Key Frame Extraction Analysis Based on Optimized Convolution Neural Network (OCNN) using Intensity Feature Selection (IFS)

Dr. T. Prabakaran
Associate Professor
Department of CSE
Joginpally B.R. Engineering College,
India
prabaakar.t@gmail.com

Loveleen Kumar
Assistant Professor,
Department of Computer Science and Engineering,
Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur,
Rajasthan
loveleentak@gmail.com

Dr. Ashabharathi S
Visvesvaraya technological University
India
sashabharathi@gmail.com

Dr Prabhavathi S
Professor
RYMEC Ballari
India
prabkir@rymec.in

Dr Maneesh Vilas Deshpande
Assistant Professor
Department of Computer Science
Tai Golwalkar Mahavidyalaya Ramtek
India
maneeshvdeshpande@rediffmail.com

Mochammad Fahlevi
Department of Management,
BINUS Online Learning,
Bina Nusantara University,
Indonesia 11480
mochammad.fahlevi@binus.ac.id

2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS) | 978-1-6654-7657-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICTACS56270.2022.9988474

Abstract-The multimedia is playing role of timing frames in videos. The representation frame shows the intention on video definition. The keyframes the important factor for extraction information from video frames. The non-related frames is a problem for finding new key exposure. In this paper, we present a new method for extracting essential frames from motion capture data using Optimized Convolution Neural Network (OCNN) and Intensity Feature Selection (IFS) for better visualisation and understanding of motion content. It first removes noise from motion capture data using the Butterworth filter, then reduces the size via principal component analysis (PCA). Finding the zero-crosses of velocity in the main components yields the initial set of crucial frames. To avoid redundancy, the first batch of important frames is divided into identical poses. Experiments are based on data access from frames in the motion capture database, and experimental results suggest that crucial frames retrieved by our method can improve motion capture visualisation and comprehension.

Keyframes: Video framing, keyframes, deep learning, feature extraction, classification, principal component analysis.

I. INTRODUCTION

Key frames are commonly utilised in non-linear browsing and video content analysis applications. It is critical to understand how to swiftly extract key frames from video. This article [1] presents a method for extracting key frames from compressed video streams. The similarity packages of nearby I-frame DC images are determined first, followed by the clustering technique, and lastly the major frames are chosen based on the clustering findings. Experiment findings show that our method can extract relevant key frames from test video files fast and easily.

The rapid growth of online videos has created an urgent need to find almost copy video. Nearby copy key frame detection is the basis for locating duplicate video. Based on sophisticated analysis in duplicate key frame detection [2], this paper enables grayscale pyramid (GSP) to enhance the global features of color maps. By creating a spatial pyramid of brightness and sheer size, the algorithm enhances the brightness of the global features and the strength of the shear-scale changes [3]. Experiments have shown that GSP is much stronger than color charts in detecting recurring key frames near light and scale transformations, and for other modifications, both are almost equally effective.

This movement allows us to obtain fully (or at least) fully closed and fully open key frames and initiate the partitioning process. Gabor filtration is done by analyzing the image structure used [4]. After this stage, use derivative techniques and interpolations to define where (if any) the decay is. Finally, the pathology is localized and some features can be extracted for each fold.

II. RELATED WORK

Key frame Extraction is a simple and effective method for accomplishing this. Because it just gives the video's main material, key frame extraction is also known as video abstraction. Because key frames are used to summarise basic video content, they are very significant in video data applications, particularly for redundant monitoring. The motion as the main feature that is calculated using the difference between the frames can be used to define key frames. To generalise video content, the suggested technique makes use of the interaction of colour channels.

Key frame extraction is one of the most important issues in video comprehension and recovery research, especially as the number of uploaded individual videos is

978-1-6654-7657-7/22/$31.00 ©2022 IEEE

increasing rapidly. Frame clustering is theoretically mature method of core frame extraction. However, the clustering process can take a long time, which seems impractical. In this paper, we propose an improved clustering method based on video features and demonstrate that it is useful in key frame extraction.

Key frame extraction is the basic process of video content analysis and retrieval. This paper proposes an effective fast method to effectively extract key frames from compressed video streams. It first calculates the homogeneity of DC images of adjacent I frames, then clusters the homogeneity using the k means algorithm and finally selects the main frames according to the clustering results. Experimental results show that our system can extract the correct key frames from the test video files and extract the key frames in less time.

Color disparity is a pressing issue in stereoscopic video. This paper presents a robust method for colour correction in stereoscopic video. To begin, we employ Scale-Invariant Feature (SIFT) based feature point matching to locate the image's matching points. The parameter model between the photos is then determined using the value of each matching point pair, and the optimal value is produced using the minimal square approach. Finally, color-based key frame extraction technology is employed to perform consistent colour correction in the parameter model of each key frame and its subsequent frame. The experimental findings suggest that the proposed method can create good editing quality for stereoscopic videos.

Key frame extraction is critical for video recovery. We present a new approach for retrieving the frequency-adaptive human motion sequence model, producing high-quality key frames, and extracting the human motion sequence key frame. First, we define an inter-frame similarity measure based on body part characteristics. The Affine Propagation Clustering Algorithm then extracts critical frames. This method begins with video information distribution, finds the optimal key frame of the video with adaption, and accelerates the system speed. Finally, the key frame-based sequence reconstruction rating was validated. Comparative experiments on the CMU database assess the effectiveness of our system.

Combining Conventional Neural Networks (CNN) with Recurrent Neural Networks (RNN) provides a powerful framework for video classification difficulties since spatiotemporal information can be successfully processed simultaneously. This research compares how CNNs and RNNs can be used to improve video classification performance when transliteration is employed to leverage transient information, employing transmission learning.

An new action template-based key frame extraction method was proposed to improve the performance of the authentication framework to effectively combine CNN and RNN by detecting the information regions of each frame and picking key frames based on the similarity between these areas. Extensive tests were carried out on the KTH and UCF-101 datasets employing Conv LSDM-based video classifiers. The test results are analysed using one-way ANOVA, which demonstrates that the proposed key frame extraction approach can be utilised to greatly increase video classification accuracy.

III. PROPOSED METHOD

A key frame method has been proposed to reduce the operating data by extracting key frames using motion analysis methods in the sample window. The proposed Optimized Convolution Neural Network (OCNN) And Intensity Feature Selection (IFS) from motion capture data for better visualization and understanding of motion content. It first uses the Butterworth filter to remove noise from motion capture data, and then performs principal component analysis (PCA) to reduce the size Calculates the motion variance in the model window without the excluded frames and in the original motion. The key factor in determining whether a frame in a sample window is a key frame choice is the variation of the operating variation. Simulation results show that this method can achieve good overall visual quality for different types of movements. Improves average square error measurement by up to 52% compared to existing key frame extraction methods (i.e. curve simplification)

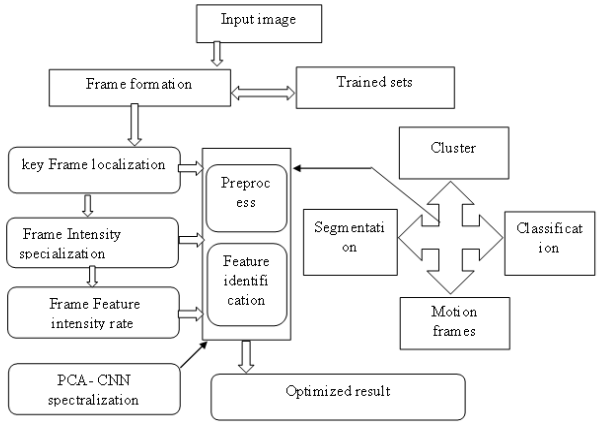

Figure1: proposed architecture (OCNN-IFS)

Creating movie codes is a difficult and expensive process that seeks to automate. Algorithms for finding the boundaries of a scene are readily available, but little work has been done on selecting individual frames to briefly represent the scene. This paper introduces a new method for automatically selecting custom key frames based on the content of the display. Figure1 shows the proposed architecture OCNN-IFS. Following a detailed description of some of the algorithms, we analyze how humans perceive the selected law as representing the scenario. Finally, we will show you how to combine these algorithms with existing ones to find the boundaries of your scenario.

3.1 Pre-Processing Query frame

The natural scene images contains the realistic part of object contents with noise like shadowing, blur contents, lightening, defected contrasting etc., in this stage preprocessing is used to remove the noisy content to resize the structural image. Using the filters are applicable to remove the non-gradient gray conversion to specify the structure of keyframe content

3.2 key Frame localization

The focus is mainly on the central area when recording and viewing. Therefore, the area outside the video frame is usually clipped before finding the area of interest. The action area is then created as the centerpiece of the video frame. This will create the greatest or least similarity between successive frames and create a video template for tracking the action area.

Keyframes can be extracted more accurately and efficiently by calculating only the difference in the running area between the frames throughout the video, minimizing the effects of possible dynamic backgrounds. Use the first two laws to calculate the average square error (MSE) for each possible area. Select the area that makes up the largest MSE as the action template for all framesKey frame extraction:

- Calculate the structural similarity measure Si between regions of interest on consecutive frames fi; fi 1.
- Compare the similarity score to thresholds T1 14 20:65; 0:90 and T2 14 20:65; 0:95 (these threshold values were established by examination of significance in action changes in our experiments): Add fi to primary list ðpf Þ

Add fi to alternative list ðaf Þ

• Repeat until the video is finished, with Npf frames extracted into pf and Naf frames extracted into af..

Extract

Key frame selection:

• Set the number of key frames ðNkf Þ.

• Find key frame ratio ðkÞ:

$$k = \begin{cases} \left\lfloor \frac{N_{pf}}{N_{kf}} \right\rfloor, & if\ N_{pf} \geq N_{kf} \\ \left\lfloor \frac{N_{af}}{N_{kf}} \right\rfloor, & otherwise \end{cases} \quad (1)$$

Return the indexes of key frames by choosing a frame from every k frame from key frame list

p_f If $N_{pf} \geq N_{kf}$, or from key frame list a_f otherwise.

3.3 Frame Intensity specialization

Background changes between successive frames reduce structural uniformity (SSIM) and different actions enhance MSE. The candidate area with the largest MSE between successive frames is assigned as an action template.

$$MSE(X,Y) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}[Y(i.j) - X(i,j)]^2 \quad (2)$$

Where m and n are the number of rows and columns in the region of interest, respectively.

$$SSIM(X,Y) = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X^2\mu_Y^2 + C_1)(\sigma_X^2\sigma_Y^2 + C_2)} \quad (3)$$

Then the best fit is found worldwide max. About color channels and templates. The sum of operations is performed on all channels, and a different average is applied to each channel.

R (x, y) is the correlation coefficient (x, y) for a single connection, where the coordinates of each pixel in the frame are represented. Is there an average pixel value in the T '(x ', y ') template? T, T '(x ', y ') returns the coordinates of each pixel in the template, as seen below:

$$T'(x',y') = T(x',y') - \frac{1}{(w.h)} \cdot \sum_{x^n y^n} T(x^n, y^n). \quad (4)$$

3.4 Frame Feature intensity rate

I' (x+x',y+y') on the other hand is the average of pixel values of a specific frame I in the region overlapped with the template, T specified as;

$$I'(x+x',y+y') = I(x+x',y+y') - \frac{1}{(w.h)} \cdot \sum_{x^n y^n} I(x+x^{"},y+y^{"}) \quad (5)$$

Where x "= 0,....., w-1andy "= 0,..., h-1 is in the template (x, y) after relocating the template's centre in the frame. This term refers to new integrations. T (x ', y ') represents the pixel value (x, y) of the template pixel, and I (x + x ', y + y ') represents the pixel value of the matching pixel position. In. Following the template matching technique, the region of interest for each frame will be customised to the location with the highest chance of maximum fitting.

$$M(t) = \sum_i \sum_j |OF_x(i,j,t)| + |OF_y(i,j,t)| \quad (6)$$

OF x I j; t) is the x component of the optical flow in pixel I as well as the frame's j, t, and y elements. Because the optical flow records all points across time, the sum equals the amount of motion across the frames. This function's slope represents the change in motion between succeeding frames, therefore local and local reflect a constant or significant function between rows.

3.5 CNN spectralization

This CNN model has been trained to predict a score based on the quality of the face in the shot. Without the need for face recognition, the key frame is chosen based on this score. For key frame recognition, the selected key frames are then transferred to an intensity-based back-end deep neural network (DNN) with optimised PCA from feature assessment.

$$L = \frac{1}{2N}\sum_{i=1}^{N}\left\|p_i^2 - t_i^2\right\|^2 \quad (7)$$

Step 1: Start the procedure.

Step 2: Then Granulate Computing takes neural construction form L images.

Step 3: After that process of image granulation with PCA, build a relationship between feature substance of key frames process 'p'.

Step 4: Then, image partitioning transfers the image into the feature vector and merges the vector space images.

Step 5: After that, Deep CCN creates feed forwards successive threshold margins in LSTM function to fix threshold features.

Step 6: Then acceptance of pooling weight to learn about the frame definitions.

Step 7: Next, the return the match case key frames relatively immense to max support.

Step 8: Attention substitution for continue frame slot (Fs)

Step 9: return Frame slot (Fs).

The proposed design aims to improve face-to-face face recognition accuracy in the video recognition system, reduce the amount of data transfer over the network and improve the real-time processing power of the face in the video recognition system. During the training phase, the CNN training utilizes the Euclidean loss function, as shown in the key extraction based on equation formation.

III. Results and Discussion

The proposed implementation results are test with image processing tool in mat lab with trained features. The performance evaluation are carried to test with sensitivity and specificity measure of precision and recall values obtains in execution stage. Optimized Convolution Neural Network (OCNN) And Intensity Feature Selection (IFS) from motion capture data for better visualization and understanding of motion content. It first uses the Butterworth filter to remove noise from motion capture data, and then performs principal component analysis (PCA) to reduce the size. We have evaluated the proposed computation is compared with different methodologies inspected previously. The collected dataset from UCI repository contain natural scene image collection VID-NAT (video nature), Multi-motion video dataset, ICDAR. The experimentation of various detection calculations was completed on different images. The performance values are evaluated by precision and recall rate with tested trained set of positive and negative values.

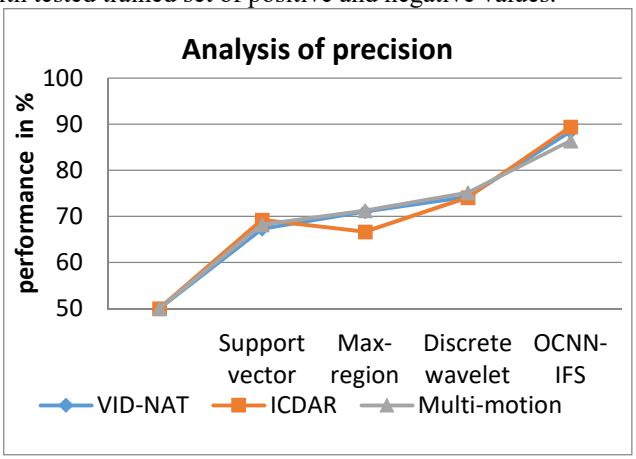

Figure 2: Analysis of a precision rate

The above figure 2 shows the observed true positive precision rates from different dataset with dissimilar methods, the proposed implementation produce higher efficient rate than other methods.

TABLE 4.1: ANALYSIS OF PRECISION RATE

	Analysis of precision in %			
Techniques dataset used	Support vector	Max-region	Discrete wavelet	OCNN-IFS
VID-NAT	67.3	71.1	74.3	88.4
ICDAR	69.3	66.7	74.1	89.5
Multi-motion video dataset	68.2	71.3	75.2	86.4

The above table 1 shows the resultant of precision rate with different image collection dataset produced by different methods. The proposed OCNN-IFS produce 88.4 % well than other methods.

The evaluation of recall observed by,

$$recall = \frac{\sum_{j=1}^{|D|} mat\ Gd(Dj)}{|D|} * false\ positive - true\ negative\ matches \ldots. (8)$$

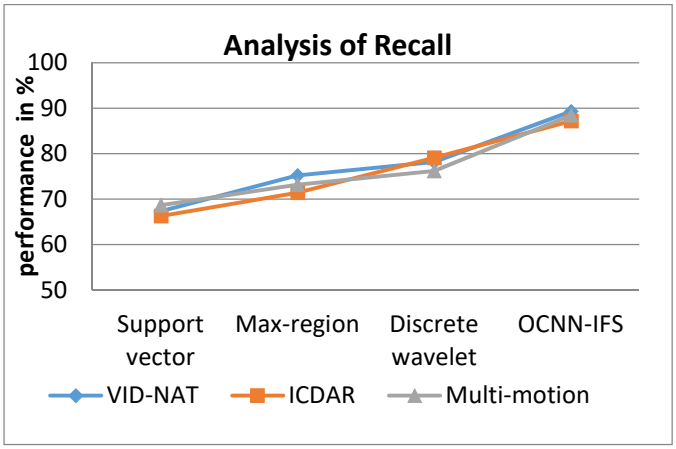

Figure 3: Analysis of recall

The above 3 shows the analysis of recall rate tested with different dataset. The collected dataset have differential tested value produced by different methods. The proposed OCNN-IFS system have higher recall rate than other methods.

TABLE 2: ANALYSIS OF RECALL

	Analysis of recall in %			
Techniques dataset used	Support vector	Max-region	Discrete wavelet	OCNN-IFS
VID-NAT	67.3	75.2	78.2	89.3
ICDAR	66.3	71.4	79.1	87.2
Multi-motion video dataset	68.7	73.2	76.2	88.4

The above table 2 shows the analysis of recall rate tested with extraction of positive negative key frames . Resultant proves the proposed system have higher efficiency of recall rate up to 88.4 % well than other methods.

$$\text{False retrieval ratio (Frr)} = \sum_{k=0}^{k=n} \times \frac{total\ dataset\ failed\ images\ (Fer)}{Total\ no\ of\ image\ rate(Fr)} \ldots\ldots\ldots (9)$$

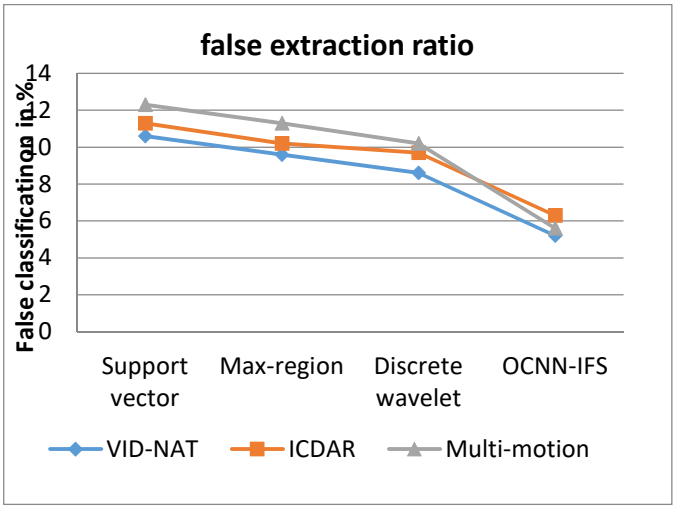

Figure 4: Analysis of false extraction

The above analysis in figure 4 shows the differential evaluation of false results compared with dissimilar methods. The implementation of new system have higher efficient result than other result.

TABLE 3: ANALYSIS OF FALSE EXTRACTION

	Analysis of false extraction in %			
Techniques dataset used	Support vector	Max-region	Discrete wavelet	OCNN-IFS
VID-NAT	10.6	9.6	8.6	5.2
ICDAR	11.3	10.2	9.7	6.3
Multi-motion video dataset	12.3	11.3	10.2	5.6

The above table 3 shows the analysis of false extraction produced by dissimilar methods tested with differential dataset, the VID-NAT, ICDAR and Multi-motion video dataset produce consecutive low false rate the proposed system OCNN-IFS produce 3.4 % low false rate.

$$\text{Time complexity (Tc)} = \sum_{k=0}^{k=n} \times \frac{\text{total images handeletd to process in dataset}}{\text{Time taken(Ts)}} \ldots\ldots (10)$$

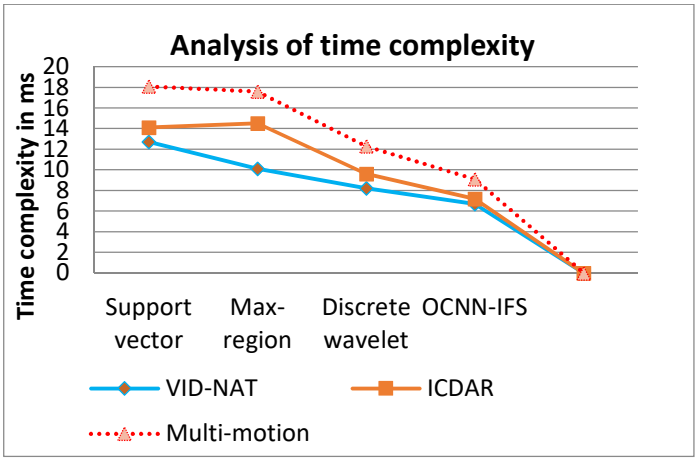

Figure 5: Analysis of time complexity

The above figure 5 shows the dissimilar methods of with differential VID-NAT, ICDAR and Multi-motion video dataset. The collected dataset represent the differential key frame formats in natural videos as well extracted proposed system with high efficiency.

Table 4: Execution of time complexity

	Execution of time evaluation (milliseconds- ms)			
Techniques dataset used	Support vector	Max-region	Discrete wavelet	OCNN-IFS
VID-NAT	12.7	10.1	8.2	6.7
ICDAR	14.1	14.5	9.6	7.2
Multi-motion video dataset	18.1	17.6	12.3	9.1

The above table 4 shows the execution state of different methods with different time taken process. The proposed system test with dataset image collected dataset VID-NAT, ICDAR and Multi-motion video dataset, the proposed OCNN-IFS system produce 6.7 (ms) higher efficiency than other methods with lower execution state.

V. CONCLUSION

The research implementation concludes the detection of Optimized Convolution Neural Network (OCNN) And Intensity Feature Selection (IFS) from motion capture data for better visualization and understanding of motion content. The features are extracted by redundant form of discriminating factors frame content based on the leaning, support objects , entities etc. The segmenting frame region are carried out using connected component feature selection with multi objective intensive access process. The proposed system higher efficiency as well precision rate 90.1 %, similar recall rate 91.6%, the false rate reduction up to 5.2 % compared to other system with lower 5.2 (ms) execution time complexity**.**

REFERENCES

[1] Y. Yang, L. Zeng and H. Leung, "Key frame Extraction from Motion Capture Data for Visualization," 2016 International Conference on Virtual Reality and Visualization (ICVRV), 2016, pp. 154-157, doi: 10.1109/ICVRV.2016.33.

[2] E. M. I. Alaoui, A. Mendez, E. Ibn-Elhaj and B. Garcia, "Key frames detection and analysis in vocal folds recordings using hierarchical motion techniques and texture information," 2009 16th IEEE International Conference on Image Processing (ICIP), 2009, pp. 653-656, doi: 10.1109/ICIP.2009.5413745.

[3] Xiaojun Guo and Fangxia Shi, "Quick extracting key frames from compressed video," 2010 2nd International Conference on Computer Engineering and Technology, 2010, pp. V4-163-V4-165, doi: 10.1109/ICCET.2010.5485659.

[4] B. F. Momin and G. B. Rupnar, "Key frame extraction in surveillance video using correlation," 2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), 2016, pp. 276-280, doi: 10.1109/ICACCCT.2016.7831645.

[5] C. Lv and Y. Huang, "Effective Key frame Extraction from Personal Video by Using Nearest Neighbor Clustering," 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2018, pp. 1-4, doi: 10.1109/CISP-BMEI.2018.8633207.

[6] F. Shi and X. Guo, "Key frame extraction based on kmeas results to adjacent DC images similarity," 2010 2nd International Conference on Signal Processing Systems, 2010, pp. V1-611-V1-613, doi: 10.1109/ICSPS.2010.5555457.

[7] C. Lü, J. Li, X. Chen and J. Pan, "Stereoscopic Video Color Correction Based on Key frame Extraction," 2013 Sixth International Symposium on Computational Intelligence and Design, 2013, pp. 250-253, doi: 10.1109/ISCID.2013.176.

[8] B. Sun, D. Kong, S. Wang and J. Li, "Key frame extraction for human motion capture data based on affinity propagation," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2018, pp. 107-112, doi: 10.1109/IEMCON.2018.8614862.

[9] M. Kim, L. Chau and W. Siu, "Key frame selection for motion capture using motion activity analysis," 2012 IEEE International Symposium on Circuits and Systems (ISCAS), 2012, pp. 612-615, doi: 10.1109/ISCAS.2012.6272106.

[10] D. Diklic, D. Petkovic and R. Danielson, "Automatic extraction of representative key frames based on scene content," Conference Record of Thirty-Second Asilomar Conference on Signals, Systems and Computers (Cat. No.98CH36284), 1998, pp. 877-881 vol.1, doi: 10.1109/ACSSC.1998.751008.