A Novel CBCD Approach Using MPEG-7 Motion Activity Descriptors

R.Roopalakshmi
Information Technology Department
National Institute of Technology Karnataka (NITK)
Surathkal, Mangalore, India - 575025
Email: r_roopalakshmi@hotmail.com

G.Ram Mohana Reddy
Information Technology Department
National Institute of Technology Karnataka (NITK)
Surathkal, Mangalore, India - 575025
Email: profgrmreddy@gmail.com

***Abstract*—Motion features contribute significant information about a video content. This paper highlights a novel CBCD (Content-Based Copy Detection) approach, by incorporating several motion activity features. First, we extract both temporal and spatial motion features to describe overall activity of a video sequence. Second, we combine these features in a feasible manner, to generate robust video fingerprints. Third, clustering based pruned search is utilized for similarity matching instead of direct searching of video fingerprints. The proposed system is tested on TRECVID-2007 data set and the results demonstrate the effectiveness of the proposed system against several transformations such as random noise, fast forward, pattern insertion, cropping and picture-inside-picture.**

***Index Terms*—Content based video copy detection, motion activity descriptor, MPEG-7, motion intensity.**

I. Introduction

Due to the exponential growth of on-line publishing activities, video content proliferation on the Internet is increasing at an impressive rate. Hence, video copy detection is essential to reduce huge piracy and copyright issues. In general, a duplicate video is defined as, a modified video sequence, derived from master video [1]. There are two approaches for detecting copies of a digital media: digital watermarking and content based video copy detection [2]. The primary task of any CBCD system is to detect video copies by utilizing content based features of the media [3]. The CBCD approaches are preferred compared to watermarking techniques because of the following key features [4], [5]: i) The video signature generation will neither destroy nor damage video content, ii) CBCD techniques are more robust than fragile watermarking techniques, iii) Signature extraction can also be done after the distribution of digital media and iv) Can detect copies, even if the original document is not watermarked.

In CBCD literature, the existing techniques are based on global and local features. Global features such as Ordinal measure [6], Color histograms [7] are compact and easy to extract, but they are less robust against region based attacks. SIFT [8] and SURF [9], are some of the popular local descriptors, which use interest points for feature extraction [10]. The local features are more robust against region based transformations, but their computational cost is high compared to global features [11].

Several researchers studied motion based video content analysis, which is used in areas such as motion based retrieval, video indexing and video characterization [12], [13]. In the past CBCD literature, motion signatures are considered as poor descriptors [14]. The reasons for this poor performance are: (a) Motion vectors are close to zero values, when they are captured at normal frame rates (25-30 fps); (b) Noisy nature of raw motion vectors and (c) Need of huge amount of information to describe the motion content. Tasdemir et al.[15] attempted to solve the CBCD task, by capturing frames at a lower rate (5fps) and also using motion vector magnitudes on a frame by frame basis. The frame based motion features may describe the content of a video clip, but they fail to provide the complete description of overall activity of a video sequence.

The main objective of this article is to develop a novel CBCD method, by fusing temporal behaviour and spatial distribution of motion activity. The main contributions of this paper are as follows:

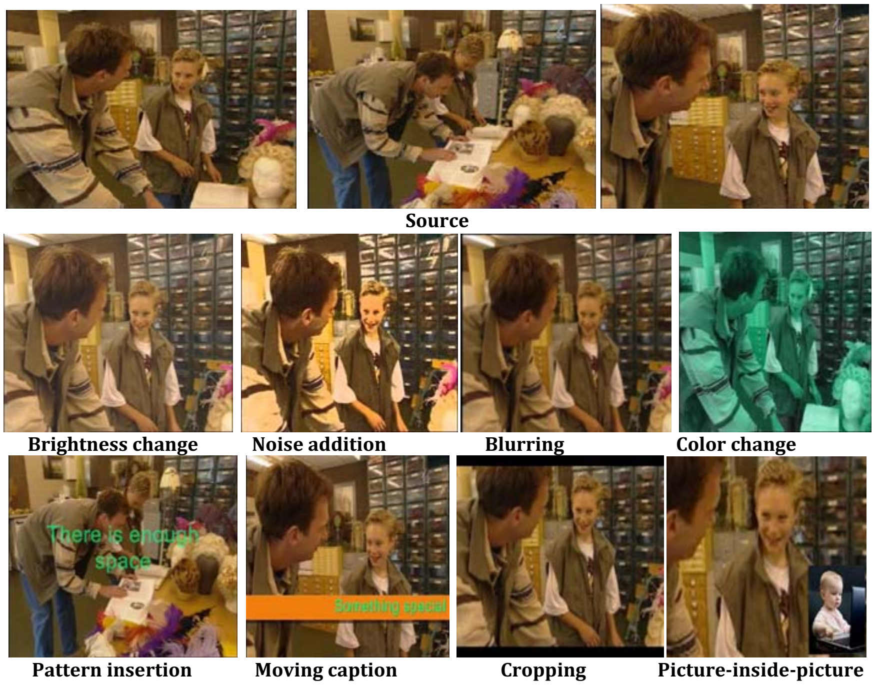
a) To describe overall activity of a video sequence, instead of traditional temporal motion vector approaches.
b) Robust motion activity features such as motion intensity, dominant direction and spatial distribution of activity are combined to achieve the CBCD task.
c) Clustering based pruned search is performed, to speed up similarity matching process.

Table I gives the video transformations used in our CBCD task and Fig.1 illustrates example frames from transformed query videos. The rest of this paper is organized as follows: Section II introduces framework of the proposed scheme along with feature extraction and matching techniques; Section III shows the experimental setup and results of the proposed scheme, followed by conclusion in Section IV.

II. Proposed framework

The block diagram of proposed copy detection framework is shown in Fig.2, and the relevant symbols are explained in Table II. The proposed framework consists of two main stages: Master video processing stage (off-line) and Query video processing stage (on-line). In the off-line stage, motion activity based features are extracted from master video frames. The features include intensity of action, spatial distribution,

978-0-7695-4589-9/11 $26.00 © 2011 IEEE
DOI 10.1109/ISM.2011.36


Source

Brightness change

Noise addition

Blurring

Color change

Pattern insertion

Moving caption

Cropping

Picture-inside-picture

Fig. 1. Example frames from transformed query videos

TABLE I
LIST OF TRANSFORMATIONS

Type	Transformations
T1	Brightness change
T2	Noise addition
T3	Blurring
T4	Color change
T5	Pattern insertion
T6	Moving caption insertion
T7	Slow motion
T8	Fast forward
T9	Cropping
T10	Picture-inside-picture

dominant direction of activity and average magnitude of motion vectors. These features are further processed and Motion Activity (MA) words are computed. Since MA words combine raw motion activity features, they comprehensively represent overall activity of video sequences. K-means clustering approach is utilized, in order to get compact & low-dimensional representation of MA words and the cluster centroids are stored as video fingerprints.

In the on-line stage, MA words are calculated, after extracting motion activity features from query video frames. The resulting MA words are clustered and cluster centroids are stored as video fingerprints. Finally clustering based similarity matching is performed for detecting video copies.

A. Fingerprint Extraction

a) MPEG-7 Motion Activity Descriptor: This Descriptor captures intensity of activity or pace of action in a video segment [16]. A 'high speed car chase' denotes a high activity sequence, whereas 'an interview scene' denotes a low activity sequence. The motion activity descriptor includes the following attributes:

$$Motion\ activity = \{I, Dir, Spatial, Temporal\} \qquad (1)$$

Intensity of activity (*I*) indicates, high or low intensity by an integer value and direction attribute (*Dir*) specifies dominant direction of activity. Spatial distribution of activity (*Spatial*) denotes number of active regions in a frame and temporal distribution attribute (*Temporal*) indicates the variation of activity over the duration of video sequence.

b) Motion Activity Features Extraction: In this work, we integrated mean motion magnitude of frames and attributes of MPEG-7 motion activity descriptor to implement the CBCD task. The reasons for this combination are: First, average motion magnitude provides better frame-level information of video clips. Second, entire activity of video sequence can be sufficiently characterized, by using different attributes of motion activity descriptor.

1) Motion Intensity: This attribute provides effective temporal description of a video shot in terms of different intensity levels [17]. The statistical properties of motion vector such as average and standard deviation, can be used to calculate intensity of motion activity. The Average Motion Vector magnitude (AMV) and Standard deviation of Motion Vector magnitude

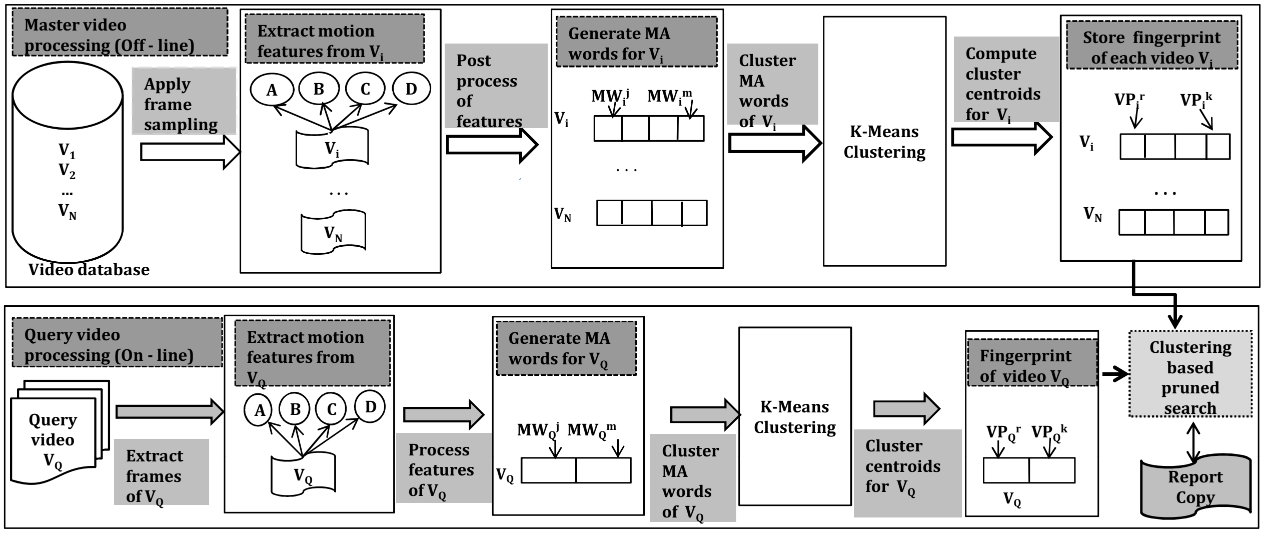

Master video processing (Off - line)
Apply frame sampling
V_1
V_2
...
V_N
Video database
Extract motion features from V_i
A
B
C
D
V_i
...
V_N
Post process of features
Generate MA words for V_i
MW_i^j
MW_i^m
V_i
...
V_N
Cluster MA words of V_i
K-Means Clustering
Compute cluster centroids for V_i
Store fingerprint of each video V_i
VP_i^r
VP_i^k
V_i
...
V_N
Query video processing (On - line)
Query video V_Q
Extract frames of V_Q
Extract motion features from V_Q
A
B
C
D
V_Q
Process features of V_Q
Generate MA words for V_Q
MW_Q^j
MW_Q^m
V_Q
Cluster MA words of V_Q
K-Means Clustering
Cluster centroids for V_Q
Fingerprint of video V_Q
VP_Q^r
VP_Q^k
V_Q
Clustering based pruned search
Report Copy

Fig. 2. The framework of proposed scheme

TABLE II
GLOSSARY OF NOTATIONS

Notation	Definition	Notation	Definition
N	Number of master video files in the database	V_Q	Query video file
V_i	i-th master video in the database	VP_i^r	r-th video fingerprint of i-th video V_i, such that r = {1, 2, 3, . . . ,k}
A	Intensity of activity	B	Spatial distribution of activity (Number of active regions of a frame)
C	Dominant direction of activity	D	Average motion vector magnitude
MW_q^j	j-th MA word of query video V_Q	VP_Q^r	r-th video fingerprint of query video V_Q
MW_i^j	j-th MA word of i-th video V_i, such that j ={1, 2, 3, .., m}	V_N	N-th master video in the database
m	Number of MA words of video V_i	k	Number of video signatures of video V_i

(SMV) of a frame are given by,

$$AMV = \frac{1}{MN} \times \sum_{i=1}^{M}\sum_{j=1}^{N} mv(i,j) \qquad (2)$$

$$SMV = \sqrt{\frac{1}{MN} \times \sum_{i=1}^{M}\sum_{j=1}^{N} |mv(i,j) - AMV|^2} \qquad (3)$$

Where $mv(i,j)$ indicate the motion vector of (i,j)-th block and MN is the frame size.

TABLE III
QUANTIZATION THRESHOLDS FOR MPEG-1 VIDEO

Activity value	Range of SMV
1	$0 \leq$ SMV <3.9
2	$3.9 \leq$ SMV < 10.7
3	$10.7 \leq$ SMV <17.1
4	$17.1 \leq$ SMV <32
5	$32 \leq$ SMV

In this article, we have used SMV of macro blocks to compute motion intensity. SMV values are quantized in the range of 1-5 as per MPEG-7 standard [16], given in Table III.

2) Spatial Distribution of Activity: This attribute indicates, whether the activity is spread across many regions or confined to one region [18]. The segmentation of frame into n×n regions, plays a significant role in predicting accurate number of active regions in a given frame. Smaller values of n may eliminate important semantic content, whereas larger values of n lead to increase in computational process.

In order to solve this discrepancy, we experimented our data set with different n values ranging from 2 to 5 and we observed that maximum accuracy rate is achieved when n=3. Thus we computed spatial distribution of activity of frames by segmenting it into 3×3 regions.

The algorithm for computing spatial distribution of activity in a frame is given by,

```
1: Calculate Spatial Activity Matrix (SAM) of each frame using the equation given by,
   $SAM = \begin{cases} magmv(i,j) & if\ magmv(i,j) \geq AMV \\ 0 & otherwise \end{cases}$   (4)
   where $magmv(i,j)$ is the magnitude of motion vector of block $(i,j)$.
2: Segment SAM of each frame into non overlapping blocks of size 3×3.
3: Compute the mean motion distribution (MMD) of $r$-th block of $k$-th frame given by,
   $MMD(r) = \frac{Sum\ of\ SAM\ values}{Size\ of\ r}$   (5)
4: Sort the regions of a frame in the ascending order of MMD values.
5: Regions with higher MMD values are considered as active regions of a given frame.
```

3) Dominant Direction of Activity: Dominant motion directions of a video clip provide significant information about its overall activity. Our goal is not to calculate the exact direction of motion of all objects, but to compute approximate dominant directions for improving the robustness of proposed CBCD system. In the proposed method, direction vector (DIR) indicates total amount of motion in four major directions including up, down, left and right given by [19],

$$DIR = \{Up, Down, Left, Right\} \tag{6}$$

Let $mv_x(k)$ and $mv_y(k)$, denote two components of motion vector of k-th block and N indicates total number of blocks, then motion in four directions are calculated as,

$$Up = \sum_{k=1}^{N}(mv_y(k)), \quad if\ mv_y \leq 0 \tag{7}$$

$$Down = \sum_{k=1}^{N}(mv_y(k)), \quad if\ mv_y > 0 \tag{8}$$

$$Left = \sum_{k=1}^{N}(mv_x(k)), \quad if\ mv_x > 0 \tag{9}$$

$$Right = \sum_{k=1}^{N}(mv_x(k)), \quad if\ mv_x \leq 0 \tag{10}$$

The highest value of DIR provides dominant direction of motion in a given frame. Direct processing of extracted raw motion activity features is tedious and computationally expensive. Hence, motion activity features are normalized and combined into high informative MA words. Since the dimension of MA words of each file is large, K-means clustering technique is used to achieve low dimensional representation of MA words.

B. Fingerprint Matching

In the proposed CBCD system, first video signatures of individual video files are grouped into clusters. In experiments, it is observed that the number of clusters for a video file ranges from 55-213. Then cluster centroids of master and query video files are compared and similarity scores are evaluated against the confidence measure. We experimented reference dataset with different confidence values ranging from 0.50 to 0.75 to reduce false positive rates. It is observed that better detection results are achieved when confidence value is 0.65 and thus this confidence value is used in the proposed copy detection task.

If R_1 and Q_1 are reference and query video clips, fp_r and fp_q are their corresponding video fingerprints. The similarity score (S) between R_1 and Q_1 is given by,

$$S(R_1, Q_1) = \sum_{i=1}^{m}\sum_{j=1}^{n}|fp_r(i) - fp_q(j)| \tag{11}$$

where m and n indicate number of video signatures of R_1 and Q_1 respectively. Here, L1-norm Euclidean distance is used to compute the similarity between two video clips. If the similarity score exceeds confidence measure, then the query video is reported as copy video.

III. Experimental Results

A. Reference Dataset

We used TRECVID-2007 Sound & Vision data set [20] for evaluating the proposed copy detection method. The video dataset includes 75 hours of video covering a wide variety of content. If the frame size and sampling rate of videos are different, then motion vectors of videos will also be different. Hence we transformed these video data into following uniform format: MPEG-1, 352×288 pixels and 5 frames/ sec. This dataset served as our reference dataset.

B. Query Construction

In our experiments, eleven video clips are selected from reference data set and one video clip is collected from non-reference dataset. Hence the query set includes twelve video clips (11+1) and duration of these clips vary from 15 to 25 seconds. By applying ten different transformations (listed in Table I) to the query set, final query video sequences are generated. As a result, there are totally 120 (12×10) video copies, which served as query clips for the proposed CBCD task. Each video copy is used to detect the corresponding video sequence in the reference dataset.

C. Evaluation Metric

A detection result is considered correct, if there is any overlap with the region from which the query is extracted. To measure the detection accuracy of the proposed scheme, we used standard performance metrics given by,

$$Precision = TP/(TP + FP) \tag{12}$$

$$Recall = TP/(TP + FN) \tag{13}$$

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (14)$$

True Positives (TP) are positive examples correctly labeled as positives. False Positives (FP) refer to negative examples incorrectly labeled as positives. False Negatives (FN) refer to positive examples incorrectly labeled as negatives. F-measure indicates the robustness and discrimination ability of a system.

D. Detection Accuracy

Table IV gives the detection results of Ordinal measure [6], Tasdemir's method [15] and proposed method for T1-T5 transformations. The results from Table IV demonstrate the better detection accuracy of the proposed method when compared to reference methods.

TABLE IV
COPY DETECTION RESULTS FOR T1-T5 TRANSFORMATIONS

Transformations		Ordinal	Tasdemir's	Proposed
Type	**Metric**	**Measure** (%)	**Method** (%)	**Method** (%)
T1	P	56.93	70.15	82.85
	R	50.19	68.09	75.00
	F-M	53.34	69.10	78.72
T2	P	41.69	42.17	50.00
	R	40.48	41.18	46.15
	F-M	41.07	41.66	47.99
T3	P	56.86	55.81	57.14
	R	79.14	72.56	92.30
	F-M	66.15	63.09	70.58
T4	P	59.26	60.01	63.63
	R	67.79	69.27	84.83
	F-M	63.23	64.30	72.71
T5	P	79.68	79.82	82.85
	R	80.24	79.47	85.29
	F-M	79.95	79.64	84.05

The Ordinal measure [6] is a popular global descriptor, which is extracted as follows: partitioning the image into N blocks; sorting the blocks according to their average intensity level and ranking order of blocks are considered as Ordinal signatures. Tasdemir et al.'s method [15] uses average motion vector magnitudes of frames as video signatures for the copy detection task.

For T3 (Blurring) transformation, Ordinal measure performs well (66.15%) compared to Tasdemir's method (63.09%), because of its global descriptive properties. But the proposed method (70.58%) outperforms Ordinal measure, since it uses spatial and temporal motion activity features. For T5 (Pattern insertion) transformation, both Ordinal measure and Tasdemir's methods give very similar scores (79.95% & 79.64%), which are less than that of the proposed method.

Table V shows the copy detection results of proposed and reference methods for T6-T10 transformations. The results from Table V demonstrate the detection efficiency of proposed approach when compared to reference methods.

TABLE V
COPY DETECTION RESULTS FOR T6-T10 TRANSFORMATIONS

Transformations		Ordinal	Tasdemir's	Proposed
Type	**Metric**	**Measure** (%)	**Method** (%)	**Method** (%)
T6	P	70.64	74.58	82.50
	R	71.15	72.94	84.61
	F-M	70.89	73.75	83.54
T7	P	71.35	71.87	75.00
	R	59.18	70.35	87.80
	F-M	64.69	71.10	80.89
T8	P	60.10	62.71	65.85
	R	61.54	69.64	84.37
	F-M	66.15	63.09	70.58
T9	P	68.29	65.83	80.00
	R	73.58	69.98	82.75
	F-M	70.83	67.84	81.35
T10	P	61.54	60.17	**100.00**
	R	43.67	66.73	74.54
	F-M	51.09	63.28	85.41

For T10 (Picture-inside-picture) transformation, Ordinal measure gives poor Recall rate (43.67%) when compared to proposed and Tasdemir's methods. But the proposed approach provides better Recall (74.54%), Precision (100%) rates when compared to that of Ordinal measure(43.67% & 61.54%) and Tasdemir methods (66.73% & 60.17%). The reason for the better performance of proposed method is inclusion of dominant direction of activity as one of the feature descriptors for the CBCD task.

IV. CONCLUSION

This article proposes a novel CBCD approach by integrating several motion activity features. The experimental results prove that, the proposed CBCD method improves detection accuracy by 13.9% when compared to that of reference methods. The detection results also demonstrate the effectiveness of the proposed method against various video transformations.

Our future work will be targeted at,

a) To incorporate audio fingerprints to the proposed method.
b) To improve the robustness of proposed system against complex transformations such as camcording.
c) To use multidimensional data structures to improve the scalability of K-means clustering, which is used in proposed copy detection method.
d) To perform comparative study of computational efficiency of proposed and reference methods.

ACKNOWLEDGMENT

This work is supported by the Department of Science&Technology of Government of India under Research Grant No. SR/WOS-A/ET-48/2010.

REFERENCES

[1] Chih-Yi Chiu and H.M.Wang, "Time-Series linear search for video copies based on compact signature manipulation and containment relation modeling," IEEE Transactions on Circuits and Systems for Video Technology, vol.20, no.11, 2010.

[2] Anindya Sarkar, Vishwarkarma Singh, Pratim Ghosh, Bangalore S. Manjunath, and Ambuj Singh, "Efficient and robust detection of duplicate videos in a large database," IEEE Transactions on Circuits and Systems for Video Technology, vol. 20, no. 6, 2010.

[3] C. Y. Chiu, H. M. Wang, and C. S. Chen, "Fast min-hashing indexing and robust spatio-temporal matching for detecting video copies," ACM Transactions Multimedia Computing Communications and Applications, vol. 6, no. 2, 1–23, 2010.

[4] R.Roopalakshmi and G.Ram Mohana Reddy, "Recent trends in content based video copy detection," in proc. of IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore, India, 2010, DOI:10.1109/ICCIC.2010.5705802

[5] Roopalakshmi.R and Ram Mohana Reddy.G, "Compact and efficient CBCD scheme based on integrated color features", in proc. of IEEE International Conference on Recent Trends in Information Technology (ICRTIT 2011), Chennai, India, 2011, DOI:10.1109/ICRTIT.2011.5972370

[6] Xian-Sheng Hua, Xian Chen, Hong-Jiang Zhang, "Robust video signature based on ordinal measure," in proc. of IEEE International Conference on Image Processing (ICIP), vol. 1, pp. 685-688, 2004.

[7] H. T. Shen, X. Zhou, Z. Huang, J. Shao, and X. Zhou,"UQLIPS: A real-time near-duplicate video clip detection system," in proc. of VLDB, pp. 1374-1377,2007.

[8] David G.Lowe, "Distinctive image features from scale-invariant key points," International Journal of Computer Vision, 91-110,2004.

[9] Herbert Bay, Tinne Tuytelaars, Luc Van Gool, "SURF: Speeded up robust Features," Computer Vision and Image Understanding, 346-359,(2008).

[10] Roopalakshmi.R and Ram Mohana Reddy.G, Efficient video copy detection using simple and effective extraction of color features, in proc.of ACC-2011 in Springer-Verlag, Part IV, CCIS 193, pp. 473-480, 2011.

[11] R.Roopalakshmi and G.Ram Mohana Reddy, "A novel approach to video copy detection using audio fingerprints and PCA," second International Conference on Ambient Systems, Networks and Technologies (ANT-2011), Niagara Falls, Canada, 2011, in press.

[12] A. Divakaran, R. Regunathan, and K. A. Peker, "Video summarization using descriptors of motion activity: A motion activity based approach to key-frame extraction from video shots," Journal of Electronic Imaging, vol. 10, pp. 909-916, October 2001.

[13] I. Koprinska and S. Carrato, "Temporal video segmentation: a survey," Signal Processing: Image Communication, Elsevier Science, 2001.

[14] A.Hampapur, Ki-Ho Hyun and R.Bolle, "Comparison of sequence matching techniques for video copy detection," in proc. of the IEEE International Conference on Multimedia and Expo (ICME-01), pp. 737-740, 2001.

[15] Kasim Tasdemir, A. Enis Cetin," Motion vector based features for content based video copy detection," in proc. of IEEE International Conference on Pattern Recognition -2010, DOI:10.1109/ICPR.2010.767

[16] Sylvie Jeannin and Ajay Divakaran, "MPEG-7 visual motion descriptors", IEEE Transactions on Circuits and Systems for Video Technology,vol.11 ,no.6, June 2001.

[17] X. Sun, D. Ajay, and B. S. Manjunath, "A motion activity descriptor and its extraction in compressed domain," in proc. of IEEE Pacific-Rim Conf. Multimedia (PCM), pp. 450-453, October 2001.

[18] Andreas Savakis, Pawel Sniatala, and Radoslaw Rudnicki ,"Real-time video annotation using MPEG-7 motion activity descriptors", in proc. of MIXDES 2003.

[19] S. Benini, L.-Q. Xu, R. Leonardi, "Using lateral ranking for motion-based video shot retrieval and dynamic content characterization," in Proc. of CBMI'05, Riga, Latvia, June 21-23, 2005.

[20] TRECVID 2010 Guidelines [Online]. Available: http://www.nlpir.nist.gov/projects/tv2010/tv2010.html