ROBUST COMMERCIAL RETRIEVAL IN VIDEO STREAMS

Jinqiao Wang[1,2]*, Lingyu Duan*[2,3]*, Qingshan Liu*[1]*, Hanqing Lu*[1]*, Jesse S. Jin*[3]

[1] National Lab of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China
{jqwang, qsliu, luhq}@nlpr.ia.ac.cn
[2] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
lingyu@i2r.a-star.edu.sg
[3] The School of Design, Communication and Information Technology,
University of Newcastle, NSW 2308, Australia
Jesse.Jin@newcastle.edu.au

ABSTRACT

TV commercial video is a kind of informative medium. To fast and robustly index and retrieve commercial videos is of interest to commercial monitor, copyright protection, and commercial management. we propose a coarse-to-fine scheme to robustly retrieve commercial videos. Different from previous work using clip or key frames-based matching, our scheme has incorporated the commercial production knowledge to search the candidate commercial positions. Color and ordinal features are extracted for locating the exact commercial positions with dynamic time warping distance. Comparison experiments were carried out over TRECVID 2006 news videos and some videos from Chinese channels. Our scheme has achieved promising simulation results.

1. INTRODUCTION

TV commercial is a form of advertising in which goods, services, and ideas are promoted via the medium of television. It is generally considered to be the most effective mass-market advertising format that is reflected by the high prices TV networks charge for commercial airtime during popular TV events. TV viewers can regard commercial as a kind of information medium, and may find useful information about products or services which they do not know and might want. With the growth of cable TV, digital video storage, and processor, an 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 are 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 commercial production knowledge to facilitate the effective representation of commercial videos. Especially we introduce the concept of FMPI (Frame Marked with Production Information) [6] to help identify a commercial. Although different commercials may share similar visual content in terms of measuring low-level features, the FMPI images are clearly different and can function as a very unique identity. Moreover, different versions (short or long) of one commercial video usually share the same FMPI even if the visual content varies much. Based on the FMPI concept, we propose a coarse-to-fine commercial retrieval framework to quickly and robustly search commercials in video streams.

The rest of this paper is organized as follows. Section 2 introduces the commercial representation with FMPI images, and presents the training of FMPI recognizer with local and global features. Section 3 discusses the detection of the candidate commercial positions with FMPI search. Key frames based matching is applied to locate the exact commercials in Section 4. Experiments results are given in Section 5. Finally, Section 6 concludes our work.

2. COMMERCIAL REPRESENTATION

Nowadays enormous money is spent on TV commercial production to capture the audiences' attention. This vast expenditure has brought us a number of high-quality TV commercials with the latest technology such as special editing effects, the most popular personalities, and the best music. Many television commercials are produced so elaborately that they can be considered as miniature miniature movies, say 30 seconds. Undoubtedly, such creative arts design has make commercial video representation fairly challenging. With traditional approaches, we would have to adjust the feature extraction methods to adapt to different commercials.

Our commercial representation resorts to commercial production rules, namely, the presence of FMPI images. As shown in Fig. 1, an FMPI image can be dealt as a kind of document image involv-

ing 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. One interesting thing is that if two commercials from different companies share the same FMPI images, some copyright problems would arise. As the FMPI images always appear as an image sequence or a shot, we may apply the FMPI image recognition to key frames only.

(a) (b) (c) (d) (e) (f)

Fig. 1. Examples of the FMPI images.

2.1. FMPI Recognizer

Our FMPI images detection approach is based on SVM learning. An FMPI image is represented by properties of color, texture, and edge features. We first divide an FMPI image into 4×4 sub-images to extract local features, then the whole image is used to extract global features. The local and global features are calculated as follows.

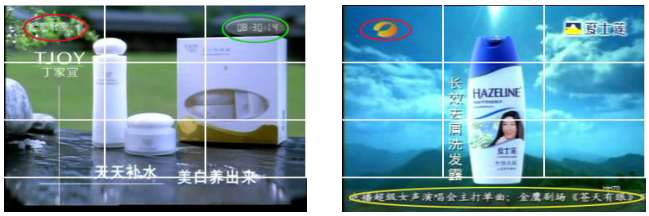
For local features, the LUV color space is used to manipulate color and it is uniformly quantized into 300 bins. Each channel is assigned with 100 bins. Three maximum bin values are selected as features from L, U, and V channels, respectively. Edges derived from an image using Canny algorithm [7] provide an accumulation of edge pixels for each sub-image, which finally acts as 16-dimensional edge density features. A set of two-dimensional Gabor filters [8] are employed to extract texture features. The filter bank comprises 4 Gabor filters that are the results of using one center frequency (i.e., one scale) and four different equidistant orientations. The application of such a filter bank to an input image results in a 4-dimensional feature vector (consisting of the magnitudes of transform coefficients) for each point of that image. The mean of feature vectors is calculated for each sub-image. A 128-dimensional feature vector is then formed to represent local features. For global features, we take into account the cues of color and edge. Three maximum bin values are selected from each color channel, which results in a 9-dimensional color feature vector for the whole image. Edges are broadly grouped into four categories: horizontal, 45 diagonal, vertical, and 135 diagonal. Edge pixels are accumulated for these categories, respectively, thus yielding 4-dimensional edge direction features. Finally, we construct a 141-dimensional visual feature vector comprising 128-dimensional local features and 13-dimensional global features.

The recognition of an FMPI image is a binary classification problem. SVM is utilized to accomplish supervised learning. C-Support Vector Classification (C-SVC) [9] is employed, and the radial basis function (RBF) kernel is used. Aiming to determine an FMPI shot, FMPI recognition is applied to shot key frames.

3. FMPI SEARCH

By searching FMPI images, we attempt to roughly locate the candidate position of commercials in video streams. The FMPI recognition is applied to key frames within each shot. The shot detection [10] is first carried out. The key frames are selected at the minima of average intensity of motion vectors. We use the probability output of SVM to determine the FMPI images; that is, the key frame with the highest probability is chosen to represent the occurrence of a commercial.

It is worthy to mention that in some channels (especially eastern TV channels), commercial videos are accompanied by the station logos, digit clocks, and moving captions as shown in Fig. 2. Towards effective feature extraction, we have to remove the logo region, clock region, and moving captions. The Width $\times$ (Height/8) region from the bottom is removed to reduce the false alarms from possible moving captions. The remaining area is partitioned into 3×3 sub-images equally. The top-left and top-right corner sub-images are not considered for feature extraction.



(a) CCTV-1 channel (b) CQTV channel

Fig. 2. Examples of overlapping logos, digit clocks and moving captions in commercial videos.

Let us consider the similarity matching between FMPI images. To build a robust color feature to distinguish different FMPI images, we quantize each FMPI image q into a fixed number r of colors, which helps to eliminate the effect of small variations within an image and to avoid a large file due to high-resolution representation [11]. Each color element is then discretized into t binary bins ($\mathbf{b} = b_1 b_2 ... b_t$) of equal or varying capacities by calculating the percentage of pixels dominating the color element. Given a frame c, the color similarity is:

$$S(q,c) = \frac{1}{\sum_{i=1}^{r} |p(\mathbf{b}_q^i) - p(\mathbf{b}_c^i)|} \quad (1)$$

where $p(\mathbf{b}_q^i)$ gives the position of the set bit within (the set of bins) $\mathbf{b}^i$ of the frame q , and each of R, G and B channels is quantized into 4 equal intervals thereby resulting in a 64 dimensional color representation, namely, $r = 64$. t is set to 7 for each color element. Compared with randomly selected key frames searching, the FMPI searching can more reduce the number of candidate commercial positions, especial in automobile commercials which have more similar shots.

4. COMMERCIAL RETRIEVAL

With FMPI search, we have obtained some candidate commercial positions in a long video sequence. Key frames based matching is subsequently carried out to exactly locate commercials. To address color distortion, frame rate change, resolution change and different versions (long or short), a set of color and ordinal features are extracted.

4.1. Color and ordinal features

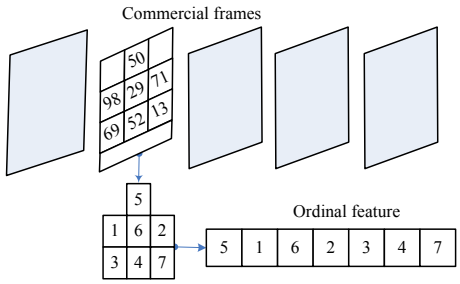


Fig. 3. Ordinal feature description.

The color feature is same as that used in FMPI search. Each key frame q_i within the commercial clip $Q = \{q_1, q_2, ..., q_n\}$ is quantized into a fixed number of colors r. Given the query commercial clip $Q = \{q_1, q_2, ..., q_n\}$ and the candidate clip $C = \{c_1, c_2, ..., c_l\}$, the color distance between key frame q_i and key frame c_j is,

$$d_c(q_i, c_j) = \sum_{k=1}^{r} |p(\mathbf{b}_{q_i}^k) - p(\mathbf{b}_{c_j}^k)| \quad (2)$$

Ordinal feature reflects the spatial information about color distribution [3]. Each frame is partitioned into 3×3 regions. Like the preprocess in Section 3, the left-up and right-up regions are removed. The remaining 7 regions are ranked by the averaged intensity and sorted in descending order. Thus a 7×1 feature vector is yielded as shown in Fig. 3. The distance of ordinal feature between q_i frame and c_j frame is calculated as,

$$d_o(q_i, c_j) = \sum_{k=1}^{m} |\pi_{q_i}^k - \pi_{c_j}^k| \quad (3)$$

where $\pi_{q_i}^k$ is the kth ordinal value of the q_i frame. m is the dimension of the ordinal feature, and m is set to 7.

The integrated distance between key frame q_i of query commercial clip Q and key frame c_j of the candidate clip C is the linear combination of the color distance and the ordinal distance, and is defined as Eq. 4.

$$d(q_i, c_j) = \omega \times d_c(q_i, c_j) + (1 - \omega) \times d_o(q_i, c_j) \quad (4)$$

where ω is the weight of the color feature, ω is set to 0.5.

4.2. Commercial Retrieval with Dynamic Time Warping

Now the commercial retrieval is reduced to the matching problem between different time series of feature. An intuitive solution to matching is to do resampling and compare the series by sample-by-sample similarity computing. The drawback of this method is that it does not produce optimal results, as it compares samples that might not correspond well. Dynamic Time Warping [12] (DTW) calculates the matching distance by recovering optimal alignments between samples from two time series. The alignment is optimal in the sense that it minimizes a cumulative distance measure consisting of local distances between aligned samples. The DTW distance is a robust measurement which allows time series to be stretched or compressed along the time-axis and is able to compare the similarity of time series with different lengths.

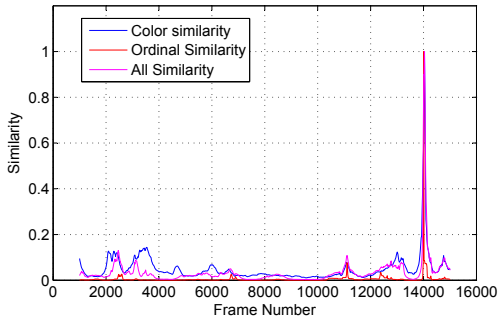


Fig. 4. Similarity curves of a query commercial clip against a long video sequence with different features.

Given the query commercial clip $Q = \{q_1, q_2, ..., q_n\}$ and the candidate clip $C = \{c_1, c_2, ..., c_l\}$, where n and l are frame number respectively. The DTW distance $D(Q, C)$ is:

$$D(Q, C) = D(n, l), \quad D(f, f) = d(q_f, c_f)$$
$$D(i, j) = d(q_i, c_j) +$$
$$\begin{cases} \min\{D(i, j-1), D(i-1, j), D(i-1, j-1)\} & i, j > f \\ \min\{D(i, j+1), D(i+1, j), D(i+1, j+1)\} & i, j < f \end{cases}$$

where $d(.,.)$ is the combined distance of color and ordinal features. f is the position of the FMPI image detected by FMPI image search. When $S = 1/D(Q, C) \geq Th$, the same commercial is claimed to be located. The detected commercial boundary is decided by the length of most similar candidate clip.

For each candidate commercial position obtained by FMPI image searching, the exact commercial position is decided by the DTW distance measure. As illustrated in Fig. 4, we search a commercial in a long video segment. Color and ordinal features are effective for commercial searching. With the combination of color and ordinal features, more noise is reduced in the searching process.

5. EXPERIMENT

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).

Fig. 5 illustrates the FMPI image recognition results with different features. "Color" and "Texture" have demonstrated individual capabilities of our color and texture features to distinguish FMPI

from Non-FMPI images. Comparatively, texture features play a more important role. The combination of color and texture features results in a significant improvement of performance. A accuracy $F1 = 90.2\%$ is obtained with color, edge and texture features.

We seek a trade-off between the recall and precision to choose an appropriate threshold in the FMPI search. As shown in Fig. 6, 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% for both category of commercials is used in our experiment.

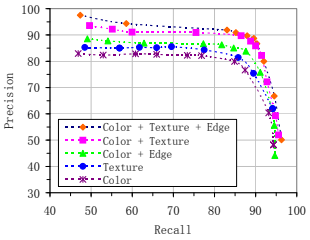


Fig. 5. The precision and recall curves of FMPI recognition with different features.

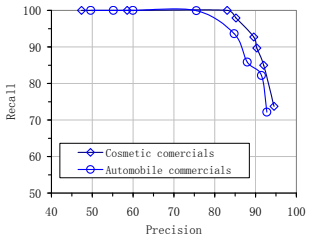


Fig. 6. The precision and recall curves of FPMI searching with different thresholds.

With the candidate positions, the color and ordinal features are used to commercial retrieval with DTW distance. 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) as shown.

$$\text{FPR} = \frac{\text{Number of false detected commercials}}{\text{Number of detected commercials}}$$

$$\text{FNR} = \frac{\text{Number of missed commercials}}{\text{Number of target commercials}}$$

We compared our algorithm with fast short video clip searching algorithm [3]. Fig. 7 and Fig. 8 illustrate the ROC curves of the two methods in two commercial categories respectively. We obtained better experimental results both in automobile commercial video and cosmetic commercial video. 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.

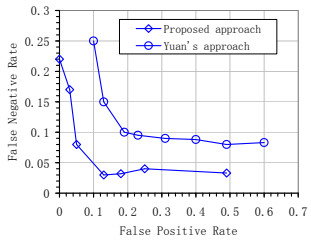


Fig. 7. The ROC curves of our method and Yuan's method for automobile commercial retrieval.

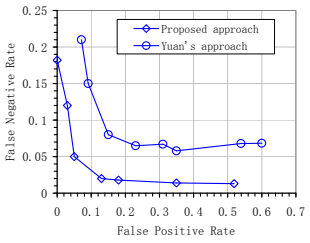


Fig. 8. The ROC curves of our method and Yuan's method for cosmetic commercial retrieval.

Our experiments were conducted on a computer with 3.0G Hz CPU and 1GB RAM. The video decoding speed is 118 frames/s. Given 12 commercial clips and 1 GB video data (185016 frames), the search process cost 42 minutes.

6. CONCLUSION

We have proposed a coarse-to-fine approach to retrieval commercials from video streams. Comparison experiments and analysis has shown our approach's advantages. Our approach's unique feature is to introduce the FMPI concept to facilitate commercial identification. When the commercial uses different versions (short or long) to adapt different purposes, the FMPI images often remain and is able to effectively represent a commercial. Such FMPI search at the coarse-level is to fast locate the candidate commercial positions from video streams. At the fine-level, DTW has make the commercial search more robust against color distortion, image resize, and frame rate changes. In the future, we will improve the search speed towards applications in large-scale video streams.

7. ACKNOWLEDGEMENT

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

8. REFERENCES

[1] Pinar Duygulu, Ming-Yu Chen, and Alexander Hauptmann, "Comparison and combination of two novel commercial detection methods," *Proc. CIVR'04*, July 2004.

[2] Juan M. Sánchez, Xavier Binefa, and Jordi Vitrià, "Shot partitioning based recognition of tv commercials," *Multimedia Tools and Applications*, pp. 223–247, Dec 2002.

[3] Junsong Yuan, Ling-Yu Duan, Qi Tian, and Changsheng Xu, "Fast and robust short video clip search using an index structure," in *Proc. ACM MIR'04*, 2004, pp. 61–68.

[4] John M. Gauch and Abhishek Shivadas, "Identification of new commercials using repeated video sequence detection," in *Proc. ICIP'05*, 2005, pp. 1252–1255.

[5] A. Shivadas and J.M. Gauch, "Real-time commercial recognition using color moments and hashing," in *Proc. ACM MIR'06*, Oct 2006.

[6] Ling-Yu Duan, Jinqiao Wang, Yantao Zheng, Jesse S. Jin, Hanqing Lu, and Changsheng Xu, "Segmentation, categorization, and identification of commercials from tv streams using multimodal analysis," in *Proc. ACM MM'06*, 2006, pp. 202–210.

[7] J. Canny, "A computational approach to edge detection," *IEEE Trans. PAMI*, vol. 8, no. 6, pp. 679–698, 1986.

[8] B.S. Manjunath and W.Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. PAMI*, vol. 18, no. 8, pp. 837–842, 1996.

[9] V. Vapnik, "The nature of statistical learning theory," *Springer-Verlag*, 1995.

[10] H.J. Zhang, A. Kankanhalli, and S. Smoliar, "Automatic partitioning of full-motion video," *ACM/Springer Multimedia Systems*, pp. 10–28, July 1993.

[11] D. S. Park, J. S. Park, T. Y. Kim, and J. H. Han, "Image indexing using weighted color histogram," in *Proc. ICIP'99*.

[12] Chih-Yi Chiu, Cheng-Hung Li, Hsiang-An Wang, Chu-Song Chen, and Lee-Feng Chien, "A time warping based approach for video copy detection," in *Proc. ICPR'06*, 2006, pp. 228–231.