\}$ is constructed for the negative set. In the next round of search, the newly reformulated query $Q_i = WH_{D^+}$ is submitted to the system. During this process, the query is expanded from Q_{i-1} to a larger set Q_i. Finally, the relevance of a retrieved shot is determined by computing the histogram correlations. Given a retrieved video shot S, a normalized keyword histogram WH_S is constructed, and the relevance measure $R(S)$ is computed in terms of vector product,

$$R(S) = VP(WH_S, WH_{D^+}) - VP(WH_S, WH_{D^-}), \quad (1)$$

where $VP(\cdot,\cdot)$ represents the inner product of the vectors. Here, vectors WH_S, WH_{D^+} and WH_{D^-} are restructured to have the same dimensions by filling the missing positions with zeros. Example of the automatic query expansion is shown in Fig.2.

2.2. Image-Based Refinement

Often, people can determine if two videos are semantically similar or not based on their visual perception. Consider one video shot of President Bush attending the APEC summit, and another shot about him giving a speech in Congress. Even though these two shots are from different stories, they both satisfy the target topic "President Bush".

In the proposed search system, we have developed an image-based refinement mechanism. In this component, we analyze the similarities between the video key-frames. Commonly, global information of the key-frames is used, e.g.,

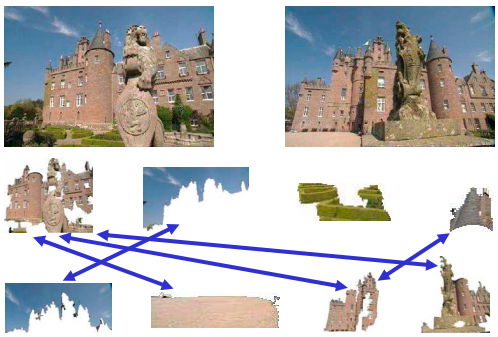

Fig. 3. A pair of example images for the region matching. Largest four regions of each image are shown. As demonstrated in the figure, many-to-many matching is allowed using the EMD measure.

global histograms of color, texture or edge orientation. These features have their innate drawbacks. They are only applicable for detecting near-duplicate images. Methods using global features fail to correlate the same object/person when different background settings are present. To overcome this problem, we use a region-based method. Given a key-frame F of the video shot, the regions, denoted by $\{I_F^1, I_F^2, \cdots, I_F^n\}$, are computed using the Mean Shift image segmentation method. For each region I_F^k, its representation is expressed as a vector, $\{T_F^k, A_F^k, V_F^k\}$, where $T_F^k = F$ is a tag of which image this region belongs to, A_F^k is the normalized area of the region, and V_F^k is the feature vector of the region. Here, we use the color moments and the edge histogram for the feature vector.

The image-based refinement is performed in a relevance feedback process. The user selects a set of relevant shots from the results returned by the previous search round. The key-frame regions of the selected shots are treated as the new visual queries for the next round. The search is based on individual regions. The returned results contain the key-frames which have the similar regions to the query regions. For example, given a query image F with multiple regions, the region-based search result is $\{(X, Y, Z), (X, Y), (W, X, Z), \cdots\}$. In this case, (X, Y, Z) are the images that have the similar regions to the first query region of F, (X, Y) are the images that have the similar regions to the second region of F, and so on. If the regions of a returned image X match the majority of the query regions of F, X is considered as a match of F. To further rank the relevance of returned images, we incorporate the Earth Mover's Distance (EMD) [9] in the image-to-image similarity computation. We model the regions in the image by the nodes in the bipartite graph, and regions from the same image are the nodes in the same partite. Here, node i in the graph is used interchangeably with region i in the image. Thus, given two images X and Y, their EMD is computed as follows,

$$EMD(X, Y) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_{ij} f_{ij}}{\sum_{j=1}^{m} f_{ij}}, \qquad (2)$$

Fig. 4. User interface of the search system.

where α_{ij} is the cost for flowing from node i to node j, f_{ij} is the flow amount from node i to node j, and m and n are the numbers of regions in images X and Y, respectively. The cost α_{ij} could be defined as the distance between the nodes. In our system, Euclidean distance is used for the region-to-region distance. The flow amount f_{ij} is computed by solving the maximum flow problem for the bipartite graph. In this graph formulation, the area A_i is used as the weight of each node i. Intuitively, EMD allows for many-to-many matching between the regions. One example of the matching is shown in Fig.3.

3. PERFORMANCE EVALUATION

We have built a web-based search system with an interactive user interface. The indexing system for text (ASR)is established using the Lucene technology [4]. To achieve fast indexing and retrieval, the regions (features) are archived in the database using the SR-tree structure [5], which is well suited for finding the nearest neighbors in the high-dimensional feature space. Fig.4 shows a screen shot of the interface. The system has been tested on a large open-benchmark dataset, which contains 140 news program videos provided by the US National Institute of Standards and Technologies (NIST) for the TRECVID 2005 forum. Each video is around 30-60 minutes long and is in MPEG-1 format. These videos were contributed by several news networks, such as CNN, NBC, CCTV, LBC, NTDTV, etc., and they cover various languages, including English, Chinese and Arabic. There are totally 43,657 video shots in the entire testing set, which are taken as the retrieval units. The ASR, machine translation transcript of Chinese and Arabic, shot boundaries and key-frames were provided as the ground truth data. We have selected ten search topics for the testing. These topics cover the categories of objects, specific persons and physical settings. They are listed in Fig.6. A snapshot of the region-based re-

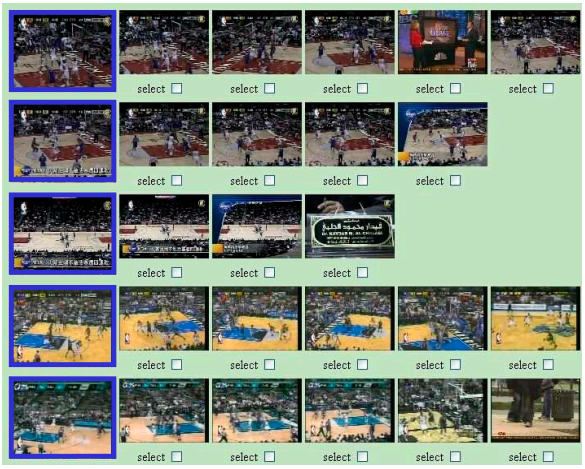

Fig. 5. Snapshot of the region-based refinement for topic "basketball game". The left column shows the query images (shown in blue boxes) , and the returned shots in each row are ranked by EMD.

finement results is also shown in Fig.5.

We perform the experiments using the following steps to demonstrate the effectiveness of each of the two refinement processes. Given a target topic, the initial query is manually formulated and submitted to the system for the search based on the ASR information only. The corresponding shots are returned along with their temporal neighboring shots. A subset of the returned shots is then labelled as "relevant" by the user, and another set of shots is considered "irrelevant". The positive and negative keyword histograms D^+ and D^- are generated based on these two sets, respectively, and they are submitted to the system as the expanded query in the next round of search. The purpose of this step is to demonstrate the effectiveness of the proposed automatic query expansion technique. Using the results of the expanded query, a set of "relevant" video shots are again selected, and their keyframes are used as the queries in the region-based refinement. The returned results in this step are ranked by their EMDs to the query images. The improvements of two proposed refinement steps are demonstrated separately in Fig.6. Based on the experimental results, the query expansion technique using automatically generated keyword histograms is able to expand the relevant set on the average of 80%, while the region-based refinement is able to further increase the number of true positives by 44%. It should be noted that the system is also efficient in terms of timing. The experiments were carried out on a Dell XPS laptop with 2.13GHz Pentium M processor and 2G RAM, and the searches of all ten topics were finished within 15 minutes.

4. CONCLUSIONS

We have presented a content-based video retrieval system. The system utilizes both the speech (ASR) and visual (image regions) content of the video shots. The contribution of the proposed system is two-fold: (1) it is able to automatically expand the search query by analyzing the keywords in the relevant shot sets, and (2) it is able to further expand the true relevant set by an image-based relevance feedback process. The proposed system has been applied to a large open-benchmark news dataset, which covers various genres, such as sports, commercials, talk shows, etc. Significant improvement in performance has been observed by using the two proposed refinement modules.

Topics	Manual Query	Automatic Query Expansion	Region-Based Refinement
Soccer game	24	44	63
Tanks/military vehicle	30	43	65
Basketball game	21	42	63
Iraqi map	15	41	51
Condoleezza Rice	25	30	39
Road with cars	45	92	145
Tennis player on court	23	36	59
Ship or boat	30	50	66
Helicopter in flight	15	17	22
People with banners	57	150	199

Fig. 6. Evaluation results of ten search topics. The numbers of relevant shots are shown separately for: (1) using the manual query only, (2) applying the automatic query expansion using keyword histograms and (3) applying the region-based refinement.

5. REFERENCES

[1] P. Browne and A. Smeaton, "Video Information Retrieval Using Objects and Ostensive Relevance Feedback", *ACM Symposium on Applied Computing*, 2004.

[2] Christiane Fellbaum, "WordNet: An Electronic Lexical Database", MIT Press.

[3] L. Hollink, M. Worring and A.T. Schreiber, "Building a Visual Ontology for Video Retrieval", *ACMMM*, 2005.

[4] http://lucene.apache.org/java/docs/

[5] N. Katayama and S. Satoh, "The SR-Tree: An Indexing Structure for High-Dimensional Nearest Neighbor Queries", *SIGMOD*, 1997.

[6] L. Kennedy, P. Natsev, S-F. Chang. "Automatic Discovery of Query Class Dependent Models for Multimodal Search". *ACMMM*, 2005.

[7] T. Lin, Chong-Wah Ngo, H.J. Zhang and Q.Y. Shi, "Integrating Color and Spatial Features for Content-based Video Retrieval", *ICIP*, 2001.

[8] M. Rautiainen and D. Doermann, "Temporal Color Correlogram for Video Retrieval", *ICPR*, 2002.

[9] Y. Rubner, C. Tomasi and L. Guibas, "A Metric for Distributions with Applications to Image Databases", *ICCV*, 1998.

[10] X-J. Wang, W.Y. Ma and X. Li, "Data-Driven Approach for Bridging the Cognitive Gap in Image Retrieval", *ICME*, 2004.

[11] R. Yan, J. Yang and A. Hauptmann, "Learning Query-Class Dependent Weights in Automatic Video Retrieval", *ACMMM*, 2004.

[12] L. Zhao, W. Qi, S.Z. Li, S.Q. Yang and H.J. Zhang, "Content-based Retrieval of Video Shot Using the Improved Nearest Feature Line Method", *ICASSP*, 2001.