IIPWHU@TRECVID 2015

Bin Xu[1], Weihang Liao[1], Zizheng Liu[1], Wentao Bao[1], Yiming Li[1], Daiqin Yang[1], Sihan Wang[2], Hongyi Liu[2], Yatong Xia[2], Yingbin Wang[2], Zhenzheng Chen[1, 2]

[1] Surveillance Event Detection (SED) / [2] Video Hyperlinking (LNK)

School of Remote Sensing and Information Engineering, Wuhan University

Wuhan, China 430079

zzchen@whu.edu.cn

In the first part of this two-part report we describe our system used in the TRECVID 2015 Surveillance Event Detection (SED) task. Details and results of Video Hyperlinking (LNK) task are described in the second part.

IIPWHU@TRECVID 2015

Surveillance Event Detection

Bin Xu, Weihang Liao, Zizheng Liu, Wentao Bao, Yiming Li, Daiqin Yang, Zhenzheng Chen*

School of Remote Sensing and Information Engineering, Wuhan University

IIPWHU@TRECVID 2015

Video Hyperlinking

Sihan Wang, Hongyi Liu, Yatong Xia, Yingbin Wang, Zhenzhong Chen*

School of Remote Sensing and Information Engineering, Wuhan University

IIPWHU@TRECVID 2015
Surveillance Event Detection

Bin Xu, Weihang Liao, Zizheng Liu, Wentao Bao, Yiming Li, Daiqin Yang, Zhenzheng Chen*

School of Remote Sensing and Information Engineering, Wuhan University

Wuhan, China 430079

zzchen@whu.edu.cn

Abstract

In this paper, we present a system based on convolutional neural network dealing with Surveillance Event Detection (SED) task in TRECVID 2015. We pay attention to the events of PersonRuns, PeopleMeet and PeopleSplitUp. In the proposed system, surveillance videos are decomposed to frame images where optical flow are applied to extract features for further processing. Optical flow images with annotations are sent to CNN to train models for classifying later. In the whole evaluation video, whether a shot contains required events or not depends on the frames classified through CNN model. The training data we used is part of Gatwick development data and we conduct the system on the 9 hour subset of the multi-camera airport surveillance domain evaluation data and a new Group Dynamic Subset using only 2 hours of this video.

1. Introduction

The wide availability of low-cost sensors and processors has greatly promoted deployments of video surveillance systems [1]. Intelligent video surveillance applications such as automatic event detection have a significant impact on home security and public security. In the last few decades, most research of human action recognition mainly experiments on controlled environment with clear background where explicit actions are performed with limited actors. However, in real-world surveillance videos, due to challenges of large variances of viewpoint, scaling, lighting, cluttered background, it is almost impossible for us to have the ideal situation. Under the circumstances, the TRECVID [2] surveillance event detection (SED) task is provided to evaluate event detection in real-world surveillance settings. In TRECVID 2015 [3], the test data is the same data that was made available to participants for previous SED evaluations, which is about 100-hour surveillance videos under five camera views from the London Gatwick International Airport with annotations of event labels. Participates' systems should output detection results for any three events in the following list: PersonRuns, CellToEar, ObjectPut, PeopleMeet, PeopleSplitUp,

Embrace, and Pointing. The main evaluation will be implemented using a 9 hour subset of the multi-camera airport surveillance domain evaluation data and a new Group Dynamic Subset using only 2 hours of this video and limited to the Embrace, PeopleMeet and PeopleSplitUp events is introduced this year.

We designed a system based on optical flow [4] and convolutional neural network dealing with surveillance event task in TRECVID 2015. We pay attention to the events of PersonRuns, PeopleMeet and PeopleSplitUp. The remainder of this paper is organized as follows. Section 2 introduce the overall retrospective system architecture, which contains preprocessing the original data, extracting feature form frame images and how the CNN works. The result of our approach performed on the TRECVID SED task is given in Section 3, and we conclude this paper in Section 4.

2. Retrospective System

2.1 Optical Flow Feature Extraction

The optical flow vector can describe the motional information of a move object, so optical flow is used as a feature. Firstly the video is decomposed into frames, and for every frame, the optical flow vector is computed. Considering the processing speed, we set a skip flag to skip most pixel. In a 9x9 neighborhood, only the mid-pixel will be considered. So for an input 720x576 frame, an 80x64 output optical flow map is generated. This map contains the motional information in this frame, and will be used as a feature in the next CNN processing.

2.2 Convolutional Neural Network

The optical flow images of parts of TRECVID 2008 dataset with available annotations will be used as training data for the convolutional neural network. We use the CNN architecture from BLVC AlexNet in Caffe [5], which is a replication of the model described in [6]. The numbers of CNN output is modified to four, corresponding to PersonRuns, PeopleMeet, PeopleSplitUp and none required events. The CNN architecture is shown in Fig. 1. The optical flow images of evaluation dataset are then sent to CNN to classify. The results of this step are the class of each frame image with corresponding possibility value.

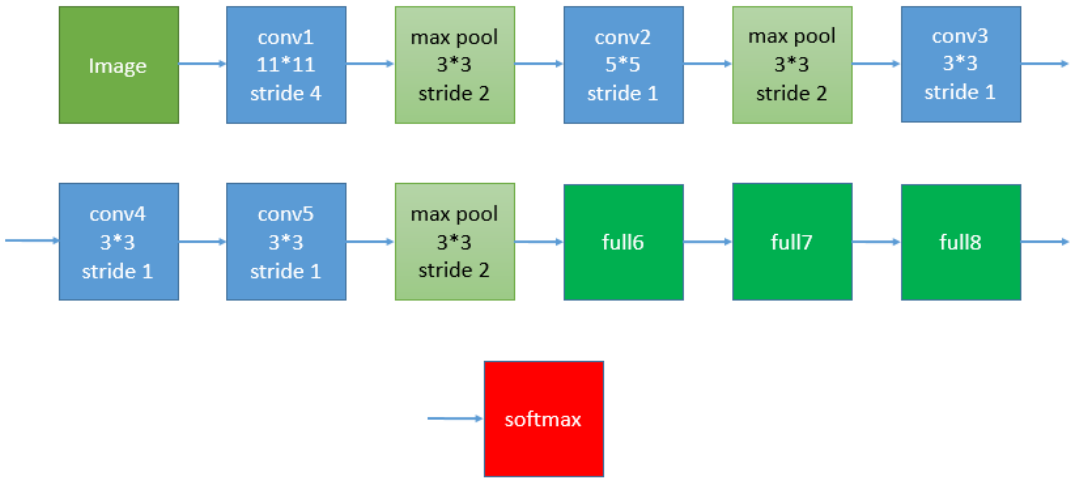

Fig. 1 The architecture of BLVC AlexNet

Each surveillance video will be segmented to shots of three seconds. In each three-second-long video, classes with possibilities of frame images are known, the mode of these images' classes will be considered as the class of the shot, with maximum possibility value being the possibility of the shot.

Thresholds of possibility value will be set to define whether the shots are considered as the corresponding event. Then the adjacent shots belonging to the same event will be combined as a consecutive event. The framework of our system can be seen in Fig. 2.

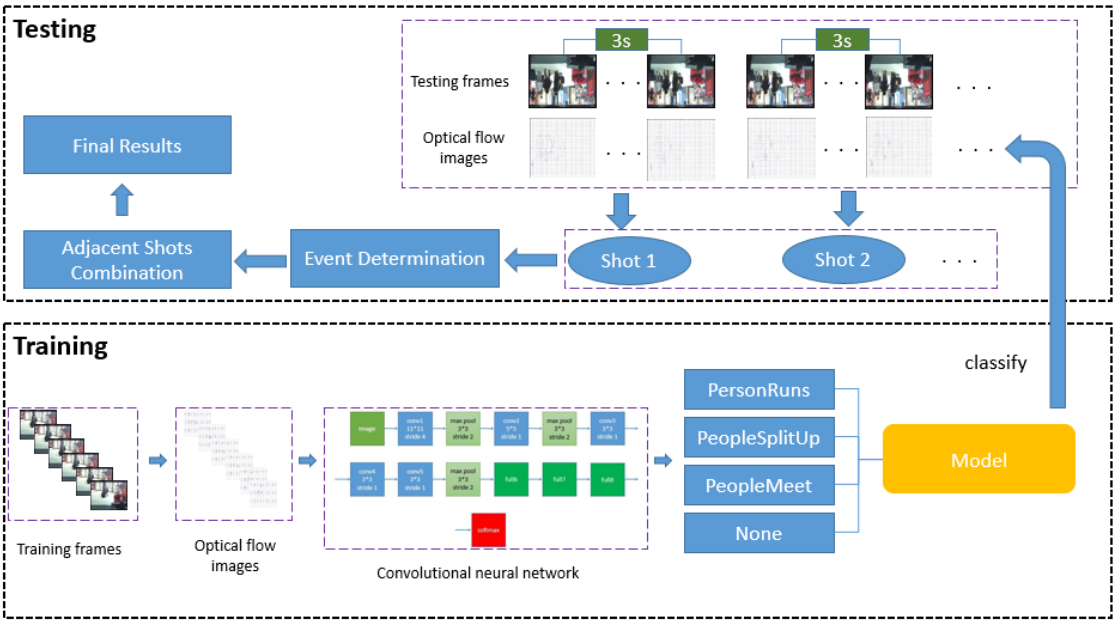

Fig. 2 The framework of system

3. Experiments

We applied our approach to the SED 2015 retrospective task. The CNN model of the system is trained on a DELL R720 server that comprises 2 Intel Xeon E5 CPUs, 64GB memory, and 24TB hard disk. This year, the main evaluation (EVAL15) will be implemented using a 9 hour subset of the multi-camera airport surveillance domain evaluation data collected by the Home Office Scientific Development Branch (HOSDB). A new Group Dynamic Subset (SUB15) using only 2 hours of this video and limited to the Embrace, PeopleMeet and PeopleSplitUp events is introduced in 2015. Table 1 and table 2 shows our main evaluation and subset results provided by NIST. Our system only contains PersonRuns, PeopleMeet and PeopleSplitUp three events.

Table 1: The actual DCR and minimum DCR of the EVAL15 result

Event	#CorDet	#FA	#Miss	ActDCR	MinDCR
PeopleMeet	36	294	220	1.0283	0.9928
PeopleSplitUp	13	144	139	0.9972	0.9884
PersonRuns	2	199	48	1.0744	1.0048

Table 2: The actual DCR and minimum DCR of the SUB15 result

Event	#CorDet	#FA	#Miss	ActDCR	MinDCR
PeopleMeet	13	81	102	1.0602	1.0060
PeopleSplitUp	11	33	86	0.9572	0.9551

4. Conclusion

In this paper we have presented the detailed implementation of our system participated in TRECVID 2015. The system is applied to event PersonRuns, PeopleMeet, PeopleSplitUp in EVAL 15 and PeopleMeet, PeopleSplitUp in SUB15. Optical flow images are generated from original dataset and used as a feature for classification. CNN models are trained to deal with the evaluation videos. The current result is not good enough and more works need to be done to improve the performance in the future.

References

[1] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognition Letters*, vol. 34, pp. 3-19, 2013.

[2] A. F. Smeaton, P. Over and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, Santa Barbara, California, USA, 2006, pp. 321-330.

[3] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton and G. Quenot, “TRECVID 2015 -- An Overview of the Goals, Tasks, Data, Evaluation

Mechanisms and Metrics," in *Proceedings of TRECVID 2015*. NIST, USA, 2015.

[4] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, pp.185 -203, 1981

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[6] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.

IIPWHU@TRECVID 2015
Video Hyperlinking

Sihan Wang, Hongyi Liu, Yatong Xia, Yingbin Wang, Zhenzhong Chen*

School of Remote Sensing and Information Engineering, Wuhan University

Wuhan, China 430079

zzchen@whu.edu.cn

Abstract

In this paper, we present a system based on Multimodal Information Fusion dealing with Video Hyperlinking (LNK) task in TRECVID 2015. We pay attention to the visual similarity, textual similarity and how the multimodal information fusion can affect the last score. In the proposed system, the textual information of videos is decomposed to segments and index is built based on these segments. The visual information of videos is decomposed to concepts, and WordNet is used to get the similarity between different concepts. After getting the textual similarity and visual similarity we use mutimodal information fusion method to calculate the last score of each result.

1. Introduction

Visual similarity, audio similarity and textual similarity are the three aspects which describe the similarity between among different videos. In our work, we concentrate on the visual similarity and textual similarity for the HyperLinking task. For the visual similarity, we employ WordNet similarity to map visual concepts with query terms. In order to compute the textual similarity, the original documents are firstly represented by vectors using bag-of-words techniques. Each document is represented by one vector where each vector elements represents the times of the words that appears in the document. After converting documents into vectors, a transformation applied for computing meaningful documents or documents similarities. In order to achieve this purpose, we use a tool named 'Lucene' [1].

We design a system for hyperlinking based on the similarities between videos and multimodal information fusion, as shown in Fig. 1. First of all, MataData is added to the database, then 'Lucene' is used to do the stemming and stopword removal. Secondly, we build index of each database. And then, before searching the index, each anchor should do the steming and stopword removal. Next, WordNet is used to calculate similarity between different concepts. Finally we combine the concept score with the Lucene score to get the final result.

Reminder of this paper is organized as follows. Section 2 introduces the different

aspects of similarities we have used and methods of fusion can be found in Section 3. The experiment results are given in Section 4, and we conclude this paper in Section 5.

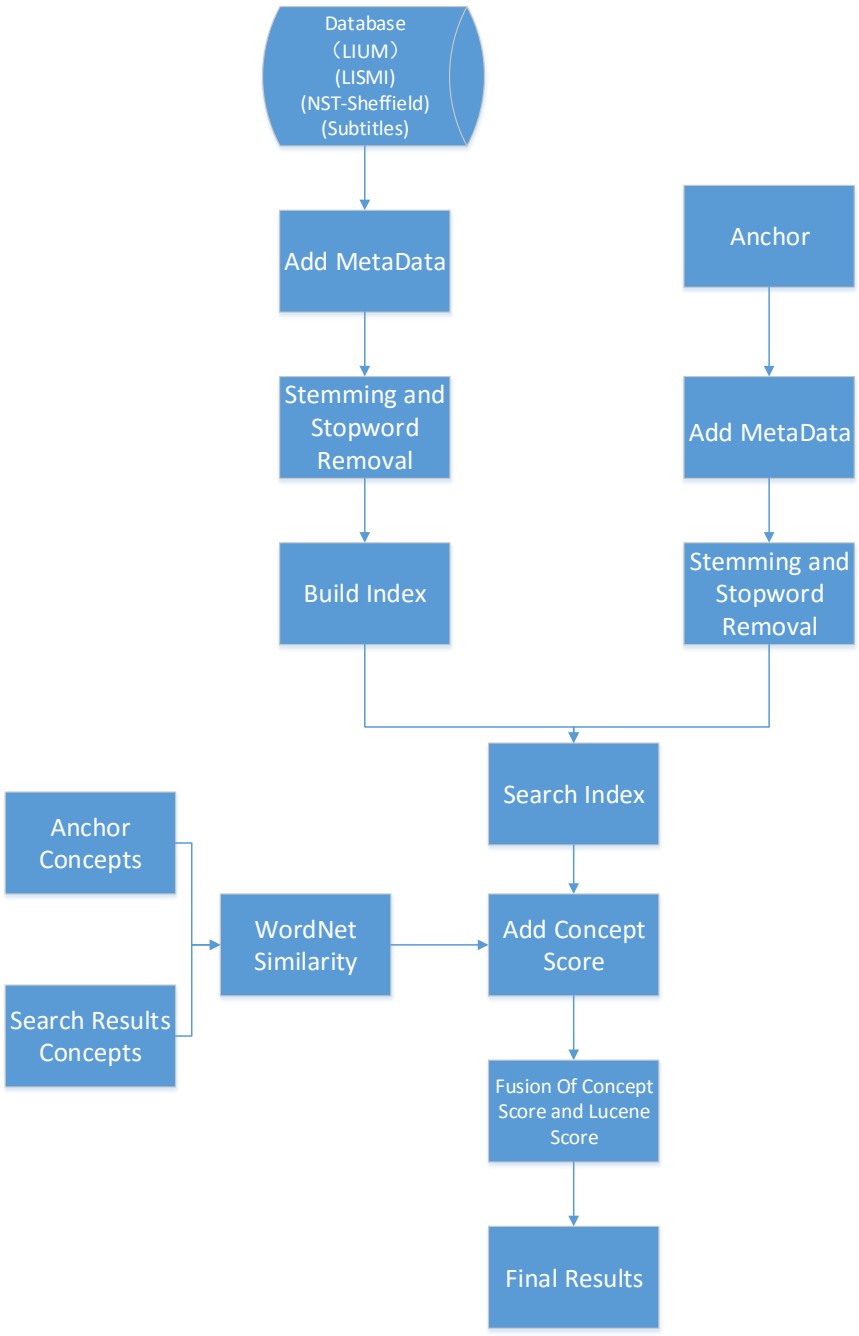

Fig. 1 The framework of system

2. Video Similarities

2.1 Visual Similarity

WordNet is a large lexial database of English, Nouns, verns, adjectives and adverbs are grouped into sets of cognitive synsets, each expressing a distinct concept. In our research, WordNet similarity is employed to map visual concepts with query terms.

2.2 Textual Similarity

2.2.1 TF-IDF

TF-IDF [2] (term frequency-inverse document frequency) is a weighting factor for information retrieval and text mining. TF-IDF is a numerical statistical method for evaluating the importance of a word for a certain document or corpus. TF-IDF value increases proportionally to the number of times which a word appears in the document, but decreases to the frequency of a word in the corpus. Our team use Lucene to achieve this method.

Apache Lucene is a high–performance, full-featured text search engine library written in Java. It is a technology suitable for nearly any application that requires full-text search. Thus we use Java programming language to build our index. The processing is shown in Fig. 2.

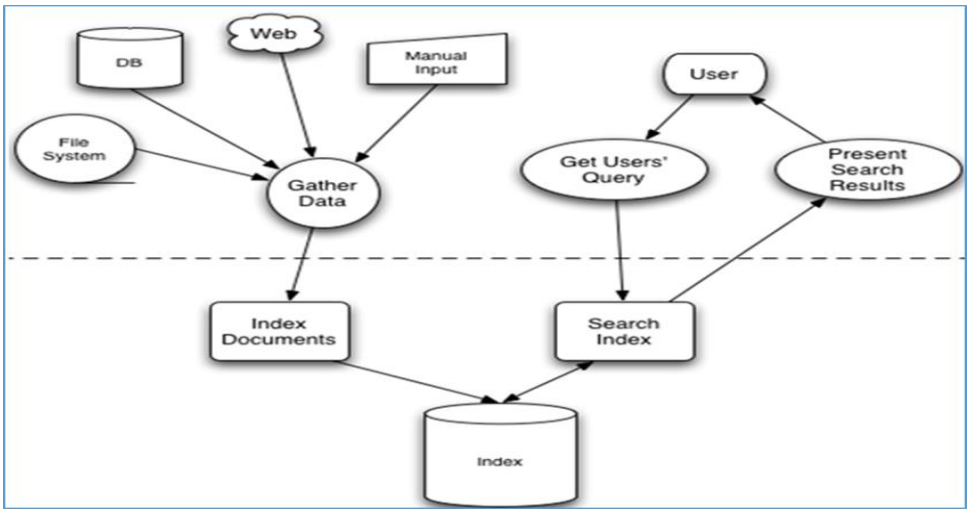

Fig. 2 The processing of Apache Lucene

2.2.2 Stopword Removal

Stop words are words which are filtered out before or after processing of natural language data. Stop words usually refer to the most common words in a language such as the, is, at, which, and so on. These words are very common in a document, so they are useless to the index and search. Before we build our index, we use a stopword list to filter out all the stop words. We use Lucene and define a stop word list to filter out the stop words.

2.2.3 Build and Search the Index

Index database is built based on segments. A video contains many segments, and a segment has it's own start time and end time. As no anchor item segment is defined, we get the segments of each anchor by comparing the start time and the end time. If the intersection is found between an anchor and a segment, then the segment can be used to search the index. However, an anchor usually contains more than one segment. By combining different segments from the same anchor into one segment, we discovered that the score have a modest increase. Besides, the duration of an anchor sometimes is very short, but a longer segment including context can improve the results significantly [3]. Thus we created segments by adding some time before and after anchor. In this way, an anchor contains more context information, and we find this method is especially helpful when the duration of an anchor is very short.

2.2.4 Threshold Adjustment of Lucene

Lucene have lots of thresholds which may have a significant influence to the final result. Through the use of various threshold, different results were returned by lucene and we analysed how the change of threshold can affect the map score. A simple example is shown in Fig. 3. Before searching the index, we set different threshold for the results, if the Lucene score of a result is lower than the threshold, the result is abandoned. Through Fig. 3, we can see that when the threshold increases, the Map score have a decrease tendency.

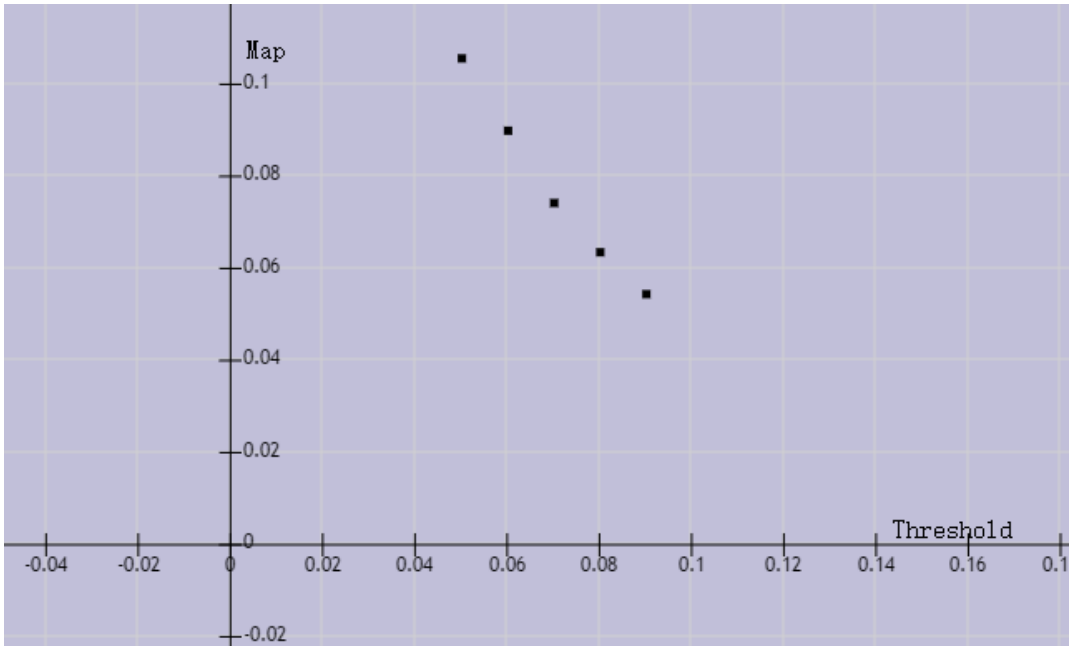

Fig. 3 The relation between threshold and map score

3. Multimodal Information Fusion

3.1 Linear Fusion

The basic fusion strategy is to linearly combine the multi-modality retrieval results together, and weight each factor according to the a priori knowledge. The weights represent the importance of each factor, and the weights may vary according to different situations and query goals, examples are provided in table 1 [4].

Table 1: Different weights of linear fusion

Video Category	w_V	w_A	w_T
Common Video	0.5	0.3	0.2
Audio-Rich Video	0.3	0.5	0.2
Text-Rich Video	0.2	0.2	0.6

3.2 Lucene MultiField Fusion

In our research, we use information from different sources to get the final textual score. For example, our index includes three fields, title, description and content [5]. Before search the index, we set the weight according to the amount of information they contained and the information importance. We try different weights in order to get the best score, as shown in table 2. We try different weights and also get different scores, finally we find that 0.1(title), 0.2(description), 0.7(content) have the highest score.

Table 2: different weights of MultiField fusion

Title	**Dscription**	**Content**
0.1	0.1	0.8
0.1	0.2	0.7
0.1	0.3	0.6
0.2	0.2	0.6
0.2	0.3	0.5
0.2	0.4	0.4

3.3 Visual and Textual Fusion

At last, we get the visual score by using wordnet and the concept information, and we also get textual score by using lucene and textual information. For the final fusion,we can use linear fusion and no-linear fusion method. We try these two methods and finally choose the linear fusion method. The weights we set for the visual score and the textual score are 0.2, 0.8.

4. Experiments

We perform experiments on the test dataset, as shown in table 3.

Table 3: Results of our experiment

num_q	all	29
videos_ret	all	39
videos_rel	all	21
avglength_ret	all	28
avglength_rel	all	83
map	all	0.43
P_5	all	0.4138
P_10	all	0.3724
P_20	all	0.3517

We submitted three results and their score are shown in table 4.

Table 4: Final score of our submissions

database	map	P_5	P_10	P_20
LIMSI	0.0995	0.5048	0.3524	0.2087
NSTSheffield	0.0864	0.4115	0.2865	0.1654
LIUM	0.0851	0.4679	0.3143	0.1812

5. Conclusion

In our system, we focus on the visual similarity, textual similarity and how the multimodal information fusion can affect the last score. The framework of our system is shown in the introduction section. In the video similarities section, we calculated the visual similarity and textual similarity seperately. WordNet is used to calculate the textual similarity and Lucene is used to build and search the index for the textual similarity. In the section 3, we use multimodal information fusion method to get the final score. And from the experiments section, we can get the following conclusions:

Multimodal information fusion plays an important role in determining the final score. Stopword removal and segments combination can increase the final score to a certain extend and threshold adjustment of Lucene may also influence the performance of our system.

References

[1] "Apache Lucene," http://lucene.apache.org/.

[2] T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization," in *Proceedings of the Fourteenth International Conference on Machine Learning*, 1997, pp. 143-151.

[3] W. Bailer and H. Stiegler, "JRS at Search and Hyperlinking of Television Content Task," 2014.

[4] M. Campbell, A. Haubold, M. Liu, A. Natsev, J. R. Smith, J. Tesic, L. Xie, R. Yan, and J. Yang, "IBM Research TRECVID-2007 Video Retrieval System," in *TRECVID*, 2007.

[5] X. Wu, C. W. Ngo, and W. L. Zhao, "Data-Driven Approaches to Community-Contributed Video Applications," *IEEE Multimedia*, vol. 17, no. 4, pp. 58-69, 2010.