COMMERCIAL VIDEO RETRIEVAL WITH VIDEO-BASED BAG OF WORDS

JINQIAO WANG, HANQING LU

National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China

LINGYU DUAN

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613

JESSE. S. JIN

The School of Design, Communication and Information Technology, University of Newcastle, NSW 2308, Australia

The rapid increasing of commercial videos in different formats, sizes and frame rate brings much difficulty for commercial video retrieval. In this paper, we present a commercial video with video-based bag of words mode and FMPI concepts, and propose a coarse-to-fine framework for robust commercial retrieval in video streams. Experimental results and comparison show the promise of the proposed algorithm.

1. Introduction

Commercial videos are pervasive in TV broadcasting and internet. Commercial video is a form of advertising in which goods, services, and ideas are promoted via the medium of television and internet. Essentially, commercial is a kind of information medium, and TV viewer may find useful information about products or services which they do not know and might want. In addition, tracing commercials from TV streams is potentially useful for competitive marketing research, ad planning, and even ad investment. With the ever-increasing channels, digital video storage, and processor, efficient and effective indexing and retrieval of TV commercials in video streams is becoming feasible and necessary for commercial monitor, copyright protection, and commercial database management.

Existing commercial retrieval approaches mainly have two categories: frame-based [1, 2] and clip-based [3, 4, 5]. In [1], visual features (color, edge, and face) are extracted from multiple key frames. The similarity of visual features is computed to detect repeated commercials. In [2], the structure of

commercials is represented by a set of key frames, and Principal Component Analysis (PCA) is used to select features for commercial recognition. Different from frame-based approaches, Clip-based methods attempt to capture unique spatial-temporal features from a sequence of frames. In [3], the ordinal pattern histogram and the cumulative color distribution histogram are extracted to capture the spatial-temporal pattern of the commercial videos. In [4], color moments are used to measure the shot-level similarity of commercial videos to identify new commercials. With subsequent moment vectors, the hashing technique is applied to video frames to detect duplicate commercials in [5].

Frame-based approaches assume a set of key frames can provide a compact representation of commercial video contents. In practice, due to the fairly dynamic content, it is difficult to come up with a unified key frame selection scheme to extract the universal features for commercial matching. For clip-based approaches, although hashing tables can accelerate the speed of retrieval, the performance would degrade with the change of frame rate or commercial length. Different from previous work, we resort to video-based bag of words and the FMPI concept (Frame Marked with Production Information) [6] to commercial retrieval. We propose a coarse-to-fine commercial retrieval framework for quickly and robustly search commercials in video streams.

2. Commercial Video Modeling with Video-based Bag of Words

2.1. *Video-based Bag of Words*

The bag of words [11] model offers a rather impoverished representation for the data, for it ignores any spatial relationships between the local or global features. Nonetheless, it has been successfully used in text domain, scene and object classification domain, because of the high discriminative power of some words and the redundancy of the language, and natural images in general. Here, we will investigate how far it can be applied to commercial videos, which contain more information including images, audio and texts.

Figure 1.Commercial video modeling with video based bag of words.

To employ the bag of words model to commercial videos, we seek to build a vocabulary of visual words, which are dependent on individual commercials, as illustrated in Fig.1. Firstly, shot detection [27] is performed. Within each shot, a set of key frames are selected at the minima of a series of average intensities

of motion vectors. That is, we use a set of key frames to represent a commercial. Secondly, we extract local regions in key frames and for each local region a descriptor is computed as visual feature vectors. The modeling of a commercial video resorts to the discovery of co-occurrence patterns from a large number of local region descriptors. Finally, we employ Earth Mover's Distance (EMD) [11] to match commercials in video streams with a coarse to fine framework.

2.2. *Feature Extraction*

Within a key frame of a commercial video, we want to extract local visual features. Local features have been shown to be powerful for their invariance to occlusions and viewpoints [7]. In this work, these features extracted in the first step should invariant to variations that are irrelevant to different versions (long or short, WMV or MPG format, and different video size). We extract local regions by random sampling, Harris [9], and SURF [8], respectively. The Harris detector locates corner-like regions, and the SURF detector extracts blob-like regions. We employ the 128-dimensional SIFT descriptor [7] to compute a gradient orientation histogram for each local region.

2.3. *Visual vocabulary*

The region descriptors are quantized according to the nearest neighbor rule. Subsequently, a visual vocabulary is constructed by applying K-means clustering to all the local descriptors extracted from training images, and those means remain as visual terms. Small clusters are pruned out as noises. Finally, each ad video can be represented as a collection of visual words.

Rather than Euclidean distance, χ^2-distance is used to measure the distance between SIFT vectors. Euclidean distance calculates the absolute differences between bins. If the absolute bin differences are small, then the Euclidean distance is small, even when bin differences are much larger than real bin values. In contrast, χ^2-distance considers the relative value of bin differences to real bin values.

3. Commercial Video Retrieval

We employ the commercial video representation with video-based bag of words model to retrieval in video streams. A coarse-to-fine scheme is proposed to robustly retrieve commercial videos in video streams. Firstly, the candidate commercial positions are detection by FMPI image search. Secondly,

commercial video matching with video based bag of words is utilized to exactly locate the commercials.

3.1. *FMPI Image Search*

FMPI is novel concept in commercial video domain. An FMPI image can be dealt as a kind of document image involving graphics (e.g., corporate symbols, logos), images (e.g., products, setting and props), texts (e.g., brand names, headlines or captions and contact information). FMPI images are used to highlight the advertised products, service, or ideas. The FMPI concept has been applied to detect individual commercial boundaries [6]. Clearly, the dynamic content and various editing effects pose some challenges in terms of shot detection and key frame selection. However, the FMPI images provide a uniform and clear pattern, which is detectable by pattern recognition. As FMPI images are different amongst different commercials, we can utilize FMPI images to help represent the commercials. As the FMPI images always appear as an image sequence, we apply the FMPI image recognition to key frames only.

Our FMPI images detection approach is based on SVM learning. An FMPI image is represented by properties of color, texture, and edge features. A 141-dimensional visual feature vector comprising 128-dimensional local features and 13-dimensional global features is constructed. The local and global features are calculated are feed into a SVM classifier. The key frame with the highest probability being FMPI is chosen to represent the occurrence of a commercial. Then FMPI image search with color histogram [10] is employed to obtained the candidate commercial positions.

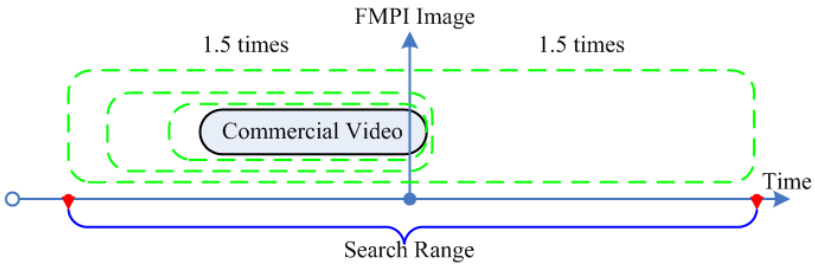

Figure 2. Commercial retrieval with video-based bag of words

3.2. *Commercial Retrieval with Video-based Bag of Words*

After the candidate positions are obtained with FMPI image search, we employ a sliding window to search for the exactly positions with video-based bag of words, as illustration in Fig.2. The total search range is 3 times length of the input commercial video. A sliding window 1.5 times of the input commercial

video is utilized to search shot by shot. After finding the most similar window, the same search strategy is employed within this window with smaller windows.

With the video-based bag of video representation of commercial videos, EMD is utilized to measure the similarity between commercial videos. For a commercial video is presented as $\{(p_1, s_1), (p_2, s_2), ..., (p_m, s_m)\}$, where m is the number of the visual vocabulary, p_i is the center of the i th cluster, and s_i is the proportional size of the cluster. The EMD between the commercial video widow $C = \{(p_1, s_1), (p_2, s_2), ..., (p_m, s_m)\}$ and the k th sliding window $W_k = \{(q_1, \omega_1), (q_2, \omega_2), ..., (q_n, \omega_n)\}$ is calculated as:

$$D(C, W_k) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d(p_i, q_j)}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}} \quad (1)$$

Where f_{ij} is a flow value that can be determined by solving a linear programming problem, and $d_{ij}(p_i, q_j)$ is the Euclidean distance between cluster centers p_i and q_j. After the commercial search with different windows, the windows with the smallest EMD are the position of commercial videos.

4. Experiments

Our experimental video data are extensively collected from TRECVID 2006 news video corpus and several Chinese TV channels. To evaluate the robustness of our algorithm, we connect all the video segments into a long segment, and choose two categories of commercials (include 50 automobile commercials and 50 cosmetic commercials) as queries. The reason why we choose the two categories is there are more similar shots that add the difficulty to retrieval. The video is in MPEG-1 format with the frame rate of 29.97 fps and the frame size of 352 × 240. The video data are changing resolution (176 × 120, 720 × 480), and resampled at different frame rate (15, 60 fps).

We seek a trade-off between the recall and precision to choose an appropriate threshold in the FMPI search. The recall is critical, for we just need to obtain the candidate positions of the commercial segments in the coarse phase. The corresponding threshold in which the recall is 100% is chosen in the experiment.

With the candidate positions, the video-based bag of words features are used to commercial retrieval with EMD. For evaluating the commercial performance, the receiver operating characteristics (ROC) curve is employed, which is based on false positive rate (FPR) and false negative rate (FNR).

Fig. 3 illustrates the relationship between the number of visual words and the performance of retrieval. For the number of visual words is 250, the average performance is 96%. In our experiment, 300 words are chosen to present a commercial video.

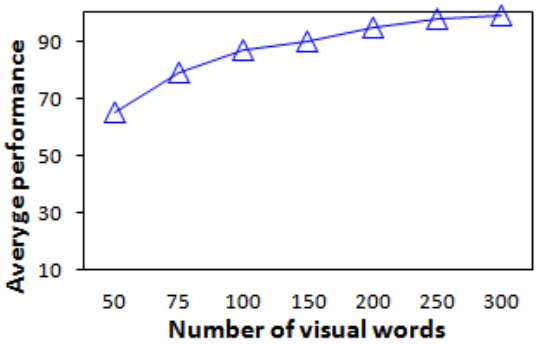

Figure 3.Commercial video retrieval with different number of visual words.

We compare our approach with Yuan's approach [3], and our previous approach with color and ordinal features. Fig.4 shows the average comparison results with the commercial database. When the frame rate changes, Yuan's method which is clip based is degraded, especially for automobile commercials which have more similar images between different commercials. But for our previous approach and video-based bag of words approach, we obtain better performance. The proposed approach obtains less noise than our previous approach, which also can be seen in Fig.4.

5. Conclusion and Future Work

In this paper, a video-base bag of words model is utilized to commercial video retrieval in video streams. FMPI is a robust indicator for a commercial video. It can reduce lots noise, and is a necessary when the commercial is got shorter or longer. Experimental and comparison results show its promise in video processing and retrieval. Next step, we will apply the video-based bag of words approach to clip retrieval and semantic concept retrieval.

6. Acknowledgement

This work is supported by National Natural Science Foundation of China (Grant No. 60475010, 60121302 and 60675003), and the 863 program No. 2006AA01Z315.

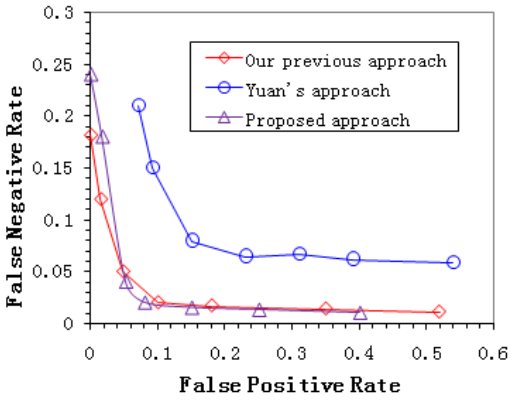

Figure 4.The ROC curves of Yuan's approach, our previous approach, and video-based bag of words approach.

References

1. Pinar Duygulu, Ming-Yu Chen, etc. Comparison and combination of two novel commercial detection methods, *Proc. CIVR'04*, July, (2004).
2. Juan M. Sanchez, Xavier Binefa, etc. Shot partitioning based recognition of tv commercials, *Multimedia Tools and Applications*, 223(2002).
3. Junsong Yuan, Ling-Yu Duan, etc. Fast and robust short video clip search using an index structure, *Proc. ACM MIR'04*, 61(2004).
4. John M. Gauch and Abhishek Shivadas, Identification of new commercials using repeated video sequence detection, *Proc. ICIP'05*, 1252(2005).
5. A. Shivadas and J.M. Gauch, Real-time commercial recognition using color moments and hashing, *Proc. ACM MIR'06*, Oct, (2006).
6. Ling-Yu Duan, JinqiaoWang, Yantao Zheng, Jesse S. Jin, Hanqing Lu, etc, Segmentation, categorization, and identification of commercials from tv streams using multimodal analysis, *Proc. ACM MM'06*.
7. D.G. Lowe. Distinctive Image Features Form Scale-invariant Keypoints, *International Journal of Computer Vision*, 60, 91(2004).
8. H. Bay, T. Tuytelaars, and L.V. Gool. SURF Speeded Up Robust Features. *Proc. ECCV'06*, 2006.
9. Harris, C. and Stephens, M. A Combined Corner and Edge Detector. *Alvey Vision Conference*, 147(1988).
10. Jinqiao Wang, Lingyu Duan, Qingshan liu, Hanqing Lu and Jesse S. Jin. Robust Commercial Retrieval in Video Streams. To appear in ICME'07, (2007).
11. Y.Rubner, C. Tomasi, and L. Guibvas. The Earth Mover's Distance as a Metric for Image Retrieval. *IJCV*, 40(2000).