FAST AND ROBUST CONTENT-BASED COPY DETECTION BASED ON QUADRANT OF LUMINANCE CENTROID AND ADAPTIVE FEATURE COMPARISON

Yusuke Uchida, Masayuki Hashimoto, Ryoichi Kawada

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

ABSTRACT

This paper proposes a fast and robust content-based copy detection scheme. Our proposal consists of a new compact feature, efficient keyframe selection and adaptive mask-based feature comparison. Firstly, a block-level luminance centroid is binarized into a 32-bit quadrant feature for fast and robust feature comparison. Subsequently, a new keyframe selection method is adopted to enhance pairwise independence between unrelated video segments in addition to choosing stable keyframes. Finally, a block-level mask-based feature comparison method is introduced to compare only stable features. Experimental results show our scheme improves recall by 0.1 at the same precision 0.9 and the processing speed in feature comparison of the proposed scheme is about twice as fast as that of conventional schemes.

Index Terms— video fingerprinting, content-based copy detection, near-duplicate detection, luminance centroid

1. INTRODUCTION

With the advancement of both computer and Internet technology, digital multimedia content is being used more widely in many applications. Video sharing services, one such application, have become very popular and have attracted a great deal of attention. One big problem with these video sharing services, however, is copyright infringement. As many people upload infringing video clips to video sharing sites without proper authorization, an automated means of detecting such clips is needed. In recent years, Content-Based Copy Detection (CBCD) technology, more generally referred to as *digital fingerprinting*, has attracted considerable research attention for this purpose. In an automated inspection system based on CBCD, content holders register copyrighted content with the operators of video sharing sites in advance. The operators extract features from the copyrighted content and store them in a database. When a user uploads a video clip, features are also extracted from the uploaded video clip in the same way and the database is searched for a match. If there is matching content in the database, the uploaded content is considered to be a copy of copyrighted content and filtered out or some other action is taken according to the content holder's intentions. Considering a CBCD system in practical terms, robustness and computational efficiency are vital because many video clips are uploaded to video sharing sites every day and these uploaded video clips are subject to various kinds of transformations such as compression, contrast changes, etc. In this paper, we focus on improving detection accuracy and speed in CBCD for efficient copyright protection.

2. RELATED WORKS

To date, many algorithms have been developed in the CBCD area. CBCD schemes can be roughly classified into two categories: one based on global features [1–4] and the other on local features [5]. Although one state-of-the-art scheme based on local features has achieved high accuracy and an efficient database search [5] in terms of content-based image retrieval, local feature detection and description [6] processes remain highly time-consuming (in the order of a few seconds per frame on an ordinary PC). In this paper, therefore, we focus on schemes based on global features for practical use. An ordinal measure (OM) [1, 2], one of the major global descriptors, has proven robust against changes in resolution or illumination. Recently, it has been revealed that gradient based features [3, 4] achieve good robustness and pairwise independence. Among them, the block-wise orientation of luminance centroid (OLC) is shown to have optimum performance [4].

3. PROPOSED APPROACH

The most significant problem of the aforementioned conventional schemes based on global features is the fact that the detection accuracy deteriorates when a copied video clip has been distorted, particularly by geometric transformations. In this paper, we overcome this problem via the following three approaches:

- Binarizing frame features into 32-bit signatures and comparing them in terms of Hamming distance, which prevents the disparity between a copied video clip and its original video segment from becoming excessive due to distortion.
- Selecting stable and distinctive keyframes in order to accomplish both robustness and pairwise independence.
- Filtering out the features that are sensitive to distortion, which improves accuracy, especially in recall.

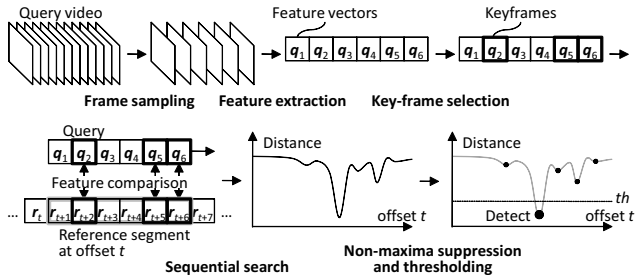

Fig. 1: Framework of the proposed scheme.

As shown in Fig. 1, the proposed scheme is briefly summarized as follows. 1) The query video is resampled at a fixed frame rate to deal with frame rate changes. 2) Robust binary features based on block-wise luminance centroid are extracted from every frame of the resampled query video. 3) Stable and distinctive keyframes are selected. 4) Reference indices are sequentially searched using a block-level mask, which enables only distinctive blocks to be compared. 5) Non-maxima suppression and thresholding are performed to determine whether the query content is infringing or not. If so, the offset of the copied segment in reference videos is estimated. In the remainder of this section, feature extraction, keyframe selection and sequential search procedures are all detailed.

3.1. Quadrant of the luminance centroid feature

We propose a binary feature based on the quadrant representation of the luminance centroid (QLC) for fast and robust retrieval. Fig. 2 shows the procedure of the QLC feature. First, each frame is divided into 4x4 blocks B_i $(1 \leq i \leq 16)$, for each of which the coordinate of the luminance centroid (x_i^c, y_i^c) is then calculated:

$$x_i^c = \frac{\sum_{(x,y)\in B_i} x \cdot I(x,y)}{\sum_{(x,y)\in B_i} I(x,y)}, \quad y_i^c = \frac{\sum_{(x,y)\in B_i} y \cdot I(x,y)}{\sum_{(x,y)\in B_i} I(x,y)}, \tag{1}$$

where $I(x,y)$ means the luminance of an image at coordinate (x,y). Subsequently, a 2-bit signature is extracted by comparing the coordinates of the luminance centroid (x_i^c, y_i^c) and the center of the block (x_i^m, y_i^m). This extreme quantization prevents features from being severely changed by distortion. Integrating these signatures from all blocks, the QLC feature $\mathbf{f}$ is created:

$$\mathbf{f} = (x_1, y_1, \cdots, x_{16}, y_{16}), \tag{2}$$

$$x_i = \begin{cases} 1 & \text{if } x_i^c \geq x_i^m \\ 0 & \text{else} \end{cases}, \quad y_i = \begin{cases} 1 & \text{if } y_i^c \geq y_i^m \\ 0 & \text{else} \end{cases}. \tag{3}$$

This quadrant representation and the mask-based feature comparison explained in section 3.3 resolve the drawbacks of the OLC feature illustrated in Fig. 3. One drawback is the fact that features θ and $\pi - \theta$ are confusing (on the left in Fig. 3) because OLC represents the orientation of luminance centroid by arcsin. This restriction is to enable OLC features to be compared in Euclid space, while the QLC feature is

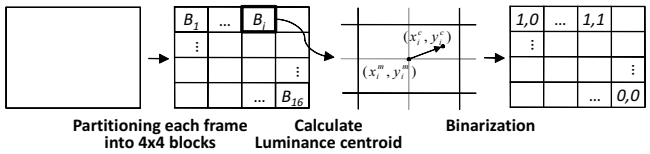

Fig. 2: Feature extraction in the proposed scheme.

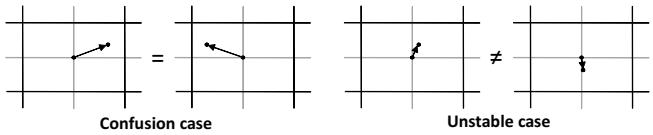

Fig. 3: Simple illustration of OLC drawbacks.

compared in the Hamming space and can distinguish the centroids illustrated on the left in Fig. 3. The other drawback is the fact that the OLC feature would change significantly by distortion, resulting in a significant performance degradation when the luminance centroid is located near the center of the block, e.g., θ changes from $\pi/2$ to $-\pi/2$ (on the right in Fig. 3). On the other hand, the QLC feature do not change significantly due to binary representation. This QLC feature is a very compact signature and requires only 32 bits per frame, while OM [1] needs 45 bits and OLC 128 bits. If properly encoded OM is applied, the bit amount can be reduced to $\lceil log_2(9!) \rceil = 19$ bits. However, decoding computation is additionally required and the resultant search speed is degraded.

3.2. Stable and distinctive keyframe selection

Let $\mathcal{Q} = (\mathbf{q}_1, \cdots, \mathbf{q}_N)$ denote a feature set extracted from a query video clip where N is the number of frames in the query video clip after resampling. Because $\mathcal{Q}$ includes redundant and noisy frames, M keyframes are selected and only keyframes are compared in a sequential search for a stable and distinctive feature comparison. A keyframe set $\mathcal{K} = (k_1, \cdots, k_M)$ is selected by maximizing the following expression:

$$\text{maximize}(A(\mathcal{K}) - \alpha B(\mathcal{K})), \tag{4}$$

$$A(\mathcal{K}) = \textstyle\sum_{j=1}^{M-1} \text{ham}(\mathbf{q}_{k_j}, \mathbf{q}_{k_{(j+1)}}), \tag{5}$$

$$B(\mathcal{K}) = \textstyle\sum_{j=1}^{M} \max(\text{ham}(\mathbf{q}_{(k_j)-1}, \mathbf{q}_{k_j}), \text{ham}(\mathbf{q}_{k_j}, \mathbf{q}_{(k_j)+1})), \tag{6}$$

where $\text{ham}(\mathbf{f}_1, \mathbf{f}_2)$ is the Hamming distance between features $\mathbf{f}_1$ and $\mathbf{f}_2$, which is defined as:

$$\text{ham}(\mathbf{f}_1, \mathbf{f}_2) = \text{bitcount}(\mathbf{f}_1 \oplus \mathbf{f}_2). \tag{7}$$

$\text{bitcount}(\cdot)$ counts the number of '1' bit, which is efficiently calculated using a lookup table for every 16 bits. This optimizing problem can be quickly solved by dynamic programming. The term $B(\mathcal{K})$ represents the distance between keyframes and their neighboring frames. By minimizing this, the keyframes become stable against the time lag between the copied video clip and the original video. This has an effect similar to the keyframe selection method described in [7], where frames are divided into sub-groups and keyframes are selected for each sub-group for which the distances to any

of the other frames in the sub-group are minimized. Furthermore, in the proposed method, the term $A(\mathcal{K})$ imposes that consecutive keyframes should be different from each other, resulting in selecting diverse keyframes and boosting pairwise independence in the feature comparison.

3.3. Mask-based feature comparison

In this section, we introduce a mask-based feature comparison method, whereby a comparison of only stable features (bits) is attempted instead of using the ordinary Hamming distance. The mask $\mathbf{m}$ is created from the coordinates of the luminance centroids:

$$\mathbf{m} = (a_1^*, b_1^*, \cdots, a_{16}^*, b_{16}^*), \tag{8}$$

$$a_i^* = \begin{cases} 0 & \text{if } |x_i^c - x_i^m| \leq \beta\sigma_x \\ 1 & \text{else} \end{cases}, \tag{9}$$

$$b_i^* = \begin{cases} 0 & \text{if } |y_i^c - y_i^m| \leq \beta\sigma_y \\ 1 & \text{else} \end{cases}, \tag{10}$$

where β is the adjustable parameter determining the mask strength and σ_x (σ_y) is the standard deviation of $x^c - x^m$ ($y^c - y^m$), which is obtained from video clips that differ from reference videos. Using mask $\mathbf{m}$, the modified Hamming distance between feature $\mathbf{f}_1$ and $\mathbf{f}_2$ is defined as

$$\mathrm{ham}^*(\mathbf{f}_1, \mathbf{f}_2, \mathbf{m}) = \mathrm{bitcount}((\mathbf{f}_1 \oplus \mathbf{f})_2 \wedge \mathbf{m}), \tag{11}$$

where $\wedge$ is an AND operator. In the proposed scheme, the mask is created from the query feature set $\mathcal{Q}$, which means no additional information is stored in the database concerning the mask-based feature comparison. Let $\mathcal{R}_t = (\mathbf{r}_{t+1}, \cdots, \mathbf{r}_{t+N})$ denote a reference feature set at offset t and $\mathcal{M} = (\mathbf{m}_1, \cdots, \mathbf{m}_N)$ denote a mask set created from the query feature set $\mathcal{Q}$. Subsequently, the distance between $\mathcal{Q}$ and $\mathcal{R}_t$ is defined as the sum of the modified Hamming distances over the keyframes with normalization by the number of bits '1' in $\mathcal{M}$:

$$\mathrm{dist}(\mathcal{Q}, \mathcal{R}_t, \mathcal{M}, \mathcal{K}) = \frac{\sum_{j\in\mathcal{K}} \mathrm{ham}^*(\mathbf{q}_j, \mathbf{r}_{j+t}, \mathbf{m}_j)}{\sum_{j\in\mathcal{K}} \mathrm{bitcount}(\mathbf{m}_j)}. \tag{12}$$

Here we simply describe $\mathrm{dist}(\mathcal{Q}, \mathcal{R}_t, \mathcal{M}, \mathcal{K})$ as $d(t)$. After calculating $d(t)$ over the reference index, non-maxima suppression with window size w and thresholding with threshold th is performed for the final results. All offsets $\hat{t}$ that satisfy the following expression are regarded as the beginning points of the copied segment:

$$d(\hat{t}) \leq \min(d(s), th) \;\; (\hat{t} - w \leq s \leq \hat{t} + w). \tag{13}$$

4. EXPERIMENTAL RESULTS

We conducted experiments to evaluate the proposed scheme in terms of robustness and computational cost. The experimental environment is described below:

- Testbed: the following experiments were tested on a machine with a Core 2 Quad 3GHz CPU and 8GB main memory using Windows XP. All experiments were performed with a single thread.
- Reference: 200 hours of video (720x480, 29.97fps) from various movie genres were used as reference videos. Features were pre-extracted and stored in the main memory for sequential search.
- Query: we created 100 queries of 60 seconds in duration, which were randomly extracted from reference videos and edited by particular transformations.
- Evaluation criteria: each scheme was evaluated in terms of accuracy (Precision-Recall curve with different thresholds) and computational cost (average processing time per query). In this paper, the allowable error of offset $\hat{t}$ is set to 2 seconds.

For query generation, the following four transformations were used:

- Cropping: crop each side by 20px.
- Contrast: enhance contrast with parameter 1.2.
- Rotation: rotate +3 degrees.
- Compression: resize to CIF and encode to DivX@256kbps.

4.1. Evaluation of the proposed scheme

In this experiment, each part of the proposed scheme was evaluated. For this purpose, the following five schemes were compared:

- OLC: OLC feature with 4x4 blocks proposed in [4].
- QLC: QLC feature proposed in this paper.
- QLC+K: QLC combined with the keyframe selection described in section 3.2.
- QLC+K2: QLC combined with the keyframe selection described in [7].
- QLC+K+M: QLC+K combined with the mask-based feature comparison described in section 3.3.

The parameters used in this experiment were set as: $N = 120$, $M = 30$, $\alpha = 2.0$, $\beta = 0.2$ and $w = 10$ (α and β were determined by preliminary experiments). Fig. 4 shows the PR curve of each scheme against each query. The common characteristics independent of the query type are described in the following. Comparing OLC and QLC, it is confirmed that the QLC feature improves performance especially in terms of recall. This improvement is mainly owing to the binary feature representation detailed in section 3.1, which prevents the QLC feature from significant change by distortion. While both keyframe selection methods (QLC+K and QLC+K2) improve performance compared with QLC, the proposed keyframe selection method is shown to be more efficient. It mainly contributes to precision because by selecting distinctive frames as keyframes, the distance between two different video clips increases. The mask-based feature comparison also improves accuracy, especially in recall. This is because by avoiding comparison of unstable bits, the distance between the copied video clip and its original video content diminishes. It is remarkable that the full proposed

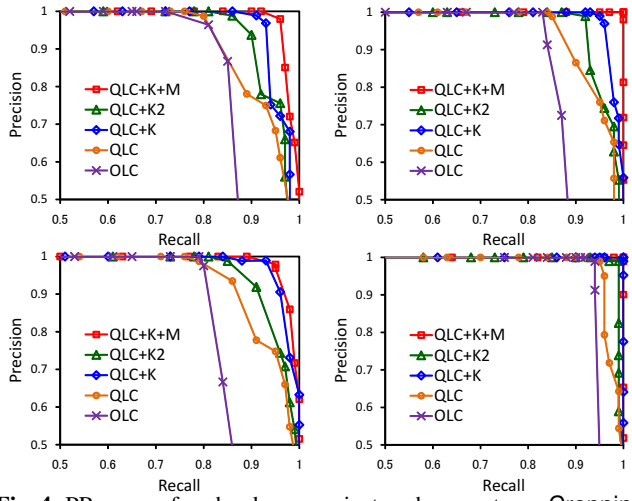

Fig. 4: PR curve of each scheme against each query type: Cropping (top-left), Contrast (top-right), Rotation (bottom-left) and Compression (bottom-right).

Table 1: Comparison of the computational cost of each scheme [sec/query].

	Feat.	Key.	Mask	Feat. comp.	Total
OLC	0.146	-	-	1.163	1.309
QLC	0.169	-	-	0.284	0.453
QLC+K	0.169	0.002	-	0.284	0.455
QLC+K2	0.169	0.001	-	0.284	0.454
QLC+K+M	0.169	0.002	0.001	0.353	0.525

scheme (QLC+K+M) achieves good robustness even against geometrical transformations (e.g., 0.96 in F-measure against the rotation query).

Table 1 shows the computational time of each scheme. The QLC feature also accelerates search speed, even if the mask-based comparison is adopted because distance calculation is implemented by bitwise operation and LUT. It is also shown that the processing overhead of the keyframe selection and mask generation in the proposed scheme is negligible.

4.2. Comparison with conventional schemes

In this section, the proposed scheme is compared with the conventional schemes described below:

- PROP: full proposed scheme.
- OM: ordinal intensity feature [1] with 3x3 block division.
- OM+K2: OM combined with keyframe selection described in [7] for fair comparison.
- OLC: orientation of luminance centroid [1] with 4x4 block division.
- OLC+K2: OLC combined with keyframe selection.

Fig. 5 shows the average PR curves of each scheme for four transformations. Although keyframe selection improves the accuracy of both the OM and OLC schemes, the proposed scheme still outperforms conventional schemes. Comparing PROP and OLC+K2 at the same precision of 0.9, the proposed scheme achieved higher performance than OLC+K2 by about 0.1 in recall.

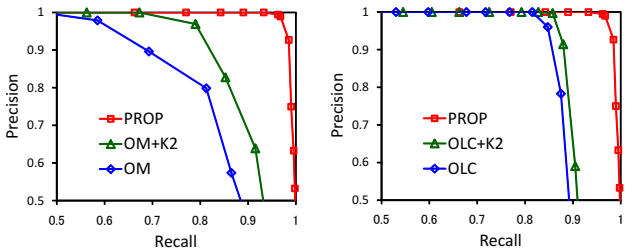

Fig. 5: Average PR curve of each scheme against four types of transformations.

Table 2: Comparison of computational cost among PROP, OM, OM+K2, OLC and OLC+K2 [sec/query].

	Feat.	Key.	Mask	Feat. comp.	Total
PROP	0.169	0.002	0.001	0.353	0.525
OM	0.025	-	-	0.638	0.663
OM+K2	0.025	0.001	-	0.638	0.664
OLC	0.146	-	-	1.163	1.309
OLC+K2	0.146	0.001	-	1.163	1.310

Computational costs are compared in Table 2. As shown in the table, though the proposed scheme requires additional cost in extracting features, it achieves the fastest processing speed. The proposed scheme is especially fast in feature comparison (about twice as fast as OM), which becomes critical when searching a huge reference database.

5. CONCLUSION

In this paper, we proposed a fast and robust content-based copy detection scheme based on the quadrant representation of luminance centroid and adaptive mask-based feature comparison for the automatic detection of infringing video clips. Experimental results show our scheme improves the recall by 0.1 at the same precision of 0.9 and the processing speed in feature comparison of the proposed scheme is about twice as fast as that of conventional schemes.

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] M. Usman and C. Kim, "Real time video copy detection under the environments of video degradation and editing," in *Proc. of ICACT*, 2008, pp. 1583–1588.

[3] S. Lee and C. D. Yoo, "Robust video fingerprinting for content-based video identification," *IEEE Trans. on CSVT*, vol. 18, no. 7, pp. 983–988, 2008.

[4] S. Lee and Y. H. Suh, "Video fingerprinting based on orientation of luminance centroid," in *Proc. of ICME*, 2009, pp. 1386–1389.

[5] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proc. of ECCV*, 2008, pp. 304–317.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[7] Z. Xu, H. Ling, F. Zou, Z. Lu, P. Li, and T. Wang, "Fast and robust video copy detection scheme using full DCT coefficients," in *Proc. of ICME*, 2009, pp. 434–437.