Format-Independent Motion Content Description based on Spatiotemporal Visual Sensitivity

Xuefeng Pan, Jintao Li, *Member*, *IEEE*, Yongdong Zhang, Sheng Tang and Lejun Yu

Abstract — *With the extensive application of digital video technology, developing format-independent motion describing method is of great significance for retrieving and searching video content in different formats. In this paper, a format-independent motion describing method for video content is proposed. The features based on visual sensitivity are extracted from spatiotemporal slice. Since the same video content gives the same visual stimuli to visual perception, the method based on this kind of visual sensitivity related feature is format-independent. The experiments show the feature proposed is sensitive to the variation of video content and robust to the variation of video format. The motion describing method proposed is format-independent*[1].

Index Terms — visual sensitivity, format-independent, motion content description, spatiotemporal slice

I. INTRODUCTION

With the extensive application of digital video technology, more and more video contents are encoded in different formats to meet different requirements. How to retrieve and search the video content in different formats is important for video application. The motion information of video content is widely used in video indexing, retrieval and annotation [3][4][5][8]. Kinds of motion descriptors have been proposed to fulfill different requirements. Some visual motion descriptors are included as a part of MPEG-7 standard [1]. Developing format-independent motion describing method is of great significance for retrieving and searching video content in different format. Nevertheless, this issue has not been extensively discussed.

Some block motion vector based describing methods have been proposed for MPEG video. A temporal and a spatial motion activity descriptor based on P frame macroblock were presented in [2]. Peker used average magnitude and average temporal derivative of motion vectors to analyze motion activity in [3]. Tan used motion vectors to estimate camera motion in [4]. Zhu classified camera motion using motion vectors to develop an efficient browsing, summarizing and retrieving system for video content in [5]. Some other motion describing methods are pixel based. Most of them are used to estimate the dominant motion of two adjacent video frames. Dufaux and Konrad proposed an efficient global motion estimating method using gradient descent over a pyramid of images in [6]. Keller used a small subset of the original image pixels and an interpolation-free formulation to decrease the computational complexity of global motion estimation in [7]. Besides these, Ngo used the structure tensor and tensor histogram to classify the camera motion types and separate the motion layers in [8].

However, the two kinds of methods are not suitable for format-independent video motion description. Firstly, the block motion vector information is hard to obtain in non-MPEG video. Secondly, the pixel based technique is not robust to pixel value variation in encoding procedure.

Three kinds of format-independent video copy detection methods were analyzed in [9]. The ordinal measure (OM) has shown superior performance than motion and color based features for format-independent video content representing. Furthermore, the OM is combined with temporal trail of frames to gain better performance in [10] by Kim. But according to visual perception theory, the motion information is very important to human visual perception. Since same video content gives same stimuli to visual perception, we think the visual sensitivity related motion features will be remained no matter the content be encoded in what format.

In this paper, we propose a format-independent motion describing method based on visual sensitivity using spatiotemporal slice DCT. The low-frequency AC DCT coefficients of slice are taken as motion description for video content. Experiments show the proposed describing method is sensitive to variation of video content and is robust to variation of video format. The comparison further indicates the motion describing method proposed in this paper is more effective than the temporal OM based method proposed in [10] in format-independent video copy detection.

The rest of this paper is organized as follows. Section 2 introduces the proposed approach in detail. Experiments and comparison results are given in Section 3. Then in Section 4, the proposed approach is concluded.

II. MOTION CONTENT DESCRIPTION BASED ON SPATIOTEMPORAL VISUAL SENSITIVITIES

Because the same video content gives the same visual stimuli no matter in what format, the visual sensitivity based features are quite stable during format variation. So this kind of features is suitable for format-independent motion describing.

[1] This work was supported in part by the Beijing Science and Technology Planning Program of China (D0106008040291), the Beijing Municipal Science & Technology Project (Z0004024040231) and the National Nature Science Foundation of China (60473002 and 60302028)

The authors are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences. P.O.Box 2704, Beijing, 100080, P. R. China.

Xuefeng Pan is also with the Graduate University of Chinese Academy of Sciences (e-mail: xfpan@ict.ac.cn).

Contributed Paper
Manuscript received April 14, 2007

A. Visual Sensitivity and Motion Stimuli

Firstly, the relation between visual sensitivity and motion stimuli is introduced. By assuming that the visual system is optimized to process natural images in real world, a relationship was generated for the power spectrum of natural time-varying image sequence (R), the visual sensitivity (K) and the noise power (N) in [11]:

$$K = \frac{(1/R)^{1/2}}{(1+N/R)^{3/2}} \tag{1}$$

If an object at distance r from the observer is moving with a relative velocity v, the power spectrum of the time-varying image sequence showing this scene is given in [12]:

$$R(f,\omega,v,r) = R_s(f)\sigma(\omega - f\cdot v/r) \tag{2}$$

Where f is spatial frequency and ω is temporal frequency; $\sigma(\omega - f\cdot v/r)$ is the Dirac delta function which is zero everywhere except for $\omega = f\cdot v/r$; and $R_s(f)$ is the spatial power spectrum of the image representing the static object, which is given by $R_s(f) = C/f^m$ according to [13]. Further research show the power spectrum of natural time-varying images sequence is [12]:

$$R(f,\omega) = \frac{C}{f^{m+1}}F(\frac{\omega}{f}) \tag{3}$$

Where C is a constant and $F(\omega/f)$ is some function of the ratio ω/f. This show the power spectrum of natural image sequence is non-separable in space and time. This is the motivation of using spatiotemporal slice in this paper.

B. Visual Sensitivity based Motion Feature Extracting via Spatiotemporal Slice DCT

In order to extract spatiotemporal coupling features, we use spatiotemporal slice to analyze video content. Spatiotemporal slice is a finite plane in the 3-D spatiotemporal image volume at a given position with a given orientation [14]. In this paper, the row at the middle of frame is used to cascade horizontal temporal slice as shown in Fig. 1, because the main subject of video content often occurs at the middle region. The column at the middle of frame also can be used to cascade vertical temporal slice.

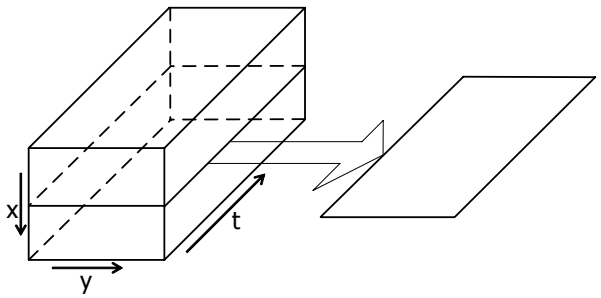

Fig. 1. Extract spatiotemporal slice from video.

As a collection of scans at the same position of every frame in time sequence, spatiotemporal slice contains spatiotemporal coupling features. The coordinate system defined as Fig. 2 is used to study the spatiotemporal coherency of slice. In which S is the spatial dimension and T is the temporal dimension.

Fig. 2. A part of spatiotemporal slice from a soccer video clip. The original slice is converted into gray image for display.

With a $W\times H$ slice, we resize it to $W\times N$. Then the slice is converted to gray image and segmented into $N\times N$ blocks. We call this kind of block *Sub* (Slice Unit Block). Then, discrete cosine transform (DCT) is used to analysis *Sub* in frequency domain to generate visual sensitivity related features. The output of a 2-D FDCT (forward DCT) is a set of coefficients which can be considered as weights of a set of standard 'basis functions' [15]. 8×8 DCT basis functions are given in Fig. 3. From the top-left, moving to the right, the functions contain increasing horizontal spatial frequency; moving down, the functions contain increasing vertical spatial frequency; moving diagonally to the right and down, the functions contain both horizontal and vertical frequencies.

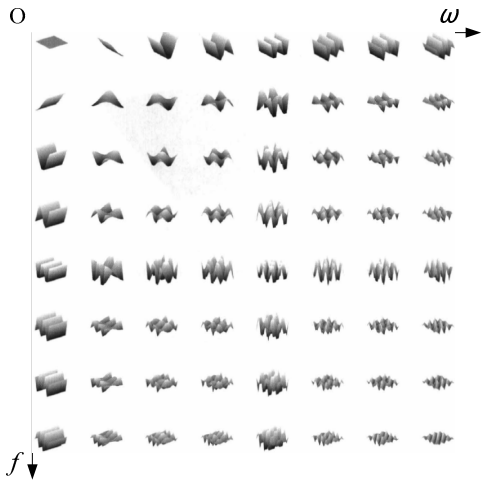

Fig. 3. 8×8 DCT basis functions [15] and the frequency in spatial and temporal dimension of *Sub*

In the DCT basis functions for *Sub*, the frequency in temporal dimension ω increases in horizontal direction, the frequency in spatial dimension f increases in vertical direction as shown in Fig. 3. Therefore, the coefficients of different DCT basis can represent the value of different ω/f. As shown in (1) and (3), the contrast sensitivity K is a function of ω/f. So the coefficients of different DCT basis for *Sub* can be used as sensitivity related features of video content.

We use two examples to illustrate how sensitivity related motion information is reflected with slice DCT. Firstly, consider the situation of a vertical bar moving from left to right as shown in Fig. 4(a). An 8×8 picture is used to portray

this situation as shown in Fig. 4(b). Let the bar move to right at the speed of one pixel per-frame (denoted as speed=1), the horizontal slice of frame sequence is shown in Fig. 4(c). When the speed is changed to two pixel per-frame (denoted as speed=2), the horizontal slice of frame sequence is as Fig. 4(d).

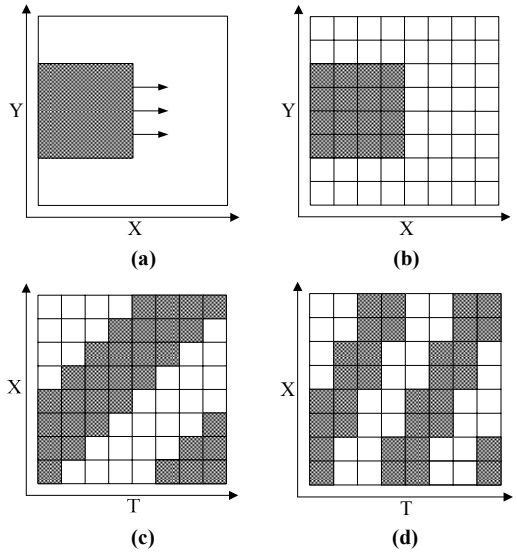

Fig. 4. A vertical bar moves from the left to the right. (a) A picture of moving process; (b) The picture of 8×8; (c) The horizontal slice when the bar moving with speed=1; (d) The horizontal slice when the bar moving with speed=2.

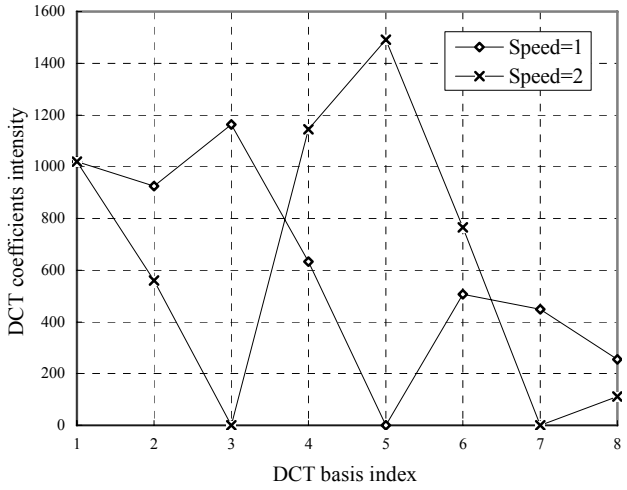

Fig. 5. The coefficients intensity of DCT basis in temporal dimension. The first DCT basis represents the DC component. The second basis represents the basic frequency. The third basis is double of the basic frequency. The fifth basis is four times of the basic frequency.

Then 2-D 8×8 DCT is used to analyze the motion information represented by the slices. The DCT coefficients intensity of every basis in temporal dimension is given in Fig. 5. When speed =1, the intensity of coefficients reaches the maximum at the double of basic frequency. When speed=2, the intensity reaches the maximum at the four times of basic frequency. That is to say ω of horizontal slice is proportional to the horizontal moving speed of same object. Because f is constant for same object, ω/f of horizontal slice is also proportional to the speed.

Then consider the situation of two vertical bar moving from left to right with speed=2 as illustrated in Fig. 6(a). When this situation is portrayed with an 8×8 picture, the spatial frequency in horizontal direction is about treble of basic frequency. The horizontal slice of frame sequence is given in Fig. 6(b). Fig. 7 represents the DCT coefficients intensity of every basis in spatial dimension. As shown in Fig. 7, the intensity reaches the maximum at the treble of basic frequency. The coefficients represent the actual spatial frequency well.

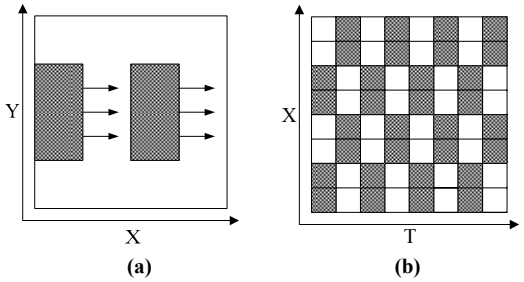

Fig. 6. Two vertical bars move from the left to the right. (a) A picture of moving process; (b) The horizontal slice when the bars moving with speed=2.

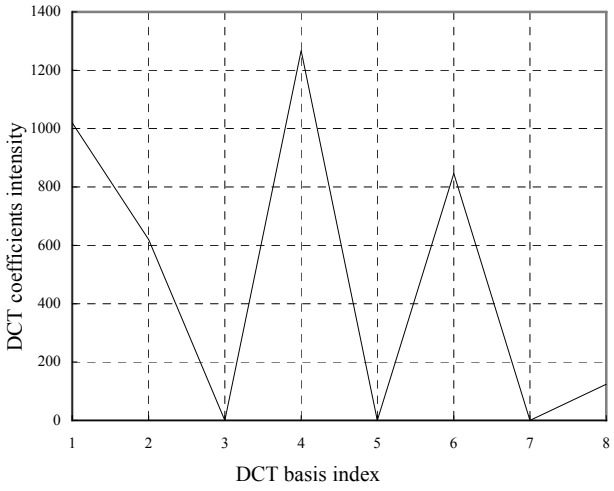

Fig. 7. The coefficients intensity of DCT basis in spatial dimension. The first DCT basis represents the DC component. The second basis represents the basic frequency. The fourth basis is treble of the basic frequency.

These two examples illustrate how DCT coefficients of slice represent the sensitivity related motion features. So we can use DCT coefficients to form format-independent features for motion content. In practice, the slice block size should be selected for convenience of DCT computing. Take this into consideration, we let N=32. An example of analyzing temporal slices blocks is given in Fig. 8.

Then a feature vector for *Sub* is derived from DCT

coefficients. Fig. 9 represents the stand deviation of the DCT coefficients derived from 8000 slice blocks of a video database which contains movie, sports, news content. As observed, the most energy of *Sub* is compacted on first 40 low-frequency AC DCT coefficients.

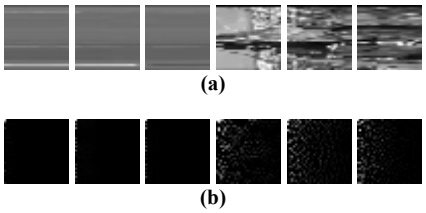

Fig. 8. An example of spatiotemporal slice blocks DCT. (a) Blocks segmented from slice, *N*=32. (b) Corresponding DCT images of blocks, *N*=32. The DCT coefficients are clamped to 0 and 255 for display.

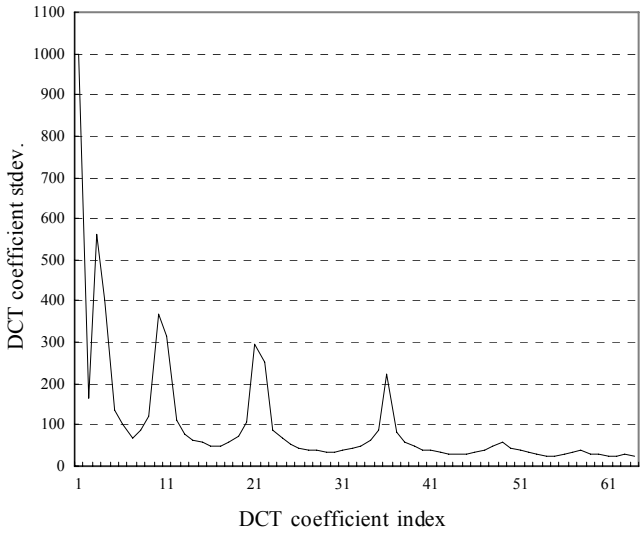

Fig. 9. The stand deviation of the DCT coefficients derived from 8000 slice blocks. The coefficients are indexed in *Zigzag* order

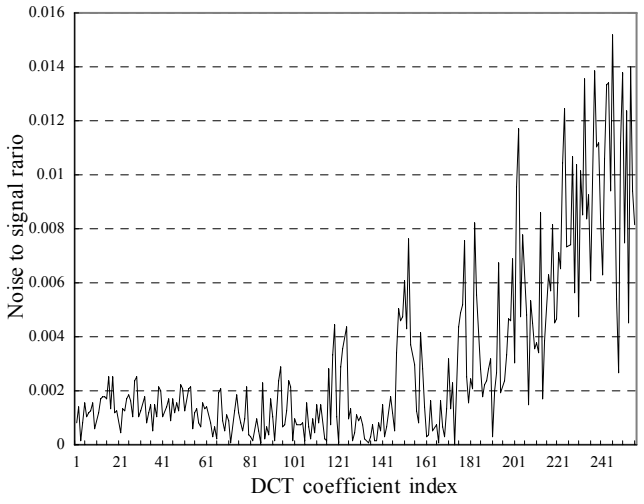

Fig. 10. The expected noise-to-signal ration for the DCT coefficients derived from 8000 slice blocks. The noise is the variance of the DCT coefficients obtained when encoding the slice blocks with JPEG-65%. The coefficients are indexed in *Zigzag* order

The estimation of the noise to signal energy ration due to JPEG compression for DCT coefficients is given in Fig. 10. We can observe that the ratio remains small for the first 40 low-frequency coefficients. So the low-frequency coefficients are not significantly affected by small distortion caused by reformatting. In this paper, a feature vector formed by the first 40 low-frequency AC DCT coefficients in *Zigzag* order are chosen for describing motion content in *Sub*. This kind of feature vector is denoted as *SubV*.

III. Experiments and Evaluation

In this section, the effectiveness of proposed method is evaluated with experiments. In order to be effective, format-independent motion describing should be sensitive to the variation of video content and robust to the variation of video format.

A. Sensitiveness and Robustness Analysis

The sensitiveness of proposed method is investigated as below. A MPEG-1 video with 19200 frames is segmented evenly into 24 clips. Then the horizontal slice of each clip is taken to calculate the similarity with all 24 clips. Denote the *SubV* sequence of *V* with *SUBV*, the similarity between two clips *V* and *V'* is evaluated by the similarity between *SUBV* and *SUBV'* computed using (4):

$$Simi(SUBV, SUBV') = 1 - \frac{\sum_{i=1}^{L} d(SubV_{V'}^{i}, SubV_{V}^{i})}{\sum_{i=1}^{L} (abs(SubV_{V}^{i}) + abs(SubV_{V}^{i}))} \quad (4)$$

Where $SubV_V^i$ denotes the *SubV* of i_{th} block of slice extracted from clip *V*. $d(\cdot)$ is the distance metric defined on *SubV* (here L_1 distance is applied), and $abs(\cdot)$ is the sum of the absolute value of elements of *SubV*. In the experiments, we set the *Sub* size as 32×32 and chose the first 40 low-frequency AC DCT coefficients of *Sub* to compose *SubV*.

Table I.
Sensitiveness Test

Similarity	Mean	Stdev.
S_{same}	1.0	0.0
S_{diff}	0.232157	0.041292

Note: S_{same}: similarity between the clips having the same content; S_{diff}: similarity between the clips having different content; Mean is the average value of the similarity; Stdev. is the standard deviation of the similarity.

The similarity between video content is assumed to obey Gaussian distribution. The *Mean* and *Stdev.* of similarities between the clips having same content and the clips having different content are given in Table I. The results show the correlation between clips with same content is much larger than the correlation between clips with different content. This shows the proposed method is sensitive to different video content.

Three video clips in MPEG-1 are used in robustness analysis. The original clips are recompressed in different formats as shown in Table II. Then we extract both horizontal

and vertical slices of the original and recompressed clips. Let $SUBV_H$ denote the *SubV* sequence of horizontal slice of V, $SUBV_V$ denote the vertical one. The similarity of every reformatted clip and original one is measured using (5) and (4):

$$Similarity(V,V') = \max(\ Simi(SUBV'_H, SUBV_H) \ , \ Simi(SUBV'_V, SUBV_V)\ ) \quad (5)$$

The proposed method shows robustness to the variation of compress format, resolution and aspect ratio even with 'pillar-box' (Fig. 11) in experiments. The average values of the similarities between reformatted clips and original clips are given in Table II. We can see the similarity of clips with the same content is larger than *Mean* + 15×*Stdev*. of S_{diff} in Table I. According to the properties of Gaussian distribution, the proposed method can tell the same content from the different content with very high probability. This shows the feature used in this paper is robust to the distortions in digitization and encoding process.

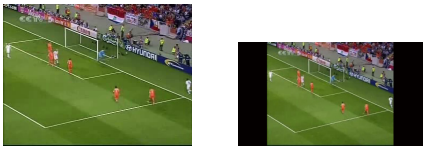

Fig. 11. Reformatted clip with 'pillar-box'

TABLE II.
ROBUSTNESS TEST

Reformatted clip	Similarity
352×288 in AVI	0.88623
320×180 in AVI	0.87302
512×288 in MPEG-1	0.90397
320×180 in AVI with 'pillar box'	0.86541
320×240 in MPEG-1 with 'pillar box'	0.87394

B. Comparisons with OM Method

The OM based method is used for comparison. According to [9], the ordinal measure (OM) gives better performance than motion based and color based features in video copy detection. Kim developed the OM method by combining temporal information to gain better performance in [10].

We use about 50 000 frame MPEG-1 videos which have a variety of content including news, sports and movie as test sequences. The videos are evenly segmented into 63 clips (original clips) and reformatted in different formats as described in Table II.

In the experiments of our method, the *Sub* size is 32×32 and the first 40 AC DCT coefficients are used. We develop the copy detection judgmental function *COPY* as below:

$$COPY < V, V' >= \begin{cases} True & if \ \ 1 - Similarity(V,V') < \tau \\ False & else \end{cases} \quad (6)$$

Where τ is a predefined threshold. Let C_d^r denote the number of correctly matched copy clips, C_d denotes the number of matched clips, C^r denote the number of actual copy clips. The *Recall* and *Precision* rate defined below are used to evaluate the effectiveness of methods:

$$Recall(\tau) = \frac{C_d^r}{C^r} \ , \ Precision(\tau) = \frac{C_d^r}{C_d} \quad (7)$$

The recall and precision rate of proposed method under different τ are given in Fig. 12. The recall and precision rate are both 100% when τ is between 0.2 and 0.35. Besides this, the recall rate decrease little when $\tau = 0.15$ and precision rate also decrease little when $\tau = 0.4$. The proposed method is very stable. The results show the feature proposed in this paper is robust to the distortions in digitization and encoding process. And the feature can distinguish different video content well.

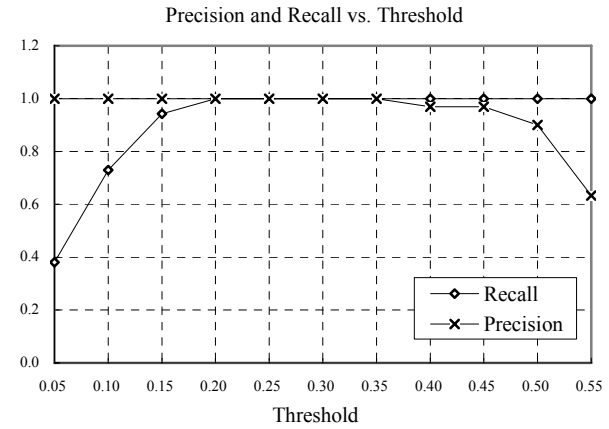

Fig. 12. The recall and precision of proposed method under different threshold τ

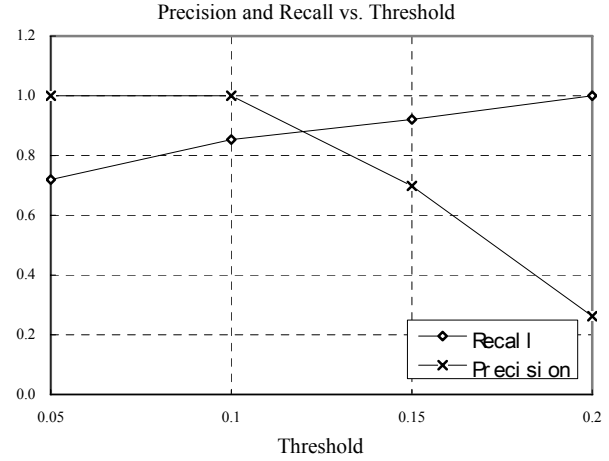

Fig. 13. The recall and precision of method proposed in [10] under different threshold τ

In experiment with the method proposed in [10], the threshold is set as 0.05, 0.1, 0.15 and 0.2. The results are given in Fig. 13. The weighting factor balancing between spatial and temporal distances is set as 0.5. According to [10],

the temporal OM method gives better performance with these settings. When threshold is 0.1 the recall and precision are 85%, 100%. When threshold is 0.2 the recall and precision are 100%, 26%. When threshold is 0.15 the recall and precision are 92%, 70%.

To compare the proposed method with temporal OM method proposed in [10], the recall and precision of both methods at different thresholds are given in Fig. 14. The result of proposed method is labeled with Pro., the result of temporal OM is labeled with [10].

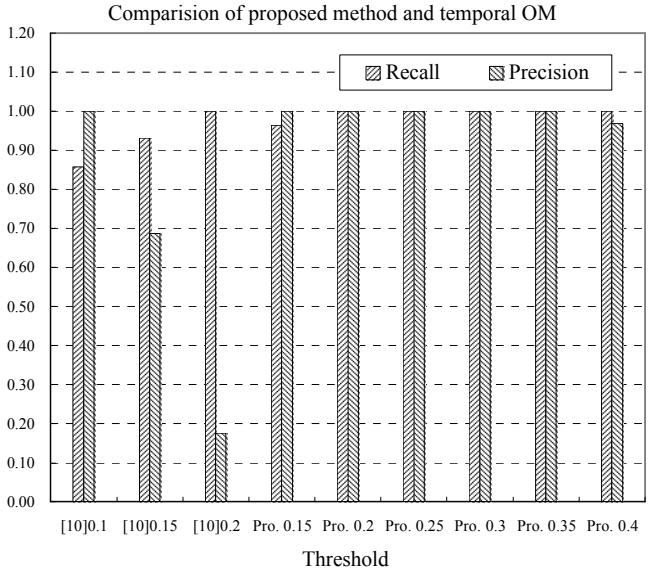

Fig. 14. Comparisons of proposed method and temporal OM method proposed in [10]. The threshold used in proposed method is labeled with Pro., the threshold used for temporal OM is labeled with [10].

From Fig. 14, when using method proposed in [10], there is not such a threshold which makes both recall and precision achieve a very sound level. It may be a tradeoff to choose 0.15 as threshold. The recall and precision decrease rapidly when threshold changes. This shows the method proposed in [10] is not very stable.

When using method proposed in this paper, the precision and recall are both 100% when τ is between 0.2 and 0.35. Besides this, the recall and precision decrease little when $\tau =$ 0.15 and 0.4. So the proposed method is much more stable than the method proposed in [10]. It can distinguish different video content in various formats well. It is an effective format-independent motion describing method for video content.

IV. Conclusions

In this paper, a format-independent motion describing method based on visual sensitivity is proposed. In proposed method, the visual sensitivity related features are analyzed using spatiotemporal slice DCT. Then the low-frequency AC DCT coefficients are taken as motion content descriptor for video. The experiments show the describing method is sensitive to the variation of video content and is robust to the variation of video format. The comparison with the video content descriptor used in [10] further indicates the effectiveness of proposed method.

References

[1] Sylvie Jeannin, Ajay Divakaran: MPEG-7 visual motion descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6), 2001, Page(s): 720-724

[2] Xinding Sun, Ajay Divakaran, B. S. Manjunath: A Motion Activity Descriptor and Its Extraction in Compressed Domain. *IEEE Pacific Rim Conference on Multimedia,* 2001: Page(s):450-457

[3] Kadir A. Peker, A. Aydin Alatan, Ali N. Akansu: Low-Level Motion Activity Features for Semantic Characterization of Video. *IEEE International Conference on Multimedia and Expo (II),* 2000

[4] Yap-Peng Tan, Drew D. Saur, Sanjeev R. Kulkarniand Peter J. Ramadge: Rapid Estimation of Camera Motion from Compressed Video with Application to Video Annotation. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(1), 2000: Page(s): 133-146

[5] Xingquan Zhu, Ahmed K. Elmagarmid, Xiangyang Xue, Lide Wu and Ann Christine Catlin: InsightVideo: Toward Hierarchical Video Content Organization for Efficient Browsing, Summarization and Retrieval. *IEEE Transactions on Multimedia*, Vol. 7, No. 4, Aug. 2005

[6] Frédéric Dufaux, Janusz Konrad: Efficient, robust, and fast global motion estimation for video coding. *IEEE Transactions on Image Processing,* 9(3) 2000: Page(s):497-501

[7] Yosi Keller, Amir Averbuch: Fast gradient methods based on global motion estimation for video compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(4) 2003: Page(s):300-309

[8] Chong-Wah Ngo, Ting-Chuen Pong, and Roland T. Chin: Motion Analysis and Segmentation through Spatio-temporal Slices Processing, *IEEE Transactions on Image Processing*, Vol. 12, No. 3, March 2003, Page(s):341-355

[9] Arun Hampapur, Ki-Ho Hyun, Ruud Bolle: Comparison of Sequence Matching Techniques for Video Copy Detection. *Proc. Storage and Retrieval for Media Databases*, Jan. 2002, Page(s): 194-201

[10] Changick Kim, Bhaskaran Vasudev: Spatiotemporal Sequence Matching for Efficient Video Copy Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 15, No. 1, January 2005, Page(s):127-132

[11] Dawei W. Dong: Spatiotemporal inseparability of natural images and visual sensitivities. *In: Computational, neural & ecological constraints of visual motion processing* (Zanker JM, Zeil J, eds, 2001), Page(s): 371-380

[12] Dawei W. Dong and Joseph J. Atick: Statistics of Natural Time-Varying Images. *Network: Computation in Neural Systems* Vol. 6(3), 1995, Page(s): 345-358

[13] A van der Schaaf, JH van Hateren: Modeling the Power Spectra of Natural Images: Statistics and Information. *Vision Research* 36, 1996, Page(s):2759-2770

[14] Peng. S. L, Medioni. G: Interpretation of image sequences by spatio-temporal analysis*, Workshop on Visual Motion*, March 1989. Page(s):344 – 351

[15] Iain E. G. Richardson: Video Codec Design, Developing Image and Video Compression Systems. Wiley, UK. 2002

Xuefeng Pan received the B.S. degree in computer science from Wuhan University, Wuhan, P. R. China in 1998. Now, he is pursuing the Ph. D. degree at Institute of Computing Technology, Chinese Academy of Sciences. His research interests include digital video processing, content-based video retrieval and computer vision.