Distribution-based Concept Selection for Concept-based Video Retrieval

Juan Cao
Institute of Computing Technology, Chinese Academy of Science
caojuan@ict.ac.cn

HongFang Jing
Institute of Computing Technology, Chinese Academy of Science
jinghongfang@ict.ac.cn

Chong-Wah Ngo
Computer Science Department, City University of Hong Kong
cwngo@cs.cityu.edu.hk

YongDong Zhang
Institute of Computing Technology, Chinese Academy of Science
zhyd@ict.ac.cn

ABSTRACT

Query-to-concept mapping plays one of the keys to concept-based video retrieval. Conventional approaches try to find concepts that are likely to co-occur in the relevant shots from the lexical or statistical aspects. However, the high probability of co-occurrence alone cannot ensure its effectiveness to distinguish the relevant shots from the irrelevant ones. In this paper, we propose distribution-based concept selection (DBCS) for query-to-concept mapping by analyzing concept score distributions of within and between relevant and irrelevant sets. In view of the imbalance between relevant and irrelevant examples, two variants of DBCS are proposed respectively by considering the two-sided and one-sided metrics of concept distributions. Specifically, the impact of positive and negative concepts toward search is explicitly considered. DBCS is found to be appropriate for both automatic and interactive video search. Using TRECVID 2008 video dataset for experiments, improvements of 50% and 34% are reported when compared to text-based and visual-based query-to-concept mapping respectively in automatic search. Meanwhile, DBCS shows continuous improvement for all rounds of user feedbacks in interactive search.

Categories and Subject Descriptors

H.3.3 [**Information storage and retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Performance, Experimentation

Keywords

Concept-based video retrieval, query-to-concept mapping, Distribution

1. INTRODUCTION

The recent advancement in high-speed computing and networking technologies has led to explosive growth of video data. How to efficiently and exactly predict the user search intentions from this vast scale of database has became an urgent but challenging issue. Among the existing search methodologies, concept-based video retrieval has emerged as a promising direction for its potential in bridging the so-called semantic gap. There has been a spurt of research attention to address this search paradigm [1][2]. A. Hauptmann's prediction of "using no more than 5000 concepts will be sufficient for accurate retrieval, despite a fairly low detection accuracy of 10% for any single concept and substantial combination errors" [1] further provides the theoretic support to the potential of using concept detectors for retrieval.

As illustrated in Fig.1, the concept-based video retrieval includes three vital parts: lexicon construction, concept detection and query-to-concept mapping. From the perspective of semantic space coverage, many concept lexicons have been proposed such as Large Scale Concept Ontology for Multimedia (LSCOM). The definition of semantic gap is also recently proposed in [4] by offering a new criterion for lexicon construction. The importance of semantic concepts for large-scale video search has long been evidenced in TREC Video Retrieval Evaluation (TRECVID) [2]. To date, query-to-concept mapping remains a challenging issue to address [3][5].

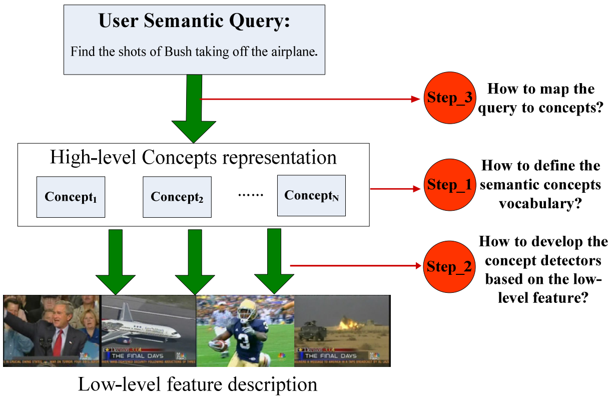
User Semantic Query:
Find the shots of Bush taking off the airplane.
High-level Concepts representation
Concept$_1$
Concept$_2$
……
Concept$_N$
Low-level feature description
Step_3
How to map the query to concepts?
Step_1
How to define the semantic concepts vocabulary?
Step_2
How to develop the concept detectors based on the low-level feature?

Figure 1: General framework of concept-based video retrieval.

There have been several works being proposed for query-to-concept mapping [2]. We can broadly summarize the existing methods into two classes: the semantic-similarity and the statistical-similarity principle.

On the semantic-similarity principle, the approaches aim to retrieve the most accurate concepts to describe the given query. Generally, these approaches compute the lexical similarity between the textual keywords and the concept description using dictionaries or other knowledge resources such as WordNet, or to infer the visual similarity by the confidence scores of concept detector models to the visual examples.

MM'09, October 19–24, 2009, Beijing, China.
Copyright 2009 ACM 978-1-60558-608-3/09/10...$10.00.

On the statistic-similarity principle, the approaches are data-driven and attempt to deduce the most possible co-occurrence concepts for a given query. The most popular statistic based method is by mining the co-occurrence patterns within the automatic speech recognition (ASR) text based on clustering or Latent Semantic Indexing (LSI) [3]. The statistical analysis approaches for visual concept-to-query mapping have not been explored as much as textual approaches [3]. Natsev et al. [3] used the standard statistical hypothesis tests such as t-score or $\chi 2$ test to evaluate the importance of each concept to the query examples based on the detector scores. Moreover, they first proposed a probabilistic local context analysis (pLCA) method by combining the concept detector results and the initial retrieval results.

Both principles can be summarized by similarity-ranking metric. The concepts selected by these methods are likely most appropriate to describe the query, but it does not mean being the most useful ones for retrieval. For example, the concept "people" is most related to the query "Find shots of one or more people walking up stairs". However, such frequent concept is not beneficial for distinguishing positive samples from large numbers of negative samples. On the other hand, some concepts are neither the most similar nor the most dissimilar ones to user query, but they exhibit considerably difference between the distribution of relevant and irrelevant sets. For example, in the topic "Find shots of a crowd of people, outdoors, filling more than half of the frame area", the concept "Dark-skinned_People" is not semantically relevant to the topic, and not the most co-occurrence concept with the topic. But in the dataset of TRECVID 2008, most of the dark-skinned-people related shots are relevant to demonstration event. Selecting such not semantically and frequently co-occurred concepts surprisingly are helpful for retrieving relevant shots, but generally ignored by the conventional approaches which maximize semantic or statistic similarity.

In this paper, we propose distribution-based concept selection (DBCS), aiming to select the most discriminative, rather than the most semantically or statistically relevant, concepts for video retrieval. The targeted concepts are those concepts whose distributions of detection score fluctuate widely between the relevant and irrelevant collections, but remain stable within both. Instead of selecting concepts by relying purely on detection scores as existing methods do, DBCS takes into account the detection variation of data collection by assessing the variance of detection scores as one of selection criteria. As a consequence, DBCS is less sensitive to performance of concept detectors, while being capable of showing stable and satisfactory retrieval performance.

The rest of this paper is organized as follows. Section 2 introduces two variants of DBCS. Section 3 reports our experimental results on automatic and interactive video search. Section 4 concludes this paper.

2. DISTRIBUTION BASED CONCEPT SELECTION (DBCS)

To present DBCS, we first introduce the following notations in the rest of the paper:

c: a concept .

s: a shot.

q: a given user query.

Q : the collection of relevant shots to the query *q*. In this paper, Q refers to the example images/videos in automatic retrieval case, or the positive samples annotated by users in the interactive retrieval.

$\bar{Q}$: the collection of irrelevant shots to the query *q*. Generally, they are pseudo negative examples sampled from the whole dataset.

The main idea of DBCS is that as the difference between $p(c|Q)$ and $p(c|\bar{Q})$ increases, the concept *c* becomes more indicative for a given query. Moreover, for all the relevant shots belonging to *Q*, if the differences between $p(c|s_i)$ and $p(c|s_j)$ ($i \neq j$) decreases, the concept becomes more reliable for the query. So DBCS aims to select the reliable and discriminative concepts.

In view of the imbalance between relevant and irrelevant datasets, two variants of DBCS are proposed respectively by considering the two-sided and one-sided metrics of concept distributions, separately called Global-DBCS and Local-DBCS.

2.1 Global-DBCS

Definition 1 *Inter-Category Variance of c* denoted as $V_{inter}(c)$ is the distribution differences of *c* between the relevant and irrelevant categories. It can be formulated as follows:

$$V_{\text{int}er}(c) = (P(c,Q) - P(c,\bar{Q}))^2 \quad , \tag{1}$$

where $P(c,Q) = \frac{1}{|Q|}\sum_{s \in Q} P(c \mid s)$, and $P(c \mid s)$ is the concept detector score of *c* for shot s. $P(c,\bar{Q})$ is defined similarly.

Large $V_{inter}(c)$ means that distributions of *c* vary greatly from *Q* to $\bar{Q}$, and thus *c* has better discriminative power. On the other hand, small V_{inter} means concept distributions is similar between *Q* and $\bar{Q}$. This gives clue that the contribution of *c* to retrieval performance is not significant.

Definition 2 *Inner-Category Variance of c in Q denoted as* $V_{inner}(c, Q)$ is the distribution difference of *c* in all shots belonging to *Q*.

$$V_{inner}(c,Q) = \frac{1}{n} \sum_{s \in Q} (P(c,s) - P(c,Q))^2 \tag{2}$$

Small $V_{inner}(c, Q)$ means that the distribution of *c* in *Q* is stable, and is reliable to represent the relevant set, while the concept with large $V_{inner}(c, Q)$ are regarded as noises for *Q*.

Based on the above analysis, a useful concept to a given query should have stable distribution within *Q* to ensure its reliability, and further have different distribution between *Q* and $\bar{Q}$ to increase its discriminability. By the definitions, Global-DBCS aims at finding out this kind of concepts with large $V_{inter}(c)$ but small $V_{inner}(c, Q)$. The score formula is defined as:

$$DBCS\_Score(c) = V_{\text{int}er}(c) / V_{inner}(c,Q) \tag{3}$$

Then the top *k* effective concepts can be selected by sorting the *DBCS_Score* in descending order.

2.2 Local-DBCS

Global-DBCS is essentially a two-sided metric where no distinction is made between positive and negative effects of concepts [1]. Specifically, positive concepts are those which if appear in shots will boost the relevancy of shots to query. Conversely, negative concepts are those which their appearances give clues to the irrelevance of shots to *Q*. To study the impact of positive and negative concepts, called P-concept and N-concept respectively, we further propose a variant of Global-DBCS called Local-DBCS.

Definition 3 *Inter-Category Variance of c to Q*, a variant of *Definition 1* which differentiates positive and negative concepts, denoted as $V_{inter}(c, Q)$ measures the distribution difference of *c* in Q to $\bar{Q}$.

$$V_{inter}(c,Q) = sign(P(c,Q) - P(c,\bar{Q})) * (P(c,Q) - P(c,\bar{Q}))^2 \tag{4}$$

Where the positive sign means the concept *c* is more likely to occur in relevant shots than in irrelevant ones, and it is a potential P-concept for *Q*. The final score for P-concept is:

$$DBCS\_Score(c,Q) = V_{inter}(c,Q) / V_{inner}(c,Q) \tag{5}$$

On the contrary, the negative sign of $V_{inter}(c,Q)$ implicates that *c* is less likely occurring in the relevant shots, and is a potential N-concept. Thus,

$$V_{inter}(c,\bar{Q}) = -V_{inter}(c,Q) \tag{6}$$

Then the N-concepts can be ranked by is defined as:

$$DBCS\_Score(c,\bar{Q}) = -DBCS\_Score(c,Q) \tag{7}$$

3. EXPERIMENTS

We evaluate the performance of DBCS separately in automatic and interactive video retrieval. In the following experiments, all the data and the ground-truth are from the 2008 TRECVID benchmark. The scores of concept detectors are from CU-VIREO374 [7]. The evaluation criterion is inferred average precision (infAP) [2]. For a given query, no more than three concepts are selected [1]. We adopt average fusion to combine the selected detectors for retrieval.

3.1 Retrieval Performance Evaluation

3.1.1 Automatic Retrieval Case

We experiment eight existing algorithms for query-to-concept mapping. A total of 48 topics provided by TRECVID 2008 dataset are tested. The details of these algorithms are:

SVM_CMVS: This method uses Multi-bag SVM as the basic classifier . We build a 10-bag SVM classifier for each query topic. The input feature to SVM is model vector composed of the detection scores of 374 detectors to the query examples of the topic. Among the 10 bags, each bag has the same positive samples formed by query examples, but different pseudo negative samples which are randomly sampled from test data.

Text-Mapping: This method firstly extracts keywords from the textual description of the queries and concepts, including the nouns and verbs. Then maps the two kinds of keywords after stemming.

Information-theoretic methods: We experiment TF-IDF, information gain (IG) and chi-square (CHI), where the input feature is detection scores of 374 detectors. Among them, IG and CHI are the most effective feature selection method [6], and TF-IDF is a method popularly used in information retrieval.

Global-DBCS and Local-DBCS: besides the provided query example images, we expand it by extracting the keyframes from the the query example video clips as the relevant shots. Then every query has about 20 relevant shots. Both methods select the top 3 concepts with the criterion that the scores of second and third concepts should be at least the half of the greatest. So the final number of selected concepts by DBCS may be less than 3.

Text+Local-DBCS: An average fusion result of the text-Mapping and Local-DBCS.

The results in Fig.2 show that the discrimilability-ranking principle (DBCS) offers a great improvement over the similarity-ranking (over SVM_CMVS 34% over Text-Mapping 50%), and outperforms the other classic information-theoretic methods. Moreover, the Local-BDCS algorithm is superior to the Global-DBCS, which verifies that the one-sided measurement metric distinguishing the positive and negative affect is more reasonable in the imbalance dataset.

Meanwhile, the fusion of DBCS and Text-Mapping achieves the best performance. Because DBCS sometimes fails when the topic's relevant shots are very few. So in this case text-mapping is a reasonable supplement to DBCS.

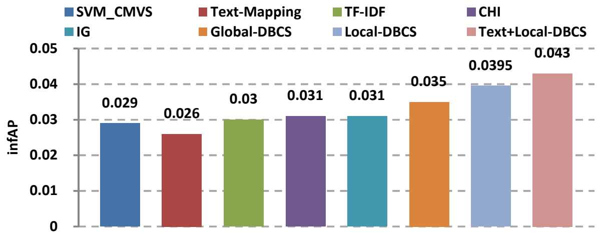

Figure 2: Concept selection algorithms evaluation in automatic video retrieval.

3.1.2 Interactive Retrieval Case

We further evaluate the performance of Local-DBCS in TV08's 48 interactive retrieval topics. We consider five rounds of feedbacks, and each round with 500 samples, where all the relevant and irrelevant shots are annotated based on the ground-truth.

The mean infAP of Local-DBCS against text-mapping after five rounds of feedbacks is shown in Fig 3. Local-DBCS shows consistently better performance than text-mapping in the whole feedbacks. Besides, the increase curve of Local-DBCS has a rapid growth in the first three rounds, because the Local-DBCS feedback algorithm can quickly find most of the potential positive samples. After that, the growth becomes muted, for the number of positive samples in the left collection is getting smaller.

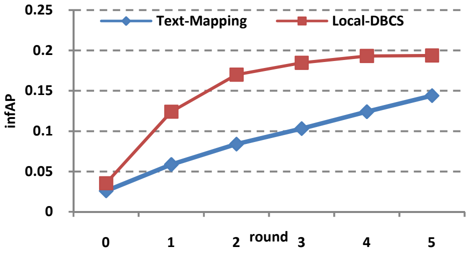

Figure 3: The performance of Local-DBCS and Text-Mapping on 48 TV08's interactive video retrieval topics.

3.2 Result Discussion

In this section, we further provide empirical insights to analyze the performance of DBCS. We regard the text-mapping and Local-DBCS separately as the baselines of similarity-ranking and discriminability-ranking metrics, and analyze their original selected concepts for two queries. Tables 1 and 2 show two examples where the selected concepts are listed in descending order. Note that in the experiment only the top-3 concepts are used for search.

Table 1: Comparison of the selected concepts by Local-DBCS and Text_Mapping methods in Topic 248.

Topic 248: Find shots of a crowd of people, outdoors, filling more than half of the camera.		
Mapping Schemes	**Local-DBCS**	**Text_Mapping**
infAP	**0.321**	**0.203**
Concept_1	Crowd	Crowd
Concept_2	People_Marching	Outdoors
Concept_3	Demonstration_Or_Protest	Person
Concept_4	Protesters	---
Concept_5	Dark-skinned_People	---

Table 2: Comparison of the selected concepts by Local-DBCS and Text_Mapping methods in Topic 261.

Topic 261: Find shots of one or more people at a table or desk, with a computer visible.		
Mapping Schemes	**Local-DBCS**	**Text_Mapping**
infAP	**0.116**	**0.012**
Concept_1	Office	Computer
Concept_2	Computer_Or_Television_Screens	Computer_Or_Television_Screens
Concept_3	Attached_Body_Parts	Person
Concept_4	Classroom	---
Concept_5	Medical_Personnel	---

Topic 248 in Table 1 shows an example where the discriminability-ranking and similarity-ranking select consistently similar concepts for retrieval. Both methods have achieved exciting performance, and select one common concept "Crowd". Local-DBCS does not select the “Outdoors” and “person”, because they not only frequently occur in the relevant set, but also in the irrelevant one. On the other hand, Local-DBCS selects the “Dark-skinned-people”, for most of dark-skinned-people videos in the datasets are related to marching event. So it has strong discriminability to distinguish the relevant shots from the irrelevant ones, and can contributes a 3.5% of improvement if the top-5 concepts are selected for search.

The topic 261 in Table 2 is another example where the discriminability-ranking and similarity-ranking select very different concepts for retrieval. Their performances are very different. Text_Mapping method selects “Computer”. Although this concept is semantically related, but its detector performance is not reliable enough to support video search. On the contrary, DBCS selects concepts based on their variances between relevant and irrelevant sets. Thus, detector performance is indeed indirectly considered during concept selection. As a result, less reliable detectors naturally have less chance to be picked by DBCS for retrieval.

4. CONCLUSION

In this paper, we propose an effective concept selection method called DBCS. This method can be used in automatic and interactive video retrieval, and outperforms the state-of-art query-to-concept mapping algorithms in both cases. The contributions of DBCS can be attributed to two factors. First, DBCS selects the most discriminative concepts, rather than the most relevant ones for retrieval. Second, DBCS considers variance of detection scores, instead of the original scores by detectors. This results in the selection of concepts with relatively robust detection performance. Both factors contribute to the significant retrieval improvement by DBCS in automatic and interactive search. Finaly, we also demonstrate that the one-sided concept selection metric like Local-DBCS is more reasonable than the two-sided method for the imbalance dataset.

5. ACKNOWLEDGMENTS

This work was supported by National Basic Research Program of China (973 Program, 2007CB311100), National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), National Nature Science Foundation of China (60873165, 60802028).

6. REFERENCES

[1] A. Hauptmann, Y. Rong, W.H. Lin, M. Christel, H.Wactlar, “Can High-Level Concepts Fill the Semantic Gap in Video Retrieval? A Case Study With Broadcast News”, *IEEE Transactions on Multimedia*, Vol. 9(5), pp: 958-966, 2007

[2] P. Over, G. Awad, T. Rose, J. Fiscus, W. Kraaij, A.F. Smeaton, TRECVID 2008--Goals, Tasks, Data, Evaluation Mechanisms and Metrics, *in Proceedings of TRECVID Workshop*, USA, 2008

[3] A. Natsev, A. Haubold, J. Teˇsi´c, L. Xie, and R. Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval: A comparative review and new approaches. *In ACM Multimedia (ACM MM)*, Sep. 2007.

[4] Y. Lu, L. Zhang, Q. Tian, W.Y. Ma. What Are the High-level Concepts with Small Semantic Gaps. *IEEE Conference on CVPR*, pp.1-8, 2008

[5] B. Huurnink，K. Hofmann，Maarten de Rijke，Assessing Concept Selection for Video Retrieval, *MIR’08*, pp. 459-466

[6] Y.M. Yang and O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. *Proceedings of ICML-97*, pp. 412-420, 1997

[7] Y.G. Jiang, A. Yanagawa, S.F. Chang, and C.W. Ngo, "CU-VIREO374:Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection", *Columbia University ADVENT Technical Report #223-2008-1*, 2008.

[8] A. Natsev, M. R. Naphade, and J. Teˇsi´c. Learning the semantics of multimedia queries and concepts from a small number of examples. *In ACM Multimedia ,* Nov. 6–11 2005.