Knowledge Base Retrieval at TRECVID 2008

David Etter

Abstract

This paper describes the Knowledge Base multimedia retrieval system for the TRECVID 2008 evaluation. Our focus this year is on query analysis and the creation of a topic knowledge base using external knowledge base information.

Index terms- clustering, topic knowledge base, fusion, multi-query-by-example, query analysis

1. Introduction

This paper describes our retrieval approach for the TREC Video Evaluation 2008. We participated in the fully automatic search task and submitted 6 runs.

Our work this year focuses on the query analysis component of our multimedia retrieval system and the generation of a topic knowledge base. The topic knowledge base enhances the initial text topic description with knowledge and context using multiple modalities.

Our 6 submitted runs are described below:

- KBVR_1: This is a run using only the ASR/MT output provided by NIST and the search topic text (Text only).
- KBVR_2: This is a run using the provided ASR/MT and a topic knowledge base constructed with 3 Wikipedia articles, 3 news articles, and 3 web pages (Text only).
- KBVR_3: This is a run using only the image features and no ASR/MT (Image only).
- KBVR_4: This is a run using image and text features with a topic knowledge base constructed using 3 Wikipedia articles and 3 MQBE image clusters.
- KBVR_5: This is a run using image and text features with a topic knowledge base constructed using 5 Wikipedia articles and 5 MQBE image clusters.
- KBVR_6: This is a run using image and text features with a topic knowledge base constructed using 5 Wikipedia articles and 20 MQBE image clusters.

2. KB System Overview

The KB multimedia retrieval system consists of four main components: feature extraction, query analysis, search, and result fusion [Figure 1]. The feature extraction component uses a number of low-level visual features to describe an image and forms the search feature space for multi-query-by-example. Query analysis involves the construction of the topic knowledge base and provides the translation from a text query topic to the multimedia feature space. The search component provides the indexing for the multimedia feature space and includes the low-level visual features and the ASR/MT features. Result fusion merges the individual ranked results from each of the 'experts' which make up our topic knowledge base.

Topic Knowledge Base

The question of how to translate a query, in the form of a text description, into a feature space which allows for effective multimedia search remains an open question. Queries to a multimedia retrieval system are often expressed in terms of a very general concept, object, or event. As an example, we can look at this year's search topics which include text descriptions such as topic 221, "a person opening a door" or topic 229, "one or more people where a body of water can be seen". The generality of these text descriptions make it difficult to understand the context of the information need and to explore the intended context of a multimedia shot.

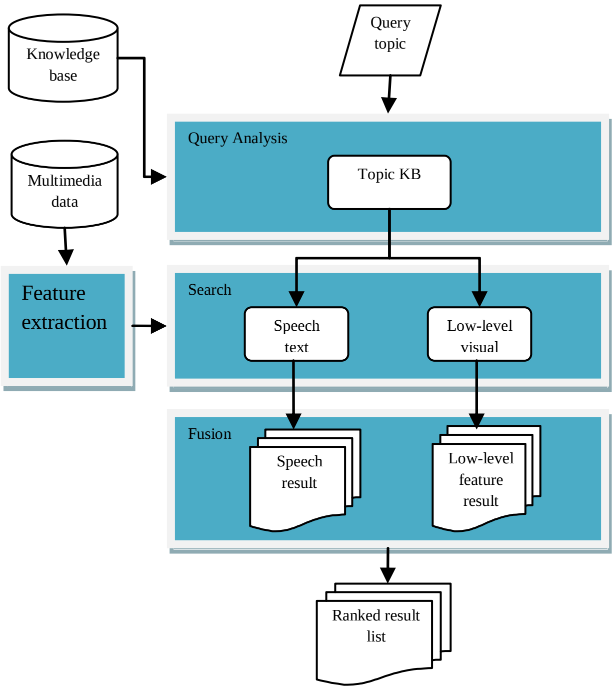

Figure 1: KB multimedia retrieval system

Beyond the challenges of providing a complete text description is the question of a visual description. The TREC evaluation [1] provides a number of sample query images or video clips as a visual query topic. This approach is the exception to a typical search, since a user will rarely have example images for each multimedia information request.

To overcome the challenges of query generality, context, and visual description, we develop a topic knowledge base [Figure 2]. Our topic knowledge base consists of three main components: concept descriptor, visual descriptor, and context descriptor. The topic knowledge base attempts to translate a text description into a multimodal knowledge base which includes specific context and examples of the information need. The topic knowledge base uses a number of external data sources to construct its visual and context related feature vectors.

Concept descriptor

The topic description represents the original information need provide by a user and is the basis of the knowledge base construction. The concept component uses an external data source which provides a detailed concept description, related concept description, and categorization. The concept descriptors for a topic are selected using a k-nearest neighbor approach to the original topic description in a term vector feature space.

Visual descriptor

The visual descriptor component consists of low-level visual feature vectors [2] representing example queries for the topic description. These example queries represent multiple context views of the query topic in the visual feature space. Sample visual queries are selected from an external image repository, where the topic description and concept descriptions are used as queries. The image repository is indexed by text metadata descriptions and similarity to a query is determined using only the text and metadata features. Large image repositories constructed using un-supervised algorithms often provide unreliable results due to the text metadata used for their indexing and result ranking. In order to overcome this limitation we construct a visual filter to help eliminate unrelated images [Figure 3].

We would like the visual descriptor to include multiple non-overlapping visual examples. The component selects the k-nearest neighbors to each of the concepts represented in the concept descriptor. After filtering “unrelated” images we select our N examples using a k-means clustering algorithm [3] on the low-level visual features images. This approach removes our duplicate and near duplicate samples and provides a variety of positive images for multi-query-by-example.

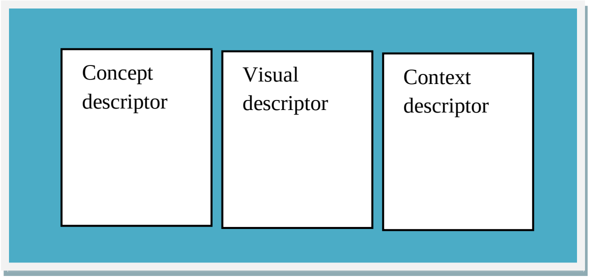

Figure 2: Topic Knowledge Base

Context descriptor

Current context provides a context related expansion of the concept feature component. This component uses a current news repository [4] to select news topics relevant to the concept features. The context features are selected using a k-nearest neighbor approach to each of the concept features, in a term vector feature space.

The three components of the topic knowledge base provide a detailed multimedia description of the original topic query. These components are 'experts' in the multimedia feature space of the query and provide separate ranked results for each component and each context of a component.

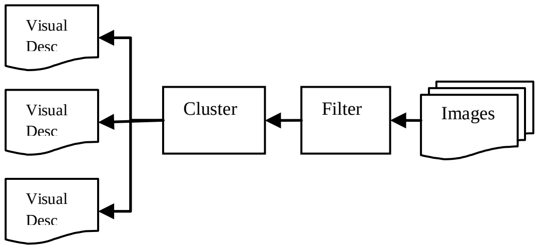

Figure 3: Visual descriptor component

Fusion

The KB system follows a late fusion model [5] where ranked result lists from each of the topic knowledge base components are combined to create a final ranked result. Weighting coefficients are applied first at the component member level and finally at the overall component.

3. Experiments

ASR-MT required run

The automatic search task requires a baseline run using only the ASR/MT [6] output provided by NIST and the search topic text. We used a document retrieval approach for run KBVR_1, where the MT corresponding to each shot is used to build our 'shot document'. To handle misalignment or overlap of ASR and shots, we use a sliding window on the MT. Our sliding-window-3 approach includes the shot before and the shot after when constructing the term vector for each shot. Term weighting uses the term frequency – inverse document frequency and topic matching uses the cosine similarity. The inferred average precision of this baseline run was .00252 and topic query 245, "Find shots of a person watching a television screen - no keyboard visible", had our best score of .016.

Topic Knowledge Base run

Our topic knowledge base runs showed promising improvements over the baseline run. Run KBVR_5 was our top performing run and was based on a topic knowledge base consisting of 5 visual descriptors, 3 context descriptors, and 5 concept descriptors. The visual descriptors were created using an initial set of 400 example images per shot topic, selected from an online image repository [7]. The images were filtered and clustered to generate our 5 visual component members. The external knowledge base for our concept descriptors was based on an XML extract of the full Wikipedia repository [8]. The extract was preprocessed to create a set of 'concept' documents, for k-nearest neighbor search using the topic queries. The five concept members were used as expanded topic queries to an online news repository, to create the context descriptors. An example is shown in Table 1 of context members returned for topic 252, "one or more people, each riding a bicycle".

KBVR_5 was our top performing run and obtained an inferred average precision of .0037. Our top performing query in this run was topic 226, "Find shots of one or more people with mostly trees and plants in the background; no road or building visible", which obtained an inferred average precision of .059.

Table 1: Topic 252 "one or more people, each riding a bicycle"

Title	Description
Bicycle-Sharing Project a Big Hit at Democratic	bike-sharing program to each of the just-completed political conventions. Leading U.S. bicycle ... Twenty-one percent of those riding at Humana ... Bikes Belong is the U.S. bicycle industry organization dedicated to putting more people ...
Beating the pump: Bicycle commuters see benefits for themselves and ...	one of a small but growing number of ... After a while, he began using his bicycle for more and more ... St. downtown, a commute of about five miles each way. She likes the recumbent bicycle because she doesn't have to wear padded riding ...

4. Conclusions

This was our second year of participation in the TRECVID evaluation [9] and our emphasis this year was on the construction of topic knowledge bases and their use in an MQBE model. Our retrieval system has matured over the 2007 evaluation by incorporating additional low-level features and implementing a MQBE component.

We plan to continue to expand our retrieval system for the 2009 Evaluation by incorporating additional low-level images features and a concept-based search component. Our research will focus on enhancing our topic knowledge base and implementing a dynamic fusion method to merge the output from each of our KB components.

5. References

[1]. *Evaluation campaigns and TRECVid.* **Smeaton, Alan F, Over, Paul and Kraaij, Wessel.** New York, NY, USA : ACM, 2006. MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval. pp. 321-330.

[2]. *Efficient use of local edge histogram descriptor.* **Park, Dong Kwon, Jeon, Yoon Seok and Won, Chee Sun.** New York, NY, USA : ACM, 2000. MULTIMEDIA '00: Proceedings of the 2000 ACM workshops on Multimedia. pp. 51-54.

[3]. *k-means++: the advantages of careful seeding.* **Arthur, David and Vassilvitskii, Sergei.** Philadelphia, PA, USA : Society for Industrial and Applied Mathematics, 2007. SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. pp. 1027-1035.

[4]. Microsoft. *Live news.* [Online] 2008. http://news.live.com.

[5]. *Early versus late fusion in semantic video analysis.* **Snoek, Cees G, Worring, Marcel and Smeulders, Arnold W.** New York, NY, USA : ACM, 2005. MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia. pp. 399-402.

[6]. *Annotation of Heterogeneous Multimedia Content Using Automatic Speech Recognition.* **Huijbregts, Marijn, Ordelman, Roeland and Jong, Franciska de.** [ed.] Bianca Falcidieno, et al. s.l. : Springer, 2007. SAMT. Vol. 4816, pp. 78-90.

[7]. Google . *Google Image Search.* [Online] 2008. http://images.google.com.

[8]. Wikimedia Foundation. *Wikipedia.* [Online] 2008. http://download.wikimedia.org.

[9]. *Etter Solutions Research Group .* **Etter, David.** Gaithersburg MD : Proceedings of TREC Video Retrieval Evaluation, 2007.