The Importance of Query-Concept-Mapping for Automatic Video Retrieval

Dong Wang, Xirong Li, Jianmin Li and Bo Zhang
State Key Lab. of Intelligent Tech. and System
Department of Computer Science and Technology
Tsinghua University, Beijing, 100084,China
{wdong01,lxr}@mails.tsinghua.edu.cn,{lijianmin,dcszb}@mail.tsinghua.edu.cn

ABSTRACT

A new video retrieval paradigm of query-by-concept emerges recently. However, it remains unclear how to exploit the detected concepts in retrieval given a multimedia query. In this paper, we point out that it is important to map the query to a few relevant concepts instead of search with all concepts. In addition, we show that solving this problem through both text and image inputs are effective for search, and it is possible to determine the number of related concepts by a language modeling approach. Experimental evidence is obtained on the automatic search task of TRECVID 2006 using a large lexicon of 311 learned semantic concept detectors.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.2.4 [**Database Management**]: Systems—*multimedia databases, query processing*

General Terms

Algorithms, Design, Experimentation

Keywords

Automatic Video Retrieval, Query-Concept-Mapping

1. INTRODUCTION

Until now, access to video has been limited to noisy text associated with the video content, whether automatically recognized speech, closed captions, or social tags. The achievements are limited since apart from its noisy nature, the succinct text usually does not elaborate on the visual obvious. Only recently, a few hundreds of semantic concepts [8] including various roles of people, objects, scenes and events are detected automatically with varied performance in [11, 4, 7]. The generic method to learn the detectors is based on generic features with a few labeled examples. Thus a new query-by-concept video search paradigm emerges as using larger concept lexicons for search. Intuitively, if queries can be automatically mapped to related semantic concepts, search performance will benefit significantly. For example, a query as "scenes with snow" will surely benefit from concept "Snow", or even "Sky" since a snowy scene is often with sky present. An important research problem rises here as how to map the query to the concepts automatically, reliably and scalably (abbreviated as QUCOM (query-concept-mapping) hereafter). Note that both the number of related concepts and their respective weights has to be determined.

A few work positively supports the usefulness of the query-by-concept paradigm for retrieval [7, 6, 10, 4, 12, 11, 5]. However, some of these results are too optimistic in using manually annotated concept indices [5], oracle selected concepts [12] or human judged relevant concepts [5]. Only the top first concept are selected in [11]. Therefore they either avoid the varied concept performance or bypass the difficult QUCOM problem. Others use a rather small lexicon ($<$40) of concept detectors [6, 10, 5]. The exception of [4] combines the concept search with other approaches, leaving the concept search performance unanswered.

In this paper, we take an initiative to evaluate both the necessity and effectiveness of QUCOM on automatic video retrieval performance, under realistic conditions using a large lexicon ($>$ 300) of concept detectors. To be more specific, we attempt to answer the following two related research questions: 1). Given different queries and a predefined concept lexicon, does there exist a number of concepts to improve video retrieval accuracy? 2). If the answer is yes, then how to design effective methods to select these relevant concepts, together with their weights to improve video retrieval accuracy? To answer these questions, we employ a video search engine using 311 learned concept detectors. The experimental results from the automatic search task of TRECVID 2006 [1] show both the necessity and effectiveness of performing QUCOM.

We organize the remainder of this paper as follows. In Section 2, we introduce the retrieval model and our solution to QUCOM in detail. Then we present the experimental setup in Section 3. We analyze results in Section 4 and conclude in Section 5.

2. SOLVING THE QUCOM PROBLEM

A fundamental difference between standard text retrieval and query-by-concept video retrieval is that in the former, the query words are explicitly given whereas in the latter, the user only, at best, implicitly specifies the semantic concepts through the queries text and/or example images. Without QUCOM, all the concepts will be used for search and the irrelevant concepts may degrade the retrieval performance. The search engine is responsible for solving the QUCOM utilizing the query input. After that, the query-by-concept paradigm almost reduces to standard text retrieval. Ideally, QUCOM should be solved on a per query basis and on-the-fly due to the real-time search need. There are two kinds of cues for solving QUCOM depending on the types of query input. Though it is straight forward and fast to link query to concepts by text matching

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

between the query text and concept description [4] or a predefined concept ontology [11], this line of research ignores the visual aspect of the concepts, which might be also important for solving QUCOM. On the other hand, query images, if provided, establish a visual link between the information need and the semantic concepts. Predicting the concepts on the query images and concatenating the scores results in a vector in the concept space where the concept vectors of video corpus reside. However, if we overlook QUCOM and simply take all concepts into account, as previous work [6, 9, 13] did, this will augment the risk of bringing in more irrelevant and even noisy ones, and may degenerate the retrieval performance.

Since query text and image(s) provide complementary information, we consider both of them for solving QUCOM under a restricted vector space model. Given a lexicon of concepts $L = \{c_i\}$ and a corpus $\mathcal{C} = \{\vec{d_i}\}$ where each shot (visual document) is represented as a concept vector $\vec{d_i} \in V$, V being the concept space $\vec{d_i}$ resides. $\vec{d_i} = [d_i^1, d_i^2, \ldots, d_i^n]^T$ is the detected concept vector in V for the i^{th} shot. Note that $d_i^j = P(c_j|d_i)$ is the estimated frequency/probability of concept c_j in d_i. A query Q can be represented as $Q = \{Q_{txt}, Q_{img}\}$, where Q_{txt} is the text input and $Q_{img} = \vec{q}$ where $\vec{q} \in V$ is the query example represented as a concept vector. A simple extension to multiple example images are given in Sec. 2.2. Then QUCOM finds a subset of relevant concepts L_s, which corresponds to a few dimensions in V, based on Q_{txt} and Q_{img} (if available). After getting L_s, the restricted vector space model ranks each shot $\vec{c}$ by its relevance to query Q, and is defined as

$$R(d, Q) := \sum_{c \in L_s} w_c(Q) w_c(d) \tag{1}$$

where L_s is the selected concept subset, and for each concept $c \in L_s$ two weights $w_c(d)$ and $w_c(Q)$ are associated with the shot and the query respectively. Two kinds of weights as $w_c(Q_{txt})$ and $w_c(Q_{img})$ can be defined depending on the information utilized. By restricting $R(d, Q)$ in the subset L_s, we essentially set a dynamic threshold to the inferred concept probability in a given query.

2.1 Solving QUCOM with Query Text

As each concept in LSCOM is associated with its textual description, we can match the query text with the textual concept descriptions to get relevant concepts. Both the description and the query text are normalized: commonly occurring words are removed, and stemming is performed. Each detector description is represented by a term vector, where the elements in the vector correspond to unique normalized terms. Similar to [11], this text match returns an ordered top-k list of concepts, with a relevance score $w_c(Q_{txt})$ associated for each concept c under the vector space model.

2.2 Solving QUCOM with Query Images

Suppose that $Q_{img} = \{\vec{q}\}$, one method for QUCOM is to rank the concepts according to $P(c|q)$ without considering the corpus $\mathcal{C}$ statistics [11]. The intuition is that more frequent concepts are more likely to be relevant. However, though simple, this method overlooks the concept performance on $\mathcal{C}$ and biases towards the most common concepts which may not be useful for search. Note that $P(c|d)$ is similar to the well-known term frequency (*tf*) in text documents, we can write $P(c|d)$ as $freq(c, d)$ to emphasis this metaphor. In this line of reasoning, we take the averaged probabilities of concept c in the corpus as the shot (document) frequency $df(c)$ for concept (term) c, by defining $freq(c) = \frac{1}{N}\sum_d freq(c, d) = \frac{1}{N}\sum_d P(c|d)$ and N is the size of the corpus. This $freq(c)$ differs from the well established document frequency (*df*) only by a normalization constant N. The $\log(\frac{1}{freq(c)})$ is exactly the inverse document frequency (*idf*) measure.

Three models can be designed as $w_c(Q_{img}) = w_c^{\cdot}(q)$ for QUCOM based on these quantities to get a ranked list of the concepts:

$$w_c^{DELTA}(q) = freq(c, q) - freq(c), c \in L, \tag{2}$$

$$w_c^{CTFIDF}(q) = freq(c, q)\log(1/freq(c)), c \in L, \tag{3}$$

$$w_c^{PMIWS}(q) = \log(freq(c, q)/freq(c)), c \in L. \tag{4}$$

As indicated in the superscript, these models in Eq.(2)-(4) correspond to the delta, *c-tf-idf* [7] and PWIMS [13] respectively. Besides concept popularity measured in $freq(c, q)$, the $freq(c)$ measures the concept specificity since concepts with smaller $freq(c)$ might be more distinctive. The three models are therefore different combination schemes leveraging differently on the two quantities. The delta function looks like Rocchio feedback and selects the concepts according to their frequency divergence from the corpus mean given the query. The essence of the *c-tf-idf* based concept selection method is to pick out concepts which maximally reduce the uncertainty of the corpus's relevance to the query [2]. The PMIWS method in Eq.(4) measures the pointwise information gain when given $freq(c, q)$ over $freq(c)$. However, the $\log(freq(c, q))$ term in Eq.(4) may suppress $freq(c, q)$ too much.

Given one query image concept vector $\vec{q}$, concepts are ranked in terms of $w_c^{\cdot}(q)$ as defined in Eq. (2)-(4). Then a top-k concepts are selected as L_s^k, a subset L_s with k concepts. If multiple query images are given as $Q_{img} = \{\vec{q_1}, \ldots, \vec{q_m}\}$, we assume that they have consistent information need, and therefore $freq(c, q) = freq(c, Q_{img}) = \frac{1}{m}\sum_{q' \in Q_{img}} P(c|q')$, where Q_{img} is the query image set and m is the number of images.

2.3 Solving QUCOM with Text and Images

Under a unified representation of the vector space model, it is quite straight forward to fuse both kinds of information by linearly combining the respective weights, defined as

$$w_c(Q) = \lambda_1 w_c(Q_{txt}) + \lambda_2 w_c(Q_{img}), \tag{5}$$

where $w_c(Q_{txt})$ and $w_c(Q_{img})$ are corresponding component weights, and λ_1, λ_2 are their respective combination factors. We set $\lambda_1 = 1/w_{max}^{txt}$ and $\lambda_2 = 1/w_{max}^{img}$ to balance the influence of text and image, where w_{max}^{txt} is the maximum weight for all concepts from the text modality and w_{max}^{img} defined similarly.

2.4 Determining the Number of Concepts

Although fixing the top-k concepts is acceptable, as shown in Sec.4.2, it is more desirable to decide how many concepts are helpful for a query. When relevance information from user feedback is available, it would be much easier. However, in automatic video retrieval without relevance information, an feasible method is to estimate a query language model (unigram distribution over concepts), and then calculate the Kullback-Leibler divergence between the query and corpus language model, given by

$$D_{KL}(Q||\mathcal{C}) = \sum_{c \in L_s} P(c|Q)\log\frac{P(c|Q)}{P(c|\mathcal{C})}, \tag{6}$$

where $c \in L_s$ is a selected concept. To further take advantage of the search result R, an additional language model $P(c|R)$ is introduced. A ratio r of $D_{KL}(R||Q)$ to $D_{KL}(R||\mathcal{C})$ seems a good indicator for measuring the search result quality and subsequently the concept number. It is reasonable to search with multiple set of different number of top concepts and select the one with the largest r. Thus, this measure provides a method to determine the number

of relevant concepts for a specific query. More formally, $P(c|Q)$, $P(c|R)$ and r are given respectively by

$$P(c|Q) = tf(c, q), \tag{7}$$

$$P(c|R) = \sum_{d \in R} P(c|d)P(d|q), \tag{8}$$

$$r = D_{KL}(R||Q)/D_{KL}(R||\mathcal{C}), \tag{9}$$

where d is a shot in R. We set R to top 10 shots throughout this study. The weight $P(d|q)$ in Eq. (8) is given by the estimated likelihood $P(q|d)$ and a Bayesian inversion with uniform prior probabilities for shots in R and a zero prior for others.

3. EXPERIMENTAL SETUP

A serial of experiments are conducted on the TRECVID 2006 (TV06) [1] search data set to evaluate both the necessity and effectiveness of QUCOM on automatic video retrieval.

The TV06 data set consists of 150-hour multilingual news video captured from US, Arabic and Chinese, with a reference of $79,484$ segmented shots as the retrieval unit. We use all 24 multimedia search queries defined in TV06 for the experiments. They express the information need of users for video search concerning people, things, events, locations, etc. and combinations of these needs. Given such a need as input, a video search engine should produce a ranked list of results without human intervention. For each topic we return a ranked list of up to 1000 results. The ground truth for all 24 topics are generated and made available by the organizers. The performance is evaluated by Average Precision (AP) on the shot level, following the TRECVID evaluation standard. To compare results across queries, Mean Average Precision (MAP) is defined as the mean AP scores involved for all queries.

The concepts are annotated and trained on another 80 hours training set. We apply the 311 concept detectors on each shot in the data set. The concept index generation process follows the state-of-the-art concept detection system and is described in detail in [7]. We omit it here due to space limitations. Then we perform experiments first to determine the performance of each concept detector for each query, and then compare different methods for solving QUCOM:

1. Exhaustively evaluating all 311 concept detectors against all 24 queries to assess the necessity of performing QUCOM. Since we need QUCOM when only a few concepts are relevant to a query.

2. Evaluating the effectiveness of using QUCOM to choose the top k concepts for search. We evaluate the effectiveness of using text, image and their combination to solve QUCOM, especially for image since multiple methods are introduced.

3. Evaluating the effectiveness of determining the number of related concepts. We determine the number of concepts relevant to a query by maximizing r. Up to 20 concepts are evaluated by setting $k_{max} = 20$ and the L_s^k with the maximum r_k is choose.

4. RESULTS AND ANALYSIS

4.1 How Many Concepts are Relevant?

We summarize the number of concepts relevant to a query in Fig. 1. Note that only the concepts with AP>0.01 are selected as relevant. We observe that though the queries are designed towards utilizing the much smaller LSCOM-Lite lexicon with 39 concepts, a number of concepts in a larger LSCOM lexicon [8] do contribute

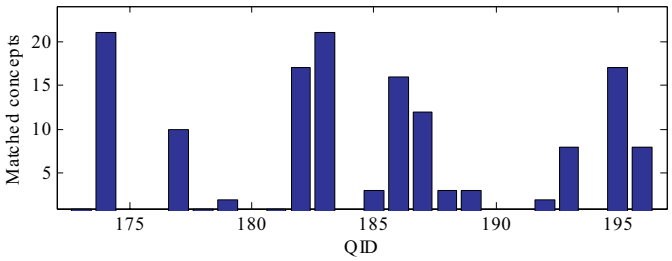

Figure 1: The number of the relevant concepts for 24 queries.

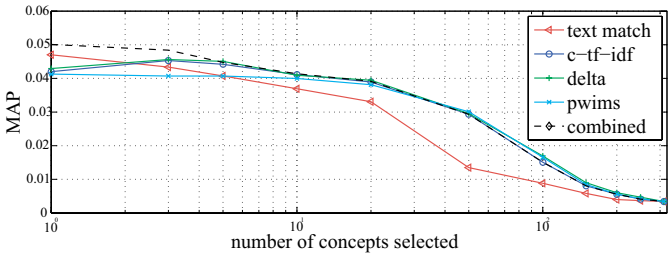

Figure 2: The impact of the selected concept number on the MAP performance of the 24 queries.

to the overall performance. However, many queries have only a few (< 10) relevant concepts, indicating the necessity to perform QUCOM. While on average 6.1 concepts are relevant to a query, the standard deviation is 7.3, meaning that a large variance of the number of relevant concepts is present across different queries. So it is also desirable to determine the number of relevant concepts.

4.2 How to Choose The Top-k Concepts?

As verified in previous study [12], the number of concepts in a lexicon correlates positively with the retrieval performance. However, as the lexicon scales up, taking many concepts for a single query without discriminating their relevance will bring in more irrelevant conepts. As a result, the performance degenerates if we fix k, the number of selected concepts, over 10 for all concepts, as shown in Fig. 2 where the x-axis shows incrementally combined concepts which are sorted in their relevance to the query and the y-axis shows their respective MAP across all queries. The resulting MAPs of the methods of text match, three visual methods and a combination of text and visual *c-tf-idf* all show that only a few top concepts are helpful for the query, indicating the importance of QUCOM, especially for large-scale concept lexicons.

Another observation from Fig. 2 is that text is good at choosing the top concept, however as the concepts selected increase, the performance of text QUCOM drops more sharply than the image QUCOM. The image QUCOM can find more relevant concepts since the maximum MAP is achieved at 3-5 concepts. Also, both *c-tf-idf* and *Delta* perform comparably while PWIMS is not satisfactory, indicating that $freq(c, q)$ is important for selecting relevant concepts. By further combining the text and image information (using *c-tf-idf*), the performance increases 6% in MAP at the top concept and robustly outperforms all others as k increases. So query images do provide additional information and should be integrated with the text whenever available.

To get a micro vision of which concept accounts for which query, we show the detailed QUCOM result from text, image (using *c-tf-idf* only) and their combination in Table 1. The MAP of using only the top concept is also shown. It can be seen that most concepts judged relevant to the queries do make sense, which shows the effectiveness of the QUCOM methods. For example, for query

Table 1: Query-Concept Mapping Results (Only the top one concept is listed due to space limitations.)

	QUCOM Strategies					
	Text		Image		Combined	
Queries	Selected Concept	AP	Selected Concept	AP	Selected Concept	AP
0173. emergency vehicles in motion	Emergency_Vehicles	0.004	Car	0.006	Emergency_Vehicles	0.004
0174. tall buildings and the top story visible	Building	0.020	Cityscape	0.018	Cityscape	0.018
0175. people leaving or entering a vehicle	Ground_Vehicles	0.004	Vehicle	0.003	Vehicle	0.003
0176. soldiers, police, or guards escorting a prisoner	Guard	0.001	Emergency_Room	0.000	Guard	0.001
0177. daytime demonstration or protest with building visible	Demonstration_Or_Protest	0.035	People_Marching	0.072	Demonstration_Or_Protest	0.035
0178. US Vice President Dick Cheney	Us_Flags	0.015	Head_of_State	0.001	Us_Flags	0.015
0179. Saddam Hussein with another persons face visible	Car_Crash	0.000	Us_Flags	0.001	Us_Flags	0.001
0180. people in uniform and in formation	Non-uniformed_Fighters	0.001	Crowd	0.000	Non-uniformed_Fighters	0.001
0181. US President George W. Bush, Jr. walking	George_Bush	0.002	Agent	0.000	George_Bush	0.002
0182. soldiers or police with weapons and military vehicles	Soldiers	0.022	Armed_Person	0.036	Soldiers	0.022
0183. water with boats or ships	Boat_Ship	0.031	Lakes	0.042	Boat_Ship	0.031
0184. people seated at a computer with display visible	Computers	0.004	Furniture	0.001	Computers	0.004
0185. people reading a newspaper	Newspaper	0.109	Furniture	0.000	Newspaper	0.109
0186. a natural scene	Beach	0.011	Waterscape_Waterfront	0.027	Beach	0.011
0187. helicopters in flight	Helicopters	0.057	Mosques	0.008	Helicopters	0.057
0188. something burning with flame visible	Road_Overpass	0.000	Smoke	0.023	Smoke	0.023
0189. people dressed in suits, seated, and with flag	Dresses	0.000	Meeting	0.006	Flags	0.039
0190. at least one person and at least 10 books	Single_Person	0.000	Flags	0.000	Person	0.000
0191. at least one adult person and at least one child	Child	0.005	First_Lady	0.002	Child	0.005
0192. a greeting by at least one kiss on the cheek	Greeting	0.023	Old_People	0.002	Greeting	0.023
0193. smokestacks, chimneys, or cooling towers with smoke	Smoke	0.002	Tower	0.017	Tower	0.017
0194. Condoleeza Rice	-	0.000	Head_And_Shoulder	0.000	Head_And_Shoulder	0.000
0195. soccer goalposts	Soccer	0.736	Soccer	0.736	Soccer	0.736
0196. scenes with snow	Snow	0.047	Sky	0.002	Snow	0.047
MAP		0.047		0.042		0.050

"**0185. people reading a newspaper**" the text-match QUCOM finds the most relevant concept, *Newspaper*. For the query "**0195. soccer goalposts**", *Soccer* can also be triggered by text-match. However, the relevant *Sports* and *Lawn* are returned only by the image QUCOM. Besides, we find certain concepts which are not explicitly related to the queries, such as *Us_Flags* for query **0179** and *Smoke* for query **0188**, are returned only by the image QUCOM. One more example comes from query **0187. helicopters in flight**. The concept *Mosques* is found to be relevant. It is no surprise if we notice that many shots are about the Iraq war, and there exists the coincidence of helicopters and mosques. So image QUCOM finds concepts which are not easily detected by text QUCOM.

Meanwhile, retrieval with QUCOM achieves the state-of-the-art performance for query-by-concept retrieval [4, 3] and is comparable with the text retrieval. Notice that the query-by-concept is clumsy at named entity search, e.g. "Cheney" and "Hussein" while text is good at dealing with them. Automatic video retrieval will surely benefit from a combination of query classification and QUCOM.

4.3 How to Determine The Number of Relevant Concepts?

Our concept number selection method based on query language model gets an MAP of 0.051 (not shown due to space limitation), and is slightly better than simply using the combined QUCOM method for top concepts and is within 85% of 0.060, the MAP of the oracle which selects the best single concept detector for all concepts. However, the average selected concept number is 2.8, much smaller than the number of 6.1 determined experimentally in Sec 4.1. This rather large gap shows that effort should still be made to get more accurate estimation of the number of relevant concepts.

5. CONCLUSIONS

One major contribution of this work is to point out the importance of QUCOM for semantic video retrieval. In addition, we have shown that text and visual information are complementary for solving QUCOM, compared a few methods in a unified vector space model, and tried to determine the relevant concepts for a query for the first time. Though preliminary, our results are encouraging and suggest a promising new line of research. Currently we are exploring new methods to determine the number of the related concepts. Possible further work includes integrating this work into the query classification framework, learning the weights for each concept in a more principled way and extending the framework to an interactive retrieval scenario.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under the grant No. 60621062 and 60605003, and the National Key Foundation R&D Projects under the grant No. 2003CB317007 and 2004CB318108.

6. REFERENCES

[1] Trecvid home page. http://www-nlpir.nist.gov/projects/trecvid.

[2] A. Aizawa. An information-theoretic perspective of tf-idf measures. *Information Processing and Management*, 39:45–65, January 2003.

[3] M. Campbell and et al. Ibm research trecvid-2006 video retrieval system. In *Proc. Of TRECVID*, 2006.

[4] S.-F. Chang, W. Hsu, W. Jiang, L. Kennedy, D. Xu, A. Yanagawa, , and E. Zavesky. Columbia university trecvid-2006 video search and high-level feature extraction. In *Proc. of TRECVID workshop*, 2007.

[5] M. G. Christel and A. G. Hauptmann. The use and utility of highlevel semantic features in video retrieval. In *Proc. of CIVR*, 2005.

[6] A. Haubold, A. P. Natsev, and M. R. Naphade. Semantic multimedia retrieval using lexical query expansion and model-based reranking. In *Proc. of ICME*, 2006.

[7] X. Li, D. Wang, J. Li, and B. Zhang. Video search in concept subspace: A text-like paradigm. In *Proc. of CIVR*, 2007.

[8] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE Multimedia Magazine*, 2006.

[9] A. P. Natsev, M. R. Naphade, and J. Tesic. Learning the semantics of multimedia queries and concepts from a small number of examples. In *Proc. of ACM Multimedia*, 2005.

[10] S.-Y. Neo, J. Zhao, M.-Y. Kan, and T.-S. Chua. Video retrieval using high level features: Exploiting query matching and confidence-based weighting. In *Proc. of CIVR*, 2006.

[11] C. G. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *IEEE Transactions on Multimedia*, 2007.

[12] C. G. Snoek and M. Worring. Are concept detector lexicons effective for video search? In *Proc. of ICME*, 2007.

[13] W. Zheng, J. Li, Z. Si, F. Lin, and B. Zhang. Using high-level semantic features in video retrieval. In *Proc. of CIVR*, 2006.