DOI: 10.24996/ijs.2019.60.7.23

Iraqi Journal of Science

ISSN: 0067-2904

Static and Dynamic Video Summarization

Abdul Amir Abdullah Karim[1], Rafal Ali Sameer*[2]

[1]Department of Computers, University of Technology, Baghdad, Iraq

[2]Department of Computers, College of Science, University of Baghdad, Baghdad, Iraq

Abstract

Video represented by a large number of frames synchronized with audio making video saving requires more storage, it's delivery slower, and computation cost expensive. Video summarization provides entire video information in minimum amount of time. This paper proposes static and dynamic video summarization methods. The proposed static video summarization method includes several steps which are extracting frames from video, keyframes selection, feature extraction and description, and matching feature descriptor with bag of visual words, and finally save frames when features matched. The proposed dynamic video summarization method includes in general extracting audio from video, calculating audio features using the average of samples in windows and find the highest average which reflects portion of video with loudest sound. The experimental results for the proposed static video summarization show that there is no redundancy between selected representative keyframes and the subjective evaluation results ensure the importance of the selected keyframes. While the experimental results for the proposed static video summarization show that all the segments of goals have been extracted to provide video summary. Static and dynamic video summarization methods done to football or soccer video type.

Keyword: segmentation, Bag of Visual Words (BoVW), summarization.

خلاصة الفيديو الثابتة والمتحركة

عبد الامير عبد الله كريم[1] ، رفل علي سمير*[2]

[1] قسم علوم الحاسوب، الجامعة التكنولوجية، بغداد، العراق

[2] قسم الحاسبات، كلية العلوم، جامعة بغداد، بغداد، العراق

الخلاصة

الفيديو يمثل عدد هائل من الصور متزامنة مع الصوت مما يجعل الفيديو يحتاج الى مساحة كبيرة للخزن وارسال الفيديو يكون بطئ ومعالجة الفيديو تحتاج وقت ومعالجة اكثر. عملية تلخيص الفيديو توفر فيديو اقصر يحمل معلومات مهمة عن الفيديو الاصلي وبوقت قصير. هذا البحث يقترح طريقتين لتلخبص الفيديو: طريقة تلخيص الفيديو بالصور و طريقة تلخيص الفيديو بالصوت والصورة. هنالك عدة مراحل لتلخيص الفيديو بالصور تتضمن استخراج الصور من الفيديو، تحديد ال keyframes، استخراج الخصائص والواصفات ومطابقتها مع حقيبة الكلمات الصورية (BoVW)، وخزن الصور او keyframes عند تطابق صفاتها مع (BoVW). الطريقة الثانية تتضمن استخراج مقاطع من الفيديو (عرض ملخص بالصوت والصورة) بعدة مراحل تتضمن استخراج الصوت وقراءة البيانات الخاصة به (header)، استخراج خصائص الصوت، حساب معدل ال samples لكل عدد ثابت من ال samples يسمى window، ايجاد اعلى معدل، واسترجاع المقطع المقابل من الفيديو الاصلي.

*Email: Rafalali@scbaghdad.edu.iq

1. Introduction

Video can be defined as a sequence of frames and sounds that makes up the video. Video is advanced multimedia that has been enabled by the availability of internet in the communication field. Increasing the use of internet videos has brought the need to summarize the videos to a smaller size to make video available for using in multiple networks easily at low cost. Summarization helps the user to determine if the video is worth watching or not [1].

Video summarization has been proposed to rapidly browse large video collections. It has also been used to efficiently index and access video content. To summarize any type of video, researchers have relied on visual features contained in frames [2].

This paper organizes as follows: section 2 presents some of the related works; section 3 presents methods used for video and image segmentation; section 4 presents SIFT features extraction and description method; section 5 presents Bag of Visual Words (BoVW) method; section 6 presents video summarization concept; section 7 presents the proposed video summarization methods; section 8 presents video summary evaluation approaches; section 9 presents experimental results; section 10 presents conclusion.

2. Relate work

There are many works created for video summarization some of them which are mostly related will be shown here In **2016**, Mohammed Hamid Al-Kubaisi, presents a thesis for video summarization by considering the power consumption and bandwidth for different videos test data through using the main parameters that measure the time, power and bandwidth before and after summarization, in addition to the power spectral density for each test. A video summarization is being applied on 8 test videos, a remarkable result regarding both power and bandwidth is achieved as the power consumption is been reduced by 80% and the bandwidth is being increased by 40% [1]. In **2016**, Marian Kogler, et.al, investigates the usefulness of local features in generating static video summaries. The proposed approach is based on bag of visual words using SIFT features. In an explorative experiment, this approach compared to summaries generated with the help of global features and concludes that the local feature based approach does not outperform the other ones, however, it seems to be more stable. Also the difference in runtime between the different approaches can be seen with extraction of SIFT features and finding of the visual words that took ten times longer than the extraction of global features [3]. In 2017, Edward Jorge Yuri, propose a dissertation for semantic video summarization that can produce meaningful and informative video summaries. The experimental results using over than 100 videos was used to achieve a stronger position about the performance of local descriptors in semantic video summarization show that the local descriptors perform better than global descriptors in video summarization [2]. In 2017, Dong-ju Jeong, et.al, present a video summarization method that is specifically for the static summary of consumer videos. Considering that the consumer videos usually have unclear shot boundaries and many low quality or meaningless frames. Experiments on videos with various lengths show that the resulting summaries closely follow the important contents of videos [4]. In 2018, Antti E. Ainasoja, et.al, focuses on the popular keyframe-based approach for video summarization. A summary is generated by temporally expanding the keyframes to key shots which are merged to a continuous dynamic video summary. Experiments verified that dynamic video based on BoW scene detection and motion analysis based keyframe selection provides a powerful processing pipeline for video summarization and they can be run in online mode [5].

3. Segmentation

The first step towards video summarization is temporal video segmentation. The aim of video segmentation is to create manageable basic elements by splitting the video stream into a set of meaningful information. Summary production starts by detecting the shot boundary or by extracting of video frames when there is no temporal analysis of the video. The video sequence is split into images, each frame is treated separately [6].

Video processing based on frame or image processing which have the same concept because every frame has the same structure of image (i.e. Video frames and digital images share canonical visual concepts) [7].

Image segmentation is defined as a method in which an image is partitioned into many parts to identify homogeneous regions such that an image is depicted into something that is easy to express and study. Image segmentation is one of the most important steps for feature extraction and analysis [8].

3.1 OTSU's Method

Otsu approach is a successful analytical and global method for image thresholding that is based on image's gray value only. Otsu method takes the best threshold **t** by searching for criterion for maximizing *between class variance* and minimizing the *within class variance*. For each value in grey image compute the weight, the mean, and the variance, the optimal threshold will equal to the lowest sum of weighted variance. Faster approach is to select the threshold with the maximum between class variance and has the minimum within class variance [9][10].

The image points in bi-level thresholding approach are split by the threshold **t** to two classes C1 and C2 where C1 gray levels range is [0, 1, ... , t] and *C2* gray levels range is [t+1, … , L-1]. The probability distributions of gray level pg1 and pg2 for C1 and C2 classes respectively are [11]:

$$pg_1(c1) = \sum_{i=0}^{t} pro \qquad \dots (1)$$

$$pg_2(c2) = \sum_{i=t+1}^{L-1} pro \qquad \dots (2)$$

Where *pro* represents probabilities of intensity.
The first class mean is m_1 and the second class mean is m_2:

$$m_1 = \sum_{i=0}^{t} i\ pro\ /pg1 \qquad \dots (3)$$

$$m_2 = \sum_{i=i+1}^{L-1} i\ pro\ /pg2 \qquad \dots (4)$$

The total mean of grey levels is denoted by m_t:

$$m_t = pg_1 \times m_1 + pg_2 \times m_2 \qquad \dots (5)$$

The variances of first and second class respectively denoted by σ_1, σ_2:

$$\sigma_1^2 = \sum_{i=0}^{t} (i - m1)^2\ pro\ /\ pg_1 \qquad \dots (6)$$

$$\sigma_2^2 = \sum_{i=i+1}^{L-1} (i - m2)^2\ pro\ /\ pg_2 \qquad \dots (7)$$

Within Class Variance σ_w:

$$\sigma_w^2 = pg_1 * \sigma_1 + pg_2 * \sigma_2 \qquad \dots (8)$$

Between Class Variance σ_b:

$$\sigma_b^2 = pg1\ (m1 - mt)^2 + pg2\ (m2 - mt)^2 \qquad \dots (9)$$

3.2 Normalization

Normalization or decorrelation of data done to bypass biasing distance or similarity measures, and to prepare data for classification algorithms. Limitation feature value for particular range can be done by linear techniques. Min-Max normalization technique used to map data to a particular range S_{MIN} to S_{MAX} but the relationship remain exist between values.

$$Norm_{(i,j)} = \left(\frac{Gray(i,j) - MinOfGray}{MaxOfGray - MinOfGray} \right) (S_{Max} - S_{Min}) + S_{Min} \qquad \dots (10)$$

Where
$Norm_{(i,j)}$ represents new normalized gray value,
S_{Max} represents the maximum value desired for the particular range,
S_{Min} represents the minimum value desired for particular range,
MinOfGray represents the minimum of the original data,
MaxOfGray represents the maximum of the original data.
When normalization techniques of data applied be careful because these techniques will move the mean, and will change the data spread [12].

3.3 Slope equation

For every line there is a slope, the slope is the changes of y to the change of x. The slope can be computed by defining two points of a line or a line equation. A line slope can be defined using Equation (11) when there are two points of a line are known:

$$m = \frac{change\ in\ y}{change\ in\ x} = \frac{\Delta y}{\Delta x} \qquad \dots (11)$$

where m refers to the slope, Δ is delta and refers to the changed magnitude, the change of x (Δx) and change of y (Δy). The general form of slope equation defined by equation (12).

$$y = mx + c \qquad \dots (12)$$

where m refers to the slope of line, c is the ordinate from the point of intersection of the line against the y axis. The lines on the plane exactly refer to one state i.e. parallel or crossed. Intersect can be perpendicular or intersect. The line is said to be parallel or perpendicular can be seen through the slope. When there are the same slopes for both lines then the two lines are said to be parallel. However, if the result of the slope is -1 then the two lines perpendicular [13].

4. Scale Invariant Feature Transform (SIFT)

SIFT is a local features detection and description algorithm, it is able to provide steady point which is invariant to translation, rotation, scaling, and lightening variation. SIFT is patent algorithm and take dense processing cost that make it too slow [14].

SIFT composed of four main stages; (a) detect keypoints, (b) localize keypoints, (c) assign orientation, and (d) describe keypoints. In the first step different scale of images created by using different value of σ in Gaussian function, then Subtract consecutive images to create DoG pyramid. After that the Gaussian image down sampled by 2 and creates DoG to down sampled image. Gaussian function is shown in equation (13) and DoG is shown in equation (14) [15][16].

$$G(m, n, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{m^2+n^2}{2\sigma^2}\right] \qquad \dots (13)$$

Where G(m, n, σ) represents changing scale Gaussian, σ represents the scale variable of the consecutive scale space, m and n represents horizontal and vertical coordinates in Gaussian window, π = 3.14

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \qquad \dots (14)$$

Where * represents the convolution operation, k represents scaling factor, G(x, y, σ) represents changing scale Gaussian function, I(x, y) represents an input image, D(x, y, σ) represents Difference of Gaussians have k times scale, x and y represents horizontal and vertical coordinates in image. Local extrema obtained by comparing every pixel after DoG with 26 other pixels (eight neighbour pixels at the current pixel's level and nine pixels in the upper level and nine pixels in the lower level). When the compared pixel is extrema pixel position and scale are saved. In the key-point localization step, low contrast points and points at edge are eliminated. Intervention point is also eliminated by using (2×2) Hessian matrix [15][17].

The descriptors build by calculating the gradient strength and orientation strength for each neighbour of a key-point. The neighbourhood of every key-point is characterized by creating 8 bins gradient and orientation histogram for 16×16 region of neighbours around key-point. The region is split up into 4×4 sub regions and each sub region have 8 directions this will produce 4×4×8= 128 dimensional vector to give a description for every key-point [17][18].

The existence of a large number of features will produce irrelevant or redundant features that increase the processing time and can also affect accuracy. The aim of feature selection is to reduce feature space dimensionality and to keep the distinctive features [19].

5. Bag of Visual Words

BoVW approach provides supervised classifiers based on visual words come from labelled images for label prediction of a unlabelled image [20]. BoVW calculated using three steps: local features extraction, features description, vocabulary generation [21].

Detected features are split to a number of clusters using the K-means clustering algorithm where each cluster will have features with similar descriptors and encodes each key-point by the index of the cluster to which it belongs this is called *vector quantization (VQ)* technique [20].

The VQ encoder encodes a given set of k-dimensional data vectors with a much smaller subset. The subset C is called a codebook and its elements C_i are called codewords, codevectors, reproducing vectors. The commonly used vector quantizers are based on nearest neighbour called Voronoi or nearest neighbour vector quantizer [22].

Every cluster defined by a visual word that represents the specific local pattern participated by the keypoints in that cluster, therefore a visual word vocabulary identifies all types of local patterns of image. The clusters number refers to the vocabulary size. The image can be defined as a bag of visual words (BoVW), or as a visual-word vector have the number of keypoints in corresponding cluster [20].

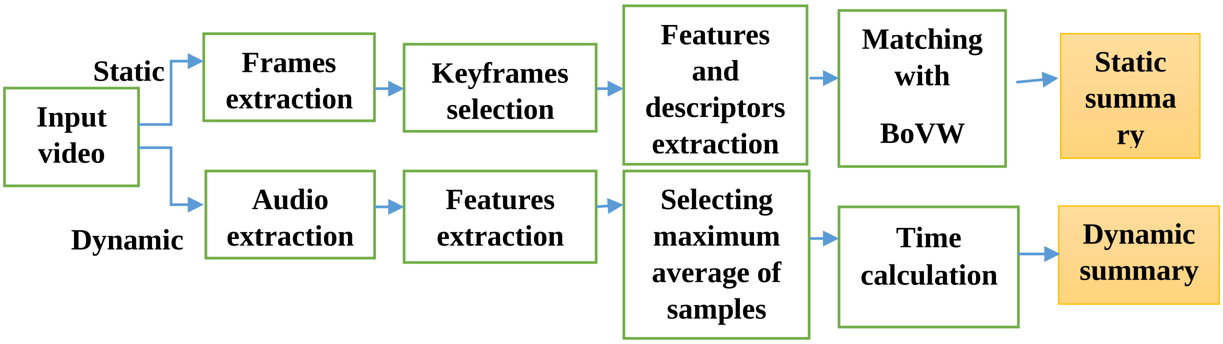
6. Video summarization

Video summarization is a technique used in creating shorter videos from original long videos [1]. A summary of video is a sequence of stable or moving images (with or without audio) presenting the concise information of the video content within minimum time possible. There are two main types of video summaries: static and dynamic video summary. Static summaries of video represented by set of keyframes taken from the original video, while dynamic summaries of video represented by of set of shots. One advantage of a dynamic summary over a static summary is the ability to include audio and motion elements that potentially enhance the amount of information conveyed by the summary. Furthermore, it is more interesting to watch a skim than a slide show of keyframes [6].

Video summarization technique has several advantages when one is using mobile phones the video summary helps save memory, increasing battery life, and reducing the cost to download a video [1].

7. The proposed video summarization methods

There are two proposed video summarization algorithms present in this section: static video summarization and dynamic video summarization.



7.1 Static video summarization method

Static video summarization gives a general idea about video using the most informative frames taken from original video. Static video summarization requires two main steps training and testing explained in the following sections with details.

7.1.1 Training process (Create Bag of Visual Word (BoVW))

In training process Bag of Visual Words (BoVW) will be created by extracting features and descriptors of 150 different images with different sizes, clustering features, and save them in database. SIFT was used in this work for image feature extraction and description. Visual feature description process based on calculating median, average, maximum, minimum, and orientation for every keypoint descriptor. Image segmentation is an important factor that affects the creation of the BoVW and matching with it. Otsu's segmentation method used in this work for image segmentation and agglomerative clustering method used for clustering purpose. SIFT feature detection method extract a large number of features that increase the processing time, hence it is important to reduce these features especially redundant features by adding preprocessing steps before the process of extracting feature is begin. Therefore grey image normalization and Otsu image segmentation have been added before feature extraction. Otsu image segmentation algorithm suffers to separate the noisy image histogram, therefore segmentation preceded by Gaussian function to blur the image and reduce any noise if there exist then followed by safe thresholding method. The proposed features selection method illustrated in algorithm (1) step (1) and an algorithm (2) step (3). The process of creating BoVW illustrate in algorithm (1):

```
Algorithm (1): "Create Bag of Visual Word (BoVW)"
Input: Training images, sigma, k.
Output: N clusters of feature description (database), N centroids.

Start
Step 1: Perform features selection:
Step 1.1: Read color image,
Step 1.2: Convert color image to gray image,
Step 1.3: Get maximum and minimum gray level values,
Step 1.4: Perform normalization equation (10)
Step 1.5: Blur gray image using Gaussian function, equation (13),
Step 1.6: Perform OTSU's image segmentation algorithm on blurred gray image,
Step 1.7: Extract features using SIFT feature detection method,
        Step 1.8: Save features position,
        Step 1.9: Extract features descriptors from gray image,
        Step 1.10: Calculate (average, median, maximum, minimum, and
        orientation) for every feature descriptor,
Step 2: Save features position extracted using SIFT from segmented image,
Step 3: Get features descriptor from gray image according to extracted features,
Step 4: Every key-point descripted by 64 integer number, calculate (median, average, maximum, minimum, and orientation) for every key-point descriptor,
Step 5: Go to step 1 until all images using for training process finished,
Step 6: Clustering features using agglomerative clustering algorithm which define the number of clusters, clustering features based on orientation, and define centroid for each cluster,
End.
```

7.1.2 Testing process (Static video summarization using object segmentation and Bag of Visual Words (BoVW))

The proposed approach for static video summarization including at first creating BoVW as discussed in section (7.1.1), then input video, frames extraction, fames reduction, frame segmentation, feature extraction, feature description, and matching with BoVW to save or ignore frame. SIFT preceded by grey image normalization and binary thresholding will be used in this work as described in section (7.1.1). Binary thresholding is done using Otsu's image segmentation method. Keyframes selection based on a proposed approach identified by slope equation and global mean for consecutive frames. The proposed approach for static video summarization illustrated in algorithm (2):

```
Algorithm (2): "Static Video Summarization using Object Segmentation and Bag of Visual Words (BoVW)"
Input: Video, BoVW, sigma, k.
Output: Representative keyframes.

Start
Step 1: Read video (soccer game video) and BoVW,
        Step 1.1: Get frames from video,
        Step 1.2: Resize frames, and save them,
Step 2: Select keyframes,
        Step 2.1: Get red, green, blue, and gray image from resized frame,
                Pixel = point from image
                Red = pixel Mod 256
                Green = ((pixel And &HFF00FF00) / 256)
                Blue = ((pixel And &HFF00000) / 65536)
                Gray = (Red + Green + Blue) /3
        Step 2.2: Calculate the global mean for resized frames red, green, blue,
        and gray component,
        Step 2.3: Calculate the slope equation (11) for two of previously calculated global
         means (here slope equation calculated using global mean of red and global mean of
         green),
        Step 2.4: If the slope is larger than one and the global mean values of red, green, blue,
        and gray component are not equal for two consecutive frames then save current frame
        as keyframe and continue,
        Step 2.5: Go to step (2.1) until resized frames finished,
Step 3: Extract and reduce features from keyframes,
Step 3.1: Read color image,
Step 3.2: Convert color image to gray image,
Step 3.3: Get maximum and minimum gray level values,
Step 3.4: Perform normalization equation (10)
Step 3.5: Blur gray image using Gaussian function, equation (13),
Step 3.6: Perform OTSU's image segmentation algorithm on blurred gray image,
Step 3.7: Extract features using SIFT feature detection method,
        Step 3.8: Save features position,
        Step 3.9: Extract features descriptors from gray image,
        Step 3.10: Calculate (average, median, maximum, minimum, and orientation) for every
        feature descriptor,
Step 4: Matching keyframe with BoVW,
        Step 4.1: Calculate keyframe descriptors centroid using k-mean,
        Step 4.2: Calculate Euclidean distance between  current keyframe
        centroid and BoVW centroids,
        Step 4.3: If keyframe descriptors match with cluster descriptors saved
        in BoVW then save keyframe, otherwise ignore current keyframe,
        Step 4.4: Go to step 3 until all keyframes finished,
Step 5: Show saved keyframes (representative keyframes),
End.
```

7.2 Dynamic video summarization method

Dynamic video summarization shows the most informative segments (visual and acoustic) in video. The proposed approach for dynamic video summarization including extract audio from video, read audio header, segment audio to N samples for each segment (N based on samples rate taken from audio header information) and the number of segments based on dividing total data size of audio by samples rate, take the maximum sample per segment, sort maximum samples and take 10 of their maximum, calculate and save time corresponding to extracted segments, extract video segments based on calculated and saved time, insert title for every segment, concatenate segments, and play segments of video. The proposed dynamic video summarization illustrates in the algorithm (3):

```
Algorithm (3): "Dynamic Video Summarization"
Input: Video, window length.
Output: Concatenated segments of video.

Start
Step 1: Read video (soccer game video),
        Step 1.1: Extract audio from video and save it as (*.wav) file,
        Step 1.2: Read audio header which include full information about audio
        (e.g. sample rate, file size, channels, etc.),
Step 2: Audio Segmentation where every segment have N samples according to window length (here window
length based on samples rate), window length = number of samples per second,
Step 3: Get the maximum sample per segment (the highest magnitude per second) and save them in list,
Step 4: Go to step 3 until the last audio segment reached,
Step 5: Sort and save maximum samples extracted from all audio segments and save their corresponding
sequence of appearance,
Step 6: Calculate time of maximum segments and save segments time in list,
Step 7: Filter segments,
        Step 7.1: Read the maximum sample saved in sorted list of maximum
        samples,
        Step 7.2: Save maximum sample from sorted maximum list and their
        corresponding time if there is 10 second at least between current
        maximum sample time and the next maximum sample time to avoid
        selecting redundant segments,
            Step 7.3: Go to step 7.1 until get 10 maximum sample,
Step 8: Extract video segment about (10 second) according to calculated time saved in previous list,
Step 9: Insert title slide for every segment and save video segments with titles,
Step 10: Concatenate video segments with titles,
Step 11: Play video summary,
End.
```

8. Video summary evaluation

The resulting video summary needs to be evaluated in order to verify the relevance of the selected “keyframes", and therefore evaluate how good the method performs. There is no standard method to evaluate performance of video summary, because of the subjectivity of the evaluation process. Consequently, there are little comparisons made between methods [2][41].

Structural SIMilarity (SSIM) method used to assess the quality of still images, that was extended to video. The SSIM was applied on video frame by frame on the luminance component and the overall video SSIM index computed as the average of the frame level quality scores [23].

The SSIM algorithm assumes that HVS is highly adapted for extracting structural information from a scene. Therefore, this algorithm attempts to model the structural information of an image. There are three steps to perform similarity measurement using SSIM algorithm which are: luminance, contrast, and structure comparisons. The luminance of each image is compared using equation (16) . The luminance comparison function $L(I_{ref}, I_{tst})$ is a function of μ_{ref} and μ_{tst}. The estimated mean intensity calculated by:

$$\mu_{ref} = \frac{1}{WH} \sum_{j=1}^{H}\sum_{i=1}^{W} I\,ref\,(i,j) \qquad \ldots (15)$$

$$L(I_{ref}, I_{tst}) = \frac{2\,\mu\,ref\,\mu\,tst + T1}{\mu^2\,ref + \mu^2\,tst + T1} \qquad \ldots (16)$$

The contrast comparison function $C(I_{ref}, I_{tst})$ is a function of σ_{ref} and σ_{tst} as in equation (18). Second standard deviation σ for each image found by:

$$\sigma_{ref} = \left(\left[\frac{1}{WH-1}\right] \sum_{j=1}^{H}\sum_{i=1}^{W}(I\,ref\,(i,j) - \mu\,ref\,)^2\right)^{\frac{1}{2}} \qquad \ldots (17)$$

$$C(I_{ref}, I_{tst}) = \frac{2\,\sigma\,ref\;\sigma\,tst + T2}{\sigma^2 ref + \sigma^2 tst + T2} \qquad \ldots (18)$$

The Structure comparison function $S(I_{ref}, I_{tst})$ is a function of $[I_{ref} - \mu_{ref}]/\sigma_{ref}$ and $[I_{tst} - \mu_{tst}]/\sigma_{tst}$.

$$S(I_{ref}, I_{tst}) = \frac{\sigma\,ref,tst + T3}{\sigma\,ref\;\sigma tst + T3} \qquad \ldots (19)$$

Where $\sigma_{ref,\,tst}$ is the correlation coefficient between the reference and test images can be estimated by:

$$\sigma_{ref,\,tst} = \frac{1}{WH-1} \sum_{j=1}^{H}\sum_{i=1}^{W}(I\,ref\,(i,j) - \mu\,ref)(I\,tst\,(i,j) - \mu\,tst) \qquad \ldots(20)$$

T_1 , T_2 , T_3 is a positive stabilizing constant chosen to prevent the denominator from becoming too small.

Finally an overall similarity measure created by combining these three comparison functions $L(I_{ref}, I_{tst})$, $C(I_{ref}, I_{tst})$, and $S(I_{ref}, I_{tst})$ and defined as follows:

$$SSIM\,(I_{ref}, I_{tst}) = [L(I_{ref}, I_{tst})]^{\alpha}\,[C(I_{ref}, I_{tst})]^{\beta}\,[S(I_{ref}, I_{tst})]^{\gamma} \qquad \ldots (21)$$

where α , β , and γ are positive constants chosen to indicate the relative importance of each component. There is a unique maximum, meaning that SSIM(x; y) = 1 if and only if x = y [23] [24].

Another popular form of similarity measure can be defined using two matrices Aij, Bij. The matrix inner product can be defined as $\sum_{i=1}^{N}\sum_{j=1}^{M} a(i,j)b(i,j)$then, the similarity test finally is given by:

$$Similarity\,(A_{ij}, B_{ij}) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} a(i,j)b(i,j)}{\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{M} a(i.j)^2}\,\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{M} b(i,j)^2}} \qquad \ldots(22)$$

When A_{ij} and B_{ij} are considered as two images matrices size (N×M), the similarity measure shows the amount of correlation between A_{ij} and B_{ij}. The similarity between two-image matrices gives its maximum value of (1) if the two images are perfectly similar [25].

Subjective methods are reliable methods to estimate the quality of video which perceived by a human observer by asking them for their opinion. Subjective measure is impractical for most applications due to the human involvement in the process. However, subjective measure provides valuable data to assess the performance of objective or automatic methods of quality assessment [23].

9. Experimental Results

Different types of videos can be used (*.mp4, *.mkv, *.mpeg, …). For static and dynamic video summarization methods. Soccer/football video used in this process downloaded from YouTube.

The first step in static video summarization method is creating BoVW. The total time required for creating BoVW in this paper was about (06:35:53). After creating BoVW read video and extract its frames using FFMPEG library for further process. The keyframes selected from extracted sequence of frames. Static video summarization method deals with cut transition and fade-in/fade-out transitions. In Table-1 general video information with processing time will be shown. Experimental results estimated using similarity measure and structure similarity (SSIM) measure between summarized frames and shown in Table-2.

Table 1-General information about static video summarization method

Video sequence	Video Length (h:m:s)	Total number of frames	Extract frames Time	Resize Time	Keyframes selection Time	Summary Time	Number of keyframes
Video 1	00:02:20	3,516	00:01:18	00:00:20	00:00:34	00:02:30	47
Video 2	00:14:41	21,161	00:02:12	00:00:48	00:02:08	00:03:21	200
Video 3	00:05:05	9,161	00:02:32	00:00:40	00:01:37	00:01:20	26
Video 4	00:10:27	15,682	00:01:55	00:00:38	00:02:00	00:02:47	160
Video 5	00:14:00	21,000	00:02:12	00:00:42	00:02:06	00:01:30	102
Video 6	00:14:00	21,000	00:02:04	00:00:40	00:02:02	00:01:40	117
Video 7	00:14:00	21,000	00:02:21	00:00:53	00:02:07	00:03:07	143
Video 8	00:14:00	21,000	00:02:10	00:00:39	00:02:10	00:02:11	150
Video 9	00:14:00	21,000	00:02:02	00:00:39	00:02:01	00:01:50	130
Video 10	00:14:00	21,000	00:02:16	00:00:37	00:02:02	00:02:16	164
Video 11	00:15:44	23,620	00:02:51	00:00:48	00:02:08	00:02:08	155
Video 12	00:15:00	22,500	00:06:24	00:01:14	00:02:11	00:06:13	129
Video 13	00:15:00	22,500	00:07:12	00:01:32	00:02:14	00:06:35	123
Video 14	00:15:00	22,500	00:07:24	00:02:21	00:02:26	00:07:21	144
Video 15	00:15:00	22,500	00:06:17	00:01:22	00:02:19	00:08:44	158
Video 16	00:15:00	22,500	00:06:14	00:01:16	00:02:16	00:05:55	108
Video 17	00:20:11	30,296	00:08:45	00:02:02	00:03:02	00:12:55	261

Table 2-The proposed static video summarization experimental results

Video sequence	Static Summary keyframes	Average similarity	Average luminance similarity	Average contrast similarity	Average structure similarity	SSIM
Video 1	10	0.7	0.7	0.7	0.1	0.1
Video 2	8	0.6	0.6	0.7	0.2	0.1
Video 3	8	0.5	0.7	0.7	0.07	0.08
Video 4	14	0.8	0.8	0.8	0.1	0.1
Video 5	10	0.8	0.8	0.8	0.1	0.1
Video 6	6	0.7	0.6	0.6	0.1	0.1
Video 7	17	0.8	0.8	0.8	0.2	0.2
Video 8	11	0.7	0.7	0.8	0.2	0.2
Video 9	11	0.8	0.8	0.7	0.2	0.2
Video 10	10	0.8	0.8	0.7	0.2	0.2

Video 11	14	0.8	0.8	0.8	0.1	0.1
Video 12	33	0.9	0.9	0.8	0.2	0.2
Video 13	30	0.9	0.8	0.8	0.3	0.2
Video 14	16	0.9	0.9	0.8	0.2	0.2
Video 15	40	0.9	0.9	0.8	0.2	0.2
Video 16	26	0.9	0.8	0.8	0.3	0.2
Video 17	16	0.9	0.9	0.8	0.2	0.2

Dynamic video summarization based on audio features. Audio extracted from video using FFMPEG library and save as (*.wav) file for further process. Processing audio based on two factors (time and amplitude) taking into account sample rate and total length of video used to find the exact position of extracted video segment. Dynamic video summarization experimental results shown in Table-3

Table 3-The proposed dynamic video summarization method experimental results

Video	Video Length (h:m:s)	Number of segments	Extract audio time	Process audio time	Insert title time	Concatenate time	Length of video summary
Video 1	00:02:20	6	00:00:02	00:01:35	00:01:40	00:00:01	00:00:32
Video 2	00:14:41	10	00:00:02	00:00:41	00:00:44	00:00:02	00:01:49
Video 3	00:05:05	This video has music sound (this video based on visual features)					
Video 4	00:10:27	18	00:00:03	00:01:20	00:01:23	00:00:03	00:03:19
Video 5	00:14:00	10	00:00:02	00:01:00	00:00:47	00:00:02	00:01:50
Video 6	00:14:00	12	00:00:02	00:00:52	00:00:52	00:00:03	00:02:12
Video 7	00:14:00	7	00:00:04	00:00:52	00:00:34	00:00:01	00:01:17
Video 8	00:14:00	11	00:00:04	00:00:44	00:00:48	00:00:02	00:02:01
Video 9	00:14:00	12	00:00:04	00:00:50	00:00:54	00:00:02	00:02:12
Video 10	00:14:00	12	00:00:04	00:00:50	00:00:54	00:00:02	00:02:12
Video 11	00:15:44	16	00:00:04	00:01:10	00:01:10	00:00:03	00:02:57
Video 12	00:15:00	9	00:00:05	00:02:12	00:02:15	00:00:02	00:01:39
Video 13	00:15:00	11	00:00:04	00:02:46	00:02:55	00:00:02	00:02:01
Video 14	00:15:00	11	00:00:08	00:02:49	00:02:58	00:00:02	00:02:01
Video 15	00:15:00	7	00:00:08	00:01:52	00:01:53	00:00:02	00:01:17
Video 16	00:15:00	10	00:00:03	00:02:25	00:02:31	00:00:02	00:01:50
Video 17	00:20:11	11	00:00:09	00:02:56	00:03:30	00:00:02	00:02:11

For every video in dynamic video summarization the time length of the original video is compared with the time length of the summarized video. The subjective measure done by 10 persons specialized in computer science especially image processing. In each run, the raters are asked to assign a score ranging from excellent to poor or to choose one of the existing answers. The questions given to the raters with a note at the beginning for their evaluation are as follows:

Note: Please view the original video first, and then assign a score to each of the summaries.

1. Is the soccer/football game boring for you, or you don't have time to see it and you need to know the results only?

1. Boring; 2. you don't have time; 3. need to know the results only; 4. none

2. What is your opinion about the results of static video summarization?

1. Excellent; 2. very good; 3. good; 4. average; 5. acceptable; 6. poor

3. What is your opinion about the results of dynamic video summarization?

1. Excellent; 2. very good; 3. good; 4. average; 5. acceptable; 6. poor

4. What is the best approach? 1. Static; 2. Dynamic

5. Can you tell me who is the winner from dynamic video Summarization? And how many goals? 1. Yes; 2. Sometimes; 3. No

6. From your sight of view does the summary pass any important frame or segment?

1. Yes; 2. Sometimes; 3. No

7. Do the keyframes have too much redundancy? 1. Yes; 2. Sometimes; 3. No

8. Was the static and dynamic video summarization algorithms interesting?

1. Yes; 2. Sometimes; 3. No

The answers of the raters was as follows according to the question sequence:

1. Ten answers were “need to know the results only”.
2. Four answers were excellent, four answers were very good, and two answers were average.
3. Six answers were excellent and four answers were very good.
4. Ten choose dynamic video summarization method.
5. Ten answers were yes.
6. Eight answers were no and two answers were sometimes.
7. Nine answers were no and one answer was sometimes.
8. Ten answers were yes.

10. Conclusion

There are two main processes for the proposed static video summarization algorithm which are training process to create BoVW and testing process to match features with BoVW. The second proposed algorithm dynamic video summarization based on audio features and show the short length of video that has the most informative segments extracted from the original video. The objective and subjective evaluation methods prove that the proposed algorithms are efficient and give important information about soccer/football video especially goals in short video length. The subjective measure shows that the dynamic video summarization is more expressiveness than static video summarization because dynamic video summarization shows both visual and acoustic information.

References

1. Al-Kubaisi M. H. **2016**. A Method for Static Video Summarization for Reducing Power Consumption and Its Impact on Bandwidth, Master Thesis, Department of computer science, Middle East University, Jordan.
2. Yuri E. J. **2017.** A New Method for Static Video Summarization Using Visual Words and Video Temporal Segmentation, Master Thesis, Department of computer science, Federal University of Ouro Preto, Brazil.
3. Fabro M. D., Schoeffmann K. and Böszörmenyi L. **2009**. Global vs. Local Feature in Video Summarization: Experimental Results. 10th Conference International Workshop of the Multimedia Metadata Community on Semantic Multimedia Database Technologies (SeMuDaTe'09) in Conjunction With the 4th International Conference on Semantic and Digital Media Technologies (SAMT), University of Klagenfurt, Austria, **539**: 1-8.
4. Jeong D., Yoo H. J. and Cho N. I. **2017**. A Static Video Summarization Method Based on the Sparse Coding of Features and Representativeness of Frames, Korea. *EURASIP Journal on Image and Video Processing,* **1**: 2-14.
5. Antti E. Ainasoja, Hietanen A., Lankinen J. and Kamarainen J. **2018**. Keyframe-based Video Summarization with Human in the Loop, Signal Processing Laboratory, In Proceedings of the 13th International Joint Conference on Computer Vision, **4**: 287-296.
6. Avila S. E., Lopes A. P., Luz A. and Araújo A. A. **2011**. VSUMM: A Mechanism Designed to Produce Static Video Summaries and a Novel Evaluation Method, Brazil. *Pattern Recognition Letters*, **32**: 56-68.
7. Song Y., Vallmitjana J., Stent A. and Jaimes A. **2015**. TVSum: SummarizingWeb Videos Using Titles, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New York, 5179-5187.
8. Sharma P. and Suji J. **2016**. A Review on Image Segmentation with its Clustering Techniques,. *International Journal of Signal Processing, Image Processing and Pattern Recognition,* **9**(5): 209-218.
9. Vala H. J. and Baxi A. **2013**. A Review on Otsu Image Segmentation Algorithm, India. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),* **2**(2): 387-389.
10. Karim A. A. and Sameer R. A. **2017**. Comparing the Main Approaches of Image Segmentation, Iraq. *Iraqi Journal of Science,* **58**(4B): 2211-2221.
11. Liu D. and Yu J. **2009**. Otsu method and K-means, China. 9th International Conference on Hybrid Intelligent Systems, **1**(1): 344-349.

12. Umbaugh S. E. **2011**. *Digital Image Processing and Analysis*. 2nd ed. CRC Press. New York. E-book.

13. Dewi A., Budiyono, C. and Riyadi, **2017**. Slope and Equation of Line: Teach and Analysis in Terms of Emotional Intelligence, International Conference on Mathematics and Science Education (ICMScE) IOP Publishing, Indonesia, 1-8.

14. He Y., Deng G., Wang Y., Wei L., Yang J., Li X. and Zhang Y. **2017**. Optimization of SIFT Algorithm for Fast-Image Feature Extraction in Line-Scanning Ophthalmoscope, China. *Optik International Journal for Light and Electron Optics*, **152**(1): 21-28.

15. El-gayar M., Soliman H. and meky N. **2013**. A Comparative Study of Image Low Level Feature Extraction Algorithms, Egypt. *Egyptian Informatics Journal*, **14**(2):175-181.

16. Panchal P. M., Panchal S. R. and Shah S. K. **2013**. A Comparison of SIFT and SURF, India. *International Journal of Innovative Research in Computer and Communication Engineering*, **1**(2): 323-327.

17. Wu J., Cui Z., Sheng V. S., Zhao P., Su D. and Gong S. **2013**. A Comparative Study of SIFT and its Variants, China. *Measurement Science Review*, **13**(3): 122-131.

18. Mendes p. j. **2013**. *Contribution to the completeness and complementarity of Local Image Features*, PhD Thesis, University of Coimbra.

19. Ghosh S., Dhamecha T. I., Keshari R., Singh R. and Vatsa M. **2015**. Feature and Keypoint Selection for Visible to Near-Infrared Face Matching, USA. IEEE 7th International Conference on Biometrics: Theory, Applications & Systems, 1-7.

20. Jun Y., Yu-Gang J., Alexander H. and Chong-Wah N. **2007**. Evaluating Bag-of-Visual-Words Representations in Scene Classification, *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval*, **2**: 197-206.

21. Mingyuan J., Christian W., Christophe G. and Atilla B. **2012**. Supervised Learning and Codebook Optimization for Bag-of-Words Models, *Springer Science Business Media*, **4**: 409-419.

22. Balwant A., and Doye D. **2012**. Speech Recognition Using Vector Quantization Through Modified K-means LBG Algorithm, Computer Engineering and Intelligent Systems, **3**:137-144.

23. Seshadrinathan K., Soundararajan R., Bovik A. C. and Cormack L. K. **2009**. Study of Subjective and Objective Quality Assessment of Video, *IEEE Transactions on Image Processing*, **19**(6): 1427-1441.

24. Dosselmann R. and Yang X. D. **2008**. A Formal Assessment of the Structural Similarity Index, Canada. *Technical Report TR-CS*, **2**: 1-14.

25. Filev P., Hadjiiski L., Sahiner B., Chan H. and Helvie M. A. **2005**. Comparison of Similarity Measures For The Task of Template Matching of Masses on Serial Mammograms, *American Association of Physicists in Medicine*, **32**(2): 515-529.