Object-Sensitive Query Analysis for Video Search*

Jingjing Liu
Dept. of Computer Science, Nankai University
Tianjin, China
nkicestone@gmail.com

Xian-Sheng Hua, Shipeng Li
Microsoft Research Asia
Beijing, China
{xshua,spli}@microsoft.com

Abstract — **This paper is concerned with the problem of improving the performance of text search baseline in video retrieval, specifically for the search tasks in TRECVID. Given a query in plain text, we first implement syntactic segmentation and semantic expansion of the query, then identify the underlying "targeted objects" which should appear in the retrieved video shots, and scale up the weights of the video shots retrieved by the query terms that represent these targeted objects. We name the approaches as "object-sensitive query analysis" for video search. Specifically, we propose a set of methods to identify the specific terms representing the "targeted objects" in a video search query, and a modified object-centric BM25 algorithm to emphasize the impact of these specific object-terms. In practice, we place the process of object-sensitive query analysis before the text search stage, and verify the effectiveness of the proposed approaches with the TRECVID 2005 and 2006 datasets. The experimental results indicate that the proposed object-sensitive approaches to query analysis bring significant improvement upon the raw text search baseline of video search.**

Keywords—video search; query analysis; object-sensitive; BM25 algorithm

Topic area—indexing and search of multimedia

I. INTRODUCTION

Video search has become an active and challenging task in recent years with the rapid growth of personal video recordings as well as online video data. Video search is defined as the retrieval of the relevant video or video segments/clips to the issued queries in forms of textual keywords/sentences or provided video/image examples or some combination of the two. With video search engines, users can easily obtain the targeted video shots that meet their interests. Fig. 1 shows an exemplary output from a video search system. The figure consists of some thumbnail examples of the top-21 ranked video shots from a video search model. The input is a query in TRECVID[1] 2005 query set: "*find shots of Condoleeza Rice*," and the retrieved video clips are from the video corpus of TRECVID 2005.

Recently, many studies on video content analysis are focused on the fusion of multimodalities for the video retrieval, which employs both visual and semantic conceptual information of video content to re-rank the search results upon the text-only search model [2][3][4][5][7]. The text search model based on text corpus associated with video contents (such as speech transcripts), which serves as the baseline of multimodal fusion, is a key to the performance of video search re-ranking. Thus, how to raise the bar of the text-based video search baseline is one of the biggest challenges in video retrieval studies.

The text retrieval problem in video search task is defined as follows. Given a query in text, we are to estimate the relevance $R(x)$ of each video shot x in the search set X ($x \in X$) to the query topic q, and order them by their relevance scores. The relevance of a shot to the text query is given by the relevance score between the associated text of the shot and the text query. Therefore, by employing a set of relevance-based text retrieval methods, the video search task can be reduced to a text retrieval problem.

Fig. 1. An exemplary video search result.

However, different from the text search problem, video search is a task relatively more content-and-visual based yet less text-relevant. The query in a text search task is often one or several keywords (e.g., a concept, a technical term); while the queries in video search are always inquiring for some visual objects, such as a person, an event and a scene. We name such objects as the "targeted objects" in a query. The query terms that represent these targeted objects, i.e., the "object-terms," should be treated with special consideration in the search stage as they indicate the users' search intention for the targeted video shots. However, most of the existing text search methods lack a strategy to emphasize specific query terms. Thus, directly applying off-the-shelf text retrieval methods to video search tasks will probably lose the underlying "visual information" of the queries, and as a result, fail to find the video shots most relevant to the users' intention of query.

---

* This work was performed while Jingjing Liu was visiting Microsoft Research Asia as a research intern.

[1] TRECVID (TREC Video Retrieval Evaluation) is a conference sponsored by NIST. The main goal of TRECVID is to promote progress in content-based retrieval from digital video via open, metrics-based evaluation. TRECVID dataset in each year consists of a query set, a video corpus, and a laboratory-style ground-truth for evaluation. http://trecvid.nist.gov/

To tackle this problem, an intuition is to analyze the video search query in a more visual-related and content-based perspective. Driven by this motivation, we examine the video corpus of TRECVID, exploit the specific characteristics of multimedia data as well as video search queries, and propose a set of query analysis approaches which can identify the "targeted objects" in a video search query and take special treatment to the query terms that represent these targeted objects. We name the proposed approaches as "*object-sensitive query analysis*" for video search.

In this paper, we will present the proposed approaches, namely *syntactic query segmentation*, *semantic query expansion*, *and object-term re-weighting*, based on the demonstration of specific characteristics of the queries in TRECVID search tasks. Evaluation of the proposed approaches on TRECVID 2005 and 2006 datasets verifies that, with the proposed object-sensitive query analysis approaches, we can effectively enhance the performance of text retrieval methods in video search and bring significant improvement upon the text-based video search baseline in TRECVID tasks.

II. PROPOSED METHODS

A. Syntactic Query Segmentation

The text topic of a query is often a sentence describing the contents of the targeted video shots, which can be taken as a text segment with complete syntactic structure. Therefore, we first conduct a syntactic analysis of the original query topic with syntactic analysis methods.

Part-Of-Speech (POS) is one of the most important indicators of the syntactic characteristics of terms, which is widely used in natural language processing area. Part-of-speech represents the syntactic property of a term, for example, a noun, a verb, an adjective, etc. In order to assess the syntactic property of query terms, we construct a POS (part-of-speech) tagging process upon query topics with an automatic POS tagging tool [11].

For example, a query topic: "*Find shots with one or more people leaving or entering a vehicle*" will be identified as: "*Find <verb> shots <noun> with <adv> one <adj> or <conj> more <adj> people <noun> leaving <verb> or <conj> entering <verb> a <preposition> vehicle <noun>*."

With these POS labels, we can do further refinement to the query topic, such as stopwords (e.g., "one," "a") removing and stemming (e.g., convert "leaving" to "leave"). Then, we segment the query topic into sub-sequences of terms by applying N-gram segmentation [8] to the query topics. The generated term sequences will be input to the search engine as different forms of the query, and the video shots retrieved by different query segments will be aggregated with different weights. Typically, the higher gram a query segment has, the higher weight the corresponding video shots will be assigned. For the exemplary query above, the video shots retrieved by the term sequence like "*people leave vehicle*" will be given a higher weight than those retrieved by "*people leave*."

B. Semantic Query Expansion

In a typical information retrieval scheme, the more occurrences a term appears within the query, the more important the query term is; therefore the higher relevance of corresponding documents retrieved by the specific term is. To expand the occurrences of query terms with supplemental information of the query, we expand a query with the video or image examples provided along with the query topic in TRECVID query set.

Fig. 2 shows an exemplary query in TRECVID query set. In the example, the sentence following the tag "*<topic>*" is the original query topic in text; while those sentences following the tags "*<video example>*" and "*<image example>*" are the descriptions of the video clips or image examples given along with the query topic. By expanding the query with the descriptions of these examples, we can not only obtain new query terms, but also gather multiple occurrences of the recurred terms in the original query topic which will be highlighted in the search stage.

<topic>Find shots with one or more people leaving or entering a vehicle
<video example1>Bushes leaving helicopter
<video example 2>Man leaves car
<image example1>Police video (policeman leaves cruiser)
<image example2>People getting into back of police van
<image example3>People getting on bus
<image example4>Soldiers move from back of large airplane to bus

Fig. 2. Example of query in TRECVID query set.

On the other hand, many queries in video search tasks contain "name entities," which refer to an object such as a person, a place, and a vehicle. For this specific type of query, we propose a query expansion method, which we name as "name entity categorization," by classifying name entities into several predefined categories and giving each name entity a label indicating its corresponding category.

Specifically, with an automatic name-entity recognition tool [1], we first identify the name entities occurring in both the text corpus associated with video data and the queries, provided by TRECVID datasets. Then, a label of generic "name entity category" such as "*<person name>*" and "*<vehicle name>*" is given to each identified name entity in both the queries and the video-associated text corpus.

With this categorization method, name entities in both the queries and the text corpus associated with video shots are tagged with the same set of labels of "name entity categories." In this way, the video shots whose associated text segments have no direct match with the original query topic but contain the same category of name entities as the query will now be retrieved by the same labels.

C. Object-Term Re-Weighting

To exploit the specific visual characteristics of video search queries, we propose an object-sensitive approach which distinguishes the query terms representing the "targeted

objects" from others representing the context or the background of the targeted objects.

Given a text topic as a video search query, we can extract a set of semantic concepts from the topic for further refinement of the targeted video shots. A semantic concept is a semantic label of the content of a video shot, e.g., "natural scene," "animal," etc. In the multimedia research area, there are many concept dictionaries available for the studies of video content analysis and retrieval. For example, the LSCOM list [10] consisting of over 800 semantic concepts of video contents developed by Columbia University, is one of the most popular concept lists, which is widely used in TRECVID. In our approach, we take such a concept list as the dictionary and find the semantic concepts of a query by detecting the direct matches of the query terms in the concept list. We take the identified semantic concepts as the "targeted objects" in the query, as these concepts represent the semantic contents of the targeted video shots.

Also, given a query topic with POS tags, we take the nouns or noun phrases in the query topic as the "targeted objects," as the nouns or noun phrases in a query topic are always the centric objects in the targeted video shots that the query is searching for. Furthermore, the nouns or noun phrases that are extracted as "targeted objects" should be treated unequally according to their different positions in the query topic as well as their contexts. Specifically, we identify the adverbs with refinement meanings (such as "at least") and take the nouns or noun phrases following these adverbs as "targeted objects."

Moreover, as the different terms of a name entity do not always occur with the same frequencies in the text corpus, we extract the underlying "targeted objects" in a query with name entities by identifying the part of name entity that is more often used, as the reference of the name entity, e.g., "*Bush*" for "*George W. Bush*".

With these object identification methods, the query terms that represent the targeted objects in a query will be identified and the corresponding video shots retrieved by these query terms will be assigned with a higher weight in the search stage, as will be explained in the next section.

III. ALGORITHMS

BM25 [6] is a ranking function to rank retrieved documents according to their relevance to a given search query, which is widely used in information retrieval studies. It is based on the probabilistic retrieval framework developed by Stephen E. Robertson et al. [6] in the 1970s and 1980s. BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms occurring in each document, regardless of the inter-relationship between the query terms within a document (e.g., their relative proximity). One of the most prominent instantiations of the function is given by:

$$relevance = \sum_{T \epsilon Q} \omega \frac{(k_1+1)tf(k_3+1)qtf}{(K+tf)(k_3+qtf)} \quad (1)$$

where Q is a query consisting of term T; tf is the occurrence frequency of the term T within the text segment, qtf is the frequency of the term T within the topic from which Q was derived, and ω is the Robertson/Sparck Jones weight [9] of T in Q.

In BM25 algorithm, all the query terms are treated equally, except that the term frequency within the query (qtf) is taken into consideration. However, when applied to video search tasks, the qtf often fails to reveal the difference of query terms, as few query terms occur more than once in the query topic in video search tasks. In our approach, we expand the query with the descriptions of the video clips/image examples given along with the query topic, as aforementioned in the above section. In this way, the frequency of the important query terms which also appear in the descriptions of the video clips/image examples will be increased, and as a result the corresponding query terms will be emphasized with the scaled-up qtf.

To take advantage of the visual characteristics of video search queries, we propose a modified qtf definition, which takes into account the semantic importance of different query terms, and emphasizes the terms which represent the "targeted objects" in the query. The modified qtf for BM25 equation is defined as:

$$qtf_{new} = \sum_i w_i * O_i(t) + qtf_{old} \quad (2)$$

where qtf_{old} represents the original term frequency within the query topic as defined in (1). $O_i(t)$ represents the indicator function which predicts whether a term t refers to a targeted object or not, determined by the aforementioned object identification methods:

$$O_i(t) = \begin{cases} 1 & if\ t\ is\ a\ targeted\ object; \\ 0 & otherwise. \end{cases} \quad (3)$$

Specifically, w_i represents the weight assigned to the object-term identified by a specific identification method. In special cases where a term is detected as the targeted object by more than one method, the weights from multiple methods are aggregated as a combined score for the query term. And in the case where the term is not detected as a targeted object by any method, the qtf_{new} will remain the same as the original query term frequency (qtf_{old}).

Based on the modified qtf definition, we propose an object-centric BM25 algorithm for video search, which is given by:

$$relevance = \sum_{T \epsilon Q} \omega \frac{(k_1+1)tf(k_3+1)(\sum w*O(j)+qtf_{old})}{(K+tf)(k_3+\sum w*O(j)+qtf_{old})} \quad (4)$$

In the modified BM25 algorithm, not only the term frequency in the query topic is counted, but also the object-related visual importance of query terms is taken into consideration. In this way, the specific query terms that represent the targeted objects in the video search query can be identified by various object identification methods and emphasized by the higher weights assigned empirically.

IV. EXPERIMENTAL RESULTS

We verify the proposed object-sensitive approaches on the datasets of TRECVID 2005 and 2006. There are 24 queries in both datasets, respectively. In TRECVID 2005 dataset, the video corpus consists of 140 videos with 45,766 shots, of a size 61.3GB. TRECVID 2006 dataset consists of 259 videos with 79,846 shots, of a size 87.7GB. In the experiments, we take the original BM25 algorithm as a baseline, and

incrementally add the proposed methods to the baseline algorithm. TABLE 1 shows the experimental results of MAP (Mean Average Precision) on all search tasks in both datasets. We also compare our experimental results with the text search baseline from Columbia University in TRECVID 2005 and 2006 [12][13].

TABLE 1. PERFORMANCE OF PROPOSED APPROACHES.

	TRECVID2005	TRECVID2006
Columbia Text Search Baseline	0.0390	0.0350
Our Baseline (BM25)	0.0384	0.0328
+ Syntactic Query Segmentation	0.0405	0.0345
+ Semantic Query Expansion	0.0434	0.0367
+ Object-Term Re-Weighting	0.0531	0.0424

From TABLE 1 we can see that the object-term re-weighting method improves the performance of text-based video search significantly. The syntactic query segmentation and the semantic query expansion methods, although have relatively smaller contribution, also improve the performance. With the proposed approaches, the video search results of MAP on TRECVID 2005 and TRECVID2006 datasets can be raised up to 0.0531 and 0.0424, respectively, which are both higher than two baselines.

Fig. 3 and Fig. 4 show the comparison of search results (Average Precision) on each individual search task between our proposed approach and that of the BM25 baseline on TREC2005 and 2006 datasets, respectively. From the results we can see that the proposed approach improves the text-based video search baseline on most queries in both datasets, regardless of specific query classes.

A complete video search system would take the text-based search results as the baseline for multimodal fusion, which merges the text search baseline model with other models that are widely used in video search studies by employing semantic conceptual information as well as visual low-level features of video shots. The details of multimodal fusion are not included here since it is not the focus of this paper.

V. CONCLUSIONS & FUTURE WORK

In this paper we addressed the problem of query analysis in text search baseline for video search. We proposed a set of object-sensitive approaches to query analysis, namely syntactic query segmentation, semantic query expansion, and object-term re-weighting. We proposed a set of methods to identify the "targeted objects" in the query which must appear in the targeted video shots, and presented a modified object-centric BM25 algorithm which emphasizes the specific object-terms. Experimental results indicate that the proposed approaches significantly improve the performance of text-based video search baseline on TRECVID 2005 and 2006 datasets.

In future work, we will tackle the problem of identifying the targeted object from multimedia data rather than only from plain text. We will also work on the object-centric modification of other relevance search algorithms than BM25. A third direction we are heading is to investigate more methods of targeted object identification to improve the object-term re-weighting approach.

Moreover, we will try to employ the object-sensitive query analysis approaches to multimodal fusion for video search tasks, to verify the effectiveness of the approaches as well as to contribute to multimodal-fusion-based video retrieval.

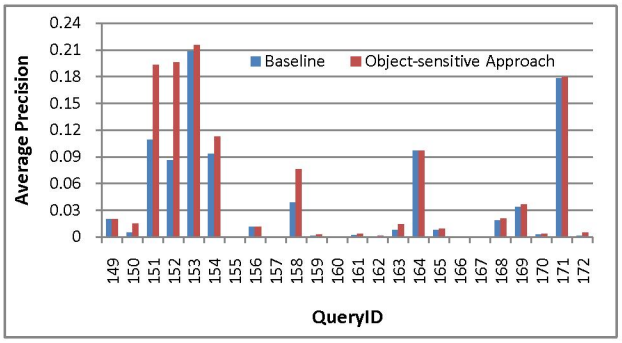

Fig. 3. Performance on each search task of TRECVID2005 dataset.

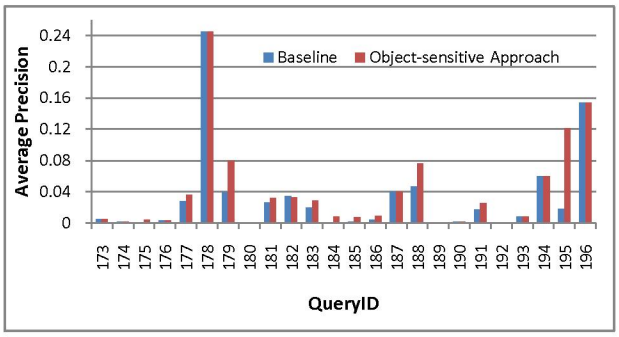

Fig. 4. Performance on each search task of TRECVID2006 dataset.

REFERENCES

[1] Alias-i. Lingpipe named entity tagger. In http://www.alias-i.com/lingpipe/.

[2] K. M. Donald and A. F. Smeaton. A comparison of score, rank and probability-based fusion methods for video shot retrieval. CIVR 2005.

[3] A. G. Hauptmann and M. G. Christel. Successful approaches in the trec video retrieval evaluations. ACM Multimedia 2004.

[4] Alexander G. Hauptmann, Wei-Hao Lin, Rong Yan, Jun Yang and Ming-Yu Chen, Extreme Video Retrieval: Joint Maximization of Human and Computer Performance, ACM Multimedia 2006.

[5] G. Iyengar, et al. Joint Visual-Text Modeling for Automatic Retrieval of Multimedia Documents, ACM Multimedia 2006.

[6] Robertson, S. E. Overview of the Okapi Projects, Journal of Documentation, Vol. 53, No. 1, 1997, pp. 3-7.

[7] Winston H. Hsu, Lyndon S. Kennedy, Shih-Fu Chang. Video Search Reranking via Information Bottleneck Principle. ACM Multimedia 2006.

[8] Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, Jenifer C. Lai. Class-Based n-gram Models of Natural Language. Computational Linguistics, 1992

[9] Robertson, S. E., and Sparck Jones, K. Relevance Weighting of Search Terms. Journal of the American Society of Information Science. 1976.

[10] LSCOM Lexicon Definitions and Annotations (Version 1.0). DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia. http://www.ee.columbia.edu/ln/dvmm/lscom/

[11] CLAWS part-of-speech tagger for English. http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/

[12] Shih-Fu Chang, et al. Columbia University TRECVID-2006 Video Search and High-Level Feature Extraction. TRECVID 2006.

[13] Shih-Fu Chang, et al. Columbia University TRECVID-2005 Video Search and High-Level Feature Extraction. TRECVID 2005.