FAST AND ACCURATE CONTENT-BASED VIDEO COPY DETECTION USING BAG-OF-GLOBAL VISUAL FEATURES

Yusuke Uchida, Koichi Takagi, Shigeyuki Sakazawa

KDDI R&D Laboratories Inc.
2-1-15 Ohara, Fujimino-shi, Saitama, Japan

ABSTRACT

In this paper, we propose a fast, accurate content-based video copy detection scheme based on bag-of-global visual features, which is characterized by (1) utilizing an efficient DCT-sign-based feature for fast detection; (2) performing multiple assignment in the temporal domain, in addition to the feature and spatial domain to ensure repeatability in segment-level matching; and (3) adopting an inverse document frequency weighting and temporal burstiness-aware scoring to emphasize distinctive visual words. Despite detection 95 times faster than real-time, the proposed system achieves a false negative rate of 0.2% against queries that are altered by non-geometric transformations without any false positives.

Index Terms— Near-duplicate detection, content-based copy detection, inverted index, visual words

1. INTRODUCTION

As digital multimedia content, computer, and Internet technologies have become ubiquitous, digital videos have been used extensively in many applications. Copyright infringement poses a significant issue for one of the applications — online video-sharing services. Because many people upload video clips to these sites without proper copyright releases, an automated system that detects copies of copyrighted video is needed. In recent years, content-based video copy detection (CBCD) technology has attracted considerable research attention for this purpose. For an automated CBCD system to be usable, it must have the following properties:

- Computationally efficient: The system must be sufficiently efficient because many video clips are uploaded to video sharing sites every day.
- Robustness (low false negative rate): The video may have been subject to editing or degradation, including the addition of captions or patterns, a change of resolution, compression, and so on. The system should detect even these altered videos robustly.
- Low false alarms: A system with too many false detections is annoying and requires ongoing operator intervention to filter out the false alarms.

To detect video copies, there are mainly two cues: visual and/or audio information. Although both cues are equally useful, we focus on visual features in this paper, especially on global visual features. Instead of trying to handle too severe alterations such as camcording, we propose an efficient and effective CBCD system to efficiently filter out the majority of infringing video clips that have not been altered by geometric transformations. The proposed system satisfies the requirements described above by (1) utilizing an efficient DCT-sign-based feature for fast detection; (2) performing multiple assignment in the temporal domain, in addition to the feature and spatial domain to ensure repeatability in segment-level matching; and (3) adopting an inverse document frequency weighting and temporal burstiness-aware scoring to emphasize distinctive visual words (VWs), resulting in suppressing false positives.

2. RELATED WORK

CBCD schemes based on visual cues are roughly classified into two categories: one based on global features [1–4] and the other on local features [5]. Although local feature-based schemes are robust against even geometric transformations such as camcording, local feature detection, description, and matching processes remain highly time consuming [3]. In this paper, therefore, we focus on global feature-based schemes for practical use. An ordinal measure (OM) [1] is one of the major global descriptors, which has proven robust against changes in resolution or illumination. In [2], OM is extended to the spatiotemporal domain to capture temporal information. In [3, 4], OM features from a video clip are summarized into compact signatures for efficient retrieval. However, it is difficult for the clip-based methods [3, 4] to detect partial copies that are embedded in unrelated video clips. In [6], the bag-of-visual words (BoVW) framework [7], which is usually used for local features, is adapted to global features, in which multiple global features are extracted from predefined windows in a keyframe. Multiple assignment in the feature domain [8] is also performed to ensure repeatability of feature matching. However, multiple assignment in the feature and spatial domain is not the optimal choice in terms of the tradeoff between repeatability and filtering rate.

3. PROPOSED CBCD SYSTEM

In this section, we describe the proposed content-based video copy detection scheme based on bag-of-global visual features. The proposed scheme consists of the following steps: feature extraction, feature quantization, indexing based on an inverted index, and searching via the voting function. It realizes fast detection by accelerating feature extraction and quantization while achieving high detection accuracy and low false alarm rate owing to multiple assignment and sophisticated scoring of VWs.

3.1. Feature extraction and multiple assignment

Multiple assignment is powerful tool to improve repeatability of VW-based feature matching by assigning multiple VWs to a single feature or keyframe [6, 8]. In this paper, multiple assignment is defined to assign multiple VWs to a single, short segment, not to a keyframe. In this section, multiple assignments in the feature, spatial, and temporal domain are introduced. First, both reference and query video clips are divided into short segments with fixed durations in the temporal domain (0.3 sec in this paper). From each of the segments, fixed number 2^{mt} of frames are subsampled at a uniform interval (multiple assignment in the temporal domain). Subsequently, these subsampled frames are divided into 2^{ms} blocks[1] (multiple assignment in the spatial domain). Finally, feature vectors are extracted from these blocks. In this paper, we adopt the DCT-sign-based feature [9] as depicted in Figure 1; each block is resized into 8x8 pixels, and 2D-DCT is performed. Top-v AC coefficients in the zigzag scan order are used as a feature vector. Subsequently, they are quantized into a v-bit binary string by taking the sign of the AC coefficients. The resulting binary strings of length v define VWs with a size of $N = 2^v$. Multiple assignment in the feature domain can be performed by toggling the most unreliable mf-bits [10]. With the multiple assignment in the feature domain, each feature is assigned to 2^{mf} VWs. The reliability of each bit is defined by the absolute value of the corresponding AC coefficient. Finally, t-th reference segment is represented by $\mathcal{R}_t = (r_{t,1}, \cdots, r_{t,w}, \cdots, r_{t,W})$, where $W(= 2^{ms})$ denotes the number of blocks and $r_{t,w}$ denotes a set of VWs associated with w-th block. We also denote s-th query segment by $\mathcal{Q}_s = (q_{s,1}, \cdots, q_{s,w}, \cdots, q_{s,W})$. The parameters introduced above (mf, ms, mt) have a considerable impact on the performance of segment-level matching as shown in Section 4.1.

3.2. Indexing and searching inverted index

For simplicity, we explain the indexing and search step only when there is a single reference video clip. This limitation is easily overcome by considering reference video identifiers or by handling many video clips as a single, long video clip. In the indexing step, for each segment of reference video, the segment and block identifiers (t, w) are stored in the vw-th list of an inverted index for all $vw \in r_{t,w}$. In the search step, segment-level matching is efficiently performed by inverted index lookups. Two segments are matched if and only if they share the same VW(s) in at least one block. The function $m(\mathcal{Q}_s, \mathcal{R}_t)$ judges whether a query segment $\mathcal{Q}_s$ is matched with a reference segment $\mathcal{R}_t$:

$$m(\mathcal{Q}_s, \mathcal{R}_t) = \begin{cases} 1 & \text{if } {}^{\exists}w \text{ s.t. } q_{s,w} \bigcap r_{t,w} \neq \emptyset \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The indexing/searching process is summarized in Figure 2.

[1] Divied into 2x1, 2x2, 4x2, and 4x4 blocks for ms = 1, 2, 3, and 4.

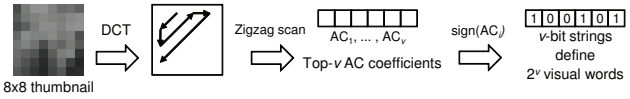

Fig. 1: Feature extraction procedure.

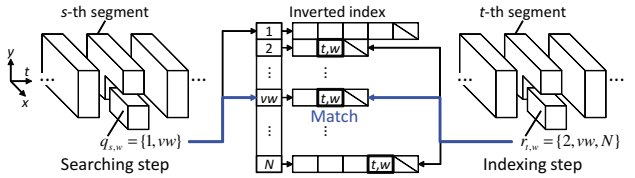

Fig. 2: Indexing and searching process using an inverted index structure.

3.3. Offset-level integration

Segment-level matching results obtained by inverted index lookups are integrated into offset-level results using a voting framework [6, 11]. Every matched segment pair $(\mathcal{Q}_s, \mathcal{R}_t)$ votes for the bin $\mathrm{B}[t-s]$ corresponding to the offset $t-s$ in a 1-D Hough space. In voting, since our scheme is based on the BoVW framework, the inverse document frequency (IDF) weighting [7] can be applicable to emphasize distinctive VWs. Though the IDF scoring has been used only for local features, experimental results in Section 4 show that it also works well for global features. Performing non-maxima suppression and thresholding to the voting table after voting, we obtain a set of offset hypotheses. Each hypothesis indicates the offset between copied segments in the query and reference clips. Each offset has segment-level matching results associated with the offset represented by a set of tuples (s, vw, w). After sorting the tuples according to a query segment identifier s, they are divided into groups to localize the copied segments. A sequence of the tuples are divided if successive two tuples (s, vw, w) and (s', vw', w') satisfy $s'-s>th$. Finally, the scores of the segmented tuples are calculated by summing up the IDF weights of VWs appearing in the tuples. Temporal burstiness-aware (TBA) scoring [6] is also adopted, in which individual VWs contribute to the score only once even if a VW is shared in consecutive query and reference segments. The beginning and ending timestamps of copied segments are calculated from min and max of s in the tuples.

4. EXPERIMENTAL EVALUATION

In this section, our CBCD system is evaluated using the TRECVID 2009 dataset. We chose the 2009 dataset rather than the most recent 2010 dataset because all queries in the more recent dataset have both video and audio, precluding evaluation of video-only queries. The 2009 dataset includes 838 reference videos (about 400 hours in total) and 1,407 query videos. Each query has been edited by the seven transformations listed in Table 1, including both photometric transformations and geometric transformations. In the framework of the TRECVID CBCD task, a CBCD system is characterized by three key performance measures[2]:

- Detection accuracy: Normalized detection cost rate (NDCR) measures the tradeoff between the cost of false negatives and false positives, and is defined by a weighted mean of the two errors.
- Localization accuracy: The accuracy of localization is measured by the F-measure, which is the harmonic mean of the precision and recall of the detected copy location relative to the true video segment.
- Efficiency: Efficiency is evaluated by the mean processing time per query.

The following experiments were performed on a machine with a Core i7 970 CPU and 24 GB of main memory.

4.1. Tradeoffs between repeatability and filtering rate

First, to confirm the effectiveness of multiple assigment in the temporal domain, we introduce two measurements to evaluate multiple assignment: repeatability (RP) and $1 - \text{filteringrate}$ ($\overline{FR}$). RP represents the probability that a query frame is matched with the groundtruth reference segment. Larger RP tends to result in low false-negative rates in detection. $\overline{FR}$ represents the rate of the number of matched reference frames against the number of all reference segments N. Smaller $\overline{FR}$ tends to result in low false-positive rates. There is a tradeoff between RP and $\overline{FR}$: a larger $\overline{FR}$ results in a larger RP in general. Similar discussions in terms of approximate nearest neighbor search are found in [12].

Table 1: Query transformations.

T2	Picture in picture
T3	Insertions of pattern
T4	Strong re-encoding
T5	Change in gamma
T6	Decrease in quality (combinations of 3 transformations from blur, gamma, frame dropping, contrast, compression, ratio, and noise)
T8	Post production (combinations of 3 transformations from crop, shift, contrast, caption, flip, insertion of pattern, and picture in picture)
T10	Combinations of 5 transformations from T2-T8

[2] http://www-nlpir.nist.gov/projects/tv2009/Evaluation-cbcd-v1.3.htm

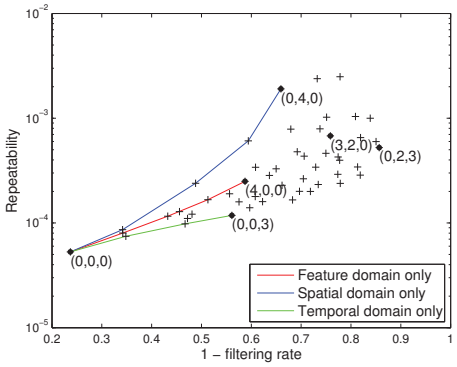

Fig. 3: The tradeoff between repeatability and filtering rate under different multiple assignment settings. From upper left to lower right, the tradeoff improves.

RP and $\overline{FR}$ are defined using Equation 1 as

$$RP = \frac{1}{S}\sum_{s=1}^{S} m(\mathcal{Q}_s, \mathcal{R}_{gt(s)}), \quad (2)$$

$$\overline{FR} = \frac{1}{ST}\sum_{s=1}^{S}\sum_{t=1}^{T} m(\mathcal{Q}_s, \mathcal{R}_t), \quad (3)$$

where $gt(s)$ denotes the identifier of the ground truth segment associated with s-th query segment.

We exhaustively evaluate multiple assignments using a subset of the reference videos and simulated queries that have been randomly extracted from the reference videos and altered by a random transformation chosen from contrast change ($\pm 50\%$), gamma change ($\pm 50\%$), and strong compression. Figure 3 shows the performance of the combination of multiple assignments parameterized by (mf, ms, mt). Abbreviating $mf + ms + mt$ to m, at most 2^m VWs are assigned to a single segment in the multiple assignment process explained in Section 3.1. Because the memory requirement increases as m increases, we set the constraint $m \leq 5$ here. It is found that multiple assignment in the temporal domain is most effective, and multiple assignment in the feature and spatial domain follows. This is mainly because multiple assignment in the feature and spatial domain always increases the number of assigned VWs at a constant rate, while the number of assigned VWs adaptively changes according to scenes in the multiple assignment in the temporal domain because repeated VWs in a segment are ignored. In noisy or dynamic scenes, a larger number of VWs is assigned to the segment, while multiple assignment in the temporal domain does nothing in a static scene.

4.2. Detection accuracy

The proposed systems are evaluated in terms of detection accuracy. Table 2 shows the resulting NDCR measures of the proposed systems for different video transformations. `Base1` and `Base2` represent the proposed system with multiple assignment defined by the parameter $(3, 2, 0)$ and $(0, 2, 3)$, respectively. These parameters are chosen to maximize repeatability among parameters w/o and w/ multiple assignment in the temporal domain. The system performing multiple as-

Table 2: NDCR scores for different systems and transformations (lower is better).

	T2	T3	T4	T5	T6	T8	T10
Base1	1.000	0.291	0.425	0.448	0.418	0.963	0.993
Base2	1.000	0.201	0.246	0.276	0.075	0.933	0.985
+IDF	1.000	0.052	0.112	0.142	0.007	0.910	0.910
+TBA	1.000	**0.007**	**0.000**	**0.000**	**0.000**	0.843	0.821
[5]	0.672	0.224	0.381	0.239	0.284	0.269	0.515
[6]-1	**0.134**	0.067	0.045	0.082	0.433	0.567	0.470
[6]-2	0.239	**0.007**	0.060	0.022	0.022	**0.231**	**0.269**

Table 3: Localization accuracy (F-measure, higher is better).

	T2	T3	T4	T5	T6	T8	T10
Prop.	0.000	**0.977**	**0.967**	**0.961**	**0.976**	0.883	0.847
[5]	0.64	0.89	0.84	0.92	0.83	0.88	0.82
[6]-1	0.937	0.938	0.936	0.940	0.939	0.936	0.941
[6]-2	**0.960**	0.952	0.949	**0.961**	0.946	**0.957**	**0.956**

signment in the temporal domain (Base2) achieved better NDCR value by improving the tradeoff between repeatability and filtering rate, as discussed in Section 4.1. We can also see that adopting IDF weighting (+IDF) and TBA scoring (+TBA) drastically improves detection accuracy. The full proposed system (+TBA) achieves an NDCR score of 0.002 on average in transformations T3 to T6, which corresponds to a false negative rate of 0.2% (one false negative against 536 positive examples) without any false positives.

In Table 2, the proposed system is also compared to the system that had achieved the best performance in TRECVID'09 CBCD task [5] and one of the state-of-the-art systems [6]. [6]-1 and [6]-2 represent a system based on global features and a system based on both local and global features described in [6], respectively. The full proposed system outperforms conventional systems in transformations T3 to T6. As the other transformations include geometric alterations, it is inherently difficult for global features to handle these transformations. Some preprocessing, such as picture in picture detection [6] or letter box detection [3], can alleviate the problems.

4.3. Evaluations based on other criteria

Table 3 shows the localization accuracy of our system and the systems described in [5] (estimated from the literature) and [6]. It is shown that our schemes have also achieved good performance on segment localization criterion in transformations T3 to T6 compared to conventional schemes, owing to the high repeatability in segment-level matching. Table 4 shows the processing time required for each step in the proposed system: decoding a query video clip, resizing query frames, feature extraction and quantization, and voting. It is shown that the most time-consuming processes are voting and decoding. In other words, the search process is as fast as decoding in our system. The system requires only 0.94 seconds to process a query video clip with a duration of 90 seconds on average, which is about 95 times faster than real-time, owing to fast feature extraction and quantization. The system in the literature [5] reported a much longer processing time of over 200 seconds, which is mainly because their system is based on local features. The system in the literature [6] requires around 15 and 121 seconds in global feature-based system and in local and global feature-based system, respectively.

Table 4: Processing time required in the proposed system [sec].

Decode	Resize	Feature	Voting	Total
0.420	0.023	0.007	0.488	0.938

5. CONCLUSION

In this paper, we proposed an efficient and effective CBCD system. The proposed system achieved a false negative rate of 0.2% against queries that were altered by non-geometric transformations without any false positives. As the proposed system is lightweight (95 times faster than real-time), it can be efficiently combined with other systems, such as local feature-based or audio feature-based systems, which are complementary to global feature-based systems.

6. REFERENCES

[1] X. Hua, X. Chen, and H. Zhang, "Robust video signature based on ordinal measure," in *Proc. of ICIP*, 2004, pp. 685–688.

[2] C. Kim and C. Vasudev, "Spatiotemporal sequence matching for efficient video copy detection," *IEEE Trans. on CSVT*, vol. 15, no. 1, pp. 127–132, 2005.

[3] S. Paisitkriangkrai, T. Mei, J. Zhang, and X. S. Hua, "Scalable clip-based near-duplicate video detection with ordinal measure," in *Proc. of CIVR*, 2010, pp. 121–128.

[4] L. Shang, L Yang, F. Wang, K. P. Chan, and X. S. Hua, "Real-time large scale near-duplicate web video retrieval," in *Proc. of MM*, 2010, pp. 531–540.

[5] Z. Liu, T. Liu, and B. Shahraray, "At&t research at trecvid 2009 content-based copy detection," in *Proc. of TRECVID*, 2009.

[6] Y. Uchida, M. Agrawal, and S. Sakazawa, "Accurate content-based video copy detection with efficient feature indexing," in *Proc. of ICMR*, 2011.

[7] J. Sivic and A. Zissermane, "Video google: A text retrieval approach to object matching in videos," in *Proc. of ICCV*, 2003, pp. 1470–1477.

[8] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. of CVPR*, 2008, pp. 1–8.

[9] J. Barr, B. Bradley, and B. T. Hannigan, "Using digital watermarks with image signatures to mitigate the threat of the copy attack," in *Proc. of ICASSP*, 2003, pp. 69–72.

[10] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *Proc. of ISMIR*, 2002, pp. 107–115.

[11] J. Law-To, L. Chen, A. Joly, and I. Laptev, "Video copy detection: a comparative study," in *Proc. of CIVR*, 2007, pp. 371–378.

[12] L. Paulevé, H. Jégou, and L. Amsaleg, "Locality sensitive hashing: A comparison of hash function types and querying mechanisms," *Pattern Recognition Letters*, vol. 31, no. 11, 2010.