Small Vocabulary with Saliency Matching for Video Copy Detection

Huamin Ren [1], Thomas B. Moeslund[1], Sheng Tang [2] and Heri Ramampiaro [3]
[1] *Visual Analysis of People Lab, Aalborg University, Aalborg, Denmark*
[2] *Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China*
[3] *Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway*

Keywords: Visual Codebook, Bag of Words, Saliency Matching, Copy Detection.

Abstract: The importance of copy detection has led to a substantial amount of research in recent years, among which Bag of visual Words (BoW) plays an important role due to its ability to effectively handling occlusion and some minor transformations. One crucial issue in BoW approaches is the size of vocabulary. BoW descriptors under a small vocabulary can be both robust and efficient, while keeping high recall rate compared with large vocabulary. However, the high false positives exists in small vocabulary also limits its application. To address this problem in small vocabulary, we propose a novel matching algorithm based on salient visual words selection. More specifically, the variation of visual words across a given video are represented as trajectories and those containing locally asymptotically stable points are selected as salient visual words. Then we attempt to measure the similarity of two videos through saliency matching merely based on the selected salient visual words to remove false positives. Our experiments show that a small codebook with saliency matching is quite competitive in video copy detection. With the incorporation of the proposed saliency matching, the precision can be improved by 30% on average compared with the state-of-the-art technique. Moreover, our proposed method is capable of detecting severe transformations, e.g. picture in picture and post production.

1 INTRODUCTION

The rapid development of the Internet technology and the dramatic increase of the available network bandwidth have led to large quantity of videos and images being uploaded to websites. Despite of the convenience, it also brings a new challenge to detect unauthorized copies and further employ control policy to protect legal copyright. Copy detection has been seen as a key technique to solve this problem.

Various approaches based on different features have been proposed for video copy detection (Law-To et al., 2007) (Kim and Vasudev, 2005) (Gengembre and Berrani, 2008) (Poullot et al., 2010) (Douze et al., 2010). One of the most promising features is based BoW feature, which is effective in handling occlusion and some minor transformations, and can easily be scaled to large datasets. The key idea behind BoW representation is to quantize the high dimensional image features to a manageable vocabulary size of visual words, which is also called codebook. This is typically achieved by grouping the low-level features collected from frames into a specified number of clusters using an unsupervised algorithm such as k-means clustering. By treating the center of each cluster as a visual word in the codebook, each extracted feature can be mapped into its closest visual word. Thus a histogram over the codebook is generated as a representation of a frame and a further retrieval scheme based on bin-to-bin matching of their histograms are used for finding similar or copy images.

Despite of the promising progress, there are still many critical problems in Bow -based approaches that need to be addressed, one of which is the size of the vocabulary (or the size of the codebook). The size of the vocabulary applied in BoW approaches has an important impact on the retrieval performance. A general observation is that more visual words will result in better performance (Liu et al., 2008) (Nistér and Stewénius, 2006) owing to a better partitioned feature space. However, large vocabulary also has two other negative effects: large storage and computation resource requirement as well as low recall risk (Nistér and Stewénius, 2006). In contrast, a small vocabulary has many significant merits, e.g. efficient in feature assignment, more compact storage requirement for BoW descriptors and easier to scale up. Even though a small vocabulary has many advantages, it also contains other negative aspects: lacking discrimination due to the tendency of more conflicts among differ-

Ren H., Moeslund T., Tang S. and Ramampiaro H..
Small Vocabulary with Saliency Matching for Video Copy Detection.
DOI: 10.5220/0004280207680773
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 768-773
ISBN: 978-989-8565-47-1

ent features into the same cluster, which will bring high false positive rates and degrade its performance in video copy detection.

To obtain robust, yet discriminative features, and degrade high false positives in retrieval results, we propose a salient BoW feature representation and further a matching scheme to trackle the problems exists in small vocabulary. To be specific, we adopt Incremental Clustering algorithm in (Ren et al., 2012) for codebook construction because of its robustness in partitioning original features and transformed features. Further, we propose a visual words selection algorithm, during which only salient visual words are selected and kept for later descriptors. Finally, we provide a matching scheme to measure BoW features, by only calculating the similarity among salient visual words.

2 RELATED WORK IN BOW FEATURE REPRESENTATION

The performance of BoW feature for copy detection is highly dependent on codebook being built and the process of quantization. Thus, we give a briefly introduction of the related work in these two topics.

Usually the codebook is built through vector quantization based on k-means clustering. However, as discussed in (Nistér and Stewénius, 2006) the size of visual codebook should normally be very large to attain a reasonable retrieval performance Some researchers have started to improve the quality of codebook by visual words selection, e.g. (Mallapragada et al., 2010) (Wang, 2007) (Zhang et al., 2010). In (Mallapragada et al., 2010), the author select a subset of visual words, using a set of images, which is defined with related or unrelated link constraints. In (Wang, 2007) the authors present a boosting feature selection approach to select the most discriminative visual words from a multi-resolution codebook. However, these algorithms are supervised methods, which largely limit their scalability in a variety of applications. Thus in (Mallapragada et al., 2010), the authors propose a DCS algorithm, which model the relationship between the data matrix and the indicator matrix using a linear function and select visual words that led to minimal fitting error as discriminative ones. Other researchers have tried to introduce auxiliary information to reduce the effect of quantization. In (Philbin et al., 2007), the authors exploit spatial information by using the feature layout to verify the consistency of the retrieved images with the query region; In (Philbin et al., 2008) (Jiang et al., 2010), the authors adopt soft assignment, which assign a feature into several nearby visual words, to bring more possible similar features into consideration.

Feature assignment is quite time-consuming especially when a large vocabulary is adopted. Therefore, some tree structures have been proposed to speed up the assignment process. The vocabulary tree (Nistér and Stewénius, 2006) proposes a hierarchical quantization, by decomposing the large-scale clustering problem into smaller-scale clustering. Nonetheless, vocabulary tree suffers from performance degradation. Approximate k-means (AKM) (Philbin et al., 2007) is proposed to replace the exact search of the closest cluster centers by approximate search to speed up the assignment. However, it is not guaranteed to converge. Robust Approximate k-means (RAKM) (Li et al., 2010) is proposed to guarantee convergence through incorporation of the closest cluster centers in the previous iterations.

Although the tree structures can speed-up feature assignment, they also need a large vocabulary to achieve a good performance. Therefore these structures still require large storage, which, in turn, limits their scalability. To address these problems, we adopt the codebook construction algorithm in (Ren et al., 2012) to build a small vocabulary, and propose a saliency matching to attain a competitive performance comparable to large vocabulary.

3 SALIENCY MATCHING

Observed by many researchers, different visual words have distinct impacts on retrieval performance. Unlike many approaches which use weighting scheme to reinforce the effect of visual words extracted from foreground, we advocate a matching scheme only considering the coordinate of salient visual words, which could measure the similarity between BoW descriptors better. Since salient visual words are only a subset of codebook, the computation cost can be greatly saved. At the same time, frames which do not contain any salient visual words are considered as noisy frames and are filtered out to improve the retrieval performance. Salient visual words selection from the codebook and saliency matching between two videos are introduced respectively in the following sections.

3.1 Salient Visual Words Selection

Video Q is denoted as a series of frames: $Q[1], Q[2], ..., Q[M]$, each of which is represented as a BoW descriptor b_Q^i, M is the number of frames in

Q. Let $b^i_Q[w]$ be the w^{th} coordinate of the BoW descriptor in the i^{th} frame.

From observation, we could find out there are some regular patterns exist in the variation of $b^i_Q[w]$. We first select a reference video and its copy video (with the same duration), then trace the variation of BoW coordinates of different visual words across the video. Six representative visual words are selected, their variations in reference video and copy video are marked as point and cross marks in Figure 1. X axis represents the frame order, y axis represents the BoW coordinates of current visual word in the video. As can be found, the two variation curves tend to have a similar trend: When BoW coordinate of the reference video approaches a peak, the corresponding BoW coordinate in the copy video is also very high, probably also achieves a peak. However, this consistency pattern doesn't exist between two different videos, as seen in Figure 2. Based on this observation, the key issue is to detect these representative visual words and then find out a proper measurement to calculate the similarity between two videos by tracing their curves.

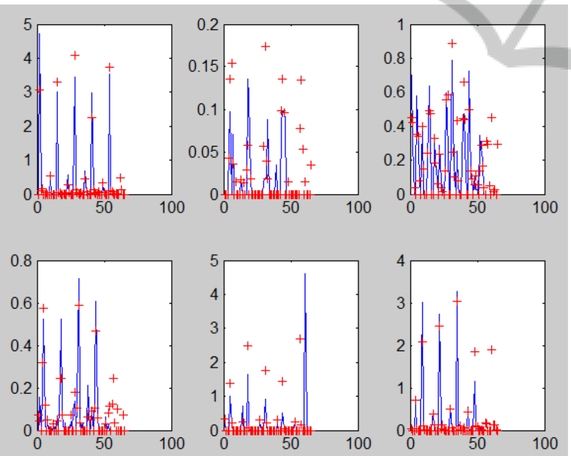


Figure 1: Comparison of BoW coordinate between copy video and its reference video.

To obtain this goal, we denote the curve in a trajectory form and then formulate the problem by tracing saliency points in the curve. By tracing $b^i_Q[w]$ across time, the variation of the w^{th} visual word in the video Q can be represented as a trajectory: $traj^w_Q = \{b^1_Q[w],...,b^M_Q[w]\}$. Hence a video could have at most K trajectories, where K is the size of the codebook. Unlike general trajectories, which trace matched key points among adjacent frames, our trajectory is built to find out *Salient Visual Words* in which certain "salient" visual content appear locally and continuously.

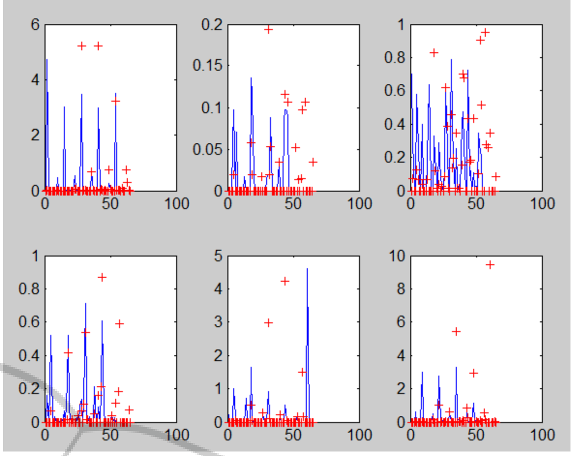


Figure 2: Comparison of BoW coordinate between two different videos.

Considering the variation of one visual word in the video, its trajectory has a locally asymptotic tendency .This phenomenon is due to the fact that the same visual word representing a specific kind of visual content which appeared in former frames, will continuously appear in the later frames. The coordinate may change under some transformations such as crop or shift. However, within a local time window, the neighborhood should have an asymptotic tendency towards one critical point. Normally this could not be a global tendency because through time some visual content may appear or disappear. Considering this local tendency, we could make an analogy of the variation of the coordinates as a dynamic system in which each point (the coordinate) changes similarly to the state in the dynamic system.

According to the manifestation of the trajectory, we define salient visual words as those whose trajectory across the video can be represented by a serie of locally asymptotically stable points, which are defined in definition 1.

Definition 1. Suppose a trajectory is represented as $\{x(t_0),x(t_1)...\}$. A point x^* is locally asymptotically stable at $t = t^*$ if it satisfies:

- x^* is stable and;
- x^* is locally attractive, e.g. there exists δ such that:

$$\|t-t^*\| < \delta \Rightarrow \lim_{t \to t^*} x(t) = x(t^*) \qquad (1)$$

In other words, the point x^* is locally asymptotically stable if x^* is locally stable and, furthermore, all solutions starting near x^* tend towards x^* as $t \to t^*$.

Then, we provide a practical method to find out these possible locally asymptotically stable points in trajectories. A stable point must be a maximum or minimum point, so the main problem is how to find stable points that all the nearby points tend to enclose.

According to the second method of Lyapunov and Lyapunovs stability theory, the original system is locally asymptotically stable if we can find a Lyapunov-candidate-function, which satisfies: 1) that the function is locally positive define and 2) time derivation of the function is locally negative semidefinite. Different from the goals in dynamic system, we aim to find out such locally asymptotically stable points assuming that the trajectory is locally asymptotically stable. Thus, we construct a candidate function and in reverse to find the points which is positive in first derivation and negative in second derivation. Actually this assumption is reasonable, because the points in the trajectory may have a fluctuation due to noise or transformations, but from a local view all the points in the neighborhood should have a tendency to a specific point. Naturally, we propose to use gradient function as a candidate function, which is approximated by first differential of BoW coordinate, due to that it can reflect the major fluctuations existing in visual content. After finding the extreme points of the gradient function, we approximate the time derivation of the gradient function by calculating, δ_t, by using second differential of BoW descriptors:

$$\delta_t = |b_Q^{i+1}[w] + b_Q^{i-1}[w] - 2 * b_Q^{i}[w]| \qquad (2)$$

Finally, we draw trajectories of salient and non-salient visual word individually in Figure 3 and 4. We select 5 videos including original video, two copy videos (by reencoding and quality degrade) and two non-reference videos. The curves for salient visual word in copy videos tend to have a high consistency with the original video, and are distinguishable from non-reference videos, while this tendency doesn't exist in non-salient trajectories. This implies that the matching of BoW descriptors on salient coordinates can provide a better solution for measuring the similarity between video and its copies.

3.2 Matching Scheme

Noisy frames are filtered out after salient visual words selection. For each left frame in the query video, we search for L nearest neighbors in the database to get matched pairs of frames, by computing the cosine distance of BoW descriptors. Note that only the coordinate of salient visual words are taken into calculation. This is because salient visual words tend to represent particular and discriminative visual content that appears or disappears through time, while non-salient visual words possibly represent background information or noise. Therefore, incorporating non-salient visual words will degrade the discriminative power and affect the reliability of similarity. A division of accumulated scores of matched frames by the number of matched frames, is used to measure the similarity between two videos.

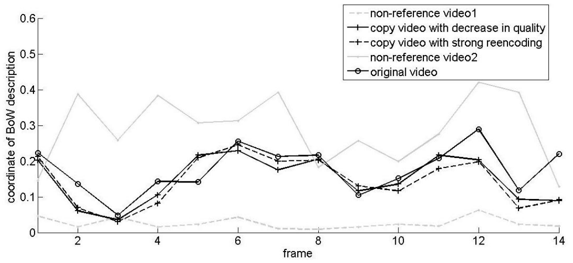


Figure 3: Salient visual word trajectories.

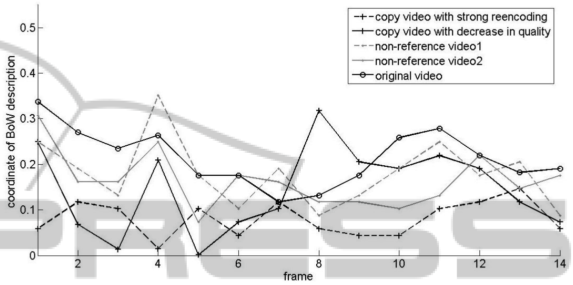


Figure 4: Non-salient visual word trajectories.

4 EXPERIMENTS

To verify the effectiveness of our saliency matching in copy detection, we compare our approach with state-of-the-art technique using a subset of TRECVID 2009 Copy Detection Dataset. There are 7 types of transformations in the dataset, including picture in picture, insertion of patterns, strong reencoding, change of Gamma, decrease in quality, post production and etc. Some of the representative transformations are shown in Figure 5, orignial images are on the left, their copy images under different transformations are shown on the right. To validate the performance under different transformations, we select two representative queries from each type. Accordingly, their reference videos are selected to form a video dataset - consisting 12 reference videos, each of which is about 30 minutes long on average.

We compare our saliency matching with the approaches proposed in (Jegou et al., 2008): HE +WGC, which is one of the state-of-the-art techniques in copy detection. We use HE + WGC to get matched pairs of frames by searching L nearest frames, accumulate the scores explained in Section 3.2 and calculate precision and recall to evaluate the performance of copy detection.

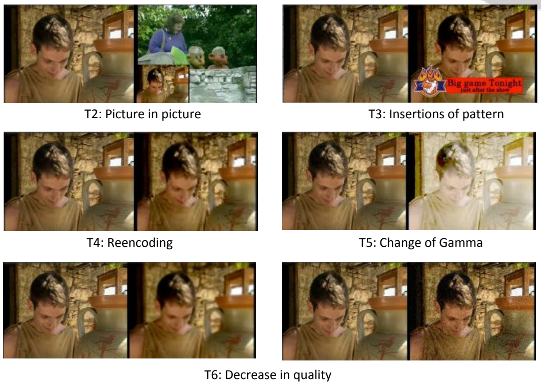
We first conduct experiments on our saliency

Table 1: Comparison experiments between TF-IDF based BoW matching and saliency matching.

TF-IDF based BoW matching			**Saliency matching**		
Precision	**Recall**	**F Score**	**Precision**	**Recall**	**F Score**
0.100	0.071	0.083	0.250	0.143	0.181
0.174	0.286	0.216	0.143	0.286	0.190
0.125	0.286	0.173	0.152	0.357	0.212
0.160	0.500	0.242	0.170	0.571	0.262
0.110	0.500	0.180	0.131	0.571	0.213
0.100	0.063	0.104	0.5	0.642	0.571
0.086	0.076	0.108	0.5	0.857	0.571

Table 2: Comparison experiments for copy detection.

Precision			**Recall**		
WGC+HE K=65	WGC+HE K=20k	SVW K=65	WGC+HE K=65	WGC+HE K=20k	SVW K=65
0.100	0.042	0.250	0.071	0.285	0.143
0.174	0.04	0.143	0.286	0.285	0.286
0.125	0.04	0.152	0.286	0.285	0.357
0.160	0.038	0.170	0.5	0.285	0.571
0.110	0.045	0.131	0.5	0.357	0.571
0.100	0.063	0.104	0.5	0.642	0.571
0.086	0.076	0.108	0.5	0.857	0.571



T2: Picture in picture

T3: Insertions of pattern

T4: Reencoding

T5: Change of Gamma

T6: Decrease in quality

Figure 5: Representative original images and copy images in TRECVID2009.

matching to get a good parameter of L. L determines the number of the matched frames. True positives may be missed under a small L, thereafter the recall is usually increased when enlarging the parameter L. However, with the increasing of L, the number of false negatives are also increased accordingly which leads to a low precision and longer processing time because of more nearest neighbours being searched. For example, when $L = 4$ the precision and recall are 0.17 and 0.57, respectively. However, when $L = 40$, they are 0.09, 0.643. With respect to execution time, after running 14 queries we found out that a query processing with $L = 4$ only needs 45.14s on average, while the processing time for $L = 40$ is 74.64s – i.e., 1.65 times longer. To get a balance between accuracy and time cost, we compute the precision and recall values by varying the values of L from 4 to 2000 and finally choose L=4.

Then, we compare saliency matching with and without frame filtering. During frame filtering, only discriminative frames are used to find L nearest neighbors to form matched pairs of frames. Discriminative frames turn out to be more contributive in copy detection, obtaining a precision and recall of 0.17 and 0.57 respectively, compared to 0.15 and 0.5 without frame filtering. The main reason is that false matched frames can be reduced and thus a higher precision is achieved.

Next, we compare our methods with TF-IDF weighting, which is a commonly used weighting scheme in BoW-based methods. To wipe off the impact of other factors, we use TF-IDF and saliency matching seperately to get matched frames, then we adopt matching scheme to compute the similarity of two videos. Determing whether a query video is a copy or not depends on threshold filtering of similarity score. By adjusting thresholds, comparative results can be seen in Table 1. As can be seen, our methods outperform TF-IDF weighting.

At last, the comparative copy detection performance between our methods and HE+WGC in terms of precision and recall is summarized in Table 2. We change the value of threshold that determines similarity between two videos, and compare the performance of the approach in (Jegou et al., 2008) using small and large vocabulary with our saliency matching using the same small vocabulary. As can be seen, our saliency matching using small vocabulary, not only outperforms HE + WGC using the same small vocabulary both in precision and recall, by 30% and 26% respectively on average, but also surpasses HE + WGC using large vocabulary by 240% and 13% respectively on average.

Our proposed approach performs significantly well in detecting some severe transformations, e.g. picture in picture and text insertion, in which new visual content appears continuously in a local time window In such cases, salient visual words can capture these discriminative visual information. However, it is not so promising in gamma transformation. Our conjecture is that saliency matching may not have obvious advantage in videos with insufficient salient visual words.

5 CONCLUSIONS

In this paper, we propose a salient visual words selection algorithm and saliency matching to measure the similarity between two videos. Due to the high consistency in salient coordinate of BoW descriptors, the number of mismatches will be reduced and the performance of BoW - based approaches with small vocabulary in copy detection can be improved by saliency matching.

Our experiments have shown that our proposed method performs well in copy detection, especially in transformations that visual content has significant changes along time. Despite the good performance in some sever transformations, e.g. picture in picture, post production and some combination of transformations, our methods do not perform well in gamma transformation, even when enlarging the number of matching frames. This is partly due to the insufficience of visual words, partly due to a loss of information during quantization in BoW features, which could be our future research directions.

REFERENCES

Douze, M., Jégou, H., Schmid, C., and Pérez, P. (2010). Compact video description for copy detection with precise temporal alignment. In *Proceedings of the 11th European conference on Computer vision: Part I*, ECCV'10. Springer-Verlag.

Gengembre, N. and Berrani, S. (2008). A probabilistic framework for fusing frame-based searches within a video copy detection system. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*. ACM.

Jegou, H., Douze, M., and Schmid, C. (2008). Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of the 10th European Conference on Computer Vision: Part I*. Springer-Verlag.

Jiang, Y., J. Yang, C. N., and Hauptmann, A. (2010). Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Transactions on Multimedia*, 12.

Kim, C. and Vasudev, B. (2005). Spatiotemporal sequence matching for efficient video copy detection. *IEEE Trans. Circuits Syst. Video Techn.*

Law-To, J., Chen, L., Joly, A., Laptev, I., Buisson, O., Gouet-Brunet, V., Boujemaa, N., and Stentiford, F. (2007). Video copy detection: a comparative study. In *CIVR*. ACM.

Li, D., Yang, L., Hua, X., and Zhang, H. (2010). Large-scale robust visual codebook construction. In *Proceedings of the international conference on Multimedia*. ACM.

Liu, D., Hua, G., Paul, A., and Tsuhan, C. (2008). Integrated feature selection and higher-order spatial feature extraction for object categorization. In *CVPR*. IEEE Computer Society.

Mallapragada, P., Jin, R., and Jain, A. (2010). Online visual vocabulary pruning using pairwise constraints. In *CVPR'10*. IEEE Computer Society.

Nistér, D. and Stewénius, H. (2006). Scalable recognition with a vocabulary tree. In *CVPR (2)*. IEEE Computer Society.

Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *CVPR*. IEEE Computer Society.

Philbin, J., Isard, M., Sivic, J., and Zisserman, A. (2008). Lost in quantization: Improving particular object retrieval in large scale image databases. In *In CVPR*.

Poullot, S., Buisson, O., and Crucianu, M. (2010). Scaling content-based video copy detection to very large databases. *Multimedia Tools Appl.*

Ren, H., Ramampiaro, H., Zhang, Y., and Lin, S. (2012). An incremental clustering based codebook construction in video copy detection. In *2012 IEEE Southwest Symposium on Image Analysis and Interpretation*. IEEE.

Wang, L. (2007). Toward a discriminative codebook: Codeword selection across multi-resolution. In *CVPR*. IEEE Computer Society.

Zhang, L., Chen, C., Bu, J., Chen, Z., Tan, S., and He, X. (2010). Discriminative codeword selection for image representation. In *ACM Multimedia*. ACM.