International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Semantic Multi-modality Fusion for Video Search

Prajakta Bhosale*[1], Nita Patil[2], S. D. Sawarkar[3]
*[1]Computer Department, Mumbai University/Datta Meghe college of Engineering, Airoli, Maharashtra, India
[2]Computer Department, Mumbai University/Datta Meghe college of Engineering, Airoli, Maharashtra, India
[3]Computer Department, Mumbai University/Datta Meghe college of Engineering, Airoli, Maharashtra, India

ABSTRACT

We collect information from different sources in different forms. Multimodality fusion can be solution for various information retrieval problems. In this paper, we propose multimodality fusion approach for video search, where search modalities are derived from a set of knowledge sources, such as text, images and videos. We break down the query modality relationship into two components that are much easier to calculate: the relationship between the query and the concept and the relevance of the concept of modality. The first can be estimated effectively online using visual mapping techniques, while the seconds can be calculated off-line based on the concept-detection precision of each modality.

Keywords: multimodality, relevance score, concept detection.

I. INTRODUCTION

A CHALLENGE of video research is the prediction of the user's search intent. The intention is often expressed using a short textual description with several words, and / or some visual examples of images and videos. A successful search system should therefore adaptively formulate a search strategy in multimodal forms, and ultimately return a set of relevant video clips. Commonly used modalities include text search, visual search and concept search. Text search attempts to match text query words to video transcripts, while visual search measures the similarity between visual query examples and target videos. In concept-based research, a large number of semantic visual concept classifiers are constructed to index video content, and an efficient search is allowed by matching textual and visual queries to semantic concepts [1]. In this multimodality search scenario, a key element of strategy planning is the dynamic assignment of merge weights to different search terms based on a query.

In this paper, we propose a new adaptive merging strategy to the query, mapping a multi-modality query to the large number of semantic concepts instead of a query class, and exploit the selected concepts to determine the merge weights. In other words, the problem of fusion is decomposed into two major stages: reasoning of the relevance of the concept-query and the relevance of the request learning by the selected concepts. Figure 1 illustrates the flow of our proposed approach. Given a query that contains a short textual description and some visual examples, the concept-to-concept mapping is first performed to derive all of the semantically and visually relevant semantic concepts.

After concept selection, in the second step, concept relevance is converted to fusion weight. The resulting fusion weights show the association between concepts and modalities, and can be easily learned from the concept detection accuracy of each modality. By considering the association between the query concept and the concept modality, the

concepts, collectively as a bridge to the query modality, are exploited to derive merge weights.

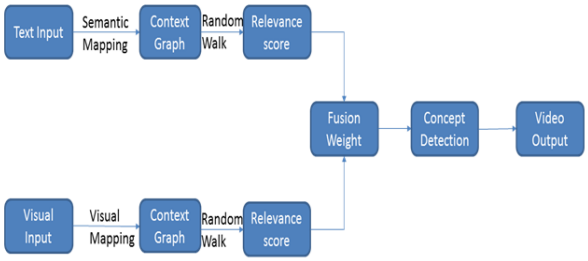

Figure 1. Multi-modality video Search Model

II. METHODS AND MATERIAL

As we have two models: 1) Text model and 2) video or image model. Methods used are for both models are as follows:

A) **Text Search:**

For text search, we use semantic mapping. Semantic mapping aims to find a set of concepts from dictionary that have the highest linguistic relatedness to the text queries. Relevance scores for text model calculated by performing random walk over context graph. Context graph are build offline for each concept.

B) Video Search:

In Second model, we have two types of input 1) Image 2) videos. To calculate relevance score, we use feature extraction. These features include static features in key frames, object features, motion features, etc.

Videos are arranged according to a descending hierarchy of video clips, scenes, shots, and frames. Video structure analysis divides videos into shot boundary detection, key frame extraction, and Feature extraction.

A shot is a consecutive sequence of frames captured by a camera action that takes place between start and stop operations, which mark the shot boundaries. In shot boundary detection, we first extract visual features from each frame, then measure similarities between frames using the extracted features, and, finally, detect shot boundaries between frames that are dissimilar.

Key frame is the frame which can represent the salient content and information of the shot. We use sequential Comparison between Frames. In these algorithms, frames subsequent to a previously extracted key frame are sequentially compared with the key frame until a frame, which is very different from the key frame, is obtained.
The static key frame features classified as color-based, texture-based, and shape-based.

Color-Based Features: Color-based features include color histograms, color moments, color correlograms, a mixture of Gaussian models, etc. Color features can be extracted from the entire image or from image blocks into which the entire image is partitioned.

Texture-Based Features: Texture features in common use include Tamura features, simultaneous autoregressive models, orientation features, wavelet transformation-based texture features, co-occurrence matrices, etc.

3) Shape-Based Features: Shape-based features that describe object shapes in the image can be extracted from object contours or regions. A common approach is to detect edges in images and then describe the distribution of the edges using a histogram.

Concept detection is done using multimodal fusion technique. We add relevance score of both model and find fusion weight. We select video with maximum weight. For group training, we are using SVM classifier for all videos.

We considered positive and negative examples for our database. Feature extraction of negative example is done on the go. SVM is trained each time you search.

III. RESULTS AND DISCUSSION

In this section, we present the various experimental results for the proposed on the datasets. We have tested proposed system on 14 videos (.mpg) and 109 key frames (.jpg). For text search, 82 words are used in dictionary.

For the retrieval purpose, any word from dictionary along with related image from database is used. As a result of text search, it gives relevance score of search based on semantic mapping.

We use the term `automatic concept selection' to describe the concept selection algorithms used in video retrieval systems to automatically translate a query to the system concept lexicon, usually returning a weighted list of concepts as a result.

For Example, Text input is 'Nature' and visual query image. It is positive example.

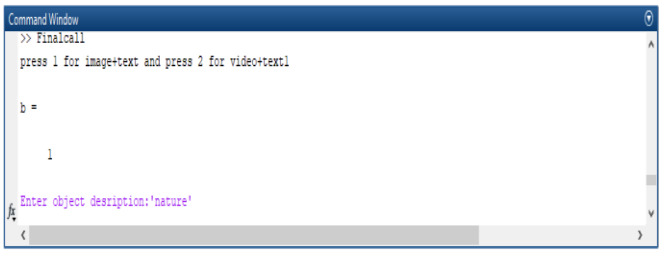

Figure 2. Text query

The result of text query is relevance score of input word as per dictionary. Here we do not consider negative input. If it is so, we stop search to provide 100% accuracy.

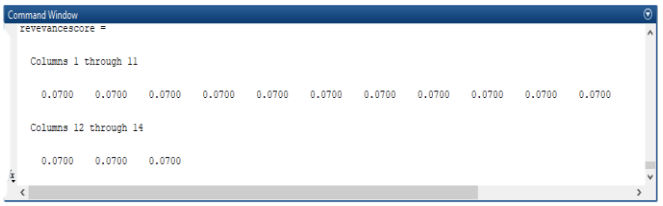

Figure 3. Relevance score for text search

Now, the second input is image as a result we directly get video.

Figure 4. Query image

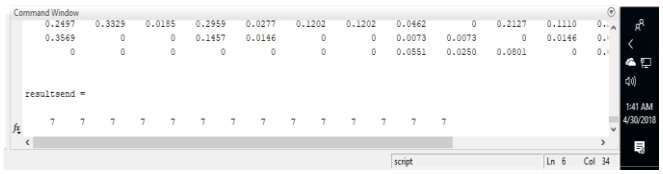

Figure 5. Final Output Video Number

Figure 6 Final Output at video Player

Our video retrieval algorithm gives 100% accuracy when the image searched is within the database and it can reach up to maximum accuracy when the image search is negative i.e. out of database search. We have also proven that the time required for the retrieval is low.

IV. CONCLUSION

We presented an approach that dynamically uses several modalities for video search, where the modality weights are calculated based on a relevance score. We have shown that modality weights can be accurately calculated for each on-the-fly query. Experimental results have suggested that semantic concepts not only can be used in the concept-based search modality, but could also be explored to determine the weights of the search terms. With our proposed approach, more appropriate query adaptive weights can be calculated without the need for

additional training requests as in many existing methods. In addition, our unique modality search experience revealed another perspective of using context information (conceptual relation), where a random walk process is imposed on a context graph to produce a better conceptual selection. We have shown that this process is useful for removing irrelevant concepts, while bringing more relevant concepts at the same time.

V. REFERENCES

[1]. Wei, X. Y., Jiang, Y. G., & Ngo, C. W. (2011). Concept-driven multi-modality fusion for video search. Circuits and Systems for Video Technology, IEEE Transactions on, 21(1), 62-73.

[2]. Hu, W., Xie, N., Li, L., Zeng, X. Maybank, (2011). A survey on visual content-based video indexing and retrieval. Systems, Man, and Cybernetics, Part C: Applications and Reviews,IEEE Transactions on, 41(6), 797-819.

[3]. M. Naphade, J. R. Smith, J. Tesi c, S.-F.Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for mul- timedia," IEEE Trans. Multimedia, vol. 13, no. 3, pp. 86–91, Jul.–Sep. 2006.

[4]. Kennedy, L.,Chang, S. F., &Natsev, A. (2008). Query-adaptive fusion for multimodal search. Proceedings of the IEEE, 96(4), 567-588.