Journal of Theoretical and Applied Information Technology
31st October 2021. Vol.99. No 20
© 2021 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

VIDEO REPRESENTATION BASED ON OPTICAL FLOW FOR DYNAMIC CONTENT ANALYSIS

[1]NARRA DHANALAKSHMI, [2]Y. MADHAVEE LATHA, [3]AVULA DAMODARAM

[1]Associate Professor, VNRVJIET, Department of ECE, India

[2]Professor. MRECW, Department of ECE, India

[3]Professor. JNTUH, Department of CSE, India

E-mail: [1]dhanalakshmi_n@vnrvjiet.in, [2]madhuvsk2003@yahoo.co.in, [3]damodarama@rediffmail.com

ABSTRACT

The efficient organization of multimedia databases challenges content -based representation to retrieve the video of interest. This paper aims to represent given video by considering its dynamic content through the analysis of optical flow. It is tended to have video segmented into overlapped sequence of frames based on gray content similarity. This step can facilitate analysis of complex video into elementary scenes. The principle involved in representing the content of video is considering the spatial movement of video content across the frames. The algorithm is designed to find the dynamic content by observing all levels of motion in the video through pyramid generation. Then, an optical flow is derived in terms of spatial and temporal information of motion regions. The histogram representation is created with both the rank and orientation of the optical flow. This kind of methodology contributes to efficient representation which enhances effective content analysis to improve the efficiency of further stages. The videos of You Tube 8M and UCF Sports data sets have used to evaluate the algorithm.

Keywords: *Temporal video segmentation, Gaussian Pyramid, Optical flow, Normalized Histogram Intersection Similarity, Video Representation*

1. INTRODUCTION

Every day, large amount of video data are being produced and uploaded to multimedia sharing databases [1]. It is not feasible to choose manual processing of multimedia data to solve various challenges in multimedia applications [2]. Understanding and representing motion content in videos remains a challenge in various domains such as computer vision, pattern recognition, machine learning etc. Hence, algorithms are needed to analyses the information about the videos to meet the user requirements[3-5]. It involves two important processes such as: feature extraction and its representation. Any problem in these processes may mislead the interpretation and decision process which in turn result in degradation of performance. Videos are information rich and also hierarchically structured. Hence, the methods are proposed for video representation towards various applications by considering individual frame, consecutive frames to extract motion clue, shot, scene and so on [6-8]. A new component based approach to represent the content in the given video is proposed in [9]. Feature extraction has been considered as significant task in many computer vision applications [10] and used Zernike phase-based descriptor for representation. There are many other low-level visual features extracted for representation such as color, corners, edges, shape, texture, interest points, and so on [11-13]. These extracted features need to be represented for analysis and histogram representation is one of such methods. [14-17] utilize color histogram model, [18-19]apply edge histograms, [20-22] discuss texture feature for representation targeted to different application scenarios. The other popular methods [23-24] are also used which uses pixel intensities for feature extraction. Motion is a unique element in the video processing to benefit many tasks such as video retrieval, object recognition, fusion, classification, video summarization, annotation, object tracking and so on. This work has two phases: Motion information extraction and content representation. Former involves detection and extraction of useful motion information from small number of consecutive video frames and later deals with efficient representation for interpreting the video content.

One wing of existing approaches will make use of low level visual features whose characteristics are not sufficient to understand the context. Other approaches are hard to obtain labeled data set for the feature of interest in the entire video. Also it suffers with the requirement of powerful hardware and more time to build the model. Hence, there is a need to balance the requirements of many computer vision and pattern recognition algorithms. As the performance highly depends on feature extraction and representation, there is a need to bridge the gap between low level features and high level semantics. In this work, the relative distance between the content of two consecutive frames is calculated as a part of feature extraction. It results in 2-channel array (u, v) referred to as optical flow field which contains information about all pixels. Then, a histogram of rank oriented optical flow is derived by calculating orientation and rank of optical flow field. This algorithm is designed by imitating the visual information identification and representation in the brain.

The rest of this paper is organized as follows: the methodology for video content representation is provided in detail first. Then, the results are discussed and finally the conclusion is presented.

2. METHODOLOGY

In this section we present our approach, illustrated in Figure 1, for representing video content in terms of histograms based on optical flow. Initially grabbing of frames from the given video is performed to manage deceleration of video representation algorithms which facilitate accessing of all frames in the correct sequence. The subsequent stages of our algorithm are explained as follows:

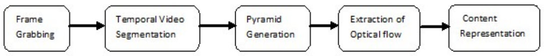


Figure 1: The process to represent content variation through histograms.

2.1 Temporal Video segmentation

The video is segmented at a finer temporal level so that complete action of an object(s) should fall into the same video segment for analysis. It can be achieved by generating the overlapped segments. The reason behind this is to deal with the content shared between the segments. These results are suitable for further analyzing the video content towards understanding multimedia databases, tasks in surveillance applications, tracking an object motion completely, matching in video information system for indexing kind of applications and so on. Temporal relation between the frames with respect to their semantic content is found by content representation in terms of histograms and then find degree of similarity between the frames using Histogram Intersection Similarity Metric. The method proposed in [25] is used to get video segments with modifications. Major modifications include use of grey information, forming directly global similarity descriptor and the use of similarity metric. The color information is converted into gray scale to get dimensionality reduction which improves processing speed without much compromising the efficiency. The gray content of each frame is distributed over 8 bins in the histogram.

Histograms of two subsequent frames are compared using Normalized Histogram Intersection Similarity Metric (NHISM) to find the similarity between the frames. Its value ranges from 0 through 1(more similar). The intersection is observed by taking smallest value of two corresponding bins of consecutive frames followed by normalization. It is described by the following formula:

$$NHISM = \frac{\Sigma_{n=0}^{7} min(I_{i,n}, I_{i+1,n})}{\Sigma_{n=0}^{7} I_{i,n}} \quad (1)$$

Where, I refers to frames in a video with i for its identification in a sequence and n represents bin index in the histogram. If the deviation between a pair of frame is more, then that frame can be termed as boundary of a segment. Forevery overlapped segment, tentative left side boundary is set for the search window and then is readjusted to the content of the current segment using NHISM. The right side boundary is to be identified as end of the current segment as the algorithm runs. The number of search windows with confirmed boundaries gives the number of temporal segments in the video.

2.2 Pyramid Generation

The algorithm for spatio-temporal content analysis can efficiently deal with varying size and motion patterns of objects by generating pyramids. Pyramid is constructed by generating a series of subsampled versions of the original frame. It facilitates to find optical flow for each pixel at different resolutions. For this, the original frame is Gaussian pre-filtered and down sampled to create the sequence of decreased resolution frames without aliasing. The Gaussian with typical standard deviation is used to get required scale by changing the amount of blurring before sampling. The first level of image resolution(NxN) is same as the original frame and subsequent images in the pyramid are scaled by a factor 1/√2 . At level ‘i’ the size of a frame(f_i) is measured as (N/2ixN/2i). The levels of the pyramid (0 < l < N) are generated by using the following equation:

$$F_l(i,j)=\Sigma_{m=-2}^{2}\Sigma_{n=-2}^{2}w(m,n)F_{l-1}(2i+m,2j+n) \quad (2)$$

The equivalent Gaussian like kernel w(m,n) is convolved with the previous images at finer levels to generate Gaussian pyramids. As this kernel is chosen as small and separable, the algorithm achieves good computational efficiency. The size of the kernel is taken as 5x5. The practical value of pyramid level is fixed to 3 because it provides optimum results. Further increase in level does not much affect the end results. This approach can well capture all speeds in a video by detecting speeds at one resolution which might be undetected at another resolution.

2.3 Extraction of Optical flow

In this step, the displacement between the successive frames in the spatial domain is estimated in terms of 2-channel array (u, v) which is also referred to as optical flow. A dense flow 2-frame polynomial expansion method, proposed by Gunner Farneback [26], deals with the spatial localization as well as inconsistent motion over sequence of frames is adopted in this work. Each image point is transformed into a set of expansion coefficients using a polynomial model by applying convolution operation. These are helpful to find pixel correspondences between the frames in spatial domain through translation. The procedure to find the optical flow is as follows: The neighborhood of subsequent frames F_n and F_{n+1} is approximated using the quadratic polynomialssuch as:

$$F_n(x)=X^TA_nx+b_n^Tx+c_n \quad (3)$$

$$F_{n+1}(x)=X^TA_{n+1}x+b_{n+1}^Tx+c_{n+1} \quad (4)$$

Find polynomial expansion coefficients A_n, b_n, c_n and A_{n+1}, b_{n+1}, c_{n+1} for the two subsequent frames F_n and F_{n+1}respectively.

$$\begin{aligned}F_{n+1}(x)=F_n(x-d)&=(x-d)^TA_n(x-d)+b_n^T(x-d)+c_n\\&=x^TA_nx-2d^TA_nx+d^TA_nd+b_n^Tx-b-n^Td+c_n\\&=x^TA_nx+(b_n-2A_nd)^Tx+d^TA_nd-b_n^Td+c_n\\&=x^TA_{n+1}x+b_{n+1}^Tx+c_{n+1}\end{aligned} \quad (5)$$

Then

$$A_{n+1}=A_n \quad (6)$$

$$b_{n+1}=b_n-2A_nd \quad (7)$$

$$c_{n+1}=d^TA_nd-b_n^Td+c_n \quad (8)$$

The displacement fields are estimated from the polynomial expansion coefficients as observed movement of point x.

$$2A_nd=-(b_{n+1}-b_n) \quad (9)$$

$$d=-\frac{1}{2}A_n^{-1}(b_{n+1}-b_n) \quad (10)$$

The above approximations are rewritten for 2-dimentional frames as:

$$A(x,y)=\frac{A_n(x,y)+A_{n+1}(x,y)}{2} \quad (11)$$

$$\Delta b(x,y)=-\frac{1}{2}(b_2(x,y)-b_1(x,y)) \quad (12)$$

$$A(x,y)d(x,y)=\Delta b(x,y) \quad (13)$$

$$d(x,y)=A(x,y)^{-1}\Delta b(x,y) \quad (14)$$

To handle all variations of motions, multi-scale displacement estimation approach is used. In this approach, the analysis begins at coarser level with the rough estimation of initial motion. As the coarser levels are of low resolution images and have small frame to frame displacements, the computational cost is low. As actual knowledge about the displacement field is not known, an initial displacement zero is given to the lowest resolution level. The basic operation involved in this approach is to perform estimation of displacement done at one level that is refined at next higher resolution pyramid level. The new polynomial expansion coefficients need to be computed every time it propagates from one level to next higher resolution level. With this the undetected large movement can be detected now.

2.4 Video Content Representation

This method of histogram representation contains the motion information which is significant to the content of video. Optical flow extraction generates 2-channel array that is (u,v). The rank(D) of pixel movement between two consecutive frames along with its orientation(O) is derived by using the following equations:

$$D=\sqrt{u^2+v^2} \quad (15)$$

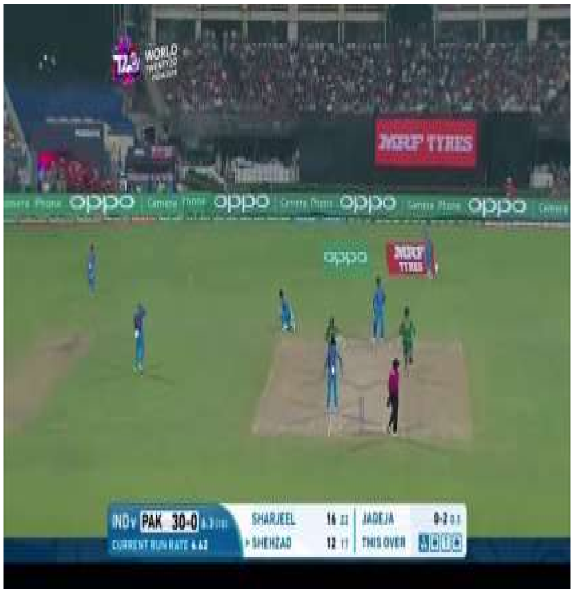
$$O=tan^{-1}\frac{v}{u} \quad (16)$$

Then the rank oriented histogram is generated by following the concept used to form HOG[23]. Few modifications have been made to the original work:(i) consider pixel displacements instead of representing pixel intensities directly (ii) Making global histogram instead of concatenating local region histograms. The global histogram representation is enough to carry optical flow information. It reduces the length of feature representation and also speeds up the processing. The histogram is generated by distributing the resultant orientations over 9 bins and bins are weighted using rank(D) of optical flow. During plotting, bin numbers are taken on x-axis which corresponds to orientation of optical flow and its rank on y-axis.

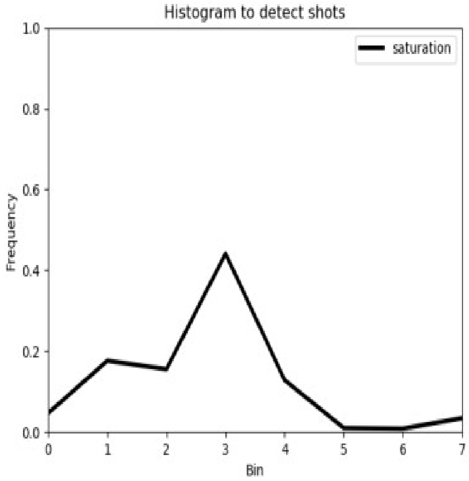
3. RESULTS AND DISCUSSIONS

The algorithm is tested over videos downloaded from the data sets: YouTube 8M and UCF Sports. Preprocessing like temporal video segmentation for scene transition is an important step. In many applications, this step may simplify the task at later stages. Example: the group of histograms which belong to a scene may easily be found. It makes understanding and analysis process less complex.

The Figure 2 (a) shows the last frame of one scene with its histogram representation in Figure 2(b). The Figure 3(a) shows the consecutive frames in next immediate scene with their histogram representations in Figure 3(b). The content similarity between the frames is found by comparing histogram plots using Normalized Histogram Intersection Similarity Metric (NHISM). If NHISM is less than threshold, then the frame transition is declared otherwise the frame is grouped within the same segment. During this process the left and right boundaries of a dynamic window are confirmed. To model the objects having quick momentum which are beyond neighborhood size, three-level pyramid is generated which is shown in Figure 4. The resolutions in an octave at three different levels starting from fine towards coarse level are 1280x720, 640x360 and 320x180 respectively.

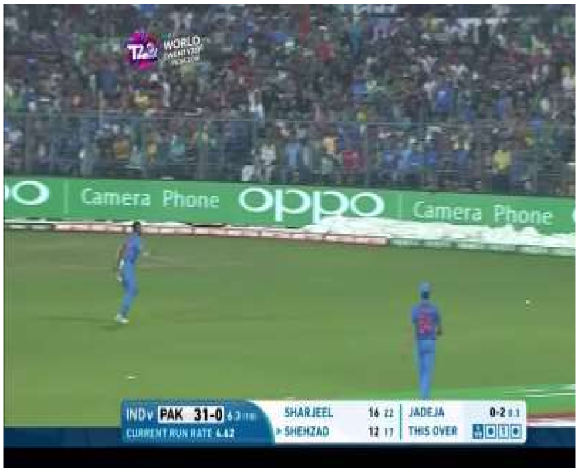


(a)

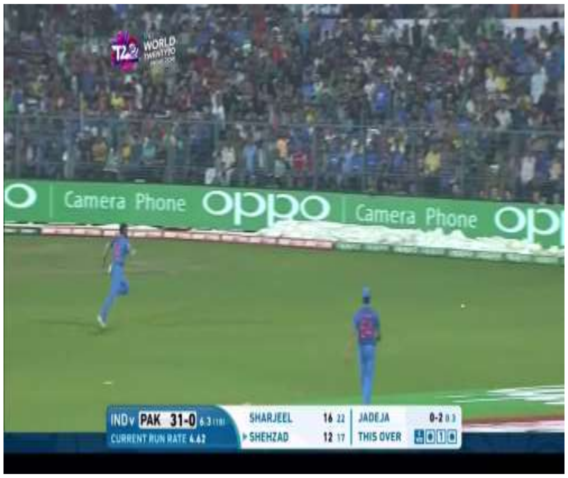


(b)

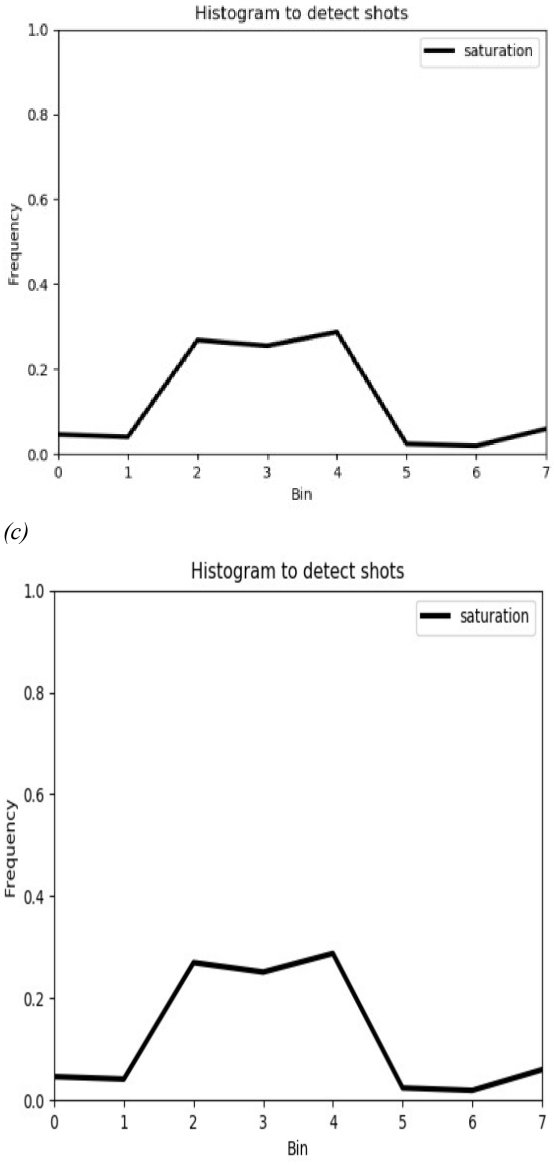
Figure 2: The last frame(a) from a scene and its histogram(b)



(a)



(b)



(c)

(d)

Figure3: The consecutive frames(a & b) at the beginning of next scene and their histograms (c & d)

The optical flow between subsequent frames Figure 5(a&b) at scene transition is shown in Figure 5(c). The corresponding rank oriented histogram is shown in Figure 5(d). The optical flow and its histogram representation between the frames from the same scene are shown in Figure 6. The optical flowof foreground objects is detected such as keeper, batsmen, and umpire. Sometimes, due to occlusions object movements could not be detected(example: bowler movement). The frames having both background and foreground information are shown in Figure 7. From this figure it is also observed that due to much variations in the audience region, it has captured lot of content variation between the frames though the wanted foreground objects move less.

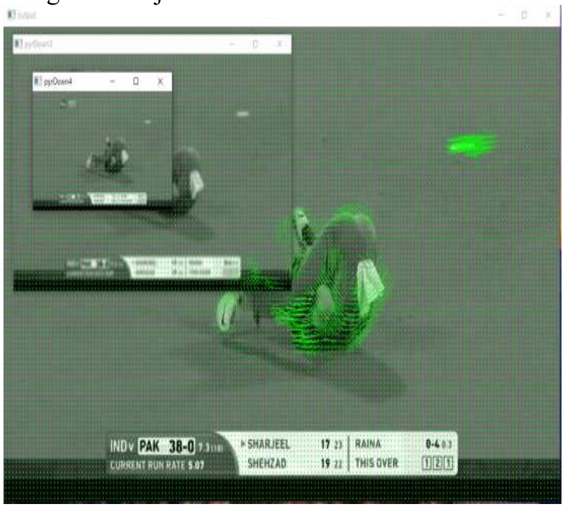
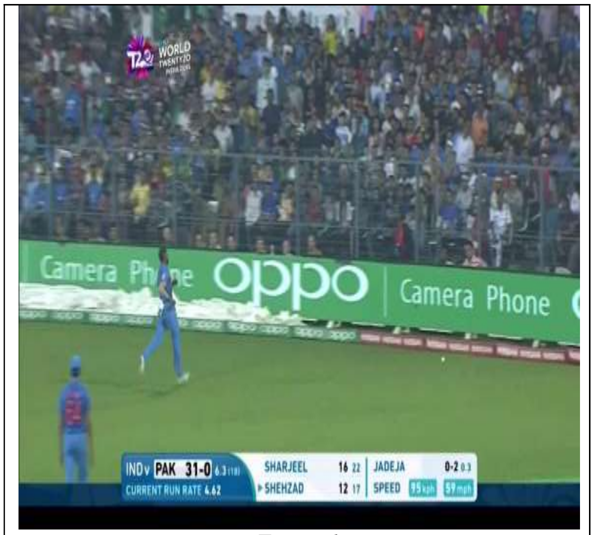
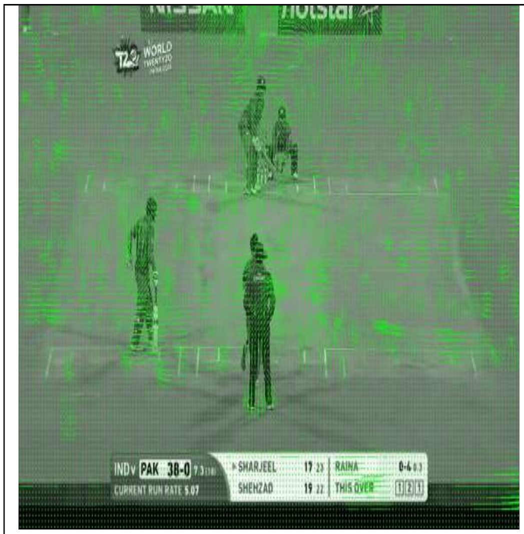


Figure4: Results for three level pyramids and obtained optical flow of an object from a video.

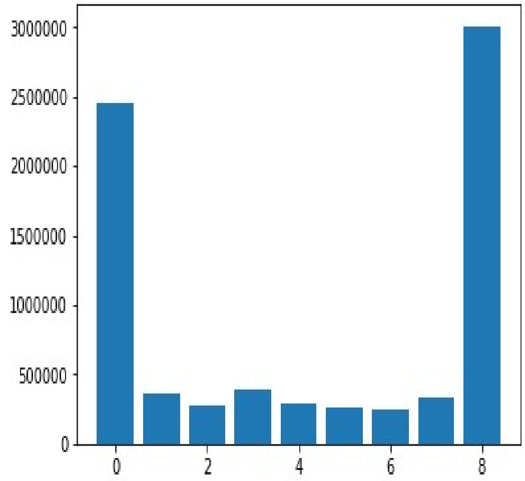


Frame-1

Frame-2



Optical flow between Frames 1 and 2

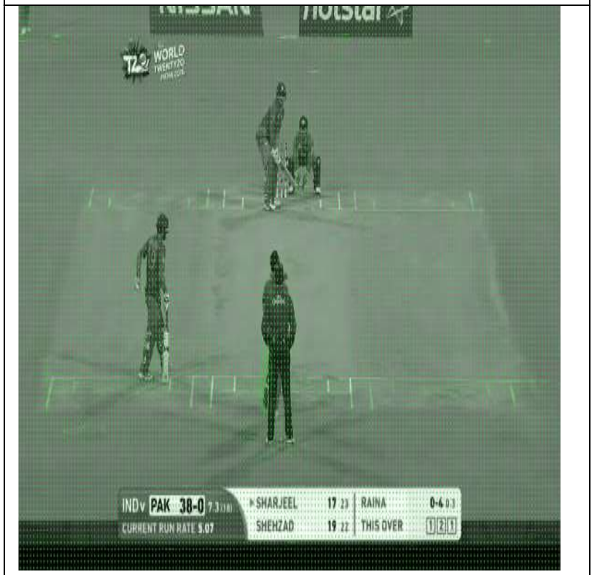


Histogram Representation

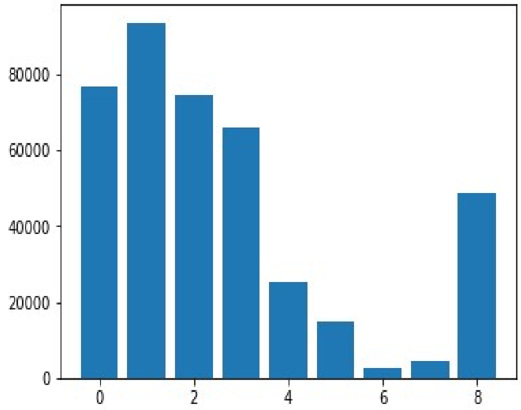
Figure. 5 Rank oriented histogram between frames at scene change.

Frame-1

Frame-2



Optical flow between Frames 1 and 2

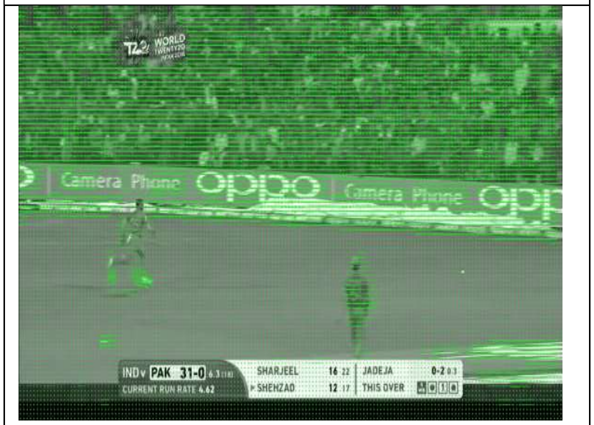


Histogram Representation

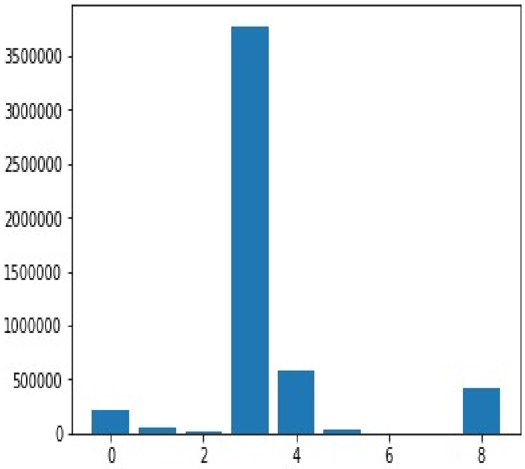
Figure. 6 Rank oriented histogram between frames within the scene.

Frame-1

Frame-2



Optical flow between Frames 1 and 2



Histogram Representation

Figure. 7 Detecting the changes in flow between two consecutive frames and its histogram.

4. CONCLUSION

This paper presents a method to represent motion information in a video by making use of key difference between the availability of unique content in images and videos. The objects and their motion can assist better understanding and analysis of video content, from which future research may benefit. As a first step, a video is temporally segmented into various overlapped segments using gray content matching between the frames. Then, the motion information is captured by computing an optical flow for every pixel in a frame. This information is represented in the form of histogram by considering both rank and orientation of optical flow. This kind of representation can better explainthe syntactic information like movement of objects in a video. This work acts as basic building block to numerous higher level tasks such as action recognition, AI-based autonomous systems, video surveillance, scene understanding and much more.

REFRENCES:

[1] Y. Xu, t. Price, F. Monrose and J. Frahm, ”Caught Red-Handed: Toward Practical VideoBased Subsequences Matching in the Presence of Real-World Transformations,” 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1397-1406, DOI: 10.1109/CVPRW.2017.182..

[2] Shen, Ling Hong, Richang Hao, Yanbin. (2020). Advance on large scale near-duplicate video retrieval. Frontiers of Computer Science. 14. DOI: 10.1007/s11704-019-8229-7..

[3]. Carlos, A. and Gomez-Uribe and Neil Hunt, “The Netflix Recommender System: Algorithms, Business Value, and Innovation”, ACM Transactions on Management Information Systems (TMIS),December 2015, Volume-6, pp. 13:1–13:19, DOI: http://dx.doi.org/10.1145/2843948.

[4]. Jie Lu, Dianshuang Wu, Mingsong Mao, Wei Wang, Guangquan Zhang, “Recommender system application developments: A survey”, Decision Support Systems, Volume 74, 2015, Pages 12-32, ISSN 0167-9236, DOI: 190 https://doi.org/10.1016/j.dss.2015.03.008.

[5]. P. Lops, M. De Gemmis, and G. Semeraro, “Content-based recommender systems: State of the art and trends”, Recommender Systems Handbook: 73–105, (2011), DOI: 10.1007/ 978-0-387-85820-3_3.

[6]. J. Carreira and A. Zisserman, ”Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4724-4733, DOI: 10.1109/CVPR.2017.502.

[7]. D. Huang et al., ”What Makes a Video a Video: Analyzing Temporal Information in Video Understanding Models and Datasets,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7366-7375, DOI: 10.1109/CVPR.2018.00769.

[8]. J. C. Stroud, D. A. Ross, C. Sun, J. Deng and R. Sukthankar, ”D3D: Distilled 3D Networks for Video Action Recognition,” 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 614-623, DOI: 10.1109/WACV45572.2020.9093274.

[9]. Adeli, Vida Fazl-Ersi, Ehsan Harati, Ahad. (2019). “A Component-based Video Content Representation for Action Recognition”, Image and Vision Computing.volume 90, 2019, 103805, ISSN 0262-8856, DOI: 10 . 1016 / j . imavis.2019.08.009.

[10]. Z. Chen and S. Sun, ”A Zernike Moment Phase-Based Descriptor for Local Image Representation and Matching,” in IEEE Transactions on Image Processing, vol. 19, no. 1,pp.205-219,Jan.2010, DOI: 10.1109/TIP.2009.2032890.

[11]. R. C. Gonzalez and R. E. Woods.Digital Image Processing, 3rd edn.Pearson, PrenticeHall, 2092007.

[12]. Jing Li, Nigel M. Allinson, “A comprehensive review of current local features for computer vision”, Neurocomputing, Volume 71, Issues 10–12, 2008, Pages 1771-1787, ISSN 0925-2312, DOI: https://doi.org/10.1016/ j.neucom.2007.11.032

[13]. Shang Liu, Xiao Bai, “Discriminative features for image classification and retrieval”, Pattern Recognition Letters, Volume 33, Issue 6, 2012, Pages 744-751, ISSN 0167-8655, DOI: https://doi.org/10.1016/j.patrec. 2011.12.008.

[14]. B. S. Manjunath, J. -. Ohm, V. V. Vasudevan and A. Yamada, ”Color and texture descriptors,” in IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, no. 6, pp. 703-715, June 2001, DOI: 10.1109/76.927424

[15]. Guang-Hai Liu, Jing-Yu Yang, “Content-based image retrieval using color difference histogram”, Pattern Recognition, Volume 46, Issue 1, 2013, Pages 188-198, ISSN 0031-3203, DOI: https : / / doi . org / 10 . 1016 / j . patcog.2012.06.001.

[16] Shan Zeng, Rui Huang, Haibing Wang, Zhen Kang, “Image retrieval using spatiograms of colors quantized by Gaussian Mixture Models”, Neurocomputing, Volume 171, 2016, Pages 673-684, ISSN 0925-2312, DOI: https://doi.org/10.1016/j.neucom.2015.07.008

[17]. S. Murala, Q. M. J. Wu, B. Raman, and R. Maheshwari, (2013)“Joint Histogram Between Color and Local Extrema Patterns for Object Tracking” Proceedings of SPIE - The International Society for Optical Engineering, 86630T (2013), DOI: 10.1117/12.2002185

[18]. Won, Chee Park, Dong Park, Soo-Jun. (2002). “Efficient Use of MPEG7 Edge Histogram Descriptor”, ETRI Journal - ETRI J. 24. 23-30. DOI:10.4218/etrij.02.0102.0103, .

[19]. Yu Wang, Yongsheng Zhao, Qiang Cai, Haisheng Li, Huaixin Yan, “A varied local edge pattern descriptor and its application to texture classification”, Journal of Visual Communication and Image Representation, Volume 34, 2016, Pages 108-117, ISSN 1047-3203,DOI:https://doi.org/10.1016/j.jvcir.2015.1 1.001.

[20]. Haralick, R.M.; Shanmugam, K. “Texture features for image classification”, IEEE Trans. Syst. Man Cybern. 1973, SMC-3, 610–621, DOI: 10.1109/TSMC.1973.4309314.

[21]. Julesz, B. “Textons, the elements of texture perception, and their interactions”. Nature 1981, 290, 91–97, DOI: 10.1038/290091a0.

[22]. Ojala, T.; Pietikainen, M.; Maenpaa, T. “Multiresolution gray-scale and rotation invariant texture classification with local binary

patterns”. IEEE Trans. Pattern Anal. Mach. Intell. 2002, 24, 971–987, DOI: 10.1109/TPAMI.2002.1017623.

[23]. Dalal, N., Triggs, B.: “Histograms of oriented gradients for human detection”. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893 (2005), DOI: 10.1109/CVPR.2005.177.

[24]. D. G. Lowe. “Distinctive image features from scale-invariant keypoints”. IJCV, 60(2):91–110, 2004, DOI: https://doi.org/10.1023/B:VISI.0000029664.99615.94.

[25]. D. Narra, Y. MadhaveeLatha and D. Avula, ”Content Based Temporal Segmentation for Video Analysis,” 2020 IEEE-HYDCON, Hyderabad, India, 2020, pp. 1-5, DOI: 10.1109/HYDCON48903.2020. 9242711.

[26]. G. Farneback , “Two-frame Motion Estimation based on Polynomial Expansion”, In Proceedings of the 13th Scandinavian Conference on Image Analysis, SCIA 2003, 363 – 370, Halmstad, Sweden, June 29-July 2, 2003, DOI: 10.1007/3-540-45103-X_50.