Key Frame Generation to Generate Activity Strip Based on Similarity Calculation

Wisnu Widiarto[1,2], Eko Mulyanto Yuniarno[2] , Mochamad Hariadi[2]
[1]*Informatics Department, Sebelas Maret University, Surakarta, Indonesia.*
[2]*Electrical Engineering Department, Sepuluh Nopember Institut of Technology, Surabaya, Indonesia.*
wisnu.widiarto@staff.uns.ac.id

***Abstract*—Management of video data is done for several purposes, such as to make the information more meaningful. Research has been conducted to manage the video in terms of detecting activity in a video. There are three stages to generate activity strip: the data source stage (preparation of the frames), the processing stage (analysis of the activity), and the final stage (the collection of key frames). The generation of activity strip is done by calculating the difference of the pixel values of two frames to detect a similarity. In this research, we used SAD (Sum of Absolute Difference) method to calculate the value of the difference of the frame. Similar frames can be grouped in the same cluster. Each cluster is considered as one frame (or multiple frames) to serve as a key frame. The key frames are used for the representation of the activity strip. A collection of activity strip will be arranged sequentially and continuously for the activity generation.**

***Index Terms*—Similarity; Key Frame; Activity Strip; Activity Generation.**

I. Introduction

The development of multimedia-based devices has an impact on the development of information and management of multimedia data and video. Video data can be managed to make information more meaningful in the field of classification, browsing and retrieval. Video management can be done using video abstraction. Video abstraction can be grouped into two basic categories [1]: the static key frame (process to select a key frame) and the dynamic video skimming (process to make a short video).

Static key frame methods are grouped into three classifications: based on sampling (randomly key frame), shot segmentation (measuring the transition), and scene segmentation (scene detection and clustering) [2]. In the first classification (based on sampling), the key frame is randomly selected from a collection of frames. The selection of this key frame can be done quickly. However, it has several problems: i) It can cause redundancy, ii) It fails to accurately represent the content of the video, and iii) It can provide inappropriate information [3].

In the second classification (based shot segmentation), the selection is done by calculating and determining the transition between the initial frame and the next frame in the video [4], while in the third classification (based scene segmentation), the key frame selection is done by making a scene detection [5] and clustering the scene in the video [6].

This research focuses on the selection of key frame based on shot segmentation. The selection of key frame based on the information activities involves determining the difference between one frame to the next frame. The differences between the two frames are calculated based on the pixel values for each frame. Similar frames are taken as a few frames are to be selected into the activity strip. The activity strip will be collected to generate activity on video. The process of calculating the similarity of the frame, selecting the activity strip and generating the activity will be described in this paper.

II. Related Work

There are several research that discuss how video management is conducted. Some of them discuss the video summary that focuses on the analysis of key frame using a segment-based statistical metric [7], fuzzy c-means clustering [8], hybrid segmentation method [9]. Meanwhile, there are some research discuss video summary based on the analysis of the video structure [10,11,12] and event detection [13,14].

Key frame research focuses on determining the shot or scene segmentation as there are only two stages to video document: shot and scene [5,15]. Shot emphasizes the visual information, while scene emphasizes the semantic information [16]. Shot is defined as sequential and continuous frame, which is accessed from a camera [2]. Scene is defined as a shot (or shot several concatenated), which follows the semantic rules. Video scene is a collection of shots that has a semantic relation and describes the storyline [17].

Research clustering focuses on determining similar frame. Having obtained the group cluster, a frame (or multiple frames) is selected to represent each group cluster [18,19,20]. Some researchers use the approach based on the value of the color histogram [21,4,11].

III. Proposed Method

This research is described using block diagram as shown in Figure 1. It is divided into three stages: Preparation (data source), Processing, and the final stage. At the preparation stage, data source is prepared and the video is divided into frames. The data source is used to analyze activity in the video.

Analysis of the activity will be done during the processing stage, which is done in three steps: calculation of similarity, classification, and selection. The calculation of similarity is applied to two frames (the initial frame and the next frame) by comparing two frames. The comparison of the two frames is calculated using parameters, such as pixel values that use the method of SAD (Sum of Absolute Difference). The pixel values of the frame consist of RGB and gray color. The classification is done to classify the frames that have a similarity. Once the classification is done, the frame of each classification is selected so that it can be used as the key frame.

Similarity is the measurement step (the processing stage): The two frames are compared to determine their similarity. Comparisons are done using the SAD (Sum of Absolute Difference) for all frames. SAD value measurement is done at the beginning of the frame and the next frame starting from the first cell (cell (1, 1)) until the last cell (cell (x, y)). Measurements are applied from the first frame to the last frame of the video.

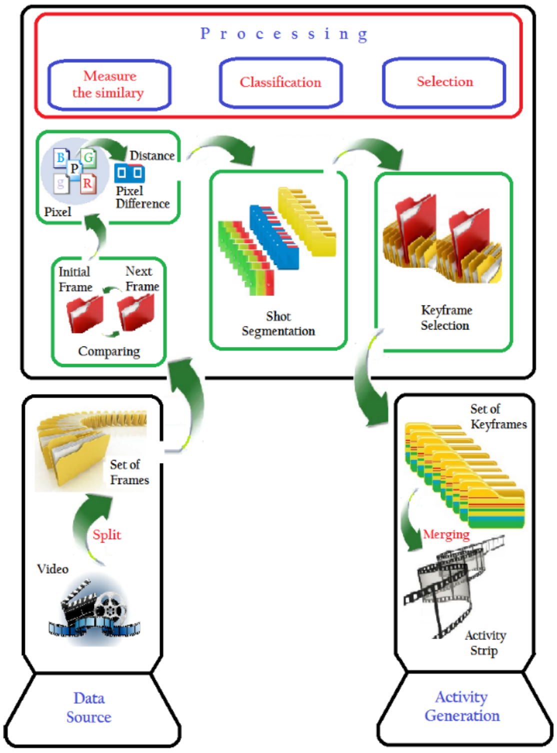


Figure 1: Block diagram of the proposed method

A. Definition 1: Sum of absolute difference:

$$SAD = \sum_{i,j \in W} |F_1(i,j) - F_2(i,j)| \qquad (1)$$

$F_1(i,j)$ is the first frame in the cell (i,j) and $F_2(i,j)$ is the second frame in the cell (i,j). SAD is the sum of absolute difference.

SAD is applied to two frames based on three primary colors (red, green and blue / RGB color). Each color is calculated based on the value of the pixel, and then we calculated the average value for each color. If all of the average value of the SAD (red, green, blue) were below the threshold, then the two frames being compared are considered Similar. If there is an average value of SAD, which is above the threshold, then the two frames are called Dissimilar.

B. Definition 2: The Similarity process:

$$((SAD_{red} < Th) \cap (SAD_{green} < Th) \cap (SAD_{blue} < Th)) \Leftrightarrow similar \qquad (2)$$

SAD_{red}, SAD_{green}, SAD_{blue} are the value of the Sum of Absolute Difference of red color, green color, blue color.

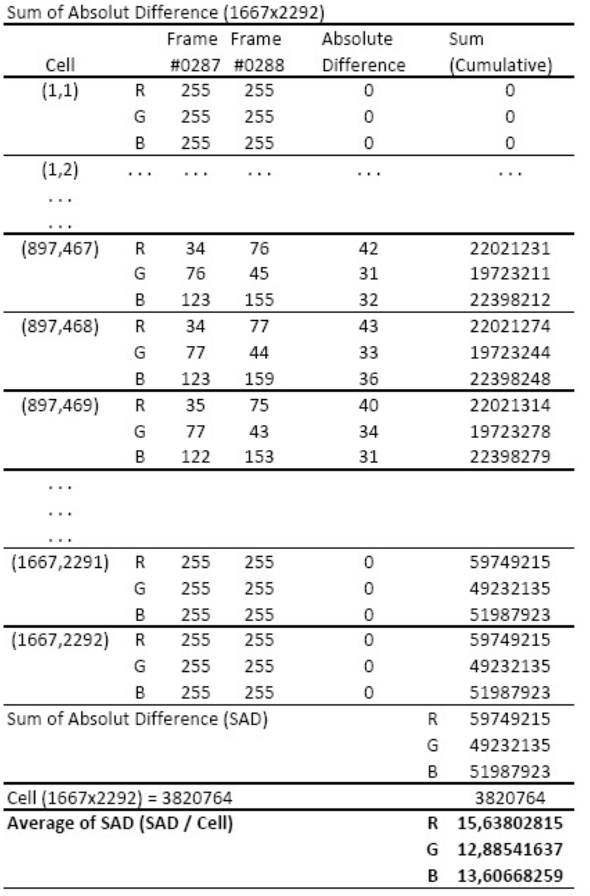
The last stage is to collect the keyframe. The keyframe collection is done sequentially and continuously. It is called an activity generation. Keyframe selected is called an activity strip.

IV. Experimental Result

The first stage of this research is to divide the video into frames. In this research, the video is divided into frames resulting in a total of 2064 frames. Each frame has a size of 1667x2292 pixels (3820764 cells). The frames are used as the data source at the stage of processing.

The second stage is processing. Similarity measurement is the first step of this stage. Measurement is done by making a comparison between the frames using the SAD and each frame is applied to any one of the three colors (RGB color).

The comparison is done between the initial frame and the subsequent frame (for example, the initial frame is #0287, and the next frame is #0288) starting from the first cell (cell (1, 1)) until the last cell (cell (1667, 2292)). For example, in the cell (897,468), frame #0287 (R = 34, G = 77, B = 123), frame #0288 (R = 77, G = 44, B = 159), The absolute difference (R = | 34-77 | = 43; G = | 77-44 | = 33; B = | 123-159 | = 36), cumulative / SAD (R = 22021274, G = 19723244, B = 22398248). In the cell (897,469), frame #0287 (R = 35, G = 77, B = 122), frame #0288 (R = 75, G = 43, B = 153), absolute difference (R = | 35-75 | = 40; G = | 77-43 | = 34; B = | 122-153 | = 31), cumulative / SAD (R = 22021314, G = 19723278, B = 22398279).



Sum of Absolut Difference (1667x2292)

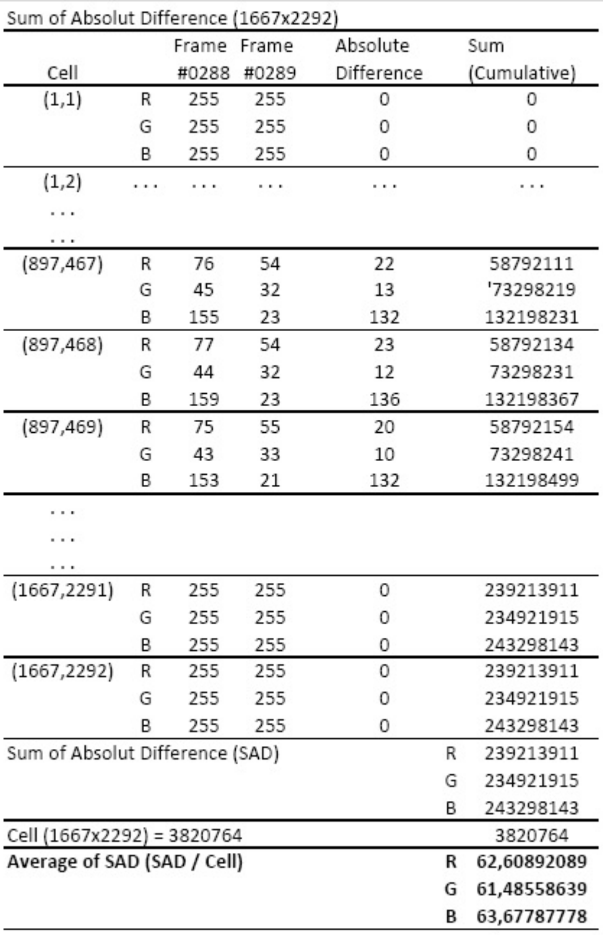
Cell		Frame #0287	Frame #0288	Absolute Difference	Sum (Cumulative)
(1,1)	R	255	255	0	0
	G	255	255	0	0
	B	255	255	0	0
(1,2)	...	...	...	...	...
...					
...					
(897,467)	R	34	76	42	22021231
	G	76	45	31	19723211
	B	123	155	32	22398212
(897,468)	R	34	77	43	22021274
	G	77	44	33	19723244
	B	123	159	36	22398248
(897,469)	R	35	75	40	22021314
	G	77	43	34	19723278
	B	122	153	31	22398279
...					
...					
...					
(1667,2291)	R	255	255	0	59749215
	G	255	255	0	49232135
	B	255	255	0	51987923
(1667,2292)	R	255	255	0	59749215
	G	255	255	0	49232135
	B	255	255	0	51987923
Sum of Absolut Difference (SAD)				R	59749215
				G	49232135
				B	51987923
Cell (1667x2292) = 3820764					3820764
Average of SAD (SAD / Cell)				**R**	**15,63802815**
				G	**12,88541637**
				B	**13,60668259**

Figure 2: SAD value of the two frames (#0287 and #0288) in the similar conditions.

The application of the calculation is done for all cells. This results in the SAD values obtained for red, green and blue are 59749215; 49232135; 51987923. The use of 1667x2292 pixel size results in 3820764 cells. The average of SAD values (R, G, B) are 15.63802815; 12.88541637; 13.60668259 (Figure 2).

Frame #0288 and #0289 are compared at the cell (1,1) to cell (1667,2292), As shwon in Figure 3, the obtained SAD values are (R, G, B) 239213911; 234921915; 243298143 and the average values of SAD (R, G, B) are 62.60892089; 61.48558639; 63.67787778. It was found that Frame #0288 and #0289 are Dissimilar, while Frame #0287 and #0288 are Similar.

The same calculation is done for all the frames, and the next step is the segmentation process. Similar frames are grouped into one segment. For the purpose of this research 14 segments were derived. In the first segment, there are 36 frames, namely from Frame #0001 to #0036, and the number of activity strip is 2 frames. The second segment contained 144 frames (frames #0037 to #0180) with 6 frames, considered as an activity strip. Similar frame capturing is done for the other segments. The results are shown in Table 1.



Sum of Absolut Difference (1667x2292)

Cell		Frame #0288	Frame #0289	Absolute Difference	Sum (Cumulative)
(1,1)	R	255	255	0	0
	G	255	255	0	0
	B	255	255	0	0
(1,2)	...	...	...	...	...
...					
...					
(897,467)	R	76	54	22	58792111
	G	45	32	13	'73298219
	B	155	23	132	132198231
(897,468)	R	77	54	23	58792134
	G	44	32	12	73298231
	B	159	23	136	132198367
(897,469)	R	75	55	20	58792154
	G	43	33	10	73298241
	B	153	21	132	132198499
...					
...					
...					
(1667,2291)	R	255	255	0	239213911
	G	255	255	0	234921915
	B	255	255	0	243298143
(1667,2292)	R	255	255	0	239213911
	G	255	255	0	234921915
	B	255	255	0	243298143
Sum of Absolut Difference (SAD)				R	239213911
				G	234921915
				B	243298143
Cell (1667x2292) = 3820764					3820764
Average of SAD (SAD / Cell)				**R**	**62,60892089**
				G	**61,48558639**
				B	**63,67787778**

Figure 3: Initial frame (#0288) and the next frame (#0289) are Dissimilar.

The selection process of the key frame into strips activity can be done by selecting the key frame of each segment. The process showed that for 25 frames are to be captured as one frame, located in a central position. The first segment has 36 frames (25 + 11). For the 25 frames, the first frame was taken from the middle frame (number 12), and the second frame consisting of 11 frames was taken the middle frame (5 number / number 30 (25 + 5)).

The second segment has 144 frames (125 + 19), which is divided into: the first 25 frames (number 12 / number 48 (36 + 12)), the second 25 frames (number 12 / number 73 (36 + 25 + 12)), the third 25 frames (number 12 / number 98 (36 + 25 + 25 + 12)), the fourth 25 frames (number 12 / number 123 (36 + 25 + 25 + 25 + 12)), the fifth 25 frames (number 12 / number 148 (36 + 25 + 25 + 25 + 25 + 12)), and the sixth 19 frames (number 9 / number 170 (36 + 25 + 25 + 25 + 25 + 25 + 9)). The same application is adopted for the other segment in order to obtain a collection of strips activity as shown in Table 2.

Table 1
Experimental Results of activity segmentation

No	Frames number	Number of Frame		Number of activity strip	
1	1-36	36	25+11	1+1	2
2	37-180	144	125+19	5+1	6
3	181-288	108	100+8	4+1	5
4	289-576	288	275+13	11+1	12
5	577-780	204	200+4	8+1	9
6	781-876	96	75+21	3+1	4
7	877-1044	168	150+18	6+1	7
8	1045-1164	120	100+20	4+1	5
9	1165-1212	48	25+23	1+1	2
10	1213-1248	36	25+11	1+1	2
11	1249-1692	444	425+19	17+1	18
12	1693-1884	192	175+17	7+1	8
13	1885-1980	96	75+21	3+1	4
14	1981-2064	84	75+9	3+1	4
		2064			88

Table 2
Experimental Results of frame number of activity strip

No	Frames number (Number of frame) (Number of activity strip)	Frame number of activity strip
1	1-36(36)(2)	12,**30**
2	37-180(144)(6)	48,73,98,123,148,**170**
3	181-288(108)(5)	192,217,242,267,**284**
4	289-576(288)(12)	300,325,350,375, 400,425,450,475, 500,525,550,**569**
5	577-780(204)(9)	588,613,638,663, 688,713,738,763,**778**
6	781-876(96)(4)	792,817,842,**865**
7	877-1044(168)(7)	888,913,938,963, 988,1013,**1035**
8	1045-1164(120)(5)	1056,1081,1106,1131, **1154**
9	1165-1212(48)(2)	1176,**1200**
10	1213-1248(36)(2)	1224,**1242**
11	1249-1692(444)(18)	1260,1285,1310,1335, 1360,1385,1410,1435, 1460,1485,1510,1535, 1560,1585,1610,1635, 1660,**1682**
12	1693-1884(192)(8)	1704,1729,1754,1779, 1804,1829,1854,**1875**
13	1885-1980(96)(4)	1896,1921,1946,**1969**
14	1981-2064(84)(4)	1992,2017,2042,**2059**
	1-2064 **(2064)(88)**	

V. CONCLUSION

The application of activity strip for this video resulted in the original video to be divided into 2064 frames. Further, it can be used to form clusters, called activity segmentation. The process of segmentation resulted in 14 segments. The segmentation activity involved selecting activity strip. The total of activity strip selected depends on the multiple number of frames in each segment. If the segment has a lot of frames, then too many strip are selected. In this research, the activity strip has as much as 88 frames. This shows that the original video that has been divided into frames (2064 frames) can be represented by selected strip (totaling 88 frames). All of the strips are reassembled in the order and continued to generate activity and represents the original video.

ACKNOWLEDGEMENT

This research was supported by Informatics Department – UNS Surakarta and Electrical Engineering Department – ITS Surabaya.

REFERENCES

[1] B.T. Truong, S. Venkatesh, Video abstraction: A systematic review and classification, ACM Transactions on Multimedia Computing Communications and Applications (TOMCCAP), vol. 3, no. 1, (2007) 1-37.

[2] W. Sabbar, A. Chergui, A. Bekkhoucha, Video summarization using shot segmentation and local motion estimation, IEEE Second International Conference on Innovative Computing Technology (INTECH), (2012) 190-193.

[3] S. Angadi, V. Naik, Entropy based fuzzy C means clustering and key frame extraction for Sports Video Summarization, 2014 Fifth International Conference on Signal and Image Processing (ICSIP), (2014) 271-279.

[4] W. Widiarto, E.M. Yuniarno, M. Hariadi, Video summarization using a key frame selection based on shot segmentation, 2015 International Conference on Science in Information Technology (ICSITech), (2015) 207-212.

[5] Z. Rasheed, M. Shah, Detection and representation of scenes in videos, IEEE Transactions on Multimedia, vol. 11, no. 6, (2005) 1097-1105.

[6] A. Girgensohn, J.S. Boreczky, Time-constrained keyframe selection technique, Multimedia tools application, vol. 11, no. 3, (2000) 347-358.

[7] M. Omidyeganeh, S. Ghaemmaghami, S. Shirmohammadi, Video keyframe analysis using a segment-based statistical metric in a visually sensitive parametric space, IEEE Transactions on image processing, vol. 20, no. 10, (2011) 2730-2737.

[8] K. Mahesh, K. Kuppusamy, A new hybrid video segmentation algorithm using fuzzy c means clustering, IJCSI International Journal of Computer Science Issues, vol. 9, issue 2, no. 1, (2012) 229-237.

[9] K. Mahesh, K. Kuppusamy, Video segmentation using hybrid segmentation method, European Journal of Scientific Research, vol. 71, no. 3, (2012) 312-326.

[10] D. Besiris, F. Fotopoulou, G. Economou, S. Fotopoulos, Video summarization by a graph-theoretic FCM based algorithm, IEEE conference publications, 15th International conference on systems, signal and image processing (IWSSIP) 2008, (2008) 511-514.

[11] Y. Song, T. Ogawa, M. Haseyama, MCMC-based scene segmentation method using structure of video, International Symposium on Communications and Information Technologies (ISCIT), (2010) 862-866.

[12] Y. Zhao, T. Wang, P. Wang, W. Hu, Y. Du, Y. Zhang, G. Xu, Scene Segmentation and Categorization Using NCuts, 2007 IEEE Conference on Computer Vision and Pattern Recognition, (2007) 1-7.

[13] M. Tavassolipour, M. Karimian, S. Kasaei, Event Detection and Summarization in Soccer Videos Using Bayesian Network and Copula, IEEE Transactions on circuits and systems for video technology, Vol. 24, No. 2, (2014) 291-304.

[14] C.L. Huang, H.C. Shih, C.Y. Chan, Semantic analysis of soccer video using dynamic Bayesian network, IEEE Transactions on Multimedia, vol. 8, no. 4, (2006) 749-760.

[15] Y. Zhu, Z. Ming, SVM-based video scene classification and segmentation, International Conference on Multimedia and Ubiquitous Engineering 2008 (MUE 2008), (2008) 407-412.

[16] R. Burget, J.K. Rai, V. Uher, J. Masek, M.K. Dutta, Supervised video scene segmentation using similarity measures, 2013 36th International Conference on Telecommunications and Signal Processing (TSP), (2013) 793-797.

[17] C.R. Huang, C.S. Chen, Video scene detection by link-constrained affinity-propagation, IEEE International comference, (2009) 2834-2837.

[18] P. Mundur, Y. Rao, Y. Yesha, Keyframe-based video summarization using Delaunay clustering, International Journal on Digital Libraries, vol. 6, no. 2, (2006) 219-232.

[19] Y. Hadi, F. Essannouni, R.O.H. Thami, Video summarization by k-medoid clustering, Proceedings of the ACM symposium on Applied Computing – SAC ’06, (2006) 1400-1401.

[20] W. Widiarto, M. Hariadi, E.M. Yuniarno, Shot segmentation of video animation to generate comic strip based on key frame selection, 2015 IEEE International Conference on Control System, Computing and Engineering (ICCSCE), (2015) 303-308.

[21] N. Haering, R.J. Qian, M.I. Sezan, A semantic event-detection approach and its application to detecting hunts in wildlife video, IEEE Transactions on Circuits System Video Technology, vol. 10, no. 6, (2000) 857-868.