International Journal of Electrical and Computer Engineering (IJECE)
Vol. 4, No. 4, August 2014, pp. 593~602
ISSN: 2088-8708

Video Summarization Based on a Fuzzy Based Incremental Clustering

Monireh Pournazari, Fariborz Mahmoudi, Amir Masoud Eftekhari Moghadam
Islamic Azad University of Qazvin, Iran

Article Info

Article history:

Received Feb 22, 2014
Revised Jun 2, 2014
Accepted Jul 24, 2014

Keyword:

Fuzzy thresholding
Incremental clustering
Mamdani fuzzy system
Video summarization

ABSTRACT

The significant development of multimedia and digital video production in recent years has led to the mass production of personal and commercial video archives. Therefore, the need for efficient tools and methods of accessing video content and information rapidly is significantly increasing. Video summarization is the removal of visual redundancy and repetitive video frames, and obtaining a short summary of the whole video so that the summary obtained effectively reflects the whole video content. Examples of these summarizations in recent years include **STIMO** and **VSUMM**. According to users' comments, in the mentioned methods, the summarization has a high rate of error in a full report of summarization and a low accuracy in non-repetitive frames production, as well as a high computation time. In this paper, in order to solve these problems, we developed a system which modeled users' and supervisors' comments. We used a fuzzy based incremental clustering by which the selection and deselection of frames are done based on fuzzy rules. The extracted rules were determined based on users' comments on the video summarization. Finally, we performed our proposed method on the video clips used in the previous methods. Produced summaries were evaluated by a qualitative method to minimize human interferences. The results obtained indicate the high accuracy of summarization and the less computation time.

Corresponding Author:

Monireh Pournazari,
Islamic Azad University of Qazvin, Iran
Email: m_pournazary@yahoo.com

1. INTRODUCTION

Accessibility of the digital video content on the web is incredibly increasing. Websites such as YouTube and iTunes Video on which people upload or download videos are successful and they are developing rapidly. In this article, instead of Tag (which is not always accessible, integrated and appropriate), we use a search tool which is based on the video content. This tool can briefly display the content to the user in order to have an idea about the video content without watching the video and the user can decide whether to download or watch the video without watching it first. In this paper, we operate based on incremental clustering. The advantage of this kind of clustering is that it can find key frames in a predetermined manner without determining the number of key frames. But instead, there is a need for an accurate determination of threshold value which summarizes the film into appropriate number of frames. Accordingly, as the previous methods were not accurate in determining the threshold and used a fixed threshold to summarize all videos, we used an expert supervisor for determining the threshold so that the fuzzy logic will determine a value as the threshold for each of the videos under study and will perform the summarization operation in the best way. Therefore, the determined threshold depends on conditions and the scene type and it is calculated separately for each film. In this paper, based on the researches conducted on feature extraction methods such as Violet, Gabor filters, color histogram, edge extraction, and content segmentation, we use the color histogram as an appropriate method for extracting features. As the required continuity between frames exists,

without additional calculations, we cluster frames and finally summarize the video content. The article is organized as follows:

In section 2, we review the recent summarization methods which have the highest percentage of summarization accuracy so far. In section 3, the proposed fuzzy thresholding is presented. In section 4, we provide the clustering method used in the article and in section 5, we present the proposed method's diagram. In section 6, we show the proposed evaluation and the results obtained. The last section of the article is the general conclusion.

2. REVIEW OF VIDEO SUMMARIZATION METHODS

A method named VSUMM was proposed by Avila et al [1] which will extract color feature from frames after the pre-sampling stage of video frames. After the removal of meaningless frames, the remaining frames will be clustered based on k-means clustering algorithm. In other words, in the first stage, the video will be broken into frames. In the second stage, color features will be extracted from video frames in the form of a color histogram in HSV color space; however, these color features are extracted from the sampled frames (1 frame per second), not all video frames. Also, in this stage, meaningless frames are removed from sampled ones. In the third stage, frames are grouped by the k-means clustering algorithm. In the next stage, after clustering, one frame is selected from each cluster as the key frame. Furini et al [2] suggested a video summarization technique called STIMO based on HSV color histograms clustering. The main argument of this article is static and dynamic summarization which is a summarization technique for dynamic and on-the-fly summaries production. This mechanism is designed for customization purposes, i.e. users can select the length of the summary and the time they have to wait for receiving the summary. STIMO does not use all audio and video information (such as closed view and user preference) and it is designed as static or dynamic in order to summarize generic videos. Dynamic summaries with understandable sounds are performed to increase readability and information transmission. STIMO core includes a procedure which calculates HSV color space distribution for all video frames/perspectives and also includes a clustering algorithm which groups similar video frames/perspectives and determines the best frame/perspective for each group according to the their color similarity. This method has a high time complexity because each input frame should be compared with all frames in the clusters and it is time consuming in more clusters and has a higher memory. Zhang et al [5] used an unsupervised clustering for extracting the key frame. At first, the video is divided into shots; then, a color histogram in HDV color space of each frame is calculated. For attribution of a frame to cluster, the similarity of the frame with the center of that cluster is calculated based on t threshold. Hanzlik et al [6] presented a method for the video summary based on the cluster-validity analysis which is performed without human supervision. At first, the whole video is grouped into clusters. Each frame is displayed by color histograms in YUV color space. A classified cluster is applied to all video frames for n times. The first time starts with a predetermined number of clusters and one cluster will be added in each time the clustering is performed. Then, the system automatically finds arbitrary combination(s) of clusters by applying cluster-validity analysis. After finding the optimized number of clusters, each cluster is displayed by a specific frame which is a new key frame for the video. A method was proposed by Gong and Liu [7] for video summarization based on the singular value decomposition (SVD). At the beginning, a set of input video frames is selected (1 out of 10 frames); then, the color histogram in RGB color space is used for video frames. In order to combine spatial information, each frame is divided into 3x3 blocks and a three-dimensional histogram is created for each block. Theses 9 histograms are incorporated to form a feature vector. Using this extracted feature vector, the feature frame matrix A (usually sparse) is created for the video. Then, SVD is performed on matrix A in order to obtain matrix V so that each vector column shows a frame in the refined feature space. After that, the nearest cluster to the refined feature space is found, the amount of this cluster's content is calculated and this value is used as a threshold for clustering the remaining frames. The system selects a frame as the key frame from each cluster which is near the center of the cluster.

3. DETERMINATION OF THE FUZZY BASED THRESHOLD VALUE

In the incremental clustering, due to incomplete information of the accurate number of clusters, we have to select a threshold value in order to select or deselect a frame to place in clusters. The determination of this threshold without considering the nature of data will lead to a high error in clustering. Also, this threshold will lead to problems in partitioning clustering due to the lack of accurate information of clusters in its input. The determination of a limitation for selecting or rejecting a frame determines the power of the produced summarization system. In this article, we determine the threshold based on information values of the time of input frame and also the frame histogram distance to the frame which is in the center of clusters. The issue of threshold determination is an issue with two inputs and one output which meets the needs of

observers and users by designing and creating 36 rules. The overall structure of the proposed method for determination of the threshold value is comprised of 5 fuzzy layers which are shown in figure 1. In figure 1, n is the number of inputs and R is the number of rules used. X_is are fuzzy inputs; in other words, the histogram difference and time difference exist between the input frame and the center of clusters. The histogram difference indicates the proximity of the color content of two input frames to the center of cluster and the time difference is related to input frames and the frame which is in the center of cluster. Accordingly, we can study video frames in terms of two aspects of histogram and time with the purpose of matching the created summary according to users' comments on time and content.

3.1. Fuzzy Diagram Structure

In the first layer (histogram difference, time difference), inputs are studied and in the second layer, the membership function of each input is calculated. In the next layer, the extracted rules are applied. These rules have been proposed based on different time modes and histogram difference in order to lead inputs toward an appropriate threshold comprehensively. Then, in the next layer, inputs are aggregated and the last layer is defuzzification stage which is performed through an averaging method. Y_is are outputs of the fuzzy system which show the desired threshold value.

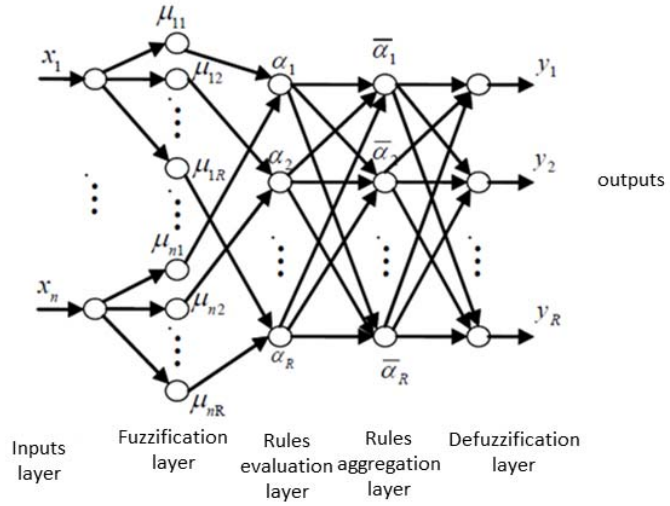

Figure 1. Fuzzy Diagram

3.2. Fuzzification of Inputs

The first step in fuzzy inference systems is receiving inputs and determining their degree of membership. In the fuzzy logic, inputs are always numerical values which are limited to the reference collection, i.e. tangible numbers for individuals and without normalization. In order to determine fuzzy intervals, we used the knowledge of maximum and minimum of input differences (histogram difference of 0 – 5, and time difference of 0 – 20). The reason of determining these values was that the highest histogram difference between input frames is not more than 5 and also, by investigating the video of collections used, it was determined that the maximum time of selecting a frame in consecutive frames is less than 20 seconds. In this article, we used the triangular fuzzy membership function. Relation (1) is the triangular membership function in which a, b and c are the locations of triangular function on X inputs. In this relation, the fuzzy inference system recieves inputs and determins the membership degree of inputs for each fuzzy set. Figures 2 and 3 show the digrams of fuzzy membership input for the two fuzzy inpus.

$$u(x;a,b,c) = \max(\min(\frac{x-a}{b-a}, \frac{c-x}{c-b}), 0) \quad (1)$$

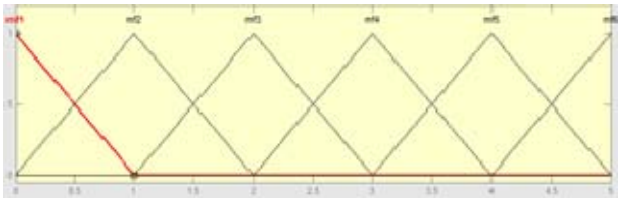

Figure 2. Membership function of histogram difference between two frames

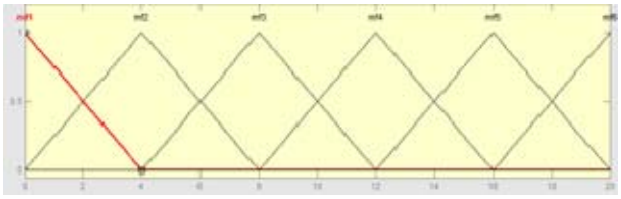

Figure 3. Membership function of time difference between two frames

All input variables should become fuzzy using membership functions. We use triangular membership functions for the fuzzification of two inputs.

3.3. Function of Fuzzy Operators

After fuzzification of inputs, the accuracy degree of each component of assumed sections (inputs) will be determined. Relation (2) is the computation function of fuzzy interface.

$$\alpha_j = \prod_{i=1}^{R} \mu_{ij} * \frac{1}{N_{adj}} \quad (2)$$

N_{adj} is the equivalent coefficient which in this article we consider it equal to n/4. μ_{ij} is the fuzzy membership coefficient.

3.4. Applying the Implication Method

The result section is a determined fuzzy set by the membership function. The process input shows a number and its output indicates a fuzzy set. According to relation (3), 90th j output is equal to:

$$\bar{\alpha}_j = \alpha_j / \sum_{i=1}^{R} \alpha_i \quad (3)$$

3.5. Outputs Aggregation

As in a fuzzy interface, decisions are made based on the evaluation of all rules, we integrate them in this layer and according to relation (4), 90th k output is equal to:

$$y_k = \sum_{j=1}^{R} w_{jk} \bar{\alpha}_j \, , k = 1,2, A \, , R \quad (4)$$

where W_{jk} is the weight of each rule which is applied to the value obtained from the assumed section. We consider the weight of each rule equal to 1.

3.6. Output Diagram and Threshold Determination

In order to regulate summarization rules, we determine a specified histogram difference and time difference for extracting the border of shots or different clusters (the clustering border is used for frames). Border determination is done as follows:

The histogram difference of various clusters is studied and the maximum different between clusters is selected. Also, determination of a valid time limit is done using the time difference between the central frame and the last one. The method of regulating input limits is shown in formula 5.

$$Intervalhistogram = MAX\ ((1/N)*\sum_{1}^{c}\sum_{1}^{c-1}(center(i)-center(j)))$$
$$Timeframemovie = MAX\ (\sum_{1}^{c} CENTER\_TIME\_FINAL\_FRAME\_CLUSTER)) \quad (5)$$

In figures 4 and 5, membership function of the threshold value and the relationship between inputs (time, histogram difference) and the threshold are selected, respectively.

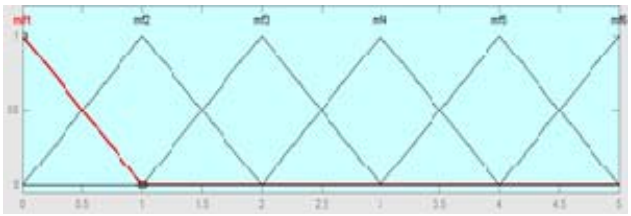

Figure 4. Membership function of the threshold value

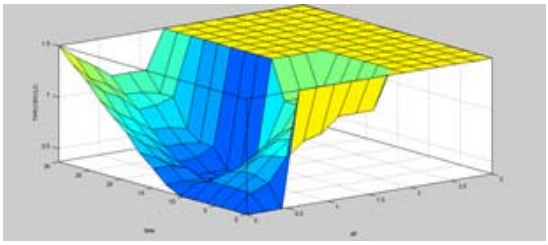

Figure 5. Ruled considered for the fuzzy system

In order to achieve the desired result for determination of the threshold and appropriate outputs which have less errors and a higher summarization percentage, we use rules and principles selected by users for a frame. There are different styles in these principles as follows:

- The selected frames must not follow each other.
- Frames must have a complete concept and meaning.
- All selected frames must show the film's concept and meaning.
- The selected frames must have less color similarity.
- They must guess similar locations in different situations.
- Time with color similarity must not be reviewed in a linear manner.
- Contentless flashes and frames must be removed.
- The selected frames with minor differences must not be placed in another cluster.

According to these principles, we regulate a relationship between time and color differences of the input frame and frames in the centers of clusters in order to satisfy general comments. Results of these rules indicate that the threshold determination significantly simulates users' comments which are shown in Figure 5.

4. CLUSTERING

In order to cluster frames whose features were extracted, we use the incremental clustering algorithm used in reference [2] which has grouped similar frames; then, one representative frame is selected in each cluster as the key frame. The extracted feature from color histogram frames is based on converting RGB color space into HSV one. This conversion occurs because HSV color space is close to the understanding of the human visual system. This operation starts with an input frame. Next, the histogram

difference and time of the new frame are calculated by the histogram difference and time of the frame which is in the center of formed clusters. Then, these two data are inserted in the rules formed in the fuzzy system and these rules establish a number for us. This number which shows the threshold value specifies that the input frame should be inserted in which cluster. If the number is more than other clusters' limits, a new cluster is formed and it is inserted in that cluster. Then, the time and histogram of the frame are recorded in the center of the new cluster. If a frame belongs to one of the previous clusters, a new center is created for that cluster according to the histogram rate of clusters and the aim is that cluster centers have the least difference with other existing frames in the same cluster so that at the end of summarization, we can select key frames without additional calculations. In this incremental clustering method, we store central frames in the memory using indexing (histogram rate and time of the input frame), we compare the histogram rate and the time of input frames with the information of central frame and evaluate their difference in order to determine that they are inserted in which fuzzy interval. One of the advantages of this method is when the time of films is long and we have a limited memory. Figure 6 shows the study of clustering. It is worth mentioning that in this incremental clustering, there is no need to compare the input frame with all clustered frames and it is compared just with one frame which is in the center of the cluster. In fact, this frame is inserted in one of the existing clusters using the determination of fuzzy high probability or a new cluster is created. With this clustering which is based on the fuzzy logic, computational speed is increasingly higher than STIMO method which is based on the incremental clustering and leads to a higher accuracy in summarization in less time, and this claim is discussed in the evaluation section. The reason of the computational complexity and the greater time of STIMO algorithm implementation is that it makes a frame to be compared with all other frames due to the purpose of clustering with less frame error which is not performed in the proposed method.

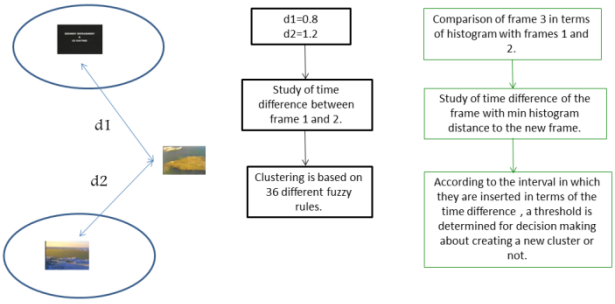

Figure 6. Study of inserting a new frame in the cluster

4.1. Pseudo Code of Fuzzy and Incremental Clustering Algorithm

In figure 7, the pseudo code relating to the extraction of fuzzy threshold and incremental algorithm is observable. In the implementation sequence, the threshold is extracted for each cluster; then, the limits specified in the incremental algorithm are used.

```
load (film)
all_image = main_frame_of_film;
%% find treashold call fuzzy function
treshold = zeros(size(all_image ));
for ii = 1 : size(all_image )
    different_histogram  = trimf(all_image (ii),[mf1, mf2, mf3, mf4, mf5, mf6]);
    different_time  = trimf([ all_image (ii)-cluster_center time] ,[mf1, mf2, mf3, mf4, mf5, mf6]);
    treshold(ii) = center_def (different_histogram , different_time );
end
%% find how many clustring find with fuzzy
    cont_fuzzy = call_fuzzy_clustring(treshold,all_image);
%%
center = zeros(cont_fuzzy,size_of_matrix_future);

for ii = 1 : size(all_image )
    image = load([all_image (ii)]);
    for jj = 1 : size(center)
        if (center(jj) - image) < MIN
            image_temp = image;
            MIN = (center(jj) - image);
        end
    end
    if (center(j) - image) < treshold(ii)
        begin
        center(j) = image_temp ;
    end
end
```

Figure 7. Pseudo code of the incremental clustring algorithm

5. DIAGRAM OF THE PROPOSED METHOD

The diagram in figure 8 shows the proposed method. In this method, the film is converted into a frame; then, features are extracted from frames and the method of fuzzy based threshold determination is implemented. At this stage, the number of clusters and threshold of each cluster are specified, and finally, the incremental algorithm is applied to the extracted feature frames.

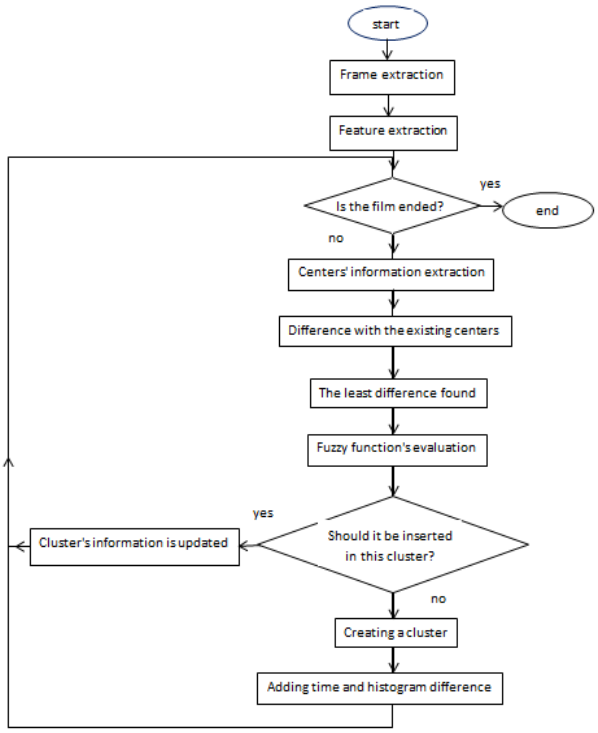

Figure 8. Diagram of the proposed method

In order to remove meaningless frames, the standard deviation of the feature vector is used so that if the feature vector of one frame has the standard deviation of zero or close to zero, it should be removed and should not enter the stage of clustering. Meaningless frames are observed as opaque video frames, frames with flash, black outputs or white flashes and also, they are created in images in the form of noise which create some problems for clustering and increase the clustering error.

6. EVALUATION

In order to evaluate the proposed method, we use the evaluation method used in reference [1] which performs the evaluation based on users' comments. This evaluation method can be observed in figure 9.

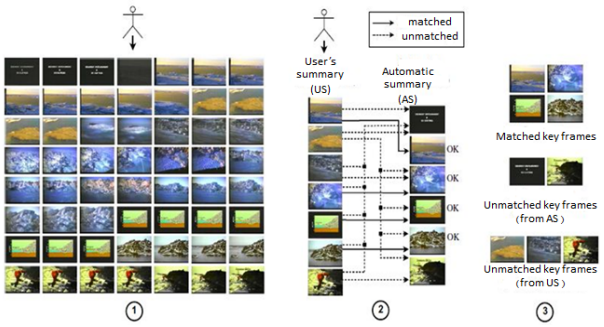

Figure 9. The value evaluation method used (from v summ)

In this evaluation, we consider a frame as the matched frame which has less difference with the key frame summarized by the user. In the related articles, this rate is considered to be 0.5 which is the best rate for confirming the difference determination.

6.1. Summarization Accuracy Rate

In this article, the evaluation was performed based on the similarity rate between summaries produced by users and the output of compared algorithms. According to the comparison of results extracted from previous methods with the proposed method, the method of this article shows the highest rate of summarization accuracy (because it was ranked based on users' output). The evaluation process is performed so that the proposed algorithm is compared with the frame summarized by users according to the conformity of the output frame. The conformity is determined through trial and error and we concluded that if the histogram difference between output frames of the algorithm and users' summary is less than 0.5, the summarization has conformity; otherwise, the summarization was done with error. Figure 10 shows this process. The accuracy rate is CUS_A which is observable in relation 6. N_{mAS} is the number of matched key frames in the produced summary through the proposed method and n_{US} is the number of key frames in users' summary.

$$CUS_A = \frac{n_{mAS}}{n_{US}} \tag{6}$$

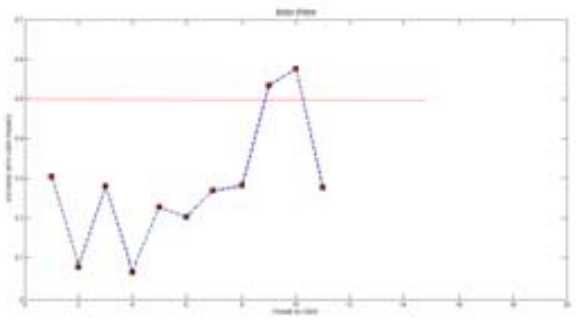

Figure 10. Difference between produced summarice

According to figure 10, we have 2 summarization errors among 11 output frames, i.e. 2 key frames produced by the proposed method are not available in the summary produced by users and are not considered by users. According to results of previous methods (STIMO, VSUMM, OV, and DT) and the proposed method on similar datasets, the proposed method has 15% superiority in the final summarization which is shown in table 1. The videos under study are 50 videos from video series of Open Video. All videos are in the format of MPEG-1, and have the resolution of 352x40 and the speed of 30 frames per second. The selected videos are divided into different groups (documentary, instructional, daily, historical and speech videos) and their time is variable from 1 to 4 minutes. 50 users have produced user summarizations and each user dealt with 5 videos, i.e. each video has 5 summaries which are produced by 5 different users. In other words, 250 video summaries were produced manually. You can see all videos and summaries of users at http://www.npdi.dcc.ufmg.br/VSUMM.

Table 1. The average of summarization accuracy rate

Summarization Methods	OV	DT	VSUMM	STIMO	Proposed Method
Mean accuracy rate	70%	53%	85%	72%	91.47%

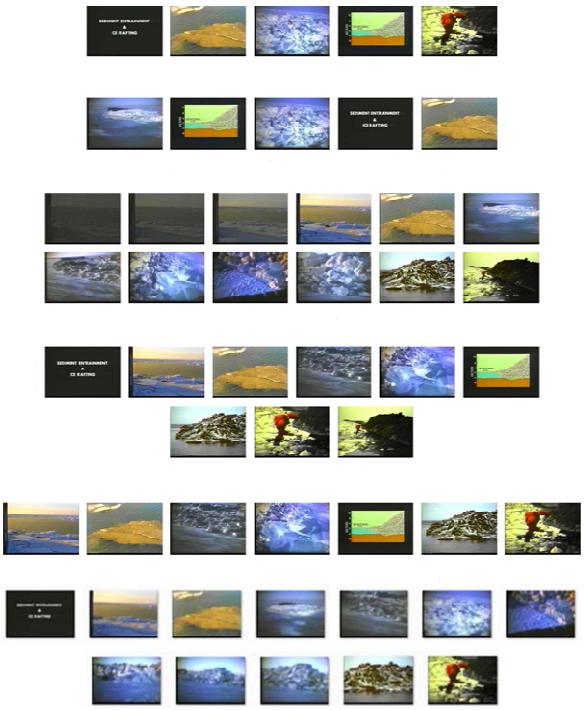

Figure 11. Video summaries of different approaches of the video Drift Ice as a Geologic Agent, segment 8 (available at Open Video Project)

6.2. Summarization Error Rate

In order to determine the summarization error, we divide unmatched frames of the proposed method and users' summarization into the number of key frames.

In other words, we should have the minimum number of unmatched frames, i.e. our algorithm's output should not have many key frames and frames provided for the output must have the maximum conformity with users' comments. According to the error of previous methods and the proposed one, in table 2 we can see that the error of the proposed method is 10% less than the previous methods. The error rate is called CUS_E which is defined as follows:

$$CUS_E = \frac{n_{\tilde{m}AS}}{n_{UT}} \tag{7}$$

where $n_{\tilde{m}AS}$ is the number of unmatched key frames in the summary produced by the proposed method.

Table 2. The average of summarization error rate

Summarization methods	OV	DT	VSUMM	STIMO	Proposed method
Mean error rate	57%	29%	38%	58%	25%

According to results of the previous methods, the proposed method has a significant superiority which is resulted from using the fuzzy expert method for the threshold determination. The key frames produced by the proposed method and other techniques are shown in figure 11 on the video of ice drift.

6.3. Comparison of Implemented Results Based on different Methods

The proposed method has a significant superiority compared to other methods due to the accuracy and error rates. This comparison was done between the proposed method and existing methods in references [1], [2], [8] and [9]. The accuracy and error rates of the aforementioned method are figures 12 and 13, respectively.

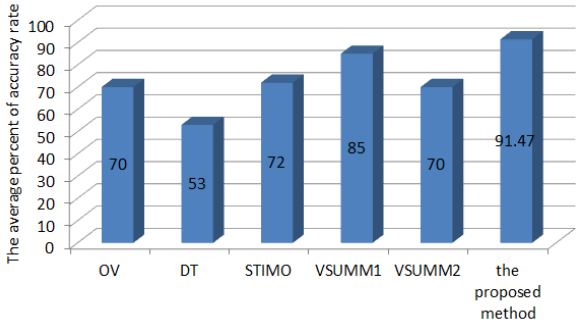

Figure 12. The comparison between the average percent of the accuracy rate of proposed methods and previous techniques

7. CONCLUSION

Video summarization is one of the most interesting issues in the film and advertising industry and different methods were designed for this issue. According to results obtained through previous methods, we cannot perform summarization with certainty. Therefore, we need an expert method which can summarize videos accurately and close to users' comments and this is possible through the fuzzy method. Furthermore, we need an appropriate computational speed in film summarization which we have achieved this purpose through improving the clustering method.

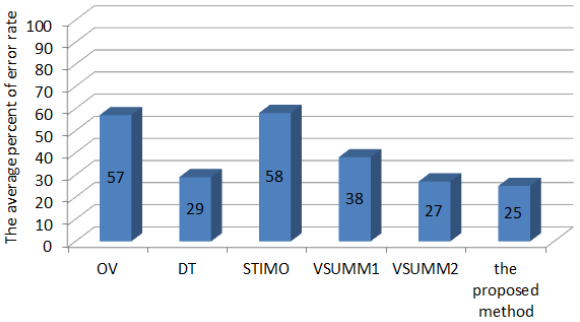

Figure 13. The comparison between the average percent of the error rate of proposed methods and previous techniques

REFERENCES

[1] Avila S., Lopes A., Luz Jr. A., Araujo A., 2011, *VSUMM: A mechanism design to produce static video summaries and a novel evaluation method*, Pattern Recognition Letters, Vol (32), pp. 56-68.
[2] Furini M., Geraci F., Montangero M., Pellegrini M., 2010, *STIMO:STIll and MOving video storyboard for the web scenario*, Multimedia Tools Appl., Vol (46), pp. 47-69.
[3] Yang Wu, Wu Yuan, Wu Zhongru, 2004, *Forecast Model Study Based on Fuzzy Neural Network*, Water Resources and Power, vol (22), pp. 63-65.
[4] Zhang X.Y. , Wang P., 2009, *Improved T-S Fuzzy Neural Network in Application of Speech Recognition System*, Computer Engineering and Applications, vol (45), pp. 246-248.
[5] Zhuang, Y., Rui, Y., Huang, T.S., Mehrotra, S., 1998, *Adaptive key frame extraction using unsupervised clustering*, In: Proc. IEEE Internet. Conf. on Image Processing (ICIP), vol. 1, pp. 866–870.
[6] Hanjalic, A., Zhang, H., 1999, *An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis*, IEEE Trans. Circuits Systems Video Technology 9 (8), 1280–1289.
[7] Gong, Y., Liu, X., 2000, *Video summarization using singular value decomposition*, In: Proc. IEEE Internat. Conf. on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, Los Alamitos, CA, USA, pp. 2174–2180.
[8] Mundur P., Rao Y., Yesha Y., 2006, *Keyframe-based video summarization using Delaunay clustering*, Int J Digit Lib, Vol (6), pp. 219-232.
[9] DeMenthon, D., Kobla, V., Doermann, D., 1998, *Video summarization by curve simplification*. In: Proc. ACM Internat. Conf. on Multimedia. NY, USA, pp. 211–218.