Motion Video Retrieval Based on Optical Flow and Haar Wavelet

Ying Chen
Department of Basic Sciences
Beijing Electronic Science and Technology Institute
Beijing, P.R. China
ychen@besti.edu.cn

Ou Wu
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
Beijing, P.R. China
wuou@nlpr.ia.ac.cn

***Abstract*—Developments in image processing techniques have made an easy retrieval for digital images. However, applications requiring content-based querying and searching of videos still remain challenging due to their huge amount of data. This paper presents an efficient algorithm based on optical flow to dig out the motion information contained in a given video shot. The corresponding indexical structure is extracted by Haar wavelet. Experimental results show that the method performs well and the extracted index is used to efficiently locate features close to the query point.**

Keywords: Optical flow, Haar wavelet, Video retrieval.

I. Introduction

Traditionally, the problem of ”video retrieval” from a large database of videos has been solved by segmenting the query video into some logical related shots and assume that the segmentation is predefined by means of boundary identification [6], [9], [16]. After segmentation, one or more key frames can be extracted [5], [8], [18]. Each key frame maps a multi-dimensional vector in some metric space and builds an index for retrieval [12], [13]. The exist video retrieval systems usually adopted a set of low-level features from different methods. Because of semantic gap, these systems had a very limited success for semantic queries. That is to say, they are efficient for simple videos, but do not work well for complex scenes since they fail to retrieval videos that match the query only partially. This inefficiency leads to the discarding of videos that may be semantically very similar to the query video. Motion is an important feature in video and some researchers woke up to this point. They [3], [4], [11], [15] tried to apply motion information to retrieve. However, it is still a challenge to find a reliable motion description of video.

Motion-based video indexing requires to extract motion features which are relevant to characterize video content. As a high-level feature, optical flow [10], [14] can easily overcome theses shortages and offers a complemental motion information to low-level features. On the other side, wavelets [1], [17] can capture both texture and shape information efficiently. The fusion of optical flow and wavelets can effect a breakthrough for semantical retrieval. In this paper, we propose a new algorithm for video retrieval by computing the optical flow field of key frames and generating the index with Haar wavelet. Experimental results show that the proposed scheme is effective in terms of accuracy and efficiency.

The rest of the paper is arranged as follows. Section 2 gives a review of the optical flow method used in this paper. Section 3 describes the steps of Haar wavelet. Section 4 discusses the retrieval model in detail. Section 5 analyzes experimental results on the retrieval of sport video shots. Section 6 concludes the paper.

II. Optical flow

Let the function $f = I(x, y, t)$ be the image brightness at point (x, y) in the image plane at time t. Assume that $I(x, y, t)$ is differentiable and the change of brightness over time for each point is constant, i.e.

$$\frac{dI}{dt} = 0. \tag{1}$$

Notice that x and y also can be regarded as the function of t respectively, by the chain rule for differentiation we have the *gradient constraint equation* [7]:

$$(\nabla I)^T \cdot \overrightarrow{v} + I_t = 0, \tag{2}$$

where

$$I_t = \frac{\partial I}{\partial t}, \tag{3}$$

$$(\nabla I)^T = (\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}), \tag{4}$$

$$\overrightarrow{v} = (\mu, \nu)^T = (\frac{dx}{dt}, \frac{dy}{dt})^T. \tag{5}$$

Equation (2) has two variables about the optical flow $\overrightarrow{v} = (\mu, \nu)^T$. To solve it, more constraints are required. Lucas and Kanade's algorithm [2] assumes that the (μ, ν) is constant for an small enough region $\mathcal{D}$ and minimizes

$$\sum_p \mathcal{W}^2(p)((\nabla I)^T \cdot \overrightarrow{v} + I_t)^2, \tag{6}$$

where $p \in \mathcal{D}$ is the pixel point and $\mathcal{W}(x)$ is the weighted function. For n points $p_i \in \mathcal{D}$ at time t, let

$$\mathcal{I} = (\nabla I(p_1), \cdots, \nabla I(p_n))^T, \tag{7}$$

978-1-61284-181-6/11/$26.00 ©2011 IEEE

$$\mathcal{I}_t = -(I_t(p_1), \cdots, I_t(p_n))^T, \tag{8}$$

and

$$\mathcal{W} = \begin{pmatrix} \mathcal{W}(p_1) & 0 & \cdots & 0 & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \mathcal{W}(p_i) & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & W(p_n) \end{pmatrix}. \tag{9}$$

By the derivative of (6) with respect to μ and ν, the $\overrightarrow{v} = (\mu, \nu)^T$ is given by

$$\mathcal{I}^T \mathcal{W}^2 \mathcal{I} \overrightarrow{v} = \mathcal{I}^T \mathcal{W}^2 \mathcal{I}_t. \tag{10}$$

Obviously, $\mathcal{I}^T \mathcal{W}^2 \mathcal{I}$ is a 2×2 matrix and equals to

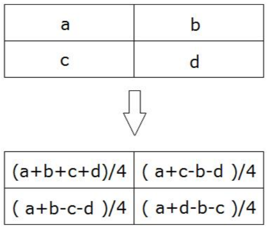
$$\begin{pmatrix} \sum \mathcal{W}^2(p_i) I_x^2(p_i) & \sum \mathcal{W}^2(p_i) I_x(p_i) I_y(p_i) \\ \sum \mathcal{W}^2(p_i) I_y(p_i) I_x(p_i) & \sum \mathcal{W}^2(p_i) I_y^2(p_i) \end{pmatrix}. \tag{11}$$

If $|\mathcal{I}^T \mathcal{W}^2 \mathcal{I}| \neq 0$, then

$$\overrightarrow{v} = (\mathcal{I}^T \mathcal{W}^2 \mathcal{I})^{-1} \mathcal{I}^T \mathcal{W}^2 \mathcal{I}_t. \tag{12}$$

III. Haar wavelet

Since the wavelet transform can be used to decompose the data into wavelet coefficients on different scales, the wavelet coefficients can be reconstructed to recover the original data.We use two-dimensional nonstandard Haar wavelet to compute the features because it enable us to accelerate the wavelet transform. It has excellent energy compaction and the wavelet representation of a function consists of a coarse overall approximation together with detail coefficients that exploit the structure of video data. In nonstandard Haar wavelet, pairwise averages and differences are applied. Fig.1 illustrates the process in detail.

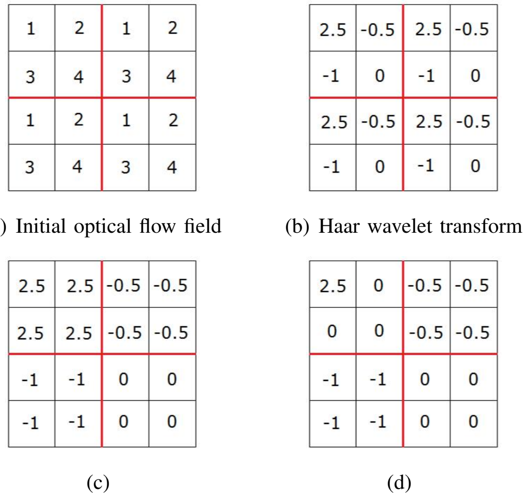


a	b
c	d

(a+b+c+d)/4	(a+c-b-d)/4
(a+b-c-d)/4	(a+d-b-c)/4

Fig. 1. Computing and distributing pairwise averages and differences

Now, let us consider the 4×4 optical flow field shown in Fig.2a. The value such as "1,2,3,4" denotes the modulus or the argument principal value of optical flow. As described in Fig.1, we obtained the wavelet transform as shown in Fig.2b. In Fig.2c, the coefficients from the top-left quadrant of 2×2 matrix are derived by transforming the corresponding four top-left quadrants of 4×4 matrix in Fig.2b, similarly, the top-right, the bottom-left, the bottom-right can be expressed like that. Finally, we choose one 2×2 quadrant to represent the whole 4×4 field. In this paper, the top-left quadrant are used. For the top-left quadrant, we repeat the step of computing and distributing pairwise and differences and Fig.2d are obtained. Simply stated, the optical flow field in Fig.2a can be expressed by a vector $(2.5, 0, 0, 0)$.



1	2	1	2
3	4	3	4
1	2	1	2
3	4	3	4

(a) Initial optical flow field

2.5	-0.5	2.5	-0.5
-1	0	-1	0
2.5	-0.5	2.5	-0.5
-1	0	-1	0

(b) Haar wavelet transform

2.5	2.5	-0.5	-0.5
2.5	2.5	-0.5	-0.5
-1	-1	0	0
-1	-1	0	0

(c)

2.5	0	-0.5	-0.5
0	0	-0.5	-0.5
-1	-1	0	0
-1	-1	0	0

(d)

Fig. 2. Representation of Haar wavelet transform for optical flow.

IV. How to retrieval

The procedure of retrieval consists of three main steps.

A. Compute optical flow

Assume that the given video shot has $N + 1$ frames with size $4w \times 4h$. N optical flow fields $\{\mathcal{G}_1, \cdots, \mathcal{G}_N\}$ will be computed by Lucas-Kanada's algorithm. Since the optical flow is a vector, we can get a matrix $\mathcal{M}_i$ about its modulus and a matrix $\mathcal{D}_i$ about its argument principal value from each $\mathcal{G}_i$.

Our algorithm focus on the motion videos. Some motion videos we called them "M-type" like explosion as shown in Fig.3a have greater magnitude information. But some motion videos we called them "D-type" like basketball match as shown in Fig.3b contain apparent direction information when one team is attacking. Obviously, $\mathcal{M}_i$ is an appropriate representation for "M-type" and $\mathcal{D}_i$ is for "D-type".

(a) Explosion

(b) Basketball match

Fig. 3. Examples for "M-type" and "D-type".

B. Construct key frame

For $\mathcal{M}_i$ and $\mathcal{D}_i$, by the Haar wavelet transform discussed in section III, vectors $m_1^i, m_2^i, \cdots, m_{wh}^i$ and $d_1^i, d_2^i, \cdots, d_{wh}^i$ are constructed. Then we define m_j and d_j as follows:

$$m_j = \frac{1}{N}\sum_{i=1}^{N} m_j^i, \tag{13}$$

$$d_j = \frac{1}{N}\sum_{i=1}^{N} d_j^i, \tag{14}$$

where $j = 1, 2, \cdots, wh$. The $2wh$ vectors $m_1, m_2, \cdots, m_{wh}$ and $d_1, d_2, \cdots, d_{wh}$ can be regarded as a representation of key frame for the given video shot.

C. Build similarity model

Let $\mathcal{V}$ and $\mathcal{V}'$ denote two video shots, based on modulus of optical flow, their feature distance is measured by:

$$Dist(\mathcal{V}, \mathcal{V}')_{\mathcal{M}} = \sum_{j=1}^{wh} \lambda_j^{\mathcal{M}} \frac{||m_j(\mathcal{V}) - m_j(\mathcal{V}')||}{||m_j(\mathcal{V}) + m_j(\mathcal{V}')||}, \tag{15}$$

where $\lambda_j^{\mathcal{M}}$'s are user specified weights. Based on argument principal value of optical flow, their feature distance is defined similarly:

$$Dist(\mathcal{V}, \mathcal{V}')_{\mathcal{D}} = \sum_{j=1}^{wh} \lambda_j^{\mathcal{D}} \frac{||d_j(\mathcal{V}) - d_j(\mathcal{V}')||}{||d_j(\mathcal{V}) + d_j(\mathcal{V}')||}, \tag{16}$$

where $\lambda_j^{\mathcal{D}}$'s are user specified weights. Then the overall feature distance can be measured by:

$$Dist(\mathcal{V}, \mathcal{V}') = \lambda Dist(\mathcal{V}, \mathcal{V}')_{\mathcal{M}} + (1-\lambda) Dist(\mathcal{V}, \mathcal{V}')_{\mathcal{D}}, \tag{17}$$

where λ is user specified weight. That is to say,the best math for the query shot is the one with the smallest overall feature distance.

V. Experiments

To evaluate the performance of the proposed method for the retrieval of video shorts, the *average normalized modified retrieval rank* (ANMRR) and the *average recall* (AR), which were developed in [19], are chosen as the benchmark indicators. The value of ANMRR determines the rank of the correct shots unretrieved and the value of AR determines the rate of the correct shots retrieved. The lower the value of the ANMRR, the better the performance. In contrast, the higher the value of the AR, the better the performance. Let $\{q_1, \cdots, q_S\}$ denote the set of the query shots. For any q_i, the set $G(q_i)$ denotes the ground truth shots and

$$K_{q_i} = \min\{4 \times \|G(q_i)\|, 2 \times \max_i \|G(q_i)\|\} \tag{18}$$

denotes the value of the threshold, where $\|\sharp\|$ is the number of the elements in the set $\sharp$. Then the set $R(q_i)$ denotes the items correctly retrieved in the top K_{q_i} results. So the value of the AR and the ANMRR are calculated as:

$$AR = \frac{1}{S}\sum_{i=1}^{S} \frac{\|R(q_i)\|}{\|G(q_i)\|}, \tag{19}$$

and

$$ANMRR = \frac{1}{S}\sum_{i=1}^{S} \frac{\sum_{m=1}^{\|G(q_i)\|} \frac{Rank(g_m)}{\|G(q_i)\|} - \frac{\|G(q_i)\|}{2} - \frac{1}{2}}{K_{q_i} - \frac{\|G(q_i)\|}{2} + \frac{1}{2}}, \tag{20}$$

where for $g_m \in G(q_i)$, the $Rank(g_m)$ equals to the order in the top K_{q_i} retrievals if $g_m \in R(q_i)$, while $K_{q_i} + 1$ if $g_m \notin R(q_i)$.

A. Experiment I

The first experiment is to compare the performance of our algorithm with different values of λ. 202 "M-type" and 198 "D-type" videos were used to test in this experiment. Fig.4 shows the ANMRR and the AR with $\lambda = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. We can see that when $\lambda = 0$, only feature distance $Dist(\mathcal{V}, \mathcal{V}')_{\mathcal{D}}$ is used. In this case, performance of "D-type" with $Dist(\mathcal{V}, \mathcal{V}')_{\mathcal{D}}$ is more efficient than "M-type". When $\lambda = 1$ and $Dist(\mathcal{V}, \mathcal{V}')_{\mathcal{M}}$ is available, we have the opposite result. All these indicate the effectiveness of $\mathcal{M}_i$ for "M-type" and $\mathcal{D}_i$ for "D-type".

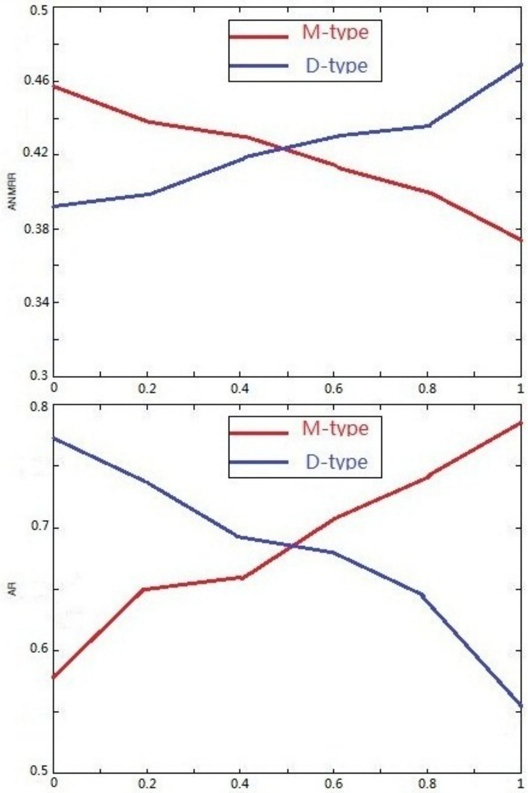


Fig. 4. Performance of "M-type" and "D-type" with different λ.

TABLE I
ANMRR VALUES AND AR VALUES WITH DIFFERENT $\lambda_j^{\mathcal{M}}$

	Increase $\lambda_j^{\mathcal{M}}$ in the left region to $T\lambda_j^{\mathcal{M}}$				Increase $\lambda_j^{\mathcal{M}}$ in the middle region to $T\lambda_j^{\mathcal{M}}$				Increase $\lambda_j^{\mathcal{M}}$ in the right region to $T\lambda_j^{\mathcal{M}}$			
	T=2		T=4		T=2		T=4		T=2		T=4	
	ANMRR	AR	ANMRR	AR	ANMRR	AR	ANMRR	AR	ANMRR	AR	ANMRR	AR
Left-explosion videos	0.3997	0.7886	0.3826*	0.8011*	0.4122	0.7761	0.4007	0.7325	0.4552	0.7551	0.4535	0.7641
Middle-explosion videos	0.4815	0.7021	0.4772	0.6953	0.3916	0.7889	0.3778*	0.7999*	0.4238	0.7446	0.4611	0.7421
Right-explosion videos	0.4771	0.6962	0.4688	0.6950	0.4816	0.6752	0.4421	0.6992	0.3995	0.7883	0.3747*	0.8118*

B. Experiment II

In this experiment, $\lambda_j^{\mathcal{M}}$ are discussed. By the way, $\lambda_j^{\mathcal{D}}$'s discussion is similar and omitted here. When we construct key frames, the Haar wavelet transform is used. However, the Haar wavelet transform is connected with region of optical flow field. Let's see the example in Fig.5. In these three video shots, the explosion occur at different region. If we increase weights of $\lambda_j^{\mathcal{M}}$ in the left location, such videos like Fig.5a should show a better performance than Fig.5b and Fig.5c.15 ”left-explosion” videos, 15 ”middle-explosion” videos and 15 ”right-explosion” videos were used to test and their results are listed in Table.I. The mark ”*” indicated the better performance and the results proved our idea. Therefore, if a prior knowledge about the query video shot is introduced by human supervision, the efficiency of retrieval will be improved.

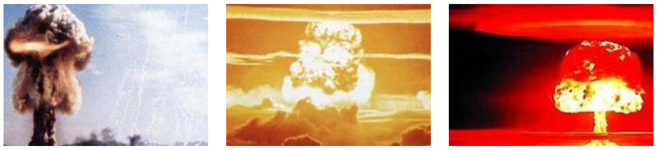

(a) Left (b) Middle (c) Right

Fig. 5. Explosion at different region.

C. Experiment III

Finally, we discuss the performance of our algorithm. Table.II listed the ANMRR and the AR values with different motion-based algorithms. It can be seen that our algorithm result in a lower ANMRR and a higher AR showed a better performance.

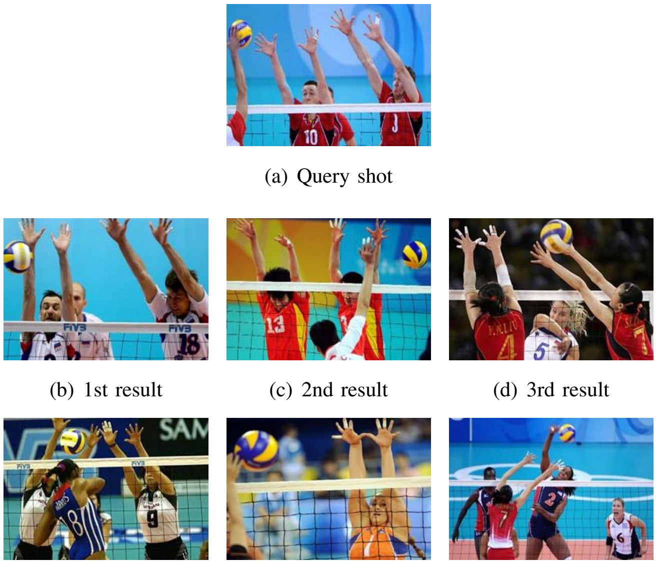
VI. CONCLUSION

In this paper, we presented an algorithm based on optical flow and Haar wavelet for video shot retrieval. More significant motion contents within a video shot can be extracted by our algorithm. Experiments have demonstrated it is really powerful. A example of query results for a given video shot is shown in Fig.6.

TABLE II
PERFORMANCE WITH DIFFERENT ALGORITHMS, WHERE THE NUMBER OF VIDEO SHOTS TO BE RETRIEVED IS 921 AND THE MARK * INDICATES THE BETTER PERFORMANCE.

	ANMMRR	AR
Algorithm in [3]	0.3980	0.7815
Algorithm in [4]	0.4002	0.8036
Algorithm in [11]	0.4211	0.7991
Algorithm in [15]	0.3992	0.7211
Our algorithm	0.3977*	0.8082*

In future works, we will have to find solutions to merge motion feature with other features, such as text, audio and so on. Actually, a video retrieval system only based on a single feature is not practical. To use more feature will lead to the detection of more semantic contents.


(a) Query shot

(b) 1st result (c) 2nd result (d) 3rd result

(e) 4th result (f) 5th result (g) 6th result

Fig. 6. Top 6 retrieval results are listed,only their first frames are here.

ACKNOWLEDGMENT

The authors would like to thank many other workmates for their valuable suggestion. This work is supported by the National Natural Science Foundation of China (Grant No. 61003115 & 60903147).

REFERENCES

[1] A.A. Nabout and B. Tibken, *Object Shape Description Using Haar-Wavelet Functions*. International Conference on Information and Communication Technologies: From Theory to Applications, Page(s): 1-6, 2008.

[2] B. Lucas and T. Kanade, *An iterative image registration technique with an application to stereo vision*. In Proc. International Joint Conference on Artificial Intelligence, Page(s): 674-679, 1981.

[3] B. Tahayna, M. Belkhatir, and S. Alhashmi, *Motion information for video retrieval*. IEEE International Conference on Multimedia and Expo, Page(s): 870-873, 2009.

[4] C.W. Su, H.Y.M. Liao, H.R. Tyan, C.W. Lin, D.Y. Chen, and K.C. Fan, *Motion Flow-Based Video Retrieval*. IEEE Transactions on Multimedia, Volume: 9, Issue: 6, Page(s): 1193-1201, 2007.

[5] D.P. Mukherjee, S.K. Das, and S. Saha, *Key Frame Estimation in Video Using Randomness Measure of Feature Point Pattern*. IEEE Transactions on Circuits and Systems for Video Technology, Volume: 17, Issue: 5, Page(s): 612-620, 2007.

[6] Herong Zheng, Zhi Liu, and Xiaofeng Wang, *Research on the Video Segmentation Method with Integrated Multi-features Based on GMM*. International Conference on Digital Image Processing, Page(s): 62-66, 2009.

[7] B.K.P. Horn and B. Schunck *Determining Optical Flow*. Artificial Intelligence, Page(s): 185-203, 1981.

[8] Rong Pan, Yumin Tian, and Zhong Wang, *Key-frame extraction based on clustering*. IEEE International Conference on Progress in Informatics and Computing, Volume: 2, Page(s): 867-871, 2010.

[9] L.S. Silva and J. Scharcanski, *Video Segmentation Based on Motion Coherence of Particles in a Video Sequence*. IEEE Transactions on Image Processing, Volume: 19, Issue: 4, Page(s): 1036-1049, 2010.

[10] Shiquan Wang, Kaiqi Huang, and Tieniu Tan; *A compact optical flow-based motion representation for real-time action recognition in surveillance scenes* . IEEE International Conference on Image Processing, Page(s): 1121-1124, 2009.

[11] Tianli Yu and Yujin Zhang, *Retrieval of video clips using global motion information* . Electronics Letters, Volume: 37, Issue: 14, Page(s): 893-895, 2001.

[12] T.L. Le, A. Boucher, M. Thonnat, and F. Bremond, *A framework for surveillance video indexing and retrieval*. International Workshop on Content-Based Multimedia Indexing, Volume: 19, Issue: 4, Page(s): 338-345, 2008.

[13] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, *A Survey on Visual Content-Based Video Indexing and Retrieval* . IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Issue: 99, Page(s): 1-23, 2011.

[14] Xiangyu Wang, S. Ramanathan, M. Kankanhalli, *A robust framework for aligning lecture slides with video* . IEEE International Conference on Image Processing, Page(s): 249-252, 2009.

[15] Xiaoming Chen, Zheming Lu, and Zhen Li, *Video retrieval using VQ-based global motion features*. International Conference on Visual Information Engineering, Page(s): 577-581, 2008.

[16] Xiaomu Song and Guoliang Fan, *Selecting Salient Frames for Spatiotemporal Video Modeling and Segmentation*. IEEE Transactions on Image Processing, Volume: 16, Issue: 12, Page(s): 3035-3046, 2007.

[17] Xin Wang, *Moving window-based double haar wavelet transform for image processing*. IEEE Transactions on Image Processing, Volume: 15, Issue: 9, Page(s): 2771-2779, 2006.

[18] Zhonghua Sun, Kebin Jia, and Hexin Chen, *Video Key Frame Extraction Based on Spatial-Temporal Color Distribution*. International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Page(s): 196-199, 2008.

[19] *MPEG-7 visual part of experimentation Model (XM) version 2.0*. MPEG-7 Output Document ISO/MPEG, Dec., 1999.