A Rotation Invariant Descriptor for Robust Video Copy Detection

Shuqiang Jiang[1,2], Li Su[3], Qingming Huang[3], Peng Cui[1,2], Zhipeng Wu[3]

[1]Key Lab of Intell. Info. Process., Chinese Academy of Sciences, Beijing 100190, China
[2]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[3]Graduate School of Chinese Academy of Sciences, Beijing, 100049, China
E-mail: { sqjiang, lsu, qmhuang, pcui, zpwu }@jdl.ac.cn

Abstract. A large amount of videos on the Internet are generated from authorized sources by various kinds of transformations. Many works are proposed for robust description of video, which lead to satisfying matching qualities on Content Based Copy Detection (CBCD) issue. However, the trade-off of efficiency and effectiveness is still a problem among the state-of-the-art CBCD approaches. In this paper, we propose a novel frame-level descriptor for video. Firstly, each selected frame is partitioned into certain rings. Then the Histogram of Oriented Gradient (HOG) and the Relative Mean Intensity (RMI) are calculated as the original features. We finally fuse these two features by summing HOGs with RMIs as the corresponding weights. The proposed descriptor is succinct in concept, compact in structure, robust for rotation like transformations and fast to compute. Experiments on the CIVR'07 Copy Detection Corpus and the Video Transformation Corpus show improved performances both on matching quality and executive time compared to the pervious approaches.

Keywords: CBCD, frame-level descriptor, rotation invariant, HOG

1 Introduction

Recently, online video websites such as YouTube, Yahoo!, Google, etc. have taken the fancy of the users for the convenience of browsing large amount of videos for free. Meanwhile, many illegal copies without authorities are uploaded. To protect intellectual properties, Content Based Copy Detection (CBCD) issue aroused researchers' great interest. As the TRECVID'08 CBCD evaluation plan [1] describes, a copy is not an exact duplicate but a segment derived from original document, with some transformations such as cropping, recoding, flipping, inserting pictures, etc. Therefore, a copy usually differs from its corresponding resource in both format and content, which makes CBCD a challenging task.

Many approaches towards CBCD have obtained satisfying results. It can be concluded that an effective descriptor is the key point in CBCD systems. As [2] pointed out, an effective descriptor must have the following two advantages: *robustness* and *discriminability*, which make the descriptor invariant to various transformations generated from the original source. Besides, facing to the exponential growth of digital video resources, CBCD approaches call for a *fast* and *compact* video descriptor, which is important in a user-oriented system.

Numerous descriptors have been proposed to meet these advantages mentioned above. Previous methods are mostly based on video keyframes. Researchers study physical global features, such as color moment and histogram [3, 4] in keyframes to deal with large corpus. These global features are compact and simple, but they suffer from serious problems like brightness changes and fail in more complex tasks. Recently, local features especially Local Interest Points (LIPs) are brought forth as effective methods to describe frames and images. In [5], Zhao *et al.* match LIPs with PCA-SIFT description and introduce fast filtering procedure. Wu *et al.* treat every keyframe as a Bag of visual Words (BoW), which are regarded as the clustering centers of LIPs [6]. Ngo *et al.* detect and localize partial near-duplicates based on LIPs matching between the frames rearranged by time stamps [7]. Local feature description of keyframes has shown its invariant property to many kinds of transformations, however, there are still some problems: a) Extraction and matching process of LIPs is particularly time consuming; b) There exist many mismatched LIPs which will greatly degrade the final performance; c) All these approaches depend on a robust keyframe extraction scheme.

On noticing these problems that the keyframe based approaches faced with, researchers propose some non-keyframe approaches instead. Kim *et al.* extract ordinal signatures from the video clip and propose a spatio-temporal sequence matching solution [8]. Wu *et al.* introduce self-similarity matrix for robust copy detection [9]. In [2], Yeh *et al.* contribute a frame-level descriptor which encoded the internal structure of a video frame by computing the pair-wise correlations between pre-indexed blocks. These descriptors are compact in structure and retain the most relevant information of a frame/clip. Besides, they are suitable to be integrated into a fast copy detection scheme.

Motivated by the non-keyframe approaches, in this paper, we propose a novel frame-level descriptor which combines the Histogram of Oriented Gradient (HOG) [10] and the Relative Mean Intensity (RMI) together by means of a weighting scheme. A frame is partitioned into rings which are invariant for the transformations such as rotation and flipping. Besides, instead of treating each frame as a whole, using a series of rings can save the local patterns and further make the descriptor more discriminative. RMI of each ring represents the bottom physical feature of a frame, and HOG, which is well-known for counting occurrences of orientation in localized portions of an image, has been improved in this paper with RMI as the weight. Compared with existing representations of video, the proposed descriptor offers the following advantages:

- **Succinct** in concept: combination of two naive features;

- **Compact** in structure: encodes each frame at a certain sample rate;
- **Invariant** for common transformations: lighting, flipping, rotation, etc.;
- **Fast** in extraction and matching procedure.

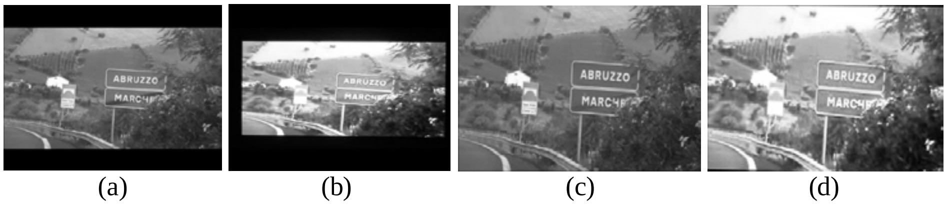
The remaining of the paper is organized as follows. In section 2, we detail the extraction and matching process of the proposed descriptor. Section 3 shows the experimental results on CIVR'07 corpus and Video Transformation Corpus. Finally, we conclude the paper with future work in section 4.

2 Descriptor Extraction

2.1 Preprocessing

2.1.1. Frame Border Removal

Border is a common trick made in a copy [12]. For each frame, we are interested in the significant content without borders. Besides, the intensities of the border are useless in frame analysis, as shown in Fig. 1. We adopt a simple method, which removes the first few lines of each direction (left, right, top, bottom) whose sum of intensity is less than a threshold (20% of the maximum in this paper). Fig. 1 (c) and (d) show the results of frames after border removal of (a) and (b).



(a) (b) (c) (d)

Fig. 1. Frames Border removal. (a) and (b) are frames from the source video and query video. (c) and (d) are results of boder removal for (a) and (b).

2.1.2. Frame Resizing

As [1] describes, copies may have different sizes with the original source. Here we employ a linear interpolation process to resize the query frames to the same size with its reference. This process is necessary because different sizes may cause different forms of the descriptors.

2.2. Frame Description

2.2.1. Extraction Process

The descriptor is extracted by encoding the pixel information of each frame. For a given video, with overlaps, we segment it into clips, which are the basic processing units in our approach. Fig. 2 shows an example. For each frame *Fr* in a video clip, we divide it into n_L rings. In Fig. 3 (a), the area between two white circles is called a ring. A ring reserves the RGB intensities of symmetrical positions in a frame, which makes it invariant for rotation and flipping.

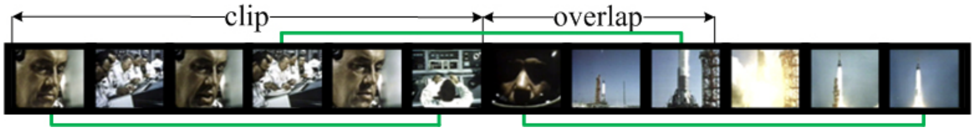


Fig. 2. Video clips and overlaps.

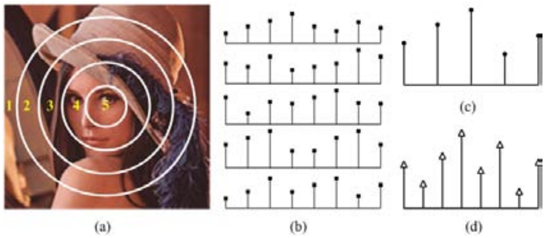


Fig. 3. Illustration of the extraction process, n_L=5, n_G=12.(a) is the original frame with 5 rings; (b) HOG of each ring; (c) is the relative mean intensity of each ring; (d) is the final descriptor.

Next, for the i^{th} ring of a frame, the relative mean intensity (RMI) is calculated:

$$RMI(i) = \sum_{p \in ring(i)} p(x,y) \Big/ \sum_{p \in Fr} p(x,y) \qquad (1)$$

where *p(x,y)* stands for the intensity of point *(x,y)*.

From equation (1), it is easily concluded that RMI is a global feature of each ring. It represents the *intra*-ring information and can help maintain a similar *inter*-ring relationship between the query video and the reference. Besides, it is not sensitive to some complex brightness changes.

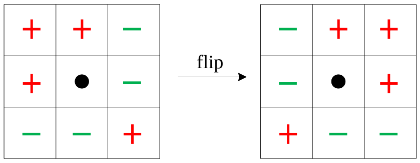


Fig. 4. Flipping frames with different gradients maps

To describe the local distribution of a frame, we adopt HOG. HOG is a widely used feature for object detection, especially for human detection [11]. For each point of a ring, n_G (chosen as even number) gradient orientations are calculated. HOGs of each ring are illustrated in Fig. 3 (b). As can be seen from Fig. 4, if the query is flipped from the reference, the gradient orientations are opposite. To avoid this change, instead of directly using the gradient orientations, we divide their absolute values into certain number of bins. With the increasing of bin number n_G, the discriminative power of HOG increases. However, the computation complexity also rises, and it will enlarge the influence of noise. We need to combine HOG and RMI into the video description to promote the discriminability.

To combine these two features, we sum the n_L HOGs with each RMI of the same ring as the weight:

$$\begin{aligned}\mathbf{D}_{nG\times 1} &= \mathbf{HOG}_{nG\times nL} \times \mathbf{RMI}_{nL\times 1} \\ &= \begin{pmatrix} HOG_{11} & HOG_{12} & \cdots & HOG_{1,nL} \\ HOG_{21} & HOG_{22} & \cdots & HOG_{2,nL} \\ \vdots & \vdots & \ddots & \vdots \\ HOG_{nG,1} & HOG_{nG,2} & \cdots & HOG_{nG,nL} \end{pmatrix}_{nG\times nL} \times \begin{pmatrix} RMI_1 \\ RMI_2 \\ \vdots \\ RMI_{nL} \end{pmatrix}_{nL\times 1}\end{aligned} \tag{2}$$

D is the final descriptor. Different from the traditional HOG, it involves the intensity and inner distribution of each frame. From the calculating process, we find that **D** is overall a global descriptor with the length of n_G. It encodes the inner relationship of a frame and the local changes of intensities. Fig. 10 shows more examples.

2.2.2. Matching Process

To match two descriptors $\mathbf{D_1}$ and $\mathbf{D_2}$, we choose χ^2 distance as the similarity metric. In the matching process, a double-minimization process is employed in matching process [9]. In the first step, for an input query video, we find the clips with the minimal distance (maximal similarity) between descriptors in each source video. Then, we select the one with the lowest distance in the source. This process is illustrated in Fig. 5.

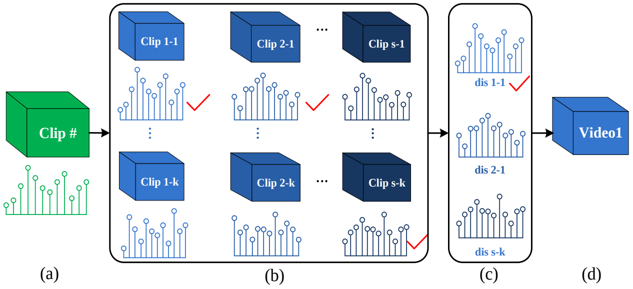


Fig. 5. A double-minimization process for matching. Each cube represents a video or a video clip, the diagram below each cube is the descriptor. (a): Query clip; (b): Finding the minimal distance in each source video; (c): Finding minimal distance among all the source videos; (d): Matching result.

3 Experimental Results

3.1 Dataset

Experiments are conducted using the CIVR'07 Copy Detection Corpus (MUSCLE VCD) [13].The CIVR'07 corpus are based on two tasks: task 1 retrieves copies of whole long videos while task 2, a much harder task, detects and locates the partial-duplicate segments from all the videos. The source data contains about 100 hours of 352×288 videos. These videos come from web, TV archives and movies, and cover documentaries, movies, sport events, TV shows and cartoons. Meanwhile, there are 15 queries for ST1 with different transformations like change of colors, blur, recording with an angle and inserting logos. There are also 3 queries for ST2 with transformations mentioned above.

The Video Transformation Corpus (VTC) [14] aims at recognizing transformations happened in video copies. It consists of ten types: Analog VCR Recording, Blur, Caption, Contrast, Picture in Picture, Crop, Ratio, Resolution Reduction, Adding Noise and Monochrome. There are 300 sources and 20 queries for each kind of transformation. Besides, we also add the CIVR'07 corpus into the mentioned transformations.

According to the evaluation plot [13], criterion of the detection scheme is defined as:

$$Matching\ \ Quality = \frac{Num\ Of\ \ Corrects}{Num\ Of\ \ Queries} \tag{3}$$

To implement fast copy detection, we down sample the frames in a fixed rate: we select one frame every 10 frames. Noticing that the adjacent frames always conserve similar content information, our sample strategy is proved to be efficient and effective by the experiments.

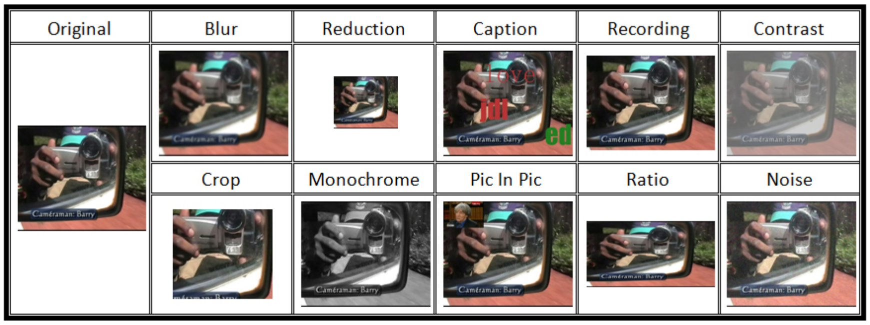


Fig. 6. Example of ten transformation types

3.2 Parameters

We choose ring number (n_L) and HOG bin number (n_G) as parameters in the experiments.

- $\boldsymbol{n_L}$: As n_L gets larger, the descriptor will possess more discriminate power. However, while the dimension increases, the descriptor is more sensitive to the border removal technique which may lower down the overall performance.
- $\boldsymbol{n_G}$: Theoretically, with more HOG bins, more information can be captured by the descriptor. However, for the existence of noise, the performance will be degraded if there are too many bins.

Table1 shows the results of different parameters on CIVR'07 corpus.

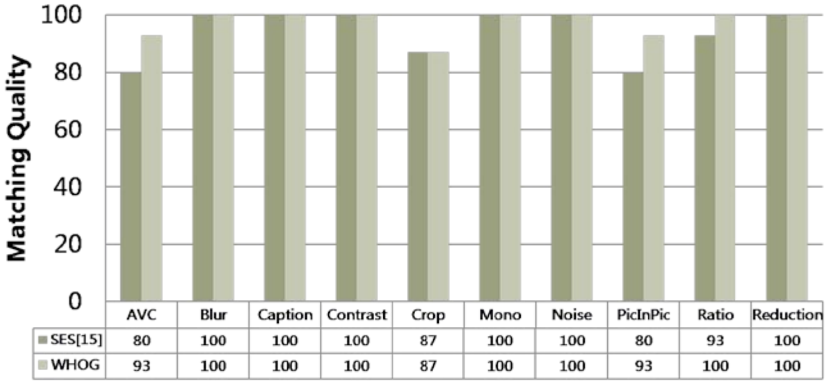
3.3 Detection Results

3.3.1 Experiment-I: Matching Qualities

Without loss of generality, we set the parameters (n_L and n_G) to be 4/8/12/16. Table 1 lists the matching qualities with the corresponding value. Table 2 shows the results of the CIVR'07 corpus with the comparison to some existing approaches and Fig. 7 shows the performance on the Video Transformation Corpus (VTC).

Table 1. Matching qualities with different n_L and n_G for CIVR'07 ST1

n_L n_G	4	8	12	16
4	93%	93%	100%	100%
8	93%	100%	100%	100%
12	93%	100%	100%	100%
16	93%	100%	100%	100%



	AVC	Blur	Caption	Contrast	Crop	Mono	Noise	PicInPic	Ratio	Reduction
■ SES[15]	80	100	100	100	87	100	100	80	93	100
■ WHOG	93	100	100	100	87	100	100	93	100	100

Fig. 7. Matching qualities of the VTC compared with our previous work [15]

Table 2. Matching qualities of the CIVR'07 corpus

Approach	ST1	ST2
CIVR'07 Teams[13]	46~87%	17%~87%
Yeh *et al*. [2]	93%	73%
Previous Work [15]	100%	80%
Ours	100%	87%

3.3.2 Experiment-II: Comparison of the proposed descriptor, HOG and RMI

In this experiment, we test the RMI, traditional HOG and the proposed descriptor in CIVR'07 ST1. According to the result, it is clear that the proposed descriptor shows improved performance than the versions only using RMI and HOG.

3.3.3 Experiment-III: Executive time

Table 3 shows the matching time of different approaches. According to the table, the proposed approach runs faster than previous works.

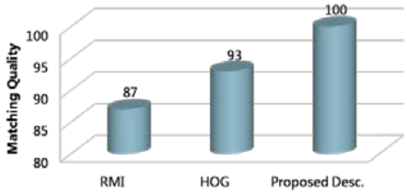


Fig. 8. Comparison of RMI, HOG and the proposed descriptor on ST1

Table 3. Executive time of different approaches (in seconds)

Approach	ST1	ST2
Yeh *et al*. [2]	1394	570
Previous work [15]	849	368
Ours	69	44

3.3.4 Experiment-IV: Similarity Metric

Besides χ^2 distance, we also try Euclidean distance, Cosine distance and Correlation Coefficient. The proposed descriptor is robust enough for different similarity metrics, among them χ^2 distance performs better and we employ it as the metric in this paper. Fig. 9 illustrates the different matching results using the above four metrics.

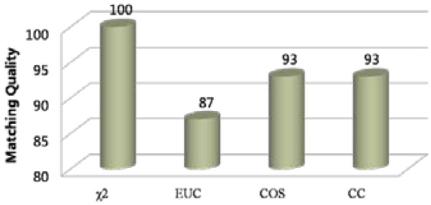


Fig. 9. Matching qualities under different measurement metrics

These experiments show that the descriptor is effective and efficient. But the length of the basic processing unit (see 2.2.1) is pre-fixed without further consideration. As long as we obtain a robust shot detection scheme, it can be easily extended to shot-based descriptor, which will help promote the segment accuracy of CIVR'07 ST2 and more complex partial near duplicated videos.

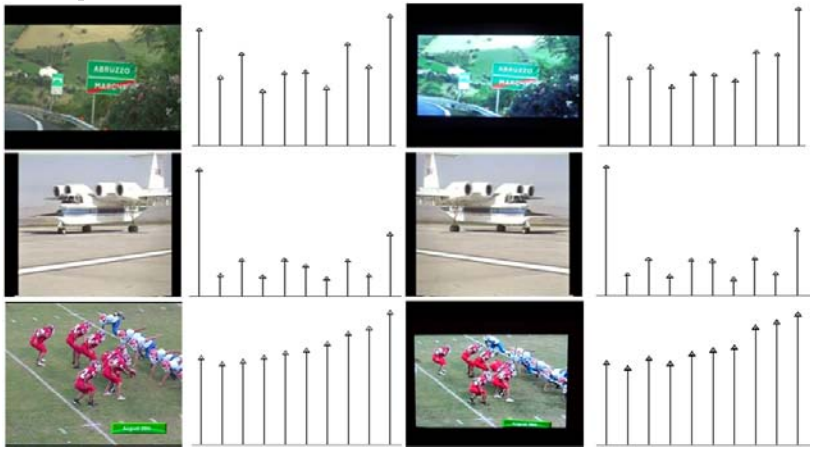


Fig. 10. Descriptors of similar frames generated by different transformations. The left columns are the sources and their descriptors, the right columns are the queries.

4 Conclusions

In this paper, we propose a novel descriptor by combining two common applied characteristics, relative mean intensity (RMI) and histogram of gradient (HOG). RMI represents the global intensity level while HOG describe the change of each pixel. As an effective descriptor, it is succinct in concept, compact in structure, robust for transformation and fast to compute. We adopt the χ^2 distance as the metric of similarities between two descriptors. Results on the CIVR'07 corpus and Video Transformation Corpus show the promotion on matching quality.

In future work, we aim to mine a similar clip-level descriptor, which can be treated as a cube of feature and it can draw correlations between adjacent frames. Another direction is to fuse this descriptor into keyframe based approach to further accelerate the CBCD framework.

Acknowledgements. This work was supported in part by National Natural Science Foundation of China: 61025011, 60833006, 61070108 and 61001177, and in part by Beijing Natural Science Foundation: 4092042.

References

1. TRECVID, http://www-nlpir.nist.gov/projects/trecvid.
2. M. Yeh and K. Cheng, “Video copy detection by fast sequence matching,” In Proc. Of ACM Int. Conf. on Multimedia, pp. 633-636, Oct. 2009.
3. A. Qamra, Y. Meng, and E. Y. Chang, “Enhanced perceptual distance functions and indexing for image replica recognition,” In IEEE Trans. Pattern Anal. Mach. Intell.,vol. 27, no. 3, pp. 379-391, Mar. 2005.
4.M. Bertini, A. D. Bimbo, and W. Nunziati, “Video clip matching using MPEG-7 descriptors and edit distance,” In Proc. of the ACM Int. Conf. on Image and Video Retrieval, pp. 133-142, 2006.
5. W. L. Zhao, C. W. Ngo, H. K. Tan, and X. Wu, “Near-duplicate keyframe identification with interest point matching and pattern learning,” In IEEE Trans. On Multimedia, vol. 9, no. 5, pp. 1037-1048, Sep. 2007.
6. X. Wu, W. L. Zhao, and C. W. Ngo, “Near-duplicate keyframe retrieval with visual keywords and semantic context,” In Proc. Of ACM Int. Conf. on Image and Video Retrieval, pp. 162-169, 2007.
7. H. Tan and C. W. Ngo, “Scalable detection of partial near-duplicate videos by visual-temporal consistency,” In Proc. of ACM Int. Conf. on Multimedia, pp. 145-154, Oct. 2009.
8. C. Kim and B. Vasudev, “Spatio-temporal sequence matching for efficient video copy detection,” In IEEE Trans. on Circuits and Systems for Video Technology, vol. 15, no. 1, pp.127-132, Jan. 2005.
9. Z. P. Wu, Q. M. Huang, and S. Q. Jiang, “Robust copy detection by mining temporal self-similarities,” In IEEE Int. Conf. on Multimedia and Expo, 2009.
10. http://en.wikipedia.org/wiki/Histogram_of_oriented_gradients
11. N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” In IEEE Int. Conf. on Computer Vision and Pattern Recognition, pp.886-893, 2005.
12. J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, and F. Stentiford, “Video copy detection: a comparative study,” In Proc. of the ACM Int. Conf. on Image and Video Retrieval, pp. 371-378, 2007.
13. MUSCLE-VCD-2007, http://www-rocq.inria.fr/imedia/civrbe-nch/index.html.
14. Z. P. Wu, S. Q. Jiang, and Q. M. Huang, “Near-Duplicate Video Matching with Transformation Recognition,” In proceedings of the ACM International Conference on Multimedia, pp. 549-552, 2009.
15. P. Cui, Z. P. Wu, S. Q. Jiang, and Q. M. Huang, “Fast Copy Detection Based on Slice Entropy Scattergraph,” In IEEE Int. Conf. on Multimedia and Expo, Singapore, pp.149-154, Jul.19-23, 2010.