ISSN:0975-9646



# A Segmentation Based Sequential Pattern Matching for Efficient Video Copy Detection

Sanket Shinde[#1], Girija Chiddarwar[#2]

[#]*Dept. of Computer Engineering, Sinhgad College of Engineering*
*Savitribai Phule Pune University, Pune, India*

***Abstract*—We propose a video copy detection system which is sufficiently robust to detect transformed versions of original videos with ability to pinpoint location of copied segments. Precisely due to the good stability and discriminative capability, SIFT feature is used for visual content description. However SIFT based feature matching has high computational cost due to large number of keypoints and high dimensionality of its feature vector. Thus to minimize computational complexity, video content is divided into number of segments depending on homogeneity of video content. SIFT features are extracted from keyframes of video segments. Then SIFT features are quantized to generate clusters. Further binary SIFT are generated for every cluster for optimized matching. To perform video segment matching with SIFT descriptors, firstly visual words matching is done at cluster level then at feature level, similarity measure is employed. In order to find video copy detection, an optimal sequential pattern matching is done which will ensure effectiveness and efficiency of the proposed video copy detection system. Our video copy detection system gives improved results in terms of accuracy and has been able to detect most of the dataset query video transformations in time-efficient manner.**

**Keywords— Copyright protection, content based video copy detection, feature extraction, visual features, SIFT feature.**

## I. INTRODUCTION

The expeditious growth of the world wide web has allowed netizens in acquiring and sharing digital media in relatively simpler way due to improvements in data transfer and processing capabilities. Due to wide use of digital devices like smart phones, cameras, more and more images and videos are produced by netizens [1] and are uploaded on the internet for business promotions or community sharing. The very easiness of video copy creation techniques instigated problem of video copyright violations so it is needed to have mechanism to protect copyright of digital videos.

Existing video copy detection techniques are mainly classified into watermarking based and content based copy detection. Each of these techniques has its own merits and drawbacks. Watermark embeds useful metadata and maintains low computational cost for copy detection operation, but watermark based copy detection does not perform well against common transformations such as rotate, blur, crop, camcording, resize, which are performed during video copy creation. If original video is distributed on video sharing sites before watermark embedding, then watermark based detection system does not have any reactive measure. Also due to video compression, possibility of vanishing watermark arises. There are many methods for watermark embedding. These watermark based schemes are based on fourier, cosine, wavelet transforms. But these transform based methods usually perform embedding of watermark into predefined set of coefficients of their corresponding domain. Thus whenever an attacker scrutinizes image and finds pattern of embedding watermark into predefined set of coefficients, he can easily remove embedded watermark.

Recently formulated Content Based Video Copy Detection (CBVCD) [2,3,4] algorithms as contrast to watermark-based methods do not rely on any watermark embedding and are invariant to most of the transformations. These CBVCD algorithms extract invariant features from the media content itself so CBVCD mechanism can be applied to probe copyright violations of digital media on the internet as an effective alternative to watermarking technique.

CBVCD algorithms first extract distinct and invariant features from the original and query videos. If same features are found in both original and query videos, then query video may be a copied version of original video.

Underlying assumption of CBVCD algorithms is that a sufficient amount of information is available in video content itself to generate its unique description; it means content itself preserves its own identity. This paper is organized as, Section II reviews variety of visual features employed by different video copy detection systems mainly categorizing features into different types depending on their extraction mechanisms. Section III describes proposed content based video copy detection system along with its algorithmic structure. Finally Section IV presents discussion of results and section V concludes this paper.

## II. RELATED WORK

For attaining both efficiency and effectiveness in video copy detection, the feature signature should adhere to two crucial properties, uniqueness and robustness. Uniqueness stipulates discriminating potential of the feature. Robustness implies potential of noise resistance means features should remain unchanged even in case of different photometric or geometric transformations. Once set of keyframes has been decided, distinct features are extracted from keyframes and used to create signature of a video. We mainly focus on visual features suitable for video copy detection, includes spatial features of keyframes, temporal features and motion features of video sequence. Spatial features of keyframes are categorized into global and local features.

### *A. Global Features*

Global features provide invariant description of a video frames rather than using only selective local features. This approach works quite well for those video frames with unique and discriminating color values. Though merits are being easy to extract and require low computational cost but global features [5,6,7,8,9,10,11] failed to differentiate between foreground and background.

Yusuke et al. [5] perform feature extraction by applying 2D-DCT on each predefined block of keyframe to get AC coefficients, this DCT-sign based feature is used as signature of both reference and query video keyframes. Gitto George Thampi et al. [6] use Daubechies wavelet transform to obtain feature descriptor from video frames. The wavelet coefficients of all frames of same segment are extracted and then mean and variance of the coefficients are computed to describe each segment of a video sequence.

Xian-Sheng Hua et al. [7] use ordinal measure for generating signature of video segment in which video frame is divided into number of blocks then for every block, average gray value is computed. Then these values are ranked in increasing order. The ranked sequence of average gray values gives ordinal measure; it incorporates rank order of blocks of video frame according to their average gray values. It is highly invariant to color degradation but not to geometric transformations.

Spatial correlation descriptor [10] uses inter-block relationship which encodes the inherent structure (pairwise correlation between blocks within video frame) forming unique descriptor for each video frame. The relationship between blocks of video frame is identified by content proximity. Original video and its transformed version will not be having similar visual features; but they preserve distinct inter-block relationship which remains invariant.

Hui Zhang et al. [11] employs Block based Gradient Histogram (BGH) which is to be extracted from set of keyframes. Firstly keyframes are divided into fixed number of blocks and for every block a multidimensional gradient histogram is generated. Set of these individual gradient histograms constitutes BGH feature for every keyframe. BGH is found to be robust against non-geometric transformations

### *B. Local Features*

Local feature based methods firstly identify points of interest from keyframes. These points of interest can be edges, corners or blobs. Once the interest point is chosen, then it is described by a local region surrounding it. A local feature represents abrupt changes in intensity values of pixel from their immediate neighborhood. It considers changes occurred in basic image properties like intensity, color values, texture. An interest point is described by obtaining values like gradient orientations from a region around that interest point. Local feature based CBCD methods [11,12,13,14] have better detection performance on various photometric and geometric transformations but only disadvantage is being high computational cost in matching.

Scale Invariant Feature Transform (SIFT) [15] employs difference of Gaussian to detect local maxima values and these interest points are described by gradient histogram based on their orientations. Hong et al. [12] use SIFT descriptor due to its good stability and discriminating ability. SIFT feature performs well among local feature category and is robust to scale variation, rotation, noise, affine transformations. Speeded-Up Robust Features (SURF) [16] feature is based on Haar wavelet responses summed up around point of interest, which give maximum value for Hessian determinant.

Hui Zhang et al. [11] use SURF feature for representing points of interest having local maxima. SURF feature has better real time performance as compared to SIFT. Hessian-Laplace feature is combination of Hessian affine detector and Laplacian of Gaussian. It employs Laplacian of Gaussian to locate scale invariant interest points on multiple scales. While at every scale, interest point attaining maximum value for both trace and determinant of Hessian matrix are selected to be affine invariant interest points. Hessian-Laplace is invariant to many transformations like scale changes, image rotation and due to detection is done at multiple scales so it is quite resilient to encoding, blurring, additive noise, camcording effects. Local feature based CBCD methods [13, 14] employ Hessian-Laplace feature along with Center-Symmetric Local Binary Patterns (CSLBP) for feature description.

### *C. Motion Features*

Color based features have difficulty in detection of camera recorded copy as frame information gets significantly distressed. This problem can be efficiently resolved by employing motion features which use motion activity in a video sequence as it remains unchanged in severe deformations. Tasdemir et al. [17] divide individual video frame into number of blocks and record motion activity between blocks of consecutive frames at reduced frame rate.

Roopalakshmi et al. [18] has implemented similar type of descriptor known as motion activity descriptor for measuring activity of a video segment whether it is highly intense or not. This motion activity descriptor derives intensity of action, major direction of motion activity, distribution of motion activity along spatial and temporal domains.

### *D. Temporal Features*

Temporal features represent variations in scene objects with respect to time domain rather than examining spatial aspect of each video frame. Shot length sequence [19] captures drastic change in consecutive frames of a video sequence. This sequence includes anchor frames which represent drastic change across consecutive frames. This sequence is computed by enlisting time length information among these anchor frames. Shot length sequence is distinctly robust feature as any separate video sequences will not be having set of successive anchor frames with similar time segment.

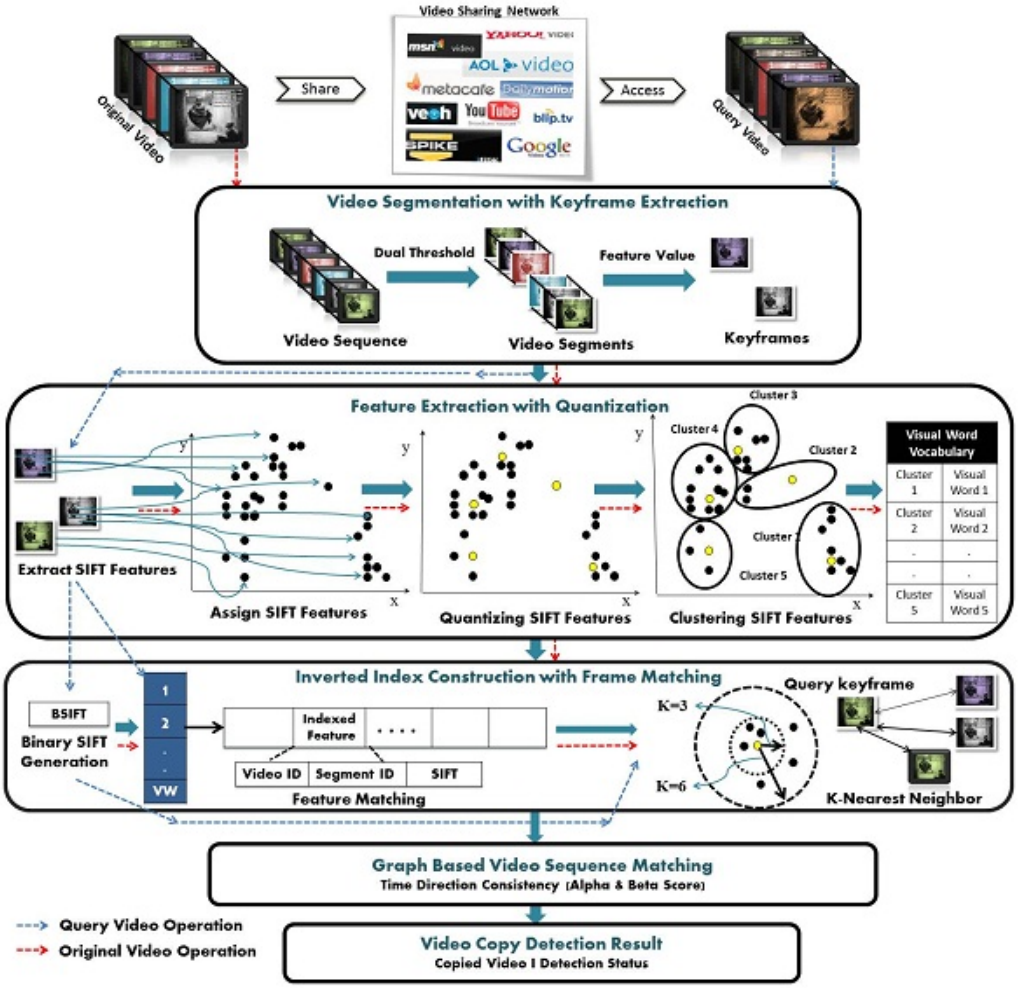


Fig. 1. Proposed content based video copy detection system

## III. PROPOSED METHODOLOGY

The architecture of proposed content based video copy detection system with local feature extraction and graph-based video sequence matching is shown in Fig 1 and is described as
sequence of stages as follows,

### A. *Video Segmentation and Keyframe Extraction*

Firstly segmentation based on changes in feature values of contiguous frames in video sequence will be performed. Input is reference videos and query video. Output will be set of segments for both reference videos and query video. Secondly extraction of keyframe from set of video frames of an each video segment will be done such that feature value of keyframe is most similar to average feature value of that video segment. Input is set of reference video segments and query video segments. Need is to represent individual video segment with keyframe. Output will be set of keyframes for both reference video segments and query video segments. Once video keyframes are generated it is needed to extract local invariant features from them.

### B. *Visual Feature Extraction*

This stage performs its operations as shown in Fig 2. Input is set of reference video and query video keyframes. Need of feature extraction is to represent individual keyframe with invariant SIFT feature description. Output is 128-dimensional SIFT feature descriptor for each keypoint extracted from reference video and query video keyframes.

*1) Perform Gaussian scale space generation:* Each keyframe is convolved with Gaussian filters at different scales to get scale space.

*2) Find scale-space extrema from DoG:* Difference of successive Gaussian-blurred keyframes is computed. Keypoints are obtained by calculating local scale space extrema of the Difference of Gaussians (DoG) scale space.

*3) Perform orientation assignment:* Each keypoint is described by histogram of gradient orientations of neighboring pixel values around it.

*4) Generate feature description:* It takes a collection of vectors in the neighborhood of each keypoint mean and consolidates this information into a set of eight vectors called the descriptor.

### C. *Feature Quantization*

This performs feature quantization of SIFT features. Bag-of-words approach is used to quantify SIFT features to a fixed number of clusters. In order to perform this operation, K-means clustering algorithm is used to identify suitable cluster for each SIFT feature while each cluster

centroid will be designated as a visual word for vocabulary generation. Input is set of SIFT features of reference video keyframes. Need is to quantify/classify individual SIFT features of reference keyframe into discrete clusters. Output is feature clusters with their respective mean/centroid.

*1) Assignment step:* Assign each SIFT descriptor to the cluster whose mean has least distance with descriptor. It means key points in each key frame are assigned to clusters which are their nearest neighbors.

*2) Update step:* Calculate new means to be centroid of data values in the new clusters.

### *D. Visual Vocabulary Generation*

This stage performs generation of vocabulary by assigning each cluster centroid as individual visual word. Each cluster centroid designated as individual visual word. All the visual words collectively generate visual word vocabulary. Input is clusters with their respective means from previous stage. Need of this stage is to generate visual word vocabulary for inverted index construction. Output is visual word vocabulary.

### *E. Binary SIFT Generation and Inverted Index Generation*

Need of this stage is to generate compact feature descriptor for fast frame matching. Output is set of binary SIFT [20] features of reference video keyframes and query video keyframes. Input is set of SIFT (128-dimensional feature descriptor) features of reference video keyframes and query video keyframes. This stage typically performs,

*1) Binary SIFT generation:* Convert a standard SIFT descriptor to a binary bit stream, binary SIFT (128-bit descriptor) as shown in Fig. 2. Binary SIFT will reduce overall computation time as does have only binary values in its feature description.

*2) Inverted index generation:* Assign entry (video id, segment id, keypoint BSIFT) to a visual word in inverted file structure if that keypoint (SIFT) belongs to that visual word/cluster. This will generate index structure to further retrieve possible matches with query video features.

### *F. Similarity based Segment Matching*

This stage returns matching results from set of segments of reference video based on similarity. Input is set of reference video keyframes and query video keyframes representing respective segments. Output is set of matching results between reference and query video segments. For query video segments and reference video segments based on similarity measurement, matching results are obtained.

### *G. Graph based Video Sequence Matching*

This stage generate the matching result graph according to the matching results and search for corresponding original video copy with help of matched result graph. Input is set of matching results between reference and query video segments. Need of this operation is to get original videos matching to query or suspicious videos. Outpu of this stage is to get maximum number of matched video segments of query video.

**procedure** FEATUREGENERATOR
- $j \leftarrow$ *video id* ▷ $j$: Identity of video
- $K_m^j \leftarrow$ *keyframe id* ▷ $K_m^j$: Keyframe identity of $m^{th}$ segment of $j^{th}$ video
- $G(s) \leftarrow GaussianKernel$
- $x \leftarrow Coordinates(x, y)$
- **for** every $m^{th}$ keyframe of $j^{th}$ video **do**
  - $L(x,s) \leftarrow G(s) \otimes K_m^j(x)$ ▷ Gaussian scale space generation
  - $DOG(x,s) \leftarrow [L(x, s+\delta s) - L(x,s)]$ ▷ Difference-of-Gaussian
  - $M_n^K \leftarrow max[DOG(x,s)]$ ▷ Scale-space Extrema
  - **for** every $n^{th}$ keypoint of $K^{th}$ keyframe **do** ▷ Orientation assignment
    - **for** every (x,y) in a neighboring region around the keypoint **do**
      - $m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2}$
      - $\theta(x,y) = \text{atan2}(L(x,y+1) - L(x,y-1), L(x+1,y) - L(x-1,y))$
      - $featuredescriptor \leftarrow generateOrientationHistogram()$
      - $DR_n^K \leftarrow featuredescriptor$ ▷ 128-dimentional feature vector
      - $(c_1, c_2, ...c_{128}) \leftarrow DR_n^K$ ▷ 128-dimentional feature vector
      - $f_m \leftarrow median(c_1, c_2, ...c_{128})$
      - **for** every $c_i$ in element set $(c_1, c_2, ...c_{128})$ **do**
        - **if** $c_i > f_m$ **then** ▷ Binary SIFT Generation
          - $b_i = 1$
        - **else** $b_i = 0$
        - **end if**
      - **end for**
      - $BR_n^K \leftarrow (b_1, b_2, ...b_{128})$ ▷ 128 dimensional binary feature vector
    - **end for**
  - **end for**
- **end for**

**end procedure**

Fig. 2. Visual SIFT feature generation with descriptor generation

## IV. RESULTS AND DISCUSSION

By applying proposed video copy detection mechanism, original video from reference videos is identified against query video. The experimental evaluation is performed in which suspicious query videos are being analyzed which are generated by applying different transformations. Result will be represented with respect to following parameters, (1) Query video, (2) Detected video: original video from video repository whose copied version is query video. (3) Detection status: check whether video is being correctly detected or not. This will finally generate detection rate in terms of number of nearest neighbor segment matches. MUSCLE-VCD-2007 [21] dataset is the benchmark used for CIVR 2007 video copy detection evaluation. MUSCLE VCD dataset can be accessible by public due to which it is being popular among video copy detection community. MUSCLE VCD dataset provides ground truth data(composed of 101 reference videos with total time length of 60 hours and 18 query videos with total time length of 4 hours, each video at 25 fps) for judging system's detection accuracy based on tasks :detecting copies(ST1) and localizing copy segments from video sequence(ST2). Due to limited resource availability, video copy detection system has been evaluated on mainly ST1 videos of MUSCLE-VCD-2007 dataset.Out of these ST1 query videos, typically following query videos have been evaluated on our video copy detection system, each query video is firstly converted from original .mpeg to .mp4 format. As per system requirement specification each query video and its corresponding orignal video has 25 frames per second, time duration of 1 Minute and resolution of 384 × 288.

Firstly basic operations for video copy detection are performed. At first, performance of video segmentation is evaluated. For every original video and query video, a video segmentation operation is performed. The Table.1 gives segmentation results for original video and its respective query video. Experiments are performed on Intel Core 2 Duo processor with 3GB RAM in Windows environment. The segmentation result in Table.1 shows that how different transformations can change possible number of video segments. For example, a query video, $STQ_1$, color changes blurring effect on Movie27, has not changed number of video segments in original video due to reason that blurring merely changes actual scenes while query video, $STQ_6$, camcording and subtitles effect on Movie76 has changed number of video
segments in original video from 13 to 76 due to reason that camcording effect has changed frame rate and subtitles changed actual frame content. The video segmentation has shown good robustness against transformations like colour changes, blurring, camcording with angle, analogic noise, change in YUV values, strong re-encoding. Video segmentation is done based on changes in feature values of contiguous video frames in video sequence. This segmentation has effectively reduced redundancy.

Table. 1. Video Segmentation Evaluation

| Original Video (25 fps/1 Minute/384×288) | Video Segments Generated | Corresponding Query Video (25 fps/1 Minute/384×288) | Video Segments Generated |
|---|---|---|---|
| Movie 27 | 68 | $STQ_1$ | 68 |
| Movie 44 | 14 | $STQ_5$ | 10 |
| Movie 76 | 13 | $STQ_6$ | 37 |
| Movie 9 | 40 | $STQ_9$ | 39 |
| Movie 21 | 16 | $STQ_{10}$ | 17 |

The implemented video segmentation performs well in terms of identifying video shots. The grid based feature comparison of consecutive video frames has generated list of video segments. Out of these video segments, keyframes have been selected to represent individual video segments. Again more the number of video segments detected in video scene, more can be accuracy of video copy detection due to fact that more segments gives minute changes in video scenes The conduction of tests showed that video segmentation approach based on feature analysis of consecutive frames shown video shots have been accurately detected which would have been quite time consuming process if done with local or global feature analysis.

The video copy detection system is evaluated for its copy detection accuracy. The results of possible segment matches to query video segments are shown in Fig.3. The results are generated in sequence as, first column shows query video in MUSCLE-VCD-2007 dataset which is tested for its originality, second column represents respective video transformations of query videos, third column shows corresponding detected video for respective query video while next column gives detection status of VCD system for whether system correctly identified query video as copied video or not and last column gives how much of detected videos have nearest neighbour match.

The graphical representation of similarity matching of video segments is showed in Fig. 3. The similarity matching is done based on maximum number of nearest segment matches to query segments. Out of the top-k (k=4) nearest neighbours obtained for every query segment, the maximum segment matches are being considered for finding out original video of which they are part of.

The video copy detection system gives good result for query videos $STQ_1$, $STQ_6$, $STQ_9$ in terms of number of segment matches out of Top-k (k=4) nearest neighbours while query videos, $STQ_5$,$STQ_{10}$ have been detected as copied versions of original videos but gave less number of segment matches in terms of nearest neighbours. So we need to perform graph based sequence matching in which we calculate individual video's average distance of segment matches thus formulated alpha score which is a fraction of number of segment matches out of top-k nearest neighbours of query video segments. Thus our video copy detection system has been able to detect most of the dataset query video transformations.

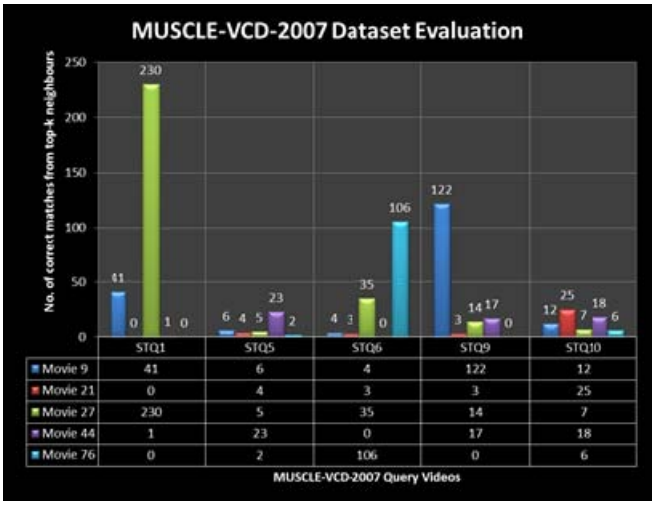


Fig. 3. Number of Segment Matches from Top-k Nearest Neighbours

## V. CONCLUSION

The proposed content based video copy detection system is able to detect video copies that are generated by applying different photometric and geometric transformations. This video copy detection system portrays the content of the video

by its features which are not sensitive to the transformations. The feature set is sufficiently robust and discriminating. The proposed video copy detection system is able to achieve video copy detection task by (1) performing dual threshold based video segmentation to reduce redundant frame matching; (2) utilizing an efficient SIFT based compact binary feature vectors for fast frame matching; (3) performing fast detection by generating compact feature description and quantization while achieving high detection accuracy and low false alarm rate owing to use of optimal graph based video sequence matching.

## REFERENCES

[1] Youtube statistics report, http://www.youtube.com/yt/press/statistics.htm

[2] R. Roopalakshmi, G. Ram Mohana Reddy, "Recent trends in content based video copy detection", in IEEE Proc.of Int. Conf. on Computational Intelligence and Computing Research, 2010.

[3] J.M. Barrios, "Content-based video copy detection", in ACM Proc. of Int.Conf. on Multimedia, pp. 1141-1142, 2009.

[4] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, F.Stentiford, "Video copy detection: a comparative study", in Proc. ACM int. conf. on Image and Video Retrieval, pp. 371-378, 2007.

[5] Yusuke Uchida, Koichi Takagi, Shigeyuki Sakazawa, "Fast and accurate content based video copy detection using bag-of-global visual features", in IEEE Proc. of ICASSP, 2012.

[6] Gitto George Thampi, D. Abraham Chandy, "Content-based video copy detection using discrete wavelet transform", in IEEE Proc.of Conf. on Information And Communication Technologies, 2013.

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

[8] Chenxia Wu, Jianke Zhu, Jiemi Zhang, "A content-based video copy detection method with randomly projected binary features", in Proc. of IEEE Computer Vision and Pattern Recog. Workshops, pp.21-26, 2012.

[9] MC Yeh, K. T. Cheng, "Video copy detection by fast sequence matching", in ACM Proc. of Int. Conf. on Image, Video Retrieval, 2009.

[10] MC Yeh, KT Cheng, "A compact, effective descriptor for video copy detection", in ACM Proc. Int. Conf. on Multimedia, pp. 633-636, 2009.

[11] Hui Zhang, Z. Zhao, A.Cai, Xiaohui Xie, "A novel framework for content-based video copy detection", in IEEE Proc. of IC-NIDC, 2010.

[12] H. Liu, H. Lu, X. Xue, "A segmentation and graph-based video sequence matching method for video copy detection", in IEEE Transactions on Knowledge and Data Engineering, pp. 679-698, 2013.

[13] M. Douze, H. Jgou, and C. Schmid, "An image-based approach to video copy detection with spatio-temporal post-ltering", in IEEE Trans. Multimedia, pp. 257-266, 2010.

[14] M. Douze, H.Jegou,C.Schmid, and P. Perez, "Compact video description for copy detection with precise temporal alignment", ECCV, 2010.

[15] D. G. Lowe, "Distinctive image features from scale invariant keypoints", in Int. Journal on Comput. Vision, pp. 91-110, 2004.

[16] Herbert Bay, Tinne Tuytelaars and Luc Van Gool, "SURF: Speeded Up Robust Feature", in Proc. of European Conf. on Computer Vision, Springer LNCS, volume 3951, part 1, pp. 404-417, 2006.

[17] Kasim Tasdemir, A. Enis etin, "Content-based video copy detection based on motion vectors estimated using a lower frame rate", in Proc. of Signal, Image and Video Processing,Springer, pp 1049-1057, 2014.

[18] R.Roopalakshmi, G.Ram Mohana Reddy, "A novel CBCD approach using MPEG-7 Motion Activity Descriptors", in IEEE Proc. of Int. Symposium on Multimedia, 2011.

[19] P. Wu, T. Thaipanich,C.-C.j. Kuo, "A suffix array approach to video copy detection in video sharing social networks", in Proc.of ICASSP, 2009.

[20] W. Zhou, H. Li, Y. Lu, M.Wang, and Q. Tian, "Visual word expansion and BSIFT verification for large-scale image search", Multimedia Syst., pp. 110, 2013.

[21] J. Law-To, A. Joly, and N. Boujemaa,"MUSCLE-VCD-2007: a live benchmark for video copy detection," 2007, http://www-rocq.inria.fr/imedia/civr-bench/.