TWO KINDS OF TIMING CUES AND THEIR USAGE IN CONCEPT DETECTION IN NEWS VIDEO

Dong Wang, Dayong Ding, Le Chen, Shen Zhang, Fuzong Lin and Bo Zhang

State Key Laboratory of Intelligent Tech. & Sys.
Department of Computer Science and Technology
Tsinghua University, Beijing, P. R. China. 100084

ABSTRACT

Two open questions remain unanswered in the Content Based Video Retrieval area. Firstly, how to find out the useful information and express them with different features? Secondly, how to fuse the heterogeneous information together to boost the retrieval performance beyond any single component? This paper presents two kinds of timing information and their usage in concept detection in news video, and a novel non-linear information fusion method to combine timing cues with other information from different sources. Experiments on TRECVID 2004 dataset show that timing cues can boost the performance when combined with other information.

1. INTRODUCTION

The multimedia data, especially the videos, is ever growing and the need to access these data with both efficiency and accuracy is more demanding and challenging than ever before. This process is known as Content Based Video Retrieval (CBVR) or sometimes as Content based Multimedia Retrieval, Compared with text retrieval that has proved to be a great success in recent years, the semantic gap between user query and search space in CBVR is greater and the user queries are more limited. Past research experience in CBIR (Content based Image Retrieval) has already shown that user queries concerning high level concepts are difficult to describe using just static low level features of color, texture and shape. In the boomed CBVR area, researchers are facing the same semantic gap as in CBIR. But fortunately, CBVR has much more information from multiple modalities and various information fusion methods to overcome the gap. The commonly used information comes from visual, aural, and textual modalities. And different kinds of features of visual aural motion text are extracted and applied [2]-[4].

But two open questions remain unanswered here. Firstly, how to find more useful hidden information and extract them as different features? Secondly, how to fuse the heterogeneous information altogether to boost the retrieval performance beyond any single component? This paper tends to discuss the two different kinds of timing information and their usage in news video concept detection. A novel non-linear fusion method is also presented to combine the timing cues with other information from different sources. Concept detection aims at finding the frequently occurred semantic concepts in video databases, e.g. "Bill Clinton", "Beach", and "Basket scored". The experiments are carried out on the Feature Extraction (synonym of concept detection) task of NIST TRECVID 2004 benchmark [1].

The two kinds of timing information we explored are:

1) Shot position. Will the shots containing the target concept occur at the same position across videos?
2) Shot clustering. Will the shots containing the target concept be temporal adjacent to each other in video?

Shot position and shot clustering have been successfully used in shot based video summarization, browsing and searching in several systems. Shot position has been used in collages for video summaries [8]. [2] shows the shots adjacent to the retrieved shot by a shot expansion option; [4] designed a 3-minute multi-document storyboard to show temporally near shots where query-words had been matched; [6] enables keyframe based temporal adjacent browsing by simply pushing a button, etc. And [5] has successfully used shot position plus text in TRECVID 2003. But when text and timing information are fused together, the performance lies somewhere in between.

However, we find that when combined with other kinds of information, timing cues can improve the performance. Following TRECVID, AP (average precision), which is defined in [7], is adopted as the main performance measure.

The paper is organized as follows. The two kinds of timing information and their usage are first introduced (Section 2). In Section 3, we discuss the information from other modalities and propose a kind of new combination method for information fusion. The experimental results are given in Section 4 and conclusion in Section 5.

2. TIME STRUCTURE ANALYSIS IN NEWS VIDEO

Compared with sports video, news video usually has more diverse content but with a relatively fixed time structure. Timing cues are strong implicit temporal structure of the news video. Typical news video may start with politic news, followed by social and weather news, and end with sports news. Based on this observation, we apply time structure analysis to news video to extract information for detecting concepts. In the following two subsections, we illustrate in details how we obtain the different kinds of timing information and their usage.

2.1 shot position

In news video, specific type of news such as sports will occur at specific positions in the video. The information can be described

by using p.d.f. (probability density function) and then be used to generate a ranked shot list (rank list). The density function can be estimated using Parzen window, which is a kind of non-parametric distribution estimation method. We choose Gaussian kernel as [5] but without bandwidth selection procedure for modeling simplicity. Instead, the bandwidth is obtained by simply averaging length of every shot containing the concept together. The estimated p.d.f. is then smoothed to eliminate zero values. Fig. 1 is the estimated p.d.f. for "Basket scored" of CNN.

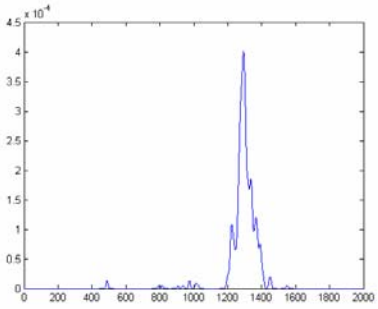

Fig. 1 Estimated p.d.f. for concept "Basket scored". The refined annotation of videos from CNN is used. The X-axis is the shot position in seconds and the Y-axis is the estimated density.

To model the shot position and length as independent event, each shot is assigned a pseudo shot length which is equal to the bandwidth used above. Given shot position p, the conditional probability of concept c, $P(c|p)$, is obtained by (1).

$$P(c \mid p) = \int_{p-L/2}^{p+L/2} pdf(x)dx \qquad (1)$$

where L is the pseudo shot length, and *p.d.f.* is the estimated p.d.f.. Then rank list based on shot timing position (TMP list), which is generated by sorting the shots by the estimated probability in descending order, is used to predict the concept occurrence. This rank list may be a weak predictor compared with other semantic information.

2.2 shot clustering

Tasty mushrooms always grow in cluster. So do the shots containing some concept.

Some concept may not have the fixed position in video, but the corresponding shots come in cluster. This is what we called shot clustering phenomenon, as Fig. 2 shows. In Fig. 2, the shots are clearly "growing in cluster". But how can we use this for concept detection?

In this paper, a novel selective shot clustering method is proposed to utilize this kind of information. And this algorithm can effectively reduce the falsely retrieved shots and improve the precision and AP measure.

In this algorithm, shots are treated as points without length and numbered in each video. A baseline rank list is chosen to construct the tree based clusters representation. The top position of the list is ranked as 1, and the second 2, etc. And three levels of nodes in the tree, which are video, shot cluster and shot, describe the list in different level. Each shot cluster k in each video contains at least one shot, and has two parameters of the start and end boundaries, SCS_k and SCE_k, which is initially defined by the cluster diameter D. That means

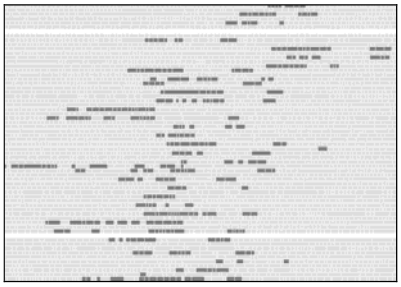

Fig.2 Shot clustering for "Basket scored". The shots are shown in time sequence and in proportion to their actual length. Only a small part of the videos is shown and shots containing the concept are in dark.

$$SCS_k = i - D/2$$
$$SCE_k = i + D/2 \qquad (2)$$

where i is the core shot of the cluster. In the tree construction process, when shot S_{ij} and cluster SC_k of the same video j satisfy $SCS_k \le i \le SCE_k$, shot S_{ij} can be absorbed by cluster k. Otherwise, a new cluster containing the shot is formed or the shot is simply discarded. After that, confidence value is propagated from the shots to clusters and to videos, and the tree is pruned to generate a more compact output list.

Four parameters are used in this algorithm. D is defined above. The other three are Cm, Sm and Im. The shot whose rank is greater than Cm cannot form new shot cluster. The shot whose rank is greater than Sm is simply discarded from the list. And Clusters with only a single shot whose rank is greater than Im will also be discarded. The detailed algorithm is shown in Fig. 3.

However the two types of information may not be directly applicable because inconsistency across the whole dataset. And this is true for the dataset of TRECVID 2004. Half of the whole set comes from CNN, where both the shot position and shot cluster cues are strong; The other half comes from ABC, where only shot cluster cue is preserved. To overcome this problem, we first split the dataset to different subsets which preserve different kinds of timing cues, and then process each subset respectively, and finally fuse the results to get the final result. Another problem is that even when all kinds of the timing cues are strong enough and used, they are not comparable in performance with the image classifiers using texture and colors. There is still space for combining timing cues with other kinds of information. We thus try to solve these two problems and to boost the performance in the next section.

3. INFORMATION FROM OTHER MODALITIES AND COMBINATIONS

The general adopted CBVR modalities such as image and text, together with their combinations with the timing cues are treated in this section.

3.1 visual and textual information representations

The low level image features are extracted both globally and in3×1 grid for keyframes of each shot. These features are then normalized using Gaussian function and selectively combined to achieve better performance. Using a SVM classifier [11], the

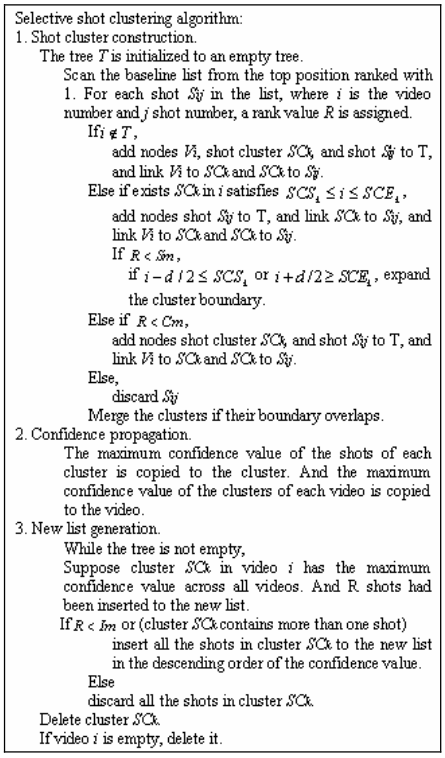

Fig. 3 shot cluster re-rank algorithm

keyframes are assigned a confidence value according to their distance to the classification boundary, and +/- sign are given by classification results. Sorted by these confidence values, the keyframe rank list (IKF list) is generated and further converted to shot rank list by taking the keyframe with maximum confidence value as the representative keyframe for the shots which have multiple keyframes.

The text modality is provided with two methods. The first one is text retrieval method. The text database is built by CC (closed caption) text of each shot which is obtained by aligning CC text with ASR text of each shot. Some hand selected keywords are chosen as query words and processed by the system developed in [8] to generate a text retrieval based rank list (TRE list). The second method is rule based text match. The CC text of each shot is selected and sorted if they can match a specific rule. These rules can not be effectively represented using simple text retrieval techniques with a bag of words. For example, in "Basket scored", when CC text contains two consecutive numbers whose range is between 60 and 120, the shot sometimes talks about a basketball game, and may be about the concept. When CC text mentions basketball teams or stars, it is very likely that the shot talks about it. Each rule can assign a score to the matched shot to measure the matched degree. Thus the second rank list is generated as the text rule match list (TRM list).

3.2 information combination using different methods

Up to now, the different modalities are treated respectively and TMP, IKF, TRE and TRM lists are generated. In this subsection, these lists are combined by different information fusion techniques to achieve better performance. The four different kinds of information fusion methods discussed in this paper are as follows:

1) rank based rank level fusion
2) confidence based measurement level fusion
3) splitting/insertion based rank level fusion
4) clustering based rank level re-rank fusion

When two lists come from different modalities and describe the same dataset, they can be fused using both 1) and 2). The rank values can be converted to "confidence like" rank scores which are determined by the corresponding rank values only. Thus same arithmetic operations can be applied to both 1) and 2). The list scores are first normalized using maximum value normalization. Then each normalized list is assigned a weighting factor in [0, 1] and combined using arithmetic operations such as plus, minus, multiply, minimal and maximum. Also a main list can be specified and only the shots in this list will be taken into account for operations.

When more than two lists come from the same modality and describe the same dataset, a variant of borda count [13] can be applied using 1). The rank lists are all chopped to a fixed length and rank score for each shot is converted to the count of the shots listed below it in a specific list. Then weighting factors are chosen for each list and voting procedure is followed to get the fused lists. No special measure is taken to solve the tie that may occur.

When two list comes from the same modality but describe different datasets, as described in section 2 to deal with the first kind of problem, a rank based list insertion algorithm 3), which is similar to [8], is adopted. The list with higher AP value is selected as a major list. Given the initial position *Pi* and insertion distance *D*, the other list is inserted into this major list starting from the initial position every *D* shots.

When one list is generated and the target concept has the shot clustering property, selective shot clustering algorithm 4) can be applied.

Table 1 summarizes the four types of information fusion methods used in this paper. The optimal fusion sequence of the lists from different modalities varies from concept to concept. This paper does not attempt to solve this general problem and will only give an example in the next section.

properties	1)	2)	3)	4)
level	rank	confidence	rank	rank
operation	arithmetic	arithmetic	insert	re-rank
linearity	linear	linear	non-linear	non-linear
dataset	same	same	different	same

Table 1 Summary of the information fusion methods used

4. EXPERIMENTS AND DISCUSSIONS

The data take up about 130 GB and are divided to development and test data. They are consisted of CNN headline News and ABC World Tonight in 1998. We experimented on one concept "Basket scored" in TRECVID 2004. We refined the annotation [9] which is done on half of the training set for the concept to eliminate inconsistencies, and labeled the other half manually. 50% of the training set is used for uni-modal classifier training, and 25% for multimodal fusion parameter estimation, the left 25% for testing. The final results are obtained on the test set.

As mentioned in section 2, the position cue is only preserved in CNN sub-dataset. Thus TMP_CNN list is obtained on sub-dataset of CNN. And we designed the fusion schema for the different lists of the mentioned modalities as Fig. 4.

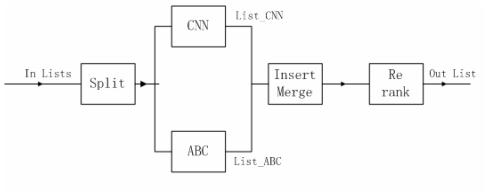

Fig.4 The basic information fusion schema.

In Fig. 4, each list is split by source using the source information from CC text. Then the sub-lists of CNN are fused to incorporate shot position information using 3). Accordingly, the MIKF list, which is generated by 20 different IKF lists using borda count of 1), and the TXT list, which is the combination of TRE and TRM lists using 2), is split to MIKF_CNN and TXT_CNN, respectively. And Table 2 shows the fusion results is encouraging. These two lists are further combined using 2) for the refined CNN sub-list. The merged list, whose generation is omitted here, is the combination of the refined CNN and the ABC sub-list using 3).

As for the final shot cluster step, we apply 4) to the merged list, and set $D=1, Cm=2000, Sm=500, \mathrm{Im}=Sm/2$ using the train set and then combine the two lists using 2). The result improved only 1.7% in Table 2. Our explanation is that the non-orthogonal timing cues are used two times, so the second one brings no surprising effects. Because when we apply 4) only to MIKF list, AP improves 15.7% on train set and 5.5% on test set.

Modalities and fusion method	AP	
	Train set	Test set
MIKF_CNN list	0.6488	0.5618
TMP_CNN list	0.1156	0.0768
2) 0.9 0.1	0.6315	0.6191
TXT_CNN list	0.1301	0.2479
TMP_CNN list	0.1156	0.0768
1) 0.6 0.4	0.3125	0.3419
Merged list	0.7851	0.6538
4) re-rank list	0.7764	0.6233
2) 0.2 0.8	0.8340	0.6649
MIKF list	0.6449	0.5453
4) re-rank list	0.7296	0.5654
2) 0.1 0.9	0.7463	0.5753

Table 2 Combination effects of timing and other information. Two lists and the fusion methods are shown in consecutive three rows. The method is shown in its sequence number with weighting parameters.

5. CONCLUSIONS AND FURTHER DIRECTIONS

In this paper, two kinds of timing information are discussed, and a novel non-linear information fusion method is presented to combine timing cues with other information for concept detection in news video.

But the fusion methods are depending on some pre-selected parameters. And these parameters vary from concept to concept. Automatic methods for choosing these parameters and methods to fuse non-orthogonal information will be addressed in future.

6. ACKNOWLEGDEMENT

The work is supported by National Natural Science Foundation of China (60135010) and (60321002), also by Supported by the Chinese National Key Foundation Research & Development Plan (2004CB318108). And special thanks go to Shaoping Ma, Min Zhang, Yiqun Liu and Le Zhao of Pattern Recognition and Intelligent Retrieval reseach group, State Key Lab of Intelligent Tech. & Sys. for providing the TMiner IR system, v2.1 in text retrieval list generation.

7. REFERENCES

[1] "The TREC Video Retrieval Track Home Page", http://www-nlpir.nist.gov/projects/trecvid/

[2] B. Adams, et al, "IBM Research TREC-2002 Video Retrieval System", *NIST Text Retrieval Conference (TREC-2002)*, November 2002.

[3] A. Amir, et al, "IBM Research TREC-2003 Video Retrieval System", *NIST TREC-2003 Video Retrieval Evaluation Conference*, November 2003.

[4] A. Hauptmann, et al, "Video Classification and Retrieval with the Informedia Digital Video Library System", *NIST Text Retrieval Conference (TREC-2002)*, November 2002.

[5] A. Hauptmann, et al, "Informedia at TRECVID 2003: Analyzing and Searching Broadcast News Video", *NIST TREC-2003 Video Retrieval Evaluation Conference*, November 2003.

[6] D. Heesh, et al, "Video Retrieval using Search and Browsing with Key Frames", *NIST TREC-2003 Video Retrieval Evaluation Conference*, November 2003.

[7] "TREC-10 Proceedings appendix on common evaluation measures", http://trec.nist.gov/pubs/trec10/appendices/measures.pdf

[8] M. Zhang, *"Study on Web Text Information Retrieval",* PhD. Thesis (In Chinese). Tsinghua Univ. Beijing China. June, 2003.

[9] C.-Y. Lin, B. L. Tseng and J. R. Smith, "Video Collaborative Annotation Forum: Establishing Ground-Truth Labels on Large Multimedia Datasets," *NIST TREC-2003 Video Retrieval Evaluation Conference*, Gaithersburg, MD, November 2003. http://www-nlpir.nist.gov/projects/tvpubs/papers/ibm.final.paper.pdf

[10] M. G. Christel, et al, "Collages as dynamic summaries for news video", *ACM Multimedia*, Juan-les-Pins, France, December, 2002.

[11] R. Collobert and S. Bengio, "SVMTorch: Support Vector Machines for Large-Scale Regression Problems", *Journal of Machine Learning Research*, 1(2001) 143-160, February 2001.

[12] J.L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System", *Speech Communication*, 37(1-2):89-108, 2002. ftp://tlp.limsi.fr/public/spcH4_limsi.ps.z

[13] D. Black, *The Theory of Committees and Elections. 2nd ed.*, Cambridge University Press. 1958, 1963.