2012 10th IAPR International Workshop on Document Analysis Systems

Similar Fragment Retrieval of Animations by a Bag-of-features Approach

Weihan Sun, Koichi Kise
Dept. Computer Science and Intelligent Systems
Osaka Prefecture University
Osaka, Japan
sunweihan@m.cs.osakafu-u.ac.jp, kise@cs.osakafu-u.ac.jp

Yoann Champeil
Dept. Eng. of Image Interaction Immersion
EISTI Engineering School
Cergy, France
yochampeil@gmail.com

***Abstract*—Serial animated cartoons, also called animations, is a kind of popular video documents that describe narratives by videos of cartoons. By using digital techniques, illegal users can divide animations into fragments and distribute them without any copyright permissions. In this paper we focus on the copyright problem of animations and try to retrieve copyright infringement fragments based on key frames. Because of the huge volumes and rapid release of animations, it is impossible to store all the episodes into the database. We propose applying bag-of-features model to retrieve similar fragments based on the visual words extracted from a limited data set. In the experiments, 12 titles of animations are employed and the similar fragments outside database are applied as queries. Our method has achieved above 98% precision at 80% recall for fragments whose durations are over 140 seconds. From the results, we show that a latest release can be retrieved based on the features of the former ones from the same title.**

***Keywords*-animation, serial animated cartoons, copyright protection, copyright infringement retrieval, similar copy, bag-of-features**

I. Introduction

Animations (Serial animated cartoons) is a kind of artwork constructed of hand-drawn or computer-made cartoons. Most of commercial animations are sourced from popular manga or comics and describe narratives by a series of episodes released regularly. In contrast to comic books, since animations can provide more vivid images through video and audio, they have a larger audience and recognition throughout the world. Traditionally, animations are distributed via television broadcasts, directly to video, or theatrically. Recently, the development of Internet bring kinds of convenient services for animations' release. Whereas the copyright problem becomes more serious, which is threatening the developments of animations. Therefore there is great demands for copyright protection.

In practice, the copyright issues are quite controversy and require the professionals' judgments. Since there are huge volumes of animations in kinds of distribution media, it is impossible to check them one by one by human beings. The purpose of our research is to apply computer techniques to detect suspicious animations instead of human beings for professionals' further judgments.

To escape the detection, illegal users often apply kinds of transformations to animations. One of the most common methods is to divide one animation into several fragments and distribute separately. Therefore, fragment detection is the problem we should consider.

For the copyright protection of video, there are two main approaches: watermarking [1] and Content Based Copy Detection (CBCD) [2]. Since watermarking is not robust for kinds of transformations, we follow the CBCD way. In this method, suspicious videos are treated as queries. Based on the database of copyrighted animations, the original ones which have been copied can be retrieved. However, in the case of animations, since they are released in serial episodes regularly for a long time, it is difficult to stored the whole series (all episodes) into the database. Therefore, we propose similar animation retrieval for this problem. Since one title of animation (one series of animations with the same title) describes a whole story, there are some common features shared by its different parts. Here, we define similar animations as the ones have the same title. In other words, the approach is to retrieve the latest release based on the former ones.

In this paper, we apply a bag-of-features approach with the visual words extracted from key frames of animations. In the experiments, we employed 12 titles of animations. The episodes with the same title but outside the database are applied to make queries. Our method achieved above 98% precision at 80% recall for fragments over 140 seconds. From the results we also show that: (1) latest releases of animations can be retrieved based on the former ones and (2) a larger data set for creating visual word dictionary can achieve a better performance.

The rest parts of this paper is arranged as follows: Section 2 provides the outline of the approach. Section 3 introduces the proposed method for similar animation retrieval. Experiments and results are shown in Section 4. Finally, Section 5 is conclusion and future work.

II. Outline of the approach

Our approach applies the bag-of-features model which is a popular method for computer version and document retrieval, such as object and scene retrieval in video [3], visual categorization [4], document categorization [5]. In our

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.79

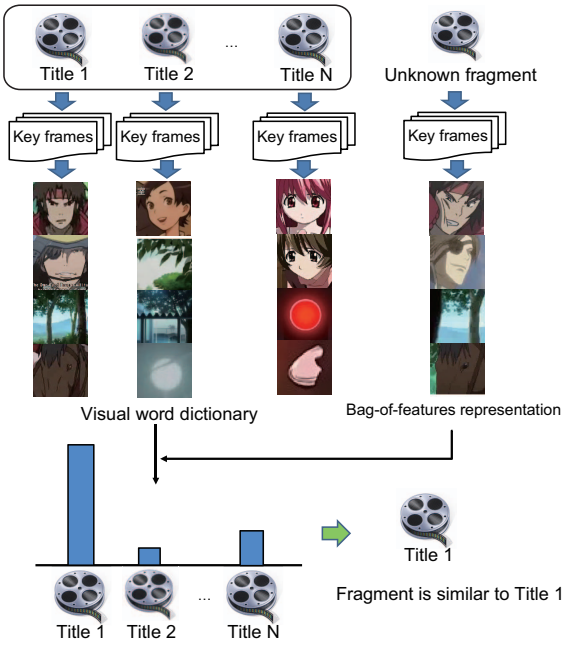


Figure 1. Outline of the approach.

previous research, we achieved similar manga retrieval using visual vocabulary based on contents of manga pages [6].

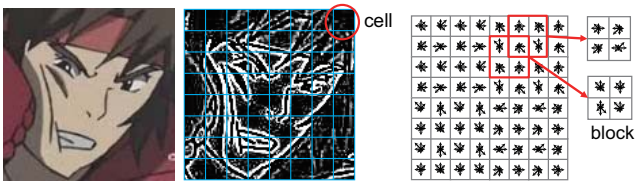
The processing for animation is shown in Fig. 1, where copyrighted animations categorized by their titles are collected, from which key frames are extracted in a certain sampling rate. Features are detected from these key frames and build a visual word dictionary. By matching the feature with the visual word dictionary, each title of animation is transformed into a bag-of-features representation. On the other hand, a suspicious animation is treated as a query. By the same method, the query is represented by a bag-of-features representation. Based on the comparison of bag-of-features representations between the query and the database, the titles of animations which are similar to the query will be reported as the result.

III. Animation Retrieval

A. Region detectors

From each key frame, two types of regions are detected: face regions of animation characters and a kind of local feature region.

For animations, characters are an indispensable parts. The narrative lines are extended around the main characters which are usually treated as discriminative features for a specific title of animation. In this research, we apply face regions of animation characters and detect them by Viola-Jones detection framework [7], which has been proved to be available for detecting faces of manga characters [8]. The detection is based on a detector trained by positive samples (images contain the object) and negative samples (non-object



(a) Detected region. (b) Cells. (c) Detected MSERs

Figure 3. HOG feature description.

images) with a machine learning method. For the detection, we collected $4,000$ face images of animation characters as positive samples and $10,000$ non-face images as negative samples to train the detector. The examples of character face detection of key frames are shown in Figs. 2(b) and (e).

In addition, MSERs (Maximally Stable Extremal Regions) [9] are also applied. It is a kind of view point covariant regions constructed by selecting stable areas from an intensity watershed image segmentation. The regions are approximately stationary as the intensity threshold is varied. By diagonalizing the covariance matrix of MSERs, we can obtain some ellipse MSERs from the image, as shown in Figs. 2(c) and (f).

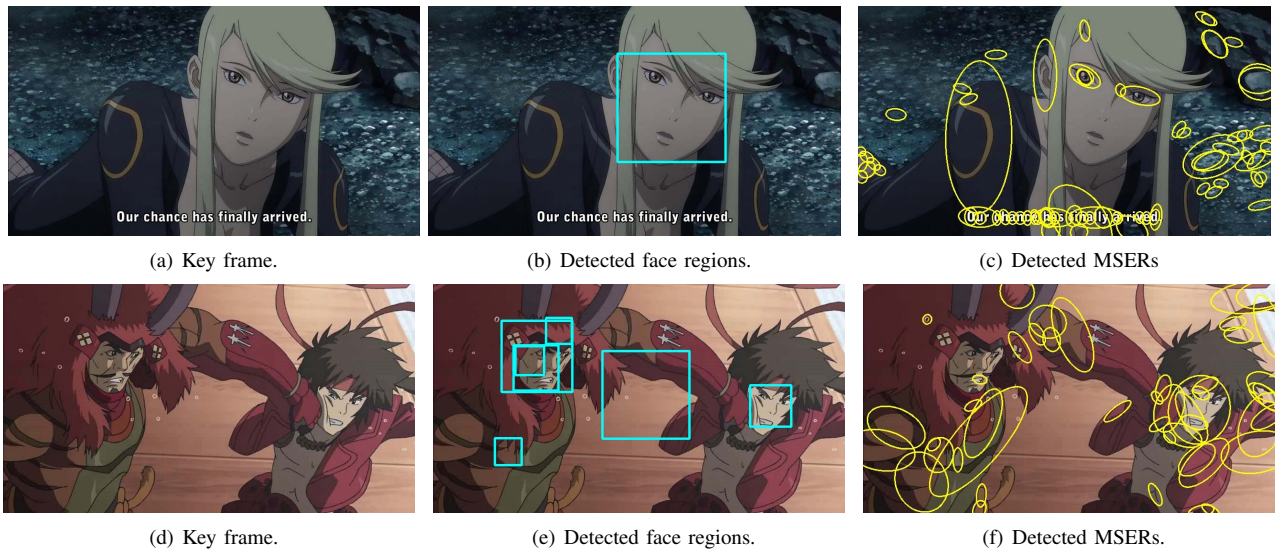
To make the regions contain more discriminative information, both face regions and MSERs are magnified as k times (in this research k is set to 1.5)

B. Feature descriptor

For the description of the detected regions, HOG (Histogram of Oriented Gradients) [10] are applied. As shown in Fig. 3, for each detected region, we calculate gradient magnitude and orientation at each pixel and divide the regions into 8×8 cells evenly. Then, the gradient orientation are quantized into 6 bins. For each cell, the gradient orientation histograms are calculated based on the gradient magnitudes. After that, cells are combined into overlapped blocks as 2×2 cells per block. By normalizing the features in blocks we obtain $6 \times 2 \times 2 \times 7 \times 7 = 1,176$ HOG features for each detected region. These features are combined into a HOG feature vector which consists of $1,176$ dimension.

C. Building a visual word dictionary

Then, the extracted feature vectors are quantized into clusters which will be applied as visual words for retrieval. For two types of regions, two visual word dictionary are prepared. The quantization is carried out by K-means in our method, and the centroids of the clusters are applied as visual words. Since the important regions are usually applied frequently, we do the clustering in the term of episodes of each title of animations and sort the clusters based on the number of regions inside. The top N clusters contain more regions are selected for the visual word dictionary. In this research, N is set to 600 for clustering of face regions, and $2,000$ for clustering of MSERs in one episode.



(a) Key frame. (b) Detected face regions. (c) Detected MSERs

(d) Key frame. (e) Detected face regions. (f) Detected MSERs.

Figure 2. Examples of detected face regions and MSERs from key frames. (Fig. 2(a) and (d) are key frames applied, Fig. 2(b) and (e) are detected face regions, Fig. 2(c) and (f) are detected MSERs. They are cited from animation "Basara".)

D. Retrieval

In the part of retrieval, by matching the feature vectors with the visual word dictionary, the titles of animations in database and the queries are represented by the bag-of-features representation.

We match the extracted feature vectors with their nearest neighbor in visual word dictionary and restrict the stable matching by Euclidean distance D between feature vector F and visual word W. If $D < T$ (T is a threshold), F is matched with W, otherwise not. The matching is more reliable for a smaller T.

To increase the speed of distance calculation, we propose applying ANN (Approximate Nearest Neighbor Searching) [11]. ANN is a method to find approximate nearest neighbors by using the k-d tree. While searching, the feature space shrunk by the factor $1/(1+\varepsilon)$ (ε is set to be 10 in this research).

In our approach, the feature vectors based on two types of regions are matched with two visual word dictionary separately and make the bag-of-features representation $\boldsymbol{V} = (F_1, F_2, ..., F_n, M_1, M_2, ..., M_m)$, where n and m are the total number of visual words based on face regions and MSERs, F_f and M_c represent weights for visual words, respectively. Both F_f and M_c are set to the frequency of regions in the cluster

$$tf_i = \frac{n_{c,d}}{n_d}$$

where $n_{c,d}$ is the number of regions of cluster c in animation d, n_d is the total number of regions in animation d. The similarity between animation d in database and the query q is calculated as

$$S_d = \frac{\boldsymbol{Q} \cdot \boldsymbol{V}_d}{\|\boldsymbol{Q}\|\|\boldsymbol{V}_d\|}$$

where $\boldsymbol{V_d}$ and $\boldsymbol{Q}$ are the bag-of-features representations of animation d and query q, respectively. According to the similarity, animations in the database are ranked and reported as results.

IV. Experiments

A. Conditions

12 titles of animations were collected. We kept their resolutions as their releases, and all the resolutions are higher than 640×480. For each title, there are five episodes and one episode lasts about 25 minutes. From these animations, the first three episodes are applied to build the three databases (DB1 contains Eps. 1, DB2 contains Eps. 1 and 2, DB3 contains Eps. 1, 2 and 3). The rests (Eps. 4 and Eps. 5) are divided into fragments of a certain duration (t) evenly as queries. Therefore, the queries are similar to the ones with the same title in the databases, but not exactly the same. The same titles as queries were treated as the right answers, thus there is one and only one right answer for each query. The results were reported by recall $R = A/B$ and precision $P = A/C$, where A is the number of right answers, B is the number of queries, and C is the number of retrieved queries. The matching of visual words is closely related to the threshold T. For a small T, since only regions with few differences were matched, we can get more reliable matching for retrieval and achieve a high precision. With the increase of T, both right and erroneous matchings are increased, and thereby the precision decreased with the increase of recall. In the experiments, T was changed and chosen for

the best performance. We extract key frames in every 2 seconds for both databases and queries. The bag-of-features methods using face regions (FR) and using MSER (MSER) are applied as well as the proposed method using both of them (FR+MSER). All experiments were done with a computer of INTEL i7-870 2.93GHz CPU and 8 GB RAM.

B. Similar fragment retrieval

First, we tested the effect of the database size for similar animation retrieval. As a benchmark, queries ($t = 70$ seconds) are applied. Based on the three databases, we built three visual word dictionaries and did retrievals using FR, MSER and FR+MSER methods. The interpolated average precisions are shown in Fig. 4. From the results, we can conclude that a larger data set can lead a better performance for all three methods.

Then, we chose the DB3 to test retrieval for fragments of different time. For the queries, t was set to 2, 20, 70, 140, 200, 400, 700 seconds. There is only 1 key frame for each query ($t = 2$ sec.), since key frames are extracted in every 2 seconds. The results are shown in Fig. 5. As t became larger, both recalls and precisions increased. Because of the stable matching methods (high precision) we applied for similar regions, the fragments with key frames contained discriminative regions will be correctly retrieved. Since not every key frame contains discriminative information for retrieval, the queries with less key frames did not achieve a high recall. The detection time increases for longer durations, since more features are applied. For queries ($t = 140$ ms), the average detection time (excluding time for feature extraction) is 161 ms for face regions and $2,521$ ms for MSERs based on DB3.

Finally, we compared the three methods based on the retrievals of 140 seconds' fragments and DB3. As shown in Fig. 6, MSERs performed better than face regions and our proposed method achieved the best performance (above 98% precision at 80% recall) by using them together.

The main reason for failures are (1) not all frames contain a discriminative information, (2) some characters are similar for shape features, as shown in Fig. 7, (3) there are some commonly used regions in animations as shown in Fig. 8.

From these results, we can see that (1) the features in the former episodes can be used for the retrieval of the latest ones (2) a larger data set applied in the database can make a better performance (3) the proposed method can achieve higher accuracy for longer fragments.

C. Discussions

For practical copy detection systems, recall means the power to find copies, and precision means how much they can reduce the manual work. For example one episode of 25 minutes are divided into 10 fragments (each fragment lasts for 150 seconds). By our proposed method, we can detect above 80% of them without many workloads for humans.

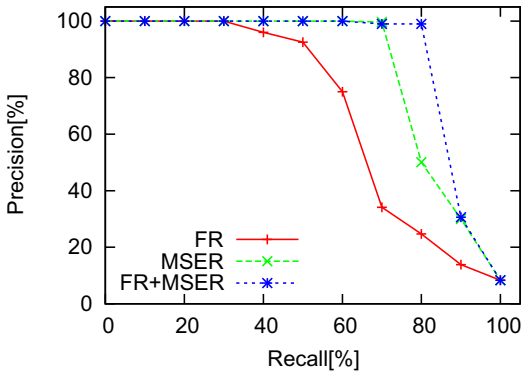
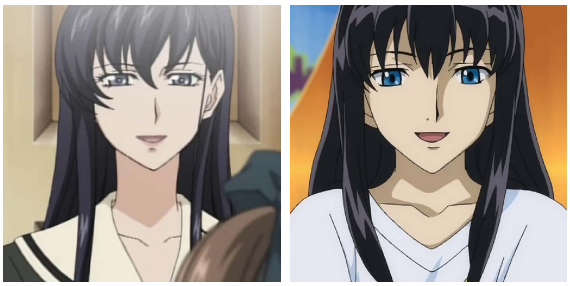


Figure 6. Interpolated average precision-recall curves of FR, MSER and FR+MSER (retrieval of 140 seconds' fragments based on DB3).



(a) (b)

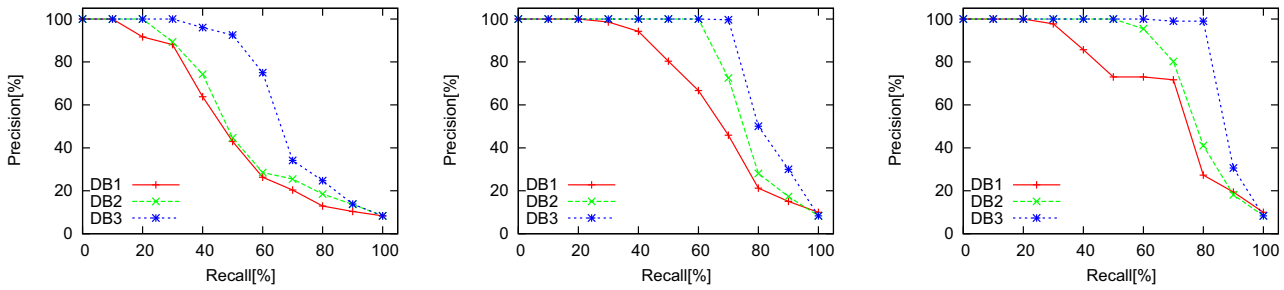
Figure 7. Example of similar characters. The color of their eyes and hair are different. (Fig. 7(a) is from Eps. 3 of "Marimite". Fig. 7(b) is from Eps. 4 of "Gundam Series".)

Figure 8. Examples of similar patterns in different animations.

For 12 titles of animations, the proposed method costs about 2.7 seconds (time of face regions plus MSERs) for detecting the fragments of 140 seconds' duration. The time will increase with more episodes applied in the database. Since huge volumes of animations need copyright protection and more episodes are required for high accuracy, the scalability is important for the method used in practice.

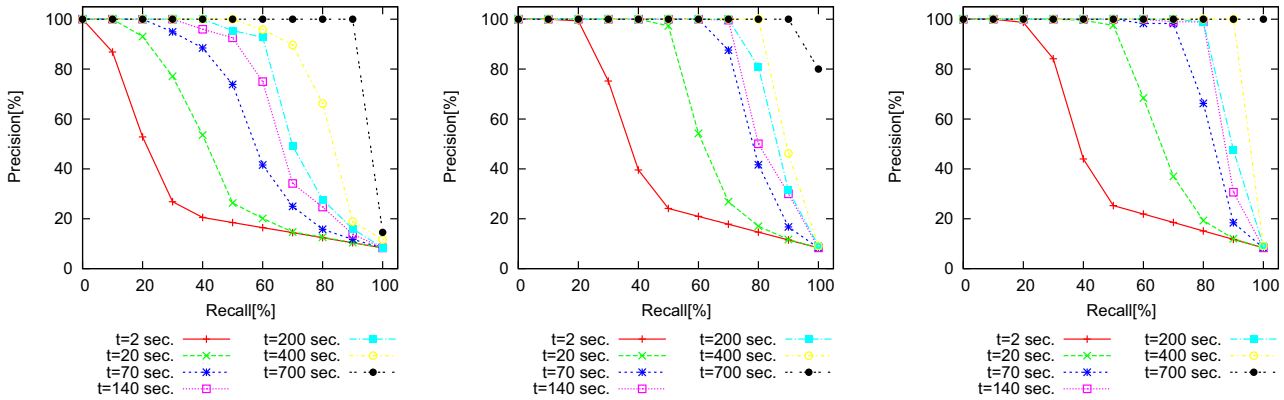
V. Conclusion

In this paper, we propose a bag-of-features method to retrieve similar fragment for copyright protection of animations. Two kinds of visual words are applied based



(a) FR. (b) MSER. (c) FR+MSER.

Figure 4. Interpolated average precision-recall curves for different data sets.



(a) FR. (b) MSER. (c) FR+MSER.

Figure 5. Interpolated average precision-recall curves for fragments of different durations.

on face regions and local feature regions detected from key frames. By the experiments, the effectiveness of the proposed method for similar animation retrieval has been proved. In addition, we show that the latest release of animation can be retrieved based on the features of the former ones from the same title and a larger data set used in the database can make a better performance.

Our future work includes (1) increase the titles of animations, (2) try other visual words to increase the performance (3) improve the scalability of the proposed method.

ACKNOWLEDGMENT

This research was supported in part by the Grant-in-Aid for Scientific Research (B)(22300062) from Japan Society for the Promotion of Science (JSPS).

REFERENCES

[1] M. Maes, T. Kalker, J.-P. Linnartz, J. Talstra, G. Depovere, and J. Haitsma, “Digital Watermarking for DVD Video Copy Protection”, IEEE Signal Processing Magazine, vol. 17, no. 5, pp 470057, Sep. 2000.

[2] J. Law-To, L. Chen, A. Joly, et al. “Video copy detection: a comparative study”, ACM International Conference on Image and Video Retrieval, pp. 371–378, 2007.

[3] J. Sivic and A. Zisserman, “Video Google: A Text Retrieval Approach to Object Matching in Videos”, IEEE International Conference on Computer Vision, pp. 1470–1477, 2003.

[4] G. Csurka, C. Bray, C. Dance and L. Fan, “Visual Categorization with Bags of Keypoints”, ECCV Workshop on Statistical Learning in Computer Vision, pp. 59–74, 2004.

[5] M. Rusiñol and J. Lladós, “Logo Spotting by a Bag-of-words Approach for Document Categorization”, International Conference on Document Analysis and Recognition, pp. 111-115, 2009.

[6] W. Sun and K. Kise, “Similar manga Retrieval Using Visual Vocabulary Based on Regions of Interest”, International Conference on Document Analysis and Recognition, pp. 1075–1079, 2011.

[7] P. Viola, M. Jones. “Robust Real-Time Face Detection”, International Journal of Computer Vision vol. 57(2), pp. 137–154, 2004.

[8] W. Sun and K. Kise, “Similar Partial Copy Detection of Line Drawings Using a Cascade Classifier and Feature Matching”, International Workshop on Computational Forensics, pp. 121–132, 2010.

[9] J. Matas, O. Chum, M. Urban and T. Pajdla, “Robust Wide Baseline Stereo from Maximally Stable Extremal Regions”, British Machine Vision Conference, pp. 384–393, 2002.

[10] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection”, IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893, 2005.

[11] S. Arya, D. Mount, R. Silverman and A. Y. Wu, “An Optimal Algorithm for Approximate Nearest Neighbor Searching”, Journal of the ACM, 45, 6, pp.891–923, 1998.