J Inf Process Syst, Vol.11, No.1, pp.76~94, March 2015
http://dx.doi.org/10.3745/JIPS.01.0006

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Viewer's Affective Feedback for Video Summarization

Majdi Dammak*, Ali Wali*, and Adel M. Alimi*

Abstract

For different reasons, many viewers like to watch a summary of films without having to waste their time. Traditionally, video film was analyzed manually to provide a summary of it, but this costs an important amount of work time. Therefore, it has become urgent to propose a tool for the automatic video summarization job. The automatic video summarization aims at extracting all of the important moments in which viewers might be interested. All summarization criteria can differ from one video to another. This paper presents how the emotional dimensions issued from real viewers can be used as an important input for computing which part is the most interesting in the total time of a film. Our results, which are based on lab experiments that were carried out, are significant and promising.

Keywords

Affective Computing, Emotion, FABO, K-NN, Motion Recognition, PCA, Video Summarization

1. Introduction

The massive explosion of multimedia streaming videos and the need for an intelligent video summarization has led to the development of a great variety of techniques in which the goal is to provide the better automatic summary. Providing the best possible analysis of a video is still an important challenge because of the great variety and diversity of videos that exist. To this end, some researchers have proposed a variety of highlight summarization methods for specific video films.

Recently, some research studies that are based on events detection have achieved interesting summarization methods. In [1], the authors propose the strategy of using a multi-support vector machine incremental learning system that is based on the Learn++ classifier for the detection of predefined events in a video. This strategy is offline and fast. In fact, any new class of events can be learned by the system from very few examples. Therefore, the extraction and synthesis of suitable video events are used for this purpose.

Some authors have relied on the analysis of the viewer's facial expressions. In [2], the authors defined some of the facial motion vectors based on twelve key points that present an affective dimensional degree of the video contents, in order to detect the personal highlights moments that can vary considerably according to each viewers. On the other hand, in [3], the authors turned to detecting the

Manuscript received December 27, 2013; first revision March 13, 2014; second revision May 13, 2014; accepted May 21, 2014; onlinefirst January 28, 2015.
Corresponding Author: Majdi Dammak (majdi.dammak@ieee.org)
* Research Groups in Intelligent Machines (REGIM), University of Sfax, Sfax 3038, Tunisia ({majdi.dammak, ali.wali, adel.alimi}@ieee.org)

important segments where a logo appears, in addition to detecting the caption region that provides information about the score of the soccer game. Moreover, in [4], the authors proposed a more unified framework for the summary of sports videos. For the semi-automatic construction of a sports video summary, they developed effective whistle, excitement, and text detection algorithms that are reliable and accurate.

In [3], the authors proposed an automatic and computational framework to analyze and summarize soccer videos using cinematic and object-based features. The framework includes some low-level soccer video processing algorithms, such as dominant color region detection and robust shot boundary detection. In the existing temporal video segmentation techniques, the authors [5] suggested a lot of overviews operating on the two compressed and uncompressed video streams.

Essentially, most of the algorithms for the uncompressed data rely on histogram comparisons, frame differences, or pixel-based blocks. However, most of the existing techniques are based on appropriate thresholding differences between successive images. Only a few machine learning approaches have tried to overcome this drawback.

In this paper, we provide report on machine learning for the film summarization system with the aim to emphasize all of the important subsequences in video films relying on the viewers' emotional feedback. With the camera fixed on the viewers, our proposed system tried to detect all of the predefined body postures and gestures defined in the Bimodal Face and Body Gesture Database (FABO), using the K-NN classifier method and it segmented the whole stream into smaller video sections. Our proposed system proved to be efficient at greatly reducing summarization workload and enhancing the accuracy of a film.

The remainder of this paper was organized as follows: Section 2 is covers related works that have been carried out on video summarization techniques. Section 3 is devoted to the background and motivation of our work. Section 4 presents the method and tools that have been used in this work. Section 5 introduces the architecture and the different algorithms that we used for our framework and in Section 6 we provide the results and in Section 7 we present comparative results and finally in Section 8 we present the conclusions of our work.

2. Related Work

The necessity of a video summarization process has become more and more important due to the increasing number of sports videos. Different techniques have been proposed in video processing literature to compress the massive number of sports videos via a summarization process.

'Complete Sports Video Summarization' is a technique proposed by [6] that can fit all the various types of users and applications. The framework that integrates highlights and reveals the need for breaks is based on a mixture of highlights, plays, and breaks [7]. Their experimental results show that fast detections of whistle sounds, the crowd's excitement, and text boxes using some slightly adjustable thresholds when applying algorithms to different videos to avoid misdetections and reduce false detections are possible.

The authors in [6] defined three types of detection models. The first is a detection of play-break by using the camera views classification. It can be used to detect the play-break transition. The second is

for highlights detection, which is based on the relay scenes in slow motion. Finally, the third model is text detection. For any sports video and after any important event, text appears on the screen. As such, it is the role of this model to detect this displayed text. This model tries to detect only the text that is displayed horizontally. This can be a limitation for other sports videos that display text in other ways.

In [8], the authors proposed an automatic content-based video summarization method using metadata for large sports video archives. They also presented the following two visualization systems for the video summary: a video clip and a video poster. These systems formulate the problems that specific to each type of video. They proposed a method for the generation of a video summary of arbitrary length while handling several videos at the same time. As a result of their experiments on baseball videos, they obtained only the significant action scenes with a recall rate of 66% and a precision rate of 83%, as compared to the summaries that are broadcast on TV.

In [9], the authors proposed a method for extracting highlights from a TV sports broadcast. The cited method does not require the modeling of domain specific events that are supposed to be interpreted by the user as highlights. Instead, it looks for highlights at places where strong excitement is provoked in the user due to the content of a video. It seems to be particularly realistic to assume that this type of independent, highlighted event induces an increase in a user's excitement. This method mimics the changes in the user's excitement by observing the temporal behavior of the domain-independent audiovisual signal properties and the editing scheme of a video. The relation between these low-level features and the evoked excitement are drawn partly from psycho-physiological research and partly from the video analyzing practice. The authors evaluated their methodology on excerpts taken from soccer games that were broadcast on TV.

In [4], the authors proposed a unified summarization scheme that integrates the highlights and play-break scenes. For the automation of the process, combing the audio and visual features provides a more accurate detection. The authors presented fast detection algorithms of whistles and excitement to take advantage of the fact that audio features are computationally cheaper than visual ones. However, due to the amount of noise in a sports audio, fast text-display detection is used to check the detected highlights. The performance of these algorithms was tested on one hour of soccer and swimming videos.

In [10], the authors presented a sports video summarization framework by using a combination of text, video, and logic analysis. Parse trees were used to analyze structured and free-style text webcasting of sports games and extracts of the games semantic events, such as goals and penalties during a soccer games. Then, the semantic events were hierarchically arranged before being passed onto a logic processing method. The logic engine receives the summary preferences from the user and subsequently parses the event hierarchy to generate the game summary according to the user's preferences. The proposed system was applied to both soccer and basketball videos. It achieved an average accuracy of 98.6% and 100% on soccer and basketball videos, respectively.

In [11], the authors presented a soccer video abstraction method based on the analysis of the audio and video streams. This method can be applied to other sports, such as rugby or American football. The main contribution of the authors is the design of an unsupervised summarization method, and more specifically, the introduction of an efficient detector of excited speech segments. An excited commentary is supposed to correspond to an interesting moment of the game. It is simultaneously characterized by an increase of the pitch (or fundamental frequency) within the voiced segments and an

increase of the energy supported by the harmonics of the pitch. The pitch is estimated from the autocorrelation function and its local increases are detected from a multi-resolution technique.

We are introducing a specific energy measure for the voiced segments. A statistical analysis of the energy measures is performed to detect the most exciting parts of the speech. A deterministic combination of excited speech detection, dominant color identification, and camera motion analysis is then performed in order to discriminate between excited speech sequences of the game and the excited speech sequences in commercials or in studio shots that are included in the processed TV programs. The method does not require a learning stage. It has been tested on seven soccer videos for a total duration of almost 20 hours.

To recognize a user's emotion feedback, many results based on different modalities (face, body gesture, etc.) in the affective computing literature. In [1], the authors provide the results for recognizing emotions based on a real-time analysis of upper body gestures using Kinect sensors. They tried to detect all of the emotional feedback from a user. To this end, they relied on three principal dimensions of body gestures to compute the user's emotions. On the one hand, they tried to track all the body movements in real time to principally compute the movement intensity and, on the other hand, they tried to detect the evolution of body motion, in order to estimate the emotions of a user.

3. Background and Motivations

When describing an emotional experience, the user often focuses on its intensity (i.e., "I felt kind of sad" or "I was very happy"). In a categorical approach, this experience type is not sufficient to characterize an emotion. An emotion is usually associated with a numerical value that represents its intensity. In psychological literature, we generally distinguish the strong emotions from low or medium intensity emotions. Their impact on an individual's behavior and cognitive abilities may indeed differ completely following due to these intensities [12].

All emotions triggered by an event (called a triggered emotion or an emotional impulse [13,14] or emotional stimulus) that differs from the emotional state. An emotional state is defined as the set of emotions felt by an individual at a given moment. A triggered emotion is a change and is used on the viewer's emotional state. It can be distinguished by its real impact on the emotional state of this viewer.

A viewer, however, may reveal a steady emotional state along most of the video show time. The movement intensity gradually decreases to zero. In emotion models, a decay function is generally defined to determine the intensity of the emotion and its value at time t to time t + 1.

In [13], the decay function is defined as the agent determines the speed decrease of the movement intensity. The study concluded that the emotion decay rate is associated with what is known as 'personality figures.'

3.1 Multimodal Emotion Recognition

In a recent paper [15], the authors presented the classification hierarchy of different human emotional channels. The first and most credible is the physiological channel that needs special sensors and causes the loss of a spontaneous interaction. It is also more expensive to use when compared to

other channels like the body, the face, and the voice.

For the visual facial expression recognition Ekman and Frieson [16] laid the foundation of the field. Their studies suggested that anger, disgust, fear, happiness, sadness, and surprise are the six basic prototypical facial expressions are universally recognized. In addition, Coulson [17] presented an experimental result on the attribution of six emotions to static body postures using computer-generated figures.

In the existing affective databases that are based on body gestures, the authors [18,19] take into account only the static and dynamic hand gestures. In their paper, [20] précised that the existing databases consist mainly of non-affective one hand gestures only, and do not take into consideration the relationship between all body parts.

To cope with the existing limitations, we used the FABO database in our research. FABO is a Bimodal Face and Body Gesture database for the automatic analysis of human nonverbal affective behavior.

The FABO database, as described on their website, was created at the Faculty of Information Technology at the University of Technology, Sydney (UTS) by Gunes and Piccardi [20] in 2005. Posed visual data was collected from volunteers in a laboratory setting by asking and directing the participants to make the required actions and movements. The FABO database contains videos of face and body expressions that were recorded by face and body cameras, simultaneously. This database is the first to date to combine facial and body displays in a truly bimodal manner, thereby making significant future progress in affective computing research possible.

3.2 Video Summarization

Video summarization is one of the solutions that has been proposed to make the browsing of multimedia easier. As creating video summaries manually is time consuming, several automatic summarization systems have been proposed in the literature. Summarizing a video sequence aims at extracting parts from the original version that are considered important. Video summarization systems generally use different rules and heuristics to extract important parts [21]. The choice of rules depends on the context (i.e., movies, news programs, sports competitions, etc.) and on the author's point of view (i.e., what is the content that is considered to be important by the author). Therefore, we cannot have standard summarization systems that generate standard summaries.

In the literature, we can distinguish two forms of summaries, which are as listed below.

- Storyboard: this is a set of key-frames extracted from the original sequence that give an overview of the whole video sequence.
- Video skim: this type of summary consists of a set of video excerpts extracted from the original sequence because they have been judged to be important. They edited together by either a cut effect or a gradual effect.

We can classify these approaches according to the form of the generated video summary, namely, the static video summarization (storyboard) and the dynamic video summarization (video skim).

The static video summary is achieved by extracting some key-frames from the original video sequences that have been judged to be important to construct a pictorial abstract at the end. The aim of this kind of summary is to help users to judge whether a given video involves the target elements. However, the main challenge for video summarization systems that generate storyboards is the issue of

how to select the suitable key-frames to construct the most informative pictorial abstract.

Although the storyboard is interesting for gaining an overview of theentire video sequence, it suffers from a lack of semantics due to the absence of the temporal and auditory information. The events chronological order and the audio help a lot in understanding the context of the original video sequence. For this reason most of there search has been focused on video skims. In the literature, we find two kinds of dynamic summaries:

- Video skims, which are used to provide users with an idea about the entire contents of the video and
- Highlights sequences, which provide users with exciting scenes and events.

In our case, we proposed the dynamic video summarization system (video skim) that relies on the classification of a viewer's emotions.

4. Tools and Methods

This section presents the different methods and tools that have been used to develop our new technique for a real-time video summarization based on the viewer's affective feedback.

4.1 The Basic PCA Approach to Face Recognition

From neuroscience to computer graphics, Principle Component Analysis (PCA) is frequently used for different forms of data analysis. PCA is a parametric method for the extraction of pertinent information from large data sets. It offers the solution of reducing a complex data set to a lower one-dimensional data with a good representation of the information for discriminative feature selection [22]. In Turk and Pentland [22]'s face recognition method, they explored the advantages of the PCA approach by using a training database of face images.

With the principle component analysis algorithm, the most important eigenvectors are detected from the training database. After calculating the eigenvectors, all face images are shown by a feature vector in the subspace that is determined by the eigenvectors. An algorithm summary is presented below.

For an image with a grey level of N×N, we used it as a N^2 one-dimensional vector. Consider a matrix X with N^2 rows and S columns. S is the number of training images.

$$X_{N^2*S} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1S} \\ x_{21} & x_{22} & \dots & x_{2S} \\ \dots & \dots & \dots & \dots \\ x_{N^2 1} & x_{N^2 2} & \dots & x_{N^2 S} \end{bmatrix}$$

Based on the PCA algorithm, we produced another matrix called the P matrix, which describes a linear transformation of every training number (the column in X) in the eigenvectors subspace, in the form of:

$$:W = PX$$

where W is the training face images projection on the subspace described by the eigenvector.

We detailed below the instructions for computing the P matrix as described in [20].

- Find the mean face vector M, as in (2), where X_i(i=1,2,…, S) represents the ith column of X: (Step 2)

$$M = \frac{1}{S}\sum_{i=1}^{S} X_i \tag{1}$$

- Subtract the mean face M from each training face X_i: (Step3)

$$H_i = X_i - M \tag{2}$$

- Compute the covariance matrix C_A, where A= [H_1, H_2... H_S]: (Step 4)

$$C_A = \frac{1}{S-1} AA^T \tag{3}$$

The largest Q eigenvectors of the covariance matrix are the vectors of the best basis for the training set X. Each eigenvector represents a column of .the matrix $P = [N^2 \times Q]$. The representation of any image I in the subspace described by the Q eigenvectors is given by the vector W_i (of length Q): (Step 5)

$$W_I = P(I - M) \tag{4}$$

For each image, a projection on the subspace will be extracted. The projection of the test image is compared with each training projection in all the classical approaches on a visible spectrum.

4.2 K-Nearest Neighbor (K-NN) Algorithm

The K-NN algorithm is amongst the simplest of all machine learning algorithms. It is a supervised classification method in a given feature space. In our case, the feature space is the eigenvector space computed with the PCA algorithm.

The number of classes, C, must be defined to construct the K-NN classifier, in order to define a labelled training set of N_{trn} samples in the feature space, with the class labels yi=1…C, and to consider the cited labelled samples in the training set as known prototypes of the C classes.

Otherwise, in the characteristics space, a norm distance must be defined for the classification. Any defined vector of W, from the characteristics space, needs to be classified in the C classes. Relying on K-NN, W is ranged within its K-nearest neighbors.

The algorithm must precede the steps mentioned below:

- For all training set prototypes, W_{tj} with j=1, 2,…, N_{trn}, define the distances $d(W, W_{tj})$.
- Sort the distances $d(W, W_{tj})$, j=1...N_{trn}, increasingly, and keep the labels of the first k prototypes (found at the first k smallest distances from W): {$y1'$, $y2'$…y_k'}.
- Assign to W the y_l' label, most frequent from the class sort array {$y1'$, $y2'$…y_k'}.

4.3 The FABO Database

The Bimodal Face and Body Gesture Database (FABO) for the Automatic Analysis of Human Nonverbal Affective Behavior, was developed at the University of Technology, Sydney (UTS) in 2005. Posed visual data was collected from volunteers in a laboratory setting by asking and directing the participants to make the required actions and movements.

Table 1. List of the face and body gestures performed for the recordings of FABO [20]

Expression	Face gesture	Body gesture	
Neutral	**No expression**	**Hands on the table, relaxed**	
Uncertainty	Lip suck Lid droop Eyes closed Eyes turn right/left/up/down	Head tilt left/right/up/down Palms up Shoulder shrug Palms up+shoulder shrug Right/left hand scratching the head/hair Right/left hand touching the right ear Right/left hand touching the right part of the nose	Right/left hand touching the chin Right/left hand touching the neck Right/left hand touching the forehead Both hands touching the forehead Right/left hand below the chin, elbow on the table Two hands behind the head
Anger	Brows lowered and drawn together Lines appear between brows Lower lid tense/may be raised Upper lid tense/lowered due to brows' action Lips are pressed together with corners straight or down or open	Open/expanded body Hands on hips/waist Closed hands/clenched fists Palm-down gesture Lift the right/left hand up	Finger point with right/left hand, shake the finger/hand Crossing the arms
Surprise	Brows raised Skin below brow stretched, not wrinkled horizontal wrinkles across forehead Eyelids opened Jaw drops open or stretching of the mouth	Right/left hand going to the head Both hands going to the head Moving the right/left hand up Two hands touching the head Two hands touching the face, mouth	Right/left hand touching the face, mouth Both hands over the head Right/left hand touching the face Self-touch two hands covering the cheeks Self-touch two hands covering the mouth Head shaking Body shift-backing
Fear	Brows raised and drawn together forehead wrinkles drawn to the center Upper eyelid is raised and lower eyelid is drawn up Mouth is open Lips are slightly tense or stretched and drawn back	Body contracted Closed body/closed hands /clenched fist Body contracted, arms around the body Self-touch (disbelief)/covering the bodyparts/arms around the body/shoulders Body shift-backing, handcovering the head	Body shift-backing, hand covering the neck Body shift-backing, hands covering the face Both hands over the head Self-touch (disbelief) covering the face with hands
Anxiety	Lip suck Lip bite	Hands pressed together in a moving sequence Tapping the tips of the fingers on the table	

	Lid droop Eyes closed Eyes turn right/left/up/down	Biting the nails Head tilt left/right/up/down	
Happiness	Corners of lips are drawn back and up Mouth parted/not with teeth exposed/not Cheeks are raised Lower eyelid shows wrinkles below it Wrinkles around the outer corners of the eyes	Body extended Hands clapping Arms lifted up or away from the body with hands made into fists	
Disgust	Upper lip is raised Lower lip is raised and pushed up to upper lip or it is lowered Nose is wrinkled Cheeks are raised Brows are lowered	Hands close to the body Body shift-backing Orientation changed/moving to the right or left Backing, hands covering the head	Backing, hands covering the neck Backing, right/left hand on the mouth Backing, move right/left hand up
Boredom	Lid droop Eyes closed Lip suck Eyes turn right/left/up/down	Body shift, change orientation, move to the right/left Hands behind the head, body shifted Hands below the chin, elbow on the table	
Sadness	Inner corners of eyebrows are drawn up Upper lid inner corner is raised Corners of the lips are drawn downwards	Contracted/closed body Dropped shoulders Bowed head Body shift-forward leaning trunk Covering the face with two hands Self-touch (disbelief)/covering the bodyparts/arms around the body/shoulders	Body extended+hands over the head Hands kept lower than their normal position, hands closed move slowly Two hands touching the head move slowly One hand touching the neck, move hands together, closed and head bent

The FABO database contains videos of face and body expressions that were recorded by the face and body cameras, simultaneously, as shown in the figures below. This database is the first to date to combine face and body displays in a truly bimodal manner, thereby enabling significant future progress to be made in affective computing research. The goal of the FABO database is to provide a data source for the research on Affective Multimodal Human Computer Interactions. This database aims to help researchers develop new techniques, technologies, and algorithms for the automatic bimodal/ multimodal recognition of human nonverbal behavior and affective states.

4.4 RGB to YCbCr Image Transformation

In our case, for all the tracked user images we need a solution that separates the user from his/her environment or background. To satisfy this need, we used the YCbCr solution, which is a way of representing colorimetric space in video. This basically results from the Hertzian transmission problems. All of the captured images by any device is the sum of its component colors. So, even in a black and white image, the signal Y (the luminance information) was created by the amount of the red, blue, and green present in the image. We sent Y, the luminance signal (black and white), plus two

chrominance pieces of information Cb (blue minus Y) and Cr (the red minus Y).The receiver can recreate the green and reproduce a colorimage. Indeed, if we have Y (red + green + blue) and Cb (Y-blue) and Cr (Y-red), we can mathematically recreate the green using the equation: Y = 0.3R + 0.6G + 0.1B. We applied a YCbCr transformation for all the images to extract the second component (Cb) of YCbCr (Fig. 1).

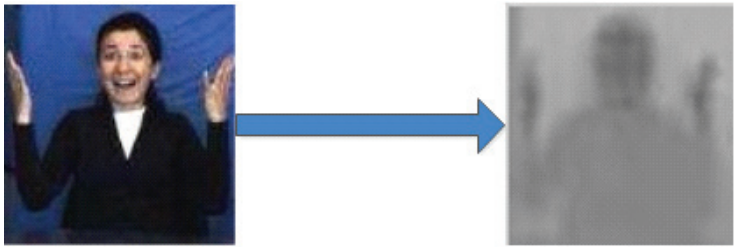

Fig. 1. RBG to YCbCr transformation.

4.5 Otsu Method

In computer vision and image processing, Otsu's method is used to perform automatic thresholding on the shape of the histogram of the image, or to transform an image to the grey levels into a binary image. The algorithm then assumes that the binary image contains only two classes of pixels, that is to say, the foreground and the background, and then calculates the optimal threshold between these two classes so that their intra-class variance is minimal.

Algorithm:

1. Calculate the histogram and the probability of each intensity level
2. Define w_i (0) and u_i (0) initial
3. Browse all possible thresholds t = 1 ... maximum intensity
   a. Update u_i and w_i
   b. Calculate O_b^2(t)
4. The desired threshold corresponds to O_b^2(t) maximum.

The weights W_i represent the probability of being in the ith class, each of which is separated by the threshold t, and U_i is the average class.

With Otsu segmentation, we computed the global threshold to obtain a binary image.

After that, we computed the complement of the obtained image, the object with the white color, and with the black, the background. Eventually we returned to the grey level image (Fig. 2).

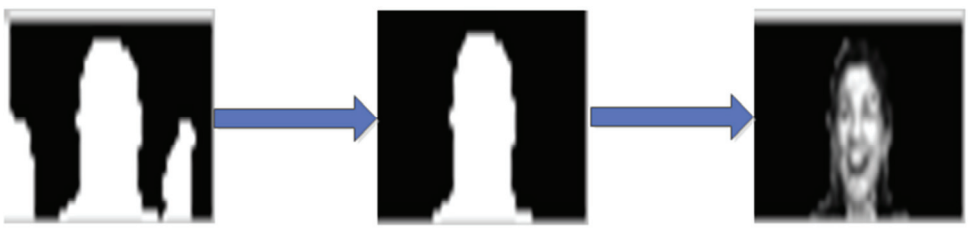

Fig. 2. Face detection.

5. Experiments

In this part of our paper we describe our framework for creating a video summarization with the emotional feedback issue from the viewers. To do so we relied on the FABO database described above. This database contains a great number of learning videos for our system.

To reduce the complexity of detecting emotions, we classified all of the training videos from the FABO database into two classes. The first class is defined by high intensity gestures and the second class involve slow intensity gestures. Therefore, we were only able to analyze half of the stored training videos.

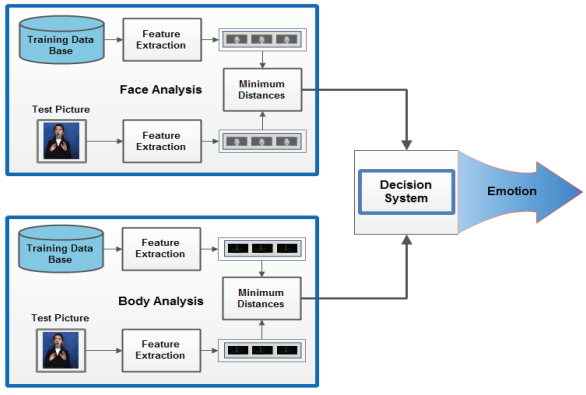

Fig. 3. Proposed architecture.

In our recent paper [1], we described the different ways to compute emotions from less to more credible. We found that body gestures are more credible then facial expressions. In fact, one of the most important dimensions for body gestures is the movement intensity. This is why we classified all of the training videos into the two classes described above. The computed movement intensity from a user defines the two classes of the video database based on the motion of all of the tracked body parts.

Our real-time emotion recognition system is structured in two parallel processes, as described in Fig. 3. One is for facial expressions and the other is for body movements, which can be detected from a user body parts tracking, using the FABO database. To this end, all of the detected frames from the viewer will be compared with the appropriate frames in our training database. The decision system will compare the two results issued from the two parallel processes to estimate the viewer's emotional state.

5.1 Body Emotion Recognition

In this section we presented our related results that were published in a previous work. We explain the way in which we computed the movement intensity for a user with the tracking of all the different points in the body skeletal provided by the Kinect software development kit. Collecting a body dynamic skeletal can reduce the complexity of our framework because the body skeletisation will be computed using the Kinect sensor in another parallel program. Our treatment was limited to the skeletisation of all the videos and the FABO database. As such, we applied the first steps as we proceeded with the facial expression recognition stage:

- YCbCr transformation – Cb component
- Otsu segmentation
- Complement of the binary image

The detection of junction points (the fingerprint recognition method and scanning the image 16 matrices) eliminates the junction points connected to the borders of the object (Fig. 4).

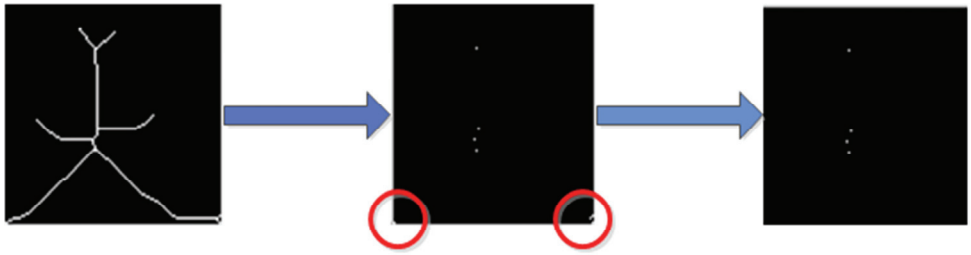

Fig. 4. Junction point detection.

We compare the features vector of the junction points with its correspondent in the training images using the K-NN classifier.

The proposed emotion recognition algorithm:

- Facial expression recognition: minimum distance
- Body emotion recognition: minimum distance
- Search for the images that have the same minimum distance for faces and bodies

5.2 Motion Attention Model

The value of the displacement vector field, which establishes the link between a pixel and another that is associated with two different frames, has a major importance for the interpretation of time-varying image sequences. The motion vector field can be used in a different way to derive the 3D motion from user or to calculate the structure of the movement. It can also be directly used for interpolation and for the noise reduction or compression of the image sequences.

The existing motion detection algorithms are based on vision techniques, such as low-level block matching; optical flow computation based on spatiotemporal gradients and Fourier methods; and high level techniques, such as image analysis, in order to extract the key features of objects, such as edges, limits, or complete objects, and to use them to solve the corresponding problems [23]. In other works, for a given frame in a video sequence, we extracted the motion field between the current frame and the next frame and calculated a set of motion characteristics.

In our case, we needed to compute the movement intensity between all of the detected frames to give more efficiency to our framework in different situations where the estimation of a user's emotion is not accurate and to eliminate all of the situations where the user can express his/her emotion without any significant intensity.

Motion intensity $\boldsymbol{I}$ is computed as the normalized magnitude of the motion vector:

$$\boldsymbol{I}(i,j) = \frac{\sqrt{dx^2_{i,j}+dy^2_{i,j}}}{MaxMag} \tag{5}$$

where $(dx^2_{i,j}, dy^2_{i,j})$ denote the two components of the motion vector, and *MaxMag* is the maximum magnitude in an MVF.

5.3 Summarization Phase

The summarized segment may contain only the important video shots from the film. The proposed system highlights the most important events during the film, for items such as goals and goal attempts, for a football game, for instance. It will do so in order to save the viewer's time and to introduce the technology for computer-based summarization into the video processing field. The proposed summarization method is based on the viewer's emotional feedback. It is not our aim to define the relation between body gestures, facial expressions, and emotions, because we relied on the results of the FABO database. Our aim is to define a multimodal video summarization framework based on the detected viewer's emotions. Our work has some laboratory limitations, in terms of the fact that we only worked with one user with a fixed background to compute the intensity of the movements.

6. Results

The presented results describe the system's performance in dealing with utterances, like anger, happiness, neutrality, or sadness, which are defined in the FABO database.

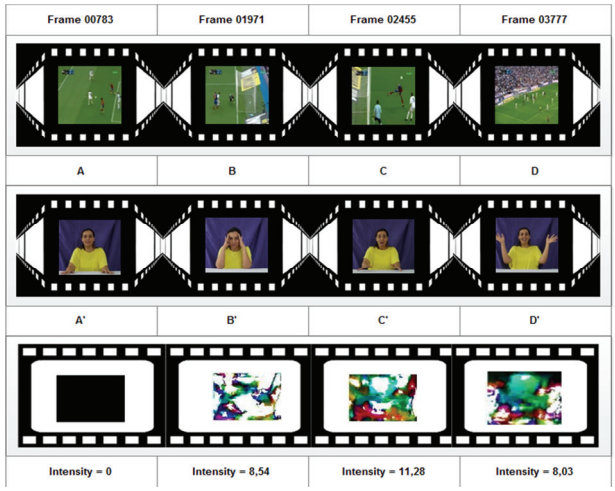

Fig. 5. The detected events.

The general results were computed with the K-NN classifiers where the fixed value of K is 3. This number is fixed due to different experiments having been conducted with different users. The decisions were made using two steps. The first step is the bimodal emotion detection for all of the detected frames and the second was the computed value of motion intensity that can provide the difference between two emotions, such as happy or very happy, according to high or low motion intensity. If no emotion from the four emotion classes was detected, the class was not defined. Fig. 5 presents the results of our system for a football video with the detected emotions and the values of the motion intensity for the different frames.

We conducted a test to determine the ability of our framework to detect all of the affective feedback that can be produced by the user, as presented in Fig. 6. We have a total of 57 cases of expressed emotion in 5 minutes of a manually summarized football video sequence. Table 1 presents the success rates of detecting the four basic emotions.

Our algorithm achieved a successful emotion detection in 49 cases with four false cases, which means a success rate of 86% for the affective feedback from a user watching the tested video.

Table 1. Table results (%)

	Neutral	Sad	Anger	Happy
Detected	91	65	83	68
Not detected	9	35	17	32

With the recognition of the user's emotions we have indexed the frames of the video, which allowed us to produce a summarized video that contains only the highlights. We defined a starting value of movement intensity so as to take into account all of the detected emotions. We did this because the user can be happy about an event in the video, thereby revealing a high intensity, and they can keep this emotion for a long period of time but will eventually shift to a lower or neutral intensity value.

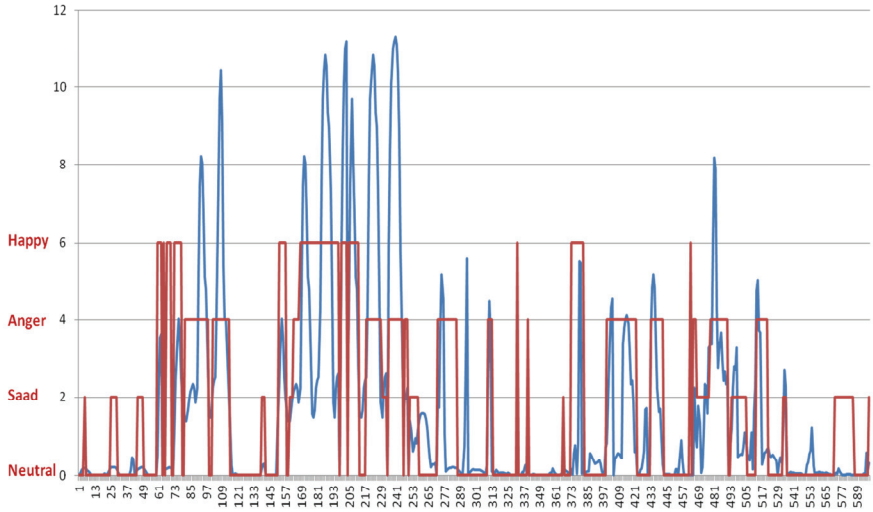

Fig. 6. Movements' intensity and recognized emotions curve (for 5 minutes and 2 frame/second).

Our system a video surveillance of viewer to detect all the social signals, so as to know, what the most important moment in the proposed film is.

The red curve is for the detected emotion and the blue one is for the movement intensity.

A video summarization scheme was developed in Fig. 5 based on the proposed user emotion recognition model described in Fig. 3. Fig. 5 shows a sequence of the detected emotions of a viewer of a football video (pictures A, B, C, and D). All of the detected frames from the viewer surveillance video have a corresponding frame in the original football video (frame 00783, frame 01971, frame 02455, and frame 03777). We have also computed the movement intensity presented (pictures A', B', C' and D'), which have the corresponding intensity of 'zero' for a neutral position with no movement and different

values for the other different motions. This allowed us to compute the intensity of the user's movements.

7. Discussion

To evaluate our results, we compared our proposed technique to some other techniques in the field. Zawbaa et al. [3] suggested a system based on all the events detected from the original video. They tried to detect the annotation on the screen as an example for a gradual logo appearance. The authors of the second work, Joho et al. [2] tried to detect all of the highlights from a video by relying on the detected emotions reflected on the viewer's face. In this case, their technique resulted neglecting the fact that a viewer can be happy for a long time over a video take play, but that this happiness is due to a specific event that occurs at a certain moment in the video. We present the results of our comparative study in Table 2 below.

Table 2. Technical comparison

	Archive	Depend on viewer	Highlight important moment	Eliminates repetition	Relation between emotions and important moment	Intensity compute
Event detection from video[3]	Yes	No	Not all	No	No	No
Highlight moment with facial emotion recognition [2]	No	Yes	Can exceeds the requested	May depend on viewer emotion	Yes	No
Our solution (based on viewer emotion and movement intensity)	Yes	Yes	Only the important for the viewer	May depend on viewer emotion and movement intensity	Yes	Yes

In Table 2, we present a comparison of our technique, which is based on bimodal emotion recognition, with two other techniques. We have introduced the dimension of the intensity of the movement in order to avoid the case that the viewers can express an emotion for a long period of time. Our technique can be applied to a real time video summarization and to video archiving, provided that we have a synchronized video focused on the viewer.

Our experimental comparison consists of two comparisons with two systems—the IM(S)^2 [21] and Parshin and Chen's system [24]. These systems are known as being the most conclusive works in the field of creating a video summarization based on the user's preferences. Since we were inspired by the method proposed by Ellouze et al. [21], we invited five users to test the system on a database composed of four different sports movies. The users were not familiar with the system, but they were regular sports video consumers. They belonged to different age categories and different backgrounds. The list of videos is presented in Table 3. The system was completely implemented by using MATLAB. The hardware platform was a PC with a 1.8 GHZ processor and 2 GB RAM.

Table 3. List of videos test

Video	Duration (min)
Football	30
Volleyball	20
Basketball	20
Handball	25

For our proposed solution we did not require any initialization action to start the video summarization, as compared to the IM(S)², in which the system displays an overview of the movie that is composed of its scenes to every user. Then, the user selected the shots that corresponded to his/her preferences. The system studied these preferences in order to generate a summary. Parshin and Chen's system is based on quantifying the preferences of the user based on some high level features, such as where the action took place, the time required for outdoor shots, the color of the human skin, the average amount of motions present the duration of the semantic segments in the shots.

Table 4. Experimental results and comparison

Video/criterion	Our system			IM(S)² [21]			Parshin and Chen [24]		
	RP	PT	LR	RP	PT	LR	RP	PT	LR
Football	4.8	4.9	4,7	4.7	4.1	4.8	2.7	3.9	4.2
Volleyball	3.9	4.8	4.6	4.3	3.6	4.2	3.4	2.3	3.2
Basketball	4.2	4.9	4.4	4.1	4.9	3.9	2.4	2.9	4.3
Handball	4.6	4.7	4.8	3.9	4.4	4.3	3.1	3.3	3.7
Average	4.375	4.825	4.625	4.25	4.25	4.3	2.9	3.1	3.85

RP=respecting the user's preferences, PT=pleasant tempo and easy to understand, LR=lack of redundancy.

In our system we paid a lot of attention to the quantity of a viewer's movements and the recognized affective feedback. We did so in order to generate the summary. The results of our system and those of Parshin and Chen's and IM(S)² are presented in Table 4 with the average values obtained for each criterion.

8. Conclusions and Futures Works

Natural human affective feedback expressions are a combination of different emotions. Using the automatic emotion recognition models, we can achieve a unique classification of the detected feedback into a specific emotional class.

In this paper, we proposed a solution to create an easier video summarization, based on the user's affective feedback, which we classified into four different classes (neutral, happy, sad, or angry) according to the motion intensity value. This additional dimension can be computed based on the user's movement and it's considered to be an important dimension.

Acknowledgement

The authors would like to acknowledge the financial support for this work that was made possible by grants from the General Direction of Scientific Research (DGRST) in Tunisia, under the ARUB

program. We would also like to acknowledge that some of the research in this paper was adapted from the Bimodal Face and Body Gesture (FABO) database for the Automatic Analysis of Human Nonverbal Affective Behavior, which was collected at the University of Technology, Sydney (UTS) by Gunes and Piccardi.

References

[1] M. Dammak, M. Ben Ammar, and A. M. Alimi, "Real-time analysis of non-verbal upper-body expressive gestures," in *Proceedings of International Conference on Multimedia Computing and Systems (ICMCS2012)*, Tangier, 2012, pp. 334-339.

[2] H. Joho, J. Staiano, N. Sebe, and J. M. Jose, "Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents," *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 505-523, 2011.

[3] H. M. Zawbaa, N. El-Bendary, A. E. Hassanien, and T. H. Kim, "Event detection based approach for soccer video summarization using machine learning," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 7, no. 2, pp. 63-80, 2012.

[4] D. Tjondronegoro, Y. P. P. Chen, and B. Pham, "Sports video summarization using highlights and play-breaks," in *Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval* (MIR2003), Berkeley, CA, 2003, pp. 201-208.

[5] Y. Qi, A. Hauptmann, and T. Liu, "Supervised classification for video shot segmentation," in *Proceedings of the International Conference on Multimedia and Expo (ICME2003),* Baltimore, MD*, 2003,* pp. 689-692.

[6] D. Tjondronegoro, Y. P. P. Chen, and Pham, "Integrating highlights for more complete sports video summarization," *IEEE Multimedia*, vol. 11, no. 4, pp. 22-37, 2004.

[7] M. G. Christel, M. A. Smith, C. R. Taylor, and D. B. Winkler, "Evolving video skims into useful multimedia abstractions," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Los Angeles, CA, 1998, pp. 171-178.

[8] Y. Takahashi, N. Nitta, and N. Babaguchi, "Video summarization for large sports video archives," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME2005)*, Amsterdam, 2005, pp. 1170-1173.

[9] A. Hanjalic, "Generic approach to highlights extraction from a sport video," in *Proceedings of the International Conference on Image Processing (ICIP 2003)*, Barcelona, Spain, 2003, pp. 1-4.

[10] M. A. Refaey, W. Abd-Almageed, and L. S. Davis, "A logic framework for sports video summarization using text-based semantic annotation," in *Proceedings of the 3rd International Workshop on Semantic Media Adaptation and Personalization (SMAP2008),* Prague, 2008, pp. 69-75.

[11] F. Coldefy and P. Bouthemy, "Unsupervised soccer video abstraction based on pitch, dominant color and camera motion analysis," in *Proceedings of the 12th Annual ACM International Conference on Multimedia*, New York, NY, 2004, pp. 268-271.

[12] G. H. Bower, "How might emotions affect learning," in *The Handbook of Emotion and Memory: Research and Theory*, S. Christianson, Ed. Hillsdale, NJ: Erlbaum, 1992, pp. 3-31.

[13] T. D. Bui, "Creating emotions and facial expressions for embodied agents," Ph.D. dissertation, University of Twente, Enschede, The Netherlands, 2004.

[14] E. A. R. Tanguy, "Emotions: the art of communication applied to virtual actors," Ph.D. dissertation, University of Bath, England, 2006.

[15] M. Dammak, M. Ben Ammar, and A. M. Alimi, "A new approach to emotion recognition," in *Proceedings of the International Conference on Innovations in Information Technology (IIT)*, Abu Dhabi, 2011, pp. 110-113.

[16] P. Ekman and W. W. Friesen, *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[17] M. Coulson, "Attributing emotion to static body postures: recognition accuracy, confusions, and viewpoint dependence," *Journal of Nonverbal Behavior*, vol. 28, no. 2, pp. 117-139, 2004.

[18] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias, "Emotion analysis in man-machine interaction systems," in *Machine Learning for Multimodal Interaction.* Heidelberg: Springer, 2005, pp. 318-328.

[19] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," *Image and Vision Computing*, vol. 24, no. 6, pp. 615-625, 2006.

[20] H. Gunes and M. Piccardi, "A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006)*, Hong Kong, 2006, pp. 1148-1153.

[21] M. Ellouze, N. Boujemaa, and A. M. Alimi, "IM(S)^2: Interactive movie summarization system," *Journal of Visual Communication and Image Representation*, vol. 21, no. 4, pp. 283-294, 2010.

[22] M. A. Turk, and A. P. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-86, 1991.

[23] R. Larsen, "Estimation of motion vector fields," in *Proceedings of the 2nd Danish Conference on Pattern Recognition and Image Analysis (DANKOMB, DSAGM yearly meeting),* 1993, pp. 37-42.

[24] V. Parshin and L. Chen, "Video summarization based on user-defined constraints and preferences," in *Computer-Assisted Information Retrieval (Recherched'Informationetses Applications [RIAO])*, Avignon, France, 2004, pp. 18-24.

Majdi Dammak http://orcid.org/0000-0002-1709-8433

He received the Master degrees in Computer Science, Engineering Information & Systems from Rouen Univ. in 2005 and 2006, respectively. During 2006, he stayed in Biology Workshop, Informatics, Statistics and Sociolinguistics in the same university in France. To develop a comparative solution of the oracle factor construction. And now he is undertaking a doctorate course as a member of the REGIM-Lab. is a Research Laboratory in Intelligent Machines located in the National Engineering School of Sfax, University of Sfax. His research interests include Image and Video Processing, Emotion Recognition, Video Summarization.

Ali Wali http://orcid.org/0000-0002-8423-7923

Assistant Professor on Computer Sciences at ISIM, University of Sfax. Got his Ph.D. in Engineering Computer Systems at National school of Engineers of Sfax, in 2013. He is member of the REsearch Groups on Intelligent Machines (REGIM). His research interests include Computer Vision and Image and video analysis. These research activities are centered around Video Events Detection and Pattern Recognition.

Adel M. Alimi http://orcid.org/0000-0002-0642-3384

He graduated in Electrical Engineering in 1990. He obtained a Ph.D. and then an HDR both in Electrical & Computer Engineering in 1995 and 2000, respectively. He is full Professor in Electrical Engineering at the University of Sfax, ENIS, since 2006.