KS-FQA: Keyframe selection based on face quality assessment for efficient face recognition in video

Sahbi Bahroun | Rahma Abed | Ezzeddine Zagrouba

Institut Supérieur d’Informatique, Research Team on Intelligent Systems in Imaging and Artificial Vision (SIIVA), LR16ES06 Laboratoire de recherche en Informatique, Modélisation et Traitement de l’Information et de la Connaissance (LIMTIC), Université de Tunis El Manar, 2 Rue Bayrouni, Ariana, Tunisia

Correspondence
Sahbi Bahroun, Institut Supérieur d’Informatique, Research Team on Intelligent Systems in Imaging and Artificial Vision (SIIVA), LR16ES06 Laboratoire de recherche en Informatique, Modélisation et Traitement de l’Information et de la Connaissance (LIMTIC), Université de Tunis El Manar, 2 Rue Bayrouni, 2080 Ariana, Tunisia.
Email: Sahbi.bahroun@isi.utm.tn

Abstract
Video is considered as one of the most useful and important forms of multimedia data, that is usually used in several applications. Despite its importance, video indexing and retrieval becomes a challenging task. In order to reduce the amount of data and keep only relevant frames, keyframe extraction becomes necessary in a content-based video retrieval (CBVR) system. In this paper, a keyframe extraction method is proposed based on the face image quality for video surveillance systems. Data is reduced by rejecting frames without faces. Then, face images are clustered by identity. After that, a set of candidate frames is selected to be proceeded. The face quality assessment is based on four metrics including pose estimation, sharpness, brightness and resolution, and the frame with the best face quality is considered as a keyframe. Experimental tests were carried on several datasets in order to prove the efficiency of authors’ method compared with state-of-the-art approaches.

1 | INTRODUCTION

Nowadays, biometrics technology, such as hand geometry, fingerprint, iris scan and face recognition are being more and more important and studied by the research community [1]. However, the most used biometric identifier is face recognition. Taking the case of a surveillance system, there is no need to ask people to place their hand or eyes in a reader (fingerprinting or iris recognition), face recognition systems take pictures of people’s faces when they enter a specified area. So that people do not feel under surveillance or their privacy being invaded [2].

Face is widely used for several applications: identification purpose, or even to distinguish other characteristics of the person such as their age, sex, ethnicity or his emotional state. That’s why, face is usually considered as the most significant biometric identifier for human recognition systems.

Video-based face recognition have rapidly overcome the image-based methods, with the arrival of inexpensive video cameras and the high processing power [3]. This is why facial recognition based on video has attracted the attention of many researchers in recent years.

In controlled environment, face recognition achieve a high accuracy rates. Unlike crowded environment, this task remains a big challenge due to: head pose variation, illumination conditions, facial expression, occlusion caused by other objects or accessories (e.g. sunglasses, scarf etc.), resolution and blurring caused by the movement of people in front of the cameras.

On the one hand, face recognition in videos is more interesting rather than from only one single-shot still image. On the second hand, the process of this huge amount of data for each video is challenging, due to the time needed to deal with all frames. In addition, faces in these frames can be either redundant or useless due to their poor quality.

To solve the problem of the huge amount of data frames in a video, we need to select some frames, face conditions are adequate for face recognition tasks. For example, in [4] the authors produce a video summarisation method based on a Convolution Recurrent Neural Network in order to extract and exploit the spatiotemporal structure of video data and achieve better performance. In this work, we will mainly interest in keyframe extraction for two reasons: first, keyframe extraction process is simple [4]. Second, keyframe extraction do not depend on the ordering of face images. By this way we are not interested on the ordering of face images, only the quality of face in these frames is considered. In [5], the authors demonstrate that extracting frames with the good face quality improve the face recognition accuracy.
Our main contribution is to define a new method for keyframe extraction based on face quality assessment. In which, we present a novel method for facial pose estimation, based on the analysis of the landmarks geometrical distribution in the face region. We have opted to use the pose metric instead of using the face symmetry, which cannot detect pitch rotation.

The proposed keyframe extraction method is based on face quality assessment using several metrics including the pose, resolution, sharpness, and brightness. Despite their importance, estimating face quality using metrics has been rarely performed for keyframe extraction from videos.

Our main goal is to integrate this module into a CBVR system. In the offline phase, and as a video pre-processing task, we extract keyframes that describe well the faces existing in the videos database. The extracted faces form a much smaller dataset than the original one. In the online phase, the system will compare the face image request with all faces in the new face dataset and returns videos in which, the requested face appears.

The rest of this paper is organised as follows. In Section 2, we introduce an overview of related works in keyframe extraction. We will describe the proposed keyframe extraction method in Section 3. Section 4 presents and analyses the experimental results obtained. Finally, Section 5 concludes the paper and opens some perspectives for future work.

2 | RELATED WORKS

Content-based video retrieval (CBVR) system has generally three phases, shown in Figure 1. Firstly, let us take a look at the offline stage, starting with the video processing step. The system extracts the pertinent elements to describe video. This description can be made by extracting a set of keyframes, shots and for scenes. Video shot is defined as a continuous set of frames having the same content. The scene is a set of shots grouped according to a well-defined criteria (similar content, movement of characters...). The last element, keyframes are the video images that describe the content of the video by eliminating any redundancy. The set of keyframes is called video summarisation. In the second step, in the offline stage, and based on the extracted elements in the first step, the indexing phase aims to extract the relevant objects to well describe the video. Those objects may be an image, particular object, motion, text, or an important event. The extracted elements will be indexed in a compact and relevant signatures and stored in a signature feature space.

Secondly, and in the online stage, the system has to calculate similarity between the query image features and each feature vector in the signature feature space. The goal is to return the most relevant videos to the image query.

Video abstraction or video summarisation is an important technique in a video processing application. The aim of video summarisation is to provide a short and representative summary for the original video [6]. Two main categories of video abstraction: static and dynamic video summarisation [7].

The static summary is a set of the most representative frames of the video. This process is also known as keyframe extraction [7]. The static summaries are considered as the simplest way to provide a video summary. Whereas, selecting these keyframes is not an easy task, specially for surveillance videos. This difficulty is due to the importance of the video content [6]. On the other hand, dynamic summaries, or video skims, is the set of the most representative video segment of the video itself. Watching a video skim is more interesting rather watching a set of keyframes, because the video skim keep audio and text information. But it is a complex task and requires different modules.
for handling all type of information required to construct the summary [6].

Since only the visual information is used for generating the static summary, this category of video summary is more simple, faster and suitable for surveillance video indexing and retrieval. Keyframes selection can be based on particular objects in the frames (movement, background, faces, colours, ...).

Most of the keyframe selection methods are global-feature-based. Those methods take into account elements that present the frame in a global manner like colour, texture, image histogram etc. On the other hand, local-feature-based methods use interest points or interest regions to generate the summary, which allows us to focus on specific objects in frames which gives us keyframes according to a particular object.

Chergui et al. [8] select one keyframe to represent each shot based on interest points. They consider that the frame having the highest number of interest points is the most informative and representative one. Gharbi et al. [9] used interest points and repeatability measurement to extract keyframes. They detect interest points in all frames (divided into shots). Then calculate a repeatability matrix for each shot. In the last stage, they form an oriented graph based on the matrix obtained. Keyframes were selected using shortest path algorithm.

In recent work [10], the authors proposed a keyframe selection method using graph modularity clustering. They generate a candidate set frames (CS) including the first frame of each shot. The other frames are chosen based on a windowing rule with a size equal to FPS (frame per second). The aim of generating these CS is to reduce the data to be proceeded.

Thanks to their ability to provide robust descriptors face several transformations (rotation, viewpoint changes...). Interest points have been successfully used in image retrieval applications [11]. Despite their importance, keyframe extraction methods based on interest point are not suitable for face recognition systems. First, those approaches do not consider the challenges to be faced to deal with face images such as the illumination’s conditions, the pose variation, facial expressions, occlusion, distance between face and camera etc. Second, using the interest points, we cannot guarantee that we will extract the most neutral face image. Interest point is not suitable to avoid the problems that can be found in video sequences such as low-resolution frames, motion blur effect, unbalanced illumination.

For these reasons, we focus on this work on keyframe extraction based on face quality assessment. The authors in [12] indicate that using face quality in video-based face identification system may improve its performance. Moreover, the face quality score is useful in image-based face recognition.

The term face quality assessment (FQA) was introduced for the first time by Griffin [13]. In which, the quality of a face image is evaluated based on the face geometry including the pose and resolution of face region, the confidence of detected eyes, illumination, facial expressions etc.

Various works related to the FAQ exist in the literature, using several face metrics. Thus, these methods follow the same strategy shown in Figure 2. Starting by detecting faces, then evaluating the face quality using different metrics. They use several metrics fused together to have a quality score. The main difference between these methods is in the choice of metrics used and the formula used to combine them into a quality score.

Fournery et al. [14] used two metrics to estimate the brightness measurement, combined with head pose, sharpness, presence of human skin and resolution in order to estimate face quality. Nasrollahi et al. [15] proposed a keyframe extraction method based on face quality assessment. The face quality assessment module is based on the use of the head pose, brightness, sharpness and face resolution. In addition, they multiplied these four metrics by different weights and combine the results together to get a final quality score for the corresponding face image. More recently, Nasrollahi et al [16] propose another keyframe extraction system based on face quality by the use of ten metrics. Pose estimation using two metrics to estimate the head pan and tilt rotation, sharpness, brightness, resolution, openness of the two eyes, their direction (gaze) and closeness of the mouth. The authors used a multilayer perceptron (MLP) to define weights. This MLP is composed of three layers having ten neurons in the input layer, where each neuron corresponds to one of the mentioned features, four neurons in the hidden layer and one neuron in the output layer to provide the quality score of the input face image. The MLP was training using 400 face images of 40 identity. Anantharajah et al. [17] proposed a quality-based frame extraction method in order to cluster faces in news video. They used face quality to select the adequate faces images. In their work, the calculation of the face quality use four metrics which are face symmetry, sharpness, contrast and brightness. Qi et al. [18] adopted symmetry, sharpness, brightness and resolution to estimate face quality. With the use of GPU acceleration to achieve a higher computing performance.

It is worth mentioning that most of the methods mentioned above use the weighting system proposed in [15]. This system has been experimentally deduced, and the weight reflects the importance of each metric. This static weights do not take into account the variations of face images. A frontal face image may be dark or with unbalanced light. So, most of the details of the face will be hidden. On the other hand, a face image with acceptable slice variation, resolution, and brightness can provide more detail. In addition, a frontal face image with low resolution or face image having high resolution with unbalanced brightness will be useless. That is why the face quality must be considered in a relative manner. In other words, by taking into account all the metrics used with the same priority.

Among all the features used in the mentioned works above, the most important one is the pose measurement. The head

**FIGURE 2**  Face quality assessment method's flowchart
pose estimation is defined as the process of detecting the orientation of the head compared to a camera. This measurement leads to evaluate how much this face is frontal. Because wide variation in pose hide most of the useful features of the face. In addition, the frontal images are more used in facial analysis systems than the rotated ones. Usually, in a video sequence, people move and look into different directions, which changes their head position. Therefore, it is important to involve the pose variation in the process of quality assessment.

The existing methods for head-pose estimation are divided into two categories [16]: local and global methods. Local methods use face components like eyebrow, eyes and lips to estimate the head-pose. But, in low-resolution images, the detection of these components is quite difficult. Global method use the whole face image to estimate the pose. The use of the global methods allows to avoid the problem of low-resolution images. In addition, we do not need to detect those small elements in a such faces, it is enough to detect the face region.

Fourney et al. [14] use three columns for the face image: columns L and R to approximate the left and right face edges. The third column C define the face's natural axis of symmetry. If the face is frontal, then C is equidistant from L and R. Otherwise, the pose score is calculated by measuring this deviation between C, L and R.

Nastrollahi et al. [15] calculate the difference between the centre of mass and the centre of region for each detected face. Whenever the rotation of the face increases the difference between these two points increases too. Therefore, they consider the face as whole and calculate the pose estimation metric using a binary image of the face region.

In some researches, the face quality is done for all video frames like in [15, 18]. dealign with all frames requires much time and memory. In addition, it may generate redundancy since the successive frame can have almost the same content. In order to exploit the 2D structure of the image, Yang et al. [19] proposed a nuclear-norm based on bilateral two-dimensional matrix regression preserving embedding (NN-MRPE) for face recognition and consider the data reduction issue as a problem of image size reduction. Others use different techniques and aims to reduce the number of frames to be treated. Gharbi et al. [9], apply first a shot detection. After that, they detect the interest point in all frames and calculate the repeatability matrix for each shot to extract the most representative frames. In [10], the authors used a windowing rule which consists of selecting one frame for each FPS. In other word, one frame per second. The authors consider that in 1 s there is no significant changes in the content of successive frames.

Despite of their simplicity, this method is not well suitable for a keyframe extraction based on face images. Using time as a selection criteria by choosing one frame per second does not guarantee that we will obtain the best face images. That is why we must choose a method that takes into account the face state: frontal, neutral, good quality etc.

In [17], authors used keyframe selection for a face clustering applications. Starting with selecting the highest quality frames to reduce the amount of data, the cluster perform with an optimal number of face images in a video.

Although several studies have used tracking to group faces into sets [17, 18] little attention has been given to face clustering to do it. Face tracking locate a target object (faces) in a video over time. While face clustering consists on grouping faces in video based on identity. Clustering face images by identity has two major applications [20]: grouping a set of unlabelled face images, and indexing large sets of face data, for more efficient search.

Face clustering using one single deep model has not been widely studied in the literature [21]. The authors attention was concentrated in clustering based on a distance between face representations. They have addressed not only face representations models, but also distance measurement. Face clustering process is evaluated based on an accuracy measurement named Pairwise F-measure. Which is based on the standard f-measure calculation but, with a slight difference in the definition of precision and recall [22]. Given a pair of images of two faces: the first one represent the user request image and the second is a face image from the face image dataset. Pairwise precision is defined as the fraction of the corrected clustered pairs over the total number of pairs of the same class. Pairwise recall is defined as the fraction of pairs correctly grouped over the total number of pairs in the same cluster.

Two major problems were faced in a clustering applications: face representation and the similarity measurement for grouping faces. Based on a face image representation and a similarity measurement, the aim of face clustering is to group these unlabelled face images based on their representation into small subsets indicating their identity. Figure 3 illustrates the face clustering process.

Recently, face representation has become more robust with the use of deep representation [23]. The most used one is the FaceNet system [24]. FaceNet is based on a deep convolutional neural network (DCNN) to extract robust features from face image. This model is trained on a triplet loss function in order to involve matching and non-matching face pairs. The learned features have been observed to be robust to several face images problems (pose variation, lighting conditions, resolution). Several researches on similarity measurement and clustering algorithms have been done. Agglomerative hierarchical start with considering each sample as a cluster. Then merge the nearest samples into the same class. Two samples are considered neighbours if they satisfies, in each iteration the distance or similarity criteria defined. More recently, the authors in [25] have proposed a rank-order distance. Based on the fact that faces having the same identity share usually their top neighbours. So, for each face, they generate a ranking order list based on sorting other faces. Then, the rank-order distance of two faces is calculated using their ranking orders.

In this study, we propose a new method to estimate the pose variation based on the method proposed in [15]. Considering the pose as the difference between the centre of mass and the centre of region, we define the centre of mass by analysing the landmark's geometrical distribution in the face region. Then, we define a new method for keyframe extraction based on face quality assessment. We select a set of candidate frame based on another criteria instead of using time criteria. Our method use
four metrics including pose, brightness, sharpness and resolution. Differing from [17, 18], we use the pose instead of using the symmetry measurement, because of the left–right difference between an image and its corresponding mirror version presents some limitations. In the one hand, it cannot detect pitch rotation variations. In the other hand, it can greatly vary due to nuisance factors such as non-frontal illumination conditions. Moreover, we use a face clustering module in order to group face images by identity rather than using a face tracker module like in [17, 18]. More details will be shown in the following section.

3 | PROPOSED METHOD

Considering the video as a set of consecutive frames, denoted as:

\[ V = \{ f_1, f_2, f_3, \ldots, f_n \} \]

where \( f_i \) is the \( i^{th} \) frame of the video and \( n \) is the number of frames. The purpose of video summarisation is to select a subset \( r \) from \( V \), having a shorter length and contains the most important frames in \( V \) [4]. The purpose of the proposed keyframe extraction method is to select a subset from \( V \) in which, the important elements to be selected are faces. In other words, the subset \( r \) will be composed of frames presenting faces of all identities presented in \( V \).

We developed a keyframe extraction method namely ‘KS-FQA’ for keyframe selection based on face quality assessment. The flowchart of the proposed method is presented in Figure 4.

First, we use a face detection method to detect faces from all the video frames. The aim of this step is to detect, find the face region in each frame of the video sequence and reject frames without faces. Then, we group the remaining frames by identities using a clustering method. After that, for each frames set of the same identity, we select a candidate set of frames. Finally, we calculate the face quality assessment (FQA) for each face image. The face image having the higher face quality is considered as a keyframe that represent well the corresponding identity.

3.1 | Face detection

In order to ensure that the selection of keyframes will be based only on the facial regions, we perform face detection in the first step. If we perform keyframe extraction in the first step,
the selection criteria and metrics used will be calculated on the whole image. Which provide erroneous results for face recognition [11]. For these reasons, we start with face detection, to make sure that the calculation of the face quality will consider only face regions.

We use the multi-task cascaded convolutional networks (MTCNN) for face detection and alignment [26]. This detector provides five landmarks, shown in Figure 5 (the two eyes, the two corners of the mouth and the nose), a confidence score that reflects the probability that the object detected is a face, and limits face boundaries in a bounding box.

We use this face detector for many reasons. First, the MTCNN allows to eliminate useless frames. The rejected frames are those in which the detector cannot detect landmarks. Like faces with a huge pose variation or an unbalanced brightness distribution, which may hides most of the details of the faces. Second, using the confidence score, we will select a set of candidate frames to be proceeded. Finally, based on the five landmarks provided by the MTCNN and their coordinates, we will calculate the head pose by a simple and efficient manner.

Figure 6 present some rejected frames by the MTCNN.

3.2 | Face clustering

To cluster faces by identity, we follow this steps: having the detecting faces in the first step, we create the face representation (embeddings) using the faceNet model. Creating a 128 dimensional vector for each face. Then, and based on this embeddings, we cluster faces using the rank-order clustering algorithm. In this kind of an agglomerative clustering technique, the embeddings are merged based on their rank-order distance.

3.3 | Candidate frame selection

In order to reduce the number of frame to be proceeded, we use the confidence score provided by the MTCNN as a criteria to select candidate frames to be keyframe. The face images having the maximum confidence score are not only the most frontal ones. Other factors are considered such as facial expressions, brightness, resolution etc.

We start by sorting the faces images based on the confidence score. To form the candidate frame set, we will choose the first five images of each faces. In other word, the five images having the face best confidence scores.

In Figure 7, we show an example of a full sequence (Figure 7(a)) and the selected five frames (Figure 7(b)).
3.4 Face quality assessment

We perform the face quality assessment (FQA) for each candidate frame. In this work, we use four metrics to estimate the face image quality: head pose, sharpness, brightness and resolution. The face image having the best quality will be used as keyframe of the corresponding face.

We choose these metrics because of their importance to estimate the face quality. First, the most frontal face is the most adequate to represent an identity. Second, and because of the movement of peoples in front of the camera, the images may be affected by noises, which leads to a low-quality face images. The third metric chosen is the brightness measurement since the unbalance quantity of bright in image can hide the face details in any position. Finally, in a low-resolution image, we cannot visualise face components. For these reasons, we assess the face quality based on these four metric. The following subsections describe the details of these metrics and their calculation methods.

3.4.1 Pose estimation score

The pose variation is the most important feature to evaluate the quality of a face. Indeed, wide variation in pose may hide most of the face details. We will follow the same strategy as [15], defining the pose as the difference between the centre of mass and the centre of region. Whenever the face rotation increases, the difference between these two points increases too.

The centre of region for each detected face is calculated using Equation (2):

\[ x_r = \frac{x_2 - x_1}{2}, \quad y_r = \frac{y_2 - y_1}{2}, \]

where \((x_1, y_1)\) are the coordinates of the top left point and \((x_2, y_2)\) are the coordinates of the right bottom point in the face region as shown in Figure 8.

The centre of mass is considered as a central point in the face. In this work, we consider that this point is the nose, defined as \((x_m, y_m)\). If the face is frontal, the coordinates of the nose are close to the centre of the region like in Figure 8(a). But when we notice a head pose variation, the distance between these two points increases too (Figure 8(b)).

Finally, we calculate the distance between the centre of mass and the centre of region for the detected faces. We calculate the pose score (PS) using the Equation (3).

\[ PS = \frac{1}{1 + \sqrt{(x_r - x_m)^2 + (y_r - y_m)^2}}. \]  

In fact, PS is the reciprocal (multiplicative inverse) of the calculated distance (to be homogeneous with the other metrics). Moreover, we added one to avoid division by zero.

Using this components, we can successfully detect the two major rotation forms by a human face: Yaw and Pitch. Which are not well detectable using a symmetry measurement. Table 1 illustrates two example of face rotation and the two detected points necessary for the pose estimation.

We distinguish the centre of the region by the symbol * and centre of mass by +.

For the yaw rotation and for a frontal face, the centre of mass and centre of region are very close. In the two side views, we note the difference between these two points increase as much as the face rotation increases too. For the pitch rotation, and even with frontal faces, our method is capable to detect a huge variation of pose in the two direction (up or down).

3.4.2 Sharpness score

People in video are often moving in front of the camera. This movement cause a blur effect on face images. Using this feature to estimate the quality of faces is useful, in order to select the most neutral face images rather than blurred images. To compute the sharpness score (SS), we start by blurring the face image by a Gaussian operator. Then, the SS is defined as the absolute difference between the original image and the blurred one, divided by the size of the face image.

\[ SS = \frac{\sum_{i=1}^{W} \sum_{j=1}^{H} |I - G(I)|}{W \cdot H}, \]

where \(I\) is the original image, \(G(I)\) is the image after using Gaussian blur. The kernel of \(G(I)\) is a five-order two dimensional Gaussian distribution matrix, with standard deviation of 1.0 in X and Y directions. \(W, H\) are the width and height of the face image region, respectively.

3.4.3 Resolution score

The resolution score (RS) is calculated by multiplying the height \(H\) and the width \(W\) of face image:

\[ RS = W \cdot H. \]
TABLE 1  Pose estimation against face rotation

| Face Rotation | Illustration | Example |
|---------------|--------------|---------|
| Yaw rotation  | ![Illustration](image1) | ![Example](image2) |
| Pitch rotation| ![Illustration](image3) | ![Example](image4) |

3.4.4  Brightness score

Face components are not well visible in dark images. That is why we need to find the face image having the best illumination distribution.

The brightness score (BS) of the face image, described by Equation (6), is equal to the summation of all pixels value divided by the image area.

$$BS = \frac{\sum_{i=1}^{W} \sum_{j=1}^{H} B_{i,j}}{W \ast H},$$

where $H$ is the height, $W$ is the width of the detected face image region and $B_{i,j}$ is the gray value of a pixel at the coordinates $i, j$. The value range of each pixel $B \in [0, 255]$.

3.5  Face quality score

After calculating the four metrics for each face image in a given sequence, the extracted features are combined into a single quality value.

First, each feature is normalised according to the maximum value of this feature for the given sequence. Second, each normalised feature will be multiplied by an associated weight referring their importance. Then, we combine all the features scores into a general face quality value. The expression of the face quality score ($Q$) is shown in Equation (7).

$$Q = \sum_{i=1}^{N} \frac{S_i}{S_{i,\text{max}}},$$

where $S$ is the score for the $i$th face metric, and $S_{i,\text{max}}$ is the maximum score value of the $i$th metric in a given sequence. In this work, $N = 4$ represents the number of metric used. The face image that has the best quality is the one that has the highest score.

4  EXPERIMENTAL RESULTS

This section is organised as follows: We start by giving an overview for the used datasets to test our system. Several datasets are used in order to face all challenges presenting in reals sequences (camera motion, occlusion, pose variations, illuminations conditions). In the second part, we will focus on the process of candidate frame selection. The third part will be devoted to prove the utility of the face detector used by evaluating the face detector used against others facial detectors usually used in the literature to detect faces in a crowded environment.

Before the evaluation of our KS-FQA method, we present the evaluation protocol and the metrics used. In the last part, we will focus on a qualitative and quantitative evaluation of the keyframe extraction method. More details are presenting in the following paragraphs.

4.1  Datasets

In order to face almost of the difficulties that can be found in real environments, we used several datasets. The first dataset is ATT [27]. This dataset contains 40 sequences from different person (10 images per person). The images are in a dark homogeneous environment with small variation of pose and wide variations in facial expressions. The second dataset is the FRI CVL dataset [28]. It contains 144 sequences of different persons. Each one have seven colour images. These images have been taken in different head rotations and facial expressions, having a good brightness distribution and resolution. This
FIGURE 9 Candidate frame set selection

dataset is changing widely in head rotation and keep the distance between the people and the camera fixed.

The last dataset is the YouTube Face Dataset [29]. This dataset contains 3,425 videos of 1595 different people. All the videos were downloaded from YouTube. An average of 2.15 videos are available for each subject.

4.2 Candidate set selection

We will extract a candidate frame set following three different scenarios:

1. Using the confidence score to select the best five frames (Figure 9(a)).
2. Choosing one frame per second for the original sequence (Figure 9(b)).
3. Eliminate frames in which no faces are detected, then select a frame per second (Figure 9(c)).

The main advantage of choosing five candidate frames for each identity, is to have the same number of each face image group to be processed, which guarantees a similar processing time for all the sequences. Moreover, by using the confidence score, the criteria of frame selection does not depend on the frame appearance in the sequence (chronological order), but only on its confidence value, which reflects the probability that the object detected is a face.

Using the FPS as a selection criteria provide frames that appears in the position n, 2n, 3n etc. (with n = FPS). That means we will have more image candidate as much as we have longer sequences. Even if we perform face detection before the selection, it reduce the frame number and as a consequence it reduce the number of candidate frames. But it is still depending on the sequence length (scenario 2) or with the number of frame having a face (scenario 3).

4.3 Utility of the face detector

In this test, we will compare the MTCNN detector against two other detectors frequently used in the literature: Dlib [30] and Viola-Jones [31].

The Viola-Jones face detector is one of the most used detectors for face detection in real-world applications. Most of today’s digital cameras use Viola-Jones.

Dlib is a library written in C++, rich in algorithm and tools that can solve complex classification problems using machine learning techniques. Among the various features offered, we are interested in methods of detecting objects or especially faces. Dlib’s facial detection algorithm is based on the histogram of oriented gradient (HoG).

The description and the test result are presented in the Table 2.

All these detectors works well in controlled environment. In other cases (crowded environment), the performance of these detectors decreases. The Viola-Jones detector detect seven faces among the 11 faces that appear in the test image. Dlib detects only five faces. On the other hand, the MTCNN detects nine faces. Although, these two face detectors are the most used in the literature, their performance remains limited in the case of a crowded environment.

The Dlib provides 68 landmarks that define face component. These landmarks are useful for face region analysis. But, this detector is limited to frontal images and super-resolution images. Moreover, in such environments, the appearance of non-frontal images or low-resolution ones is frequent.

Also, for the Viola-Jones detector, the detection of the two eyes and the region of the mouth helps to analyse the facial region. But it still have a low capacity to detect the majority of the faces in an environment encumbered, and this remains a major disadvantage. For all these argument, the MTCNN still the most suitable for real scenarios.

4.4 Evaluation protocol

The evaluation protocol used is based on subjective criteria (quality) and objectives (quantity).

Subjective evaluation consists of comparing our keyframe extraction method with the state-of-the-art methods and an expert ranking as a truth-ground provided by [33].
TABLE 2  Comparative study of face detectors

| Face Detector | Description | Example |
|---------------|-------------|---------|
| Viola-Jones [31] | The viola Jones face detector is widely used for Real-time application. The use of Haar features and AdaBoost make the training slower, but testing very fast [32]. Based on the following images, we note that this detector detects faces with different resolutions. Moreover, it locate the eyes and the mouth region. Which allows us to work with some facial components. Nevertheless, it detect only frontal faces. | ![Image](image1.jpg) |
| Dlib [30] | The Dlib present a higher performance on frontal or slightly non frontal faces. But it is sensitive to wide variation of pose and low-resolution images (the minimum size required is 80 × 80). Dlib can detect small occlusion. Besides that, Dlib detect 68 facial landmarks, that help to analyse face very well. Using this landmarks, we may determine the emotional state. | ![Image](image2.jpg) |
| MTCNN [26] | The MTCNN detect five landmarks with their coordinates: the eyes, nose and the two corner of the mouth. In addition, it detect faces having a small occlusion and small resolution. The drawback of MTCNN is the detection of the side-view faces. | ![Image](image3.jpg) |

The objective evaluation is done using our method in a face recognition system as well as in a content based video retrieval using faces.

For face recognition, the evaluation is performed by comparing the results of the accuracy rates obtained by a similar method in a first step. In other word, method that use keyframe extraction based on face quality assessment in their face recognition systems. Then, we compare our method against the most recent face recognition systems based on the deep learning methods. The second test is to evaluate our method in a content based video retrieval based on face images.

To evaluate the performance of our system, we use the following metrics:

- **Accuracy**: The accuracy define the total number of correct predictions returned within the face recognition compared to all predictions during test.
- **Precision**: Generally, precision is defined as the ratio of the number of relevant returned images divided by the sum of the returned ones. In other words, percentage of relevant images returned. Besides pairwise precision is considered as the fraction of the correctly clustered face images over the total number of face images from the same class. We note that precision is also known as positive predictive value [20] (Equation 8).

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (8)$$
where TP represent the true positive pairs, which means that the system assigns a similar images pairs to the same cluster. FP denotes the false positives pairs. In other words, assigning a dissimilar pair of images to the same class.

Recall: Recall is defined as the ratio of the correctly clustered images pairs by the total number of images of the same cluster [20]. It means that the recall measurement is the percentage of the retrieved relevant images from all returned ones (Equation 9)

\[
\text{Recall} = \frac{TP}{TP+FN}. \tag{9}
\]

In Equation (9), the FN refers to assigning two similar pairs of images to different clusters. Named false negatives pairs.

Precision/recall curve: Used to evaluate the capacity of a content based retrieval system to return relevant documents. If we return several document, we improve recall. But if we return fewer ones, we improve precision, and reduce recall. There is a trade-off between this two metrics. When the precision versus recall is closer to the top, that indicates a good performance.

### 4.5 Subjective evaluation

In this test, we compare our method with some of the state of the art methods using the foreman sequence. The first method is a local feature based keyframe extraction method [10] and the other is a face feature based [18]. A brief description of those two methods is shown below.

1. **Local feature based method [10]:** The keyframe extraction method is based on the local description using interest points and graph modularity clustering. They start by extracting some candidate keyframes based on a windowing rule with size = FPS. Then, the system detects interest points for each frame and compute the repeatability between each two frames and stocks these values in a matrix called repeatability matrix. The repeatability value reflects the similarity between frames based on their content. Finally, this matrix is modelled by an oriented graph. The keyframes are selected using graph modularity clustering.

2. **Face feature based [18]:** Use four metrics: symmetry, sharpness, brightness and resolution in order to provide a face quality score to be used as a criteria to select keyframes. The authors multiplied these four metrics by different weight, giving the higher weight to symmetry measurement and for resolution the lowest weight. Then, they sum the results up together to produce a final quality score for the corresponding face image.

Comparing the keyframe extracted by [18] and ours, we can note that the brightness distribution with the KS-FQA is better than [18]. Even if this face is not frontal, but all their components are visible (the eyes in [18] are closed). It should be noted that the extracted keyframe by [18] does not ever appear in the candidate frame for this sequence.

The ranking table presented in Table 6 shows a sequence of the CVL dataset and the resulted rankings using our KS-FQA method compared against a ground truth ranking, and the work presented in [16] that use 10 metrics and [15, 18] that use four metric and a weighted system to calculate face quality.

The rank provided by KS-FQA is almost in agreement with the ground truth. We note that we deal only with five out of seven frames. Moreover, the rejected images already have the last rank.

### 4.6 Objective evaluation

In this section, we will test our keyframe extraction method in a face recognition task and in a content based video retrieval system.

#### 4.6.1 Face verification performance

In a first step, we compare our method against similar method. In other words, methods that use keyframe extraction based on face quality assessment. Our face recognition process works as follow. We start by extraction keyframes from the sequences and put them as input for the face recognition methods. We can notice that the number of ranking is increased, while the image quality decreases.

The ranking table presented in Table 6 shows a sequence of the CVL dataset and the resulted rankings using our KS-FQA method compared against a ground truth ranking, and the work presented in [16] that use 10 metrics and [15, 18] that use four metric and a weighted system to calculate face quality.

The rank provided by KS-FQA is almost in agreement with the ground truth. We note that we deal only with five out of seven frames. Moreover, the rejected images already have the last rank.
TABLE 4  Verification accuracy of state of the art methods on YTF dataset

| Method                  | Accuracy |
|-------------------------|----------|
| Fourney et al. [14]     | 69.82%   |
| Wong et al. [12]        | 79.92%   |
| Nikitin et al. [34]     | 74.46%   |
| Anantharajah et al. [17]| 89.7 %   |
| Qi et al. [18]          | 92.6 %   |
| KS-FQA                  | 95.4     |

TABLE 5  Verification accuracy on the YTF dataset against deep methods

| Method                  | Accuracy |
|-------------------------|----------|
| DeepFace [35]           | 91.4%    |
| DeepID2+ [36]           | 93.2%    |
| FaceNet [24]            | 95.1%    |
| VGG [37]                | 97.3%    |
| Centre loss [38]        | 94.9%    |
| DFL [39]                | 96.2%    |
| KS-FQA                  | 95.4%    |

compare the face image request with all faces extracted in the first step. Similarity is measured using the euclidean distance, and face representation is defined as a 512 dimensional embedding containing coordinates in order to get more discriminate feature. Results are presenting in Table 4.

Those methods aim to get a trade off between speed and accuracy. That is why they focus on a lower computational costs metrics and formulas to estimate face quality. In our case, and for a CBVR system, this step will be performed offline. our main interest is to obtain the higher accuracy rate more than a higher speed up.

Based on the results presented in Table 4, we note that the best accuracy rates are obtained using the following four metrics: pose or symmetry to estimate how much the face is frontal, sharpness, resolution and brightness. The difference between these results lies in the formulas used to calculate metrics and how the authors combine them into one single value. We also note that the symmetry measurement does not provide frontal faces, this condition cannot be verified in all cases. Specially for pitch rotation. Moreover, the use of weights influences the results while giving more priority to metrics than others.

In Table 5, we compare our method compared to the most recent methods, that use the new technologies such as the deep learning for face recognition tasks.

The results show that the proposed method performs better than several well-known deep face models [35, 36], and its performance is comparable to the FaceNet method that use the most deeper architecture by combining 25 deep convolutions neural networks. This slight improvement is due to the use of a large discriminant embedding to characterise faces. The use of this embedding in the training set, provide a lower verification rates comparing the 128 embedding dimensionalities [24].

This weakness may be caused by a small training session. In other words, using an embedding with higher dimension: 512 require more training than smaller one. Also, it may varied due to the training set used. That is why, in our work and based on a the best face images, the use of a 512 dimension embedding provide better performance.

4.6.2  Content-based video retrieval based on face image

In this test, we will integrate our KS-FQA module in a CBVR system. Using this module in the video processing step, a face image dataset is built in the indexing part. This dataset contain only the best face image for each identity having the best quality.

To compare our module against other works, we test two other keyframe extraction methods instead of our method, in the CBVR system shown in Figure 10. The method proposed in Qi et al. [18] use four features: symmetry, sharpness, brightness and resolution. Besides, the other mentioned in Nasrollahi et al. [15] use pose, sharpness, brightness and resolution.

All these methods use almost the same metrics with a slight difference in the way of calculation. In which, we prove the usefulness of the formulas used in the quality assessment. In addition, this two methods use weighting system to calculate face quality. In spite of the fact that Qi et al. [18] use the symmetry measurement, unlike Nasrollahi et al. [15] that use the pose. This two methods use the same weight value.

TABLE 6  Ranking result for a CVL sequence

| Frames      | 1 | 2 | 3 | 4 | 5 |
|-------------|---|---|---|---|---|
| Ground-truth [33] | 6 | 4 | 1 | 5 | 6 |
| Nasrollahi et al. [16] | 7 | 4 | 1 | 5 | 6 |
| Qi et al. [18] | 6 | 4 | 5 | 2 | 7 |
| Nasrollahi et al. [15] | 7 | 5 | 3 | 4 | 6 |
| KS-FQA      | Rejected | 5 | 1 | 4 | Rejected | 2 | 3 |
Figure 11 shows that our FQA based keyframe extraction improves recall and precision against the state-of-the-art FQA methods. The precision/recall curve of our system remains at the top of the graph and close to the value 1 which means that our system is more efficient than the others. In other words, our system is more efficient to return relevant elements to the request in the first place. This efficiency is due to the basis of the descriptors chosen. Based on frontal face images.

5.1 CONCLUSION

In this paper, we present a keyframe extraction method based on face quality assessment. Starting by filtering the video frames and keeping only useful face images to group them later by identity in order to extract, for each identity the best face image based on image quality. Our method is capable to deal with several issues frequently found in videos, such as pose and illumination variations, blurring faces and low-quality face images. Thus, in the face quality assessment module, we calculate four metrics including pose estimation, sharpness, brightness and resolution.

Moreover, we proposed a novel pose estimation metric, based on the analysis of the landmarks geometric distribution within a face region. Which allows to detect both yaw and pitch face rotation. The calculated metrics are later combined together to form a single value representing the quality of a face. We do not use weights in this work in order to offer the same priority for all the metrics.

The system was evaluated on three datasets using subjective and objective tests. In the subjective evaluation, the results show a general agreement between our system ranking and a ground-truth ranking. We have addressed not only the extracted keyframes and their quality, but also the number of frames to be proceeded. The use of a candidate set reduces the number of frames to deal with.

The proposed approach proved to have superior effectiveness to state-of-the-art methods that use the keyframe extraction process. Compared to deep face recognition methods, we achieve satisfactory results compared to a list of deep methods.

In conclusion, it is evident that this study has shown the effectiveness of using only frontal faces in videos rather than using all the video frames.

On the basis of the promising findings presented in this work, we plan to assess the face quality based on a single image rather than using the complete face image set, in order to make the
video processing in real time. This will be done based on a new method for keyframes extraction based on a deep learning method.

REFERENCES

1. Chihaoui, M., et al.: A survey of 2D face recognition techniques. Comput. Syst. 5(4), 21 (2016)
2. Hassaballah, M., Aly, S.: Face recognition: challenges, achievements and future directions. IET Comput. Vis. 9, 614–626 (2015)
3. Anil, J., Suresh, L.P.: Literature survey on face and face expression recognition. International Conference on Circuit, Power and Computing Technologies, pp. 1–6. IEEE, Piscataway (2016)
4. Yuan, Z., et al.: Spatiotemporal modeling for video summarization using convolutional recurrent neural network. IEEE Access 7, 64676–64685 (2019)
5. Barf, J.R., et al.: Face recognition from video: a review. Int. J. Pattern Recogn. Artif. Intell. 26, 1266002 (2012)
6. Chamesamani, F., et al.: A study on surveillance video abstraction techniques. IEEE International Conference on Control System, Computing and Engineering (ICCSCE), pp. 470–475. IEEE, Piscataway (2015)
7. Kaur, P., Kumar, R.: Analysis of video summarization techniques. Int. J. R. Appl. Sci. Eng. Technol. (IJRASET) 6, 1157–1162 (2018)
8. Cheng, H., et al.: Video scene segmentation using the shot transition detection by local characterization of the points of interest. 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETiT). IEEE, Piscataway (2012)
9. Gharbi, H., et al.: Key frames extraction based on local features for efficient video summarization. Advanced Concepts for Intelligent Vision Systems: 17th International Conference, ACIVS. Springer, Berlin/New York (2016)
10. Gharbi, H., et al.: Key frames extraction using graph modularity clustering for efficient video summarization. 42nd IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP. IEEE, Piscataway (2017)
11. Kaayya, S., LakshmiPraya, G.G.: Multimedia indexing and retrieval: recent research work and their challenges. 3rd International Conference on Signal Processing, Communication and Networking (ICSCN), pp. 1–5. IEEE, Piscataway (2015)
12. Wang, Y., et al.: Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. Computer Vision and Pattern Recognition Workshops CVPRW, pp. 74–81. IEEE, Piscataway (2011)
13. Griffin, P.: Understanding the face image format standards. Computer Vision and Pattern Recognition Workshops (CVPRW), ANSI/NIST Workshop. IEEE, Piscataway (2005)
14. Fourney, A., Laganere, R.: Constructing face image logs that are both complete and concise. 4th Canadian Conference on Computer and Robot Vision (CRV’07), pp. 488–494. IEEE Computer Society, Los Alamitos (2007)
15. Nasrollahi, K., Thomas, B.M.: Face quality assessment system in video sequences. In: Schouten, B., et al. (eds.), European Workshop on Biometrics and Identity Management. Lecture Notes in Computer Science, vol. 5372, pp. 10–18. Springer, Berlin, Heidelberg (2008)
16. Nasrollahi, K., Moeslund, T.B.: Summarization of surveillance video sequences using face quality assessment. Int. J. Image Graph. 11, 207–233 (2011)
17. Anantharajah, K., et al.: Quality based frame selection for face clustering in news videos. International Conference on Digital Image Computing: Techniques and Applications (DICTA). IEEE, Piscataway (2013)
18. Qi, X., Liu, C.: Gpu-accelerated key frame analysis for face detection in video. 1st IEEE Workshop on Delay Sensitive Video Computing in the Cloud (DSVCC). IEEE, Piscataway (2015)
19. Deng, Y.-J., et al.: Nuclear norm-based matrix regression preserving embedding for face recognition. Neurocomputing 311, 279–290 (2018)
20. Shi, Y., et al.: Face clustering: representation and pairwise constraints. IEEE Trans. Inf. Forensics Security 13(7), 1626–1640 (2018)
21. Rajeev, R., et al.: An all-in-one convolutional neural network for face analysis. arXiv, 2016
22. Anvari, Z., Abtisors, V.: A pipeline for automated face dataset creation from unlabeled images. Proceedings of the 12th ACM International Conference on Pervasive Technologies Related to Assistive Environments. ACM, New York (2019)
23. Kulbhreshtha, P., Guha, T.: An online algorithm for constrained face clustering in videos. 25th IEEE International Conference on Image Processing (ICIP), pp. 2670–2674. IEEE, Piscataway (2018)
24. Schrö, F., et al.: Facenet: a unified embedding for face recognition and clustering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Piscataway (2015)
25. Zhu, C., et al.: A rank-order distance based clustering algorithm for face tagging. International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 481–488. IEEE, Piscataway (2011)
26. Zhang, K., et al.: Joint face detection and alignment using multi-task cascaded convolutional networks. IEEE Signal Process. Lett. 23, 1499–1503 (2016)
27. AT&T Laboratories Cambridge, Face database. https://github.com/GTDLBench/datasets/att_face_dataset/. Accessed March 2019
28. Solina, E., et al.: Color-based face detection in the 15 seconds of fame art installation. International Conference on Computer Vision/Computer Graphics Collaboration for Model-Based Imaging, Rendering, Image Analysis and Graphical Special Effects. Springer, Berlin, Heidelberg (2003)
29. Wolf, I., et al.: Face recognition in unconstrained videos with matched background similarity. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 529–534. IEEE, Piscataway (2011)
30. KING, D.: Dlib-ml : A machine learning toolkit. J. Machi. Learn. Res. 10, 1755–1758 (2009)
31. Viola, P., Jones, M.: Real-time face detection. Int. J. Comp. Vis. 10(2), 137–154 (2004)
32. Shi, Y., et al.: A review on face detection and study of Viola-Jones method. Int. J. Comp. Trends Technol. 25(1), 54–61 (2015)
33. Nasrollahi, K.: From face detection to face super-resolution using face quality assessment. PhD thesis, Faculty of Engineering and Science, Aalborg University (2009)
34. Nikitin, M., et al.: Face quality assessment for face verification in video. Proceedings of GraphiCon’2014, pp. 111–114. GraphiCon Scientific Society, Moscow (2014)
35. Taigman, Y., et al.: Deepface: closing the gap to human level performance in face verification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Piscataway (2014)
36. Sun, Y., et al.: Deeply learned face representations are sparse, selective, and robust. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Piscataway (2015)
37. Parkhi, O.M., et al.: Deep face recognition. British Machine Vision Conference. BMVA Press, Durham (2015)
38. Wen, Y., et al.: A discriminative feature learning approach for deep face recognition. European Conference on Computer Vision, pp. 499–515. Springer, Cham (2016)
39. Deng, W., et al.: Deep correlation feature learning for face verification in video. European Conference on Computer Vision, pp. 177–188. Springer, Cham (2016)
40. Ramakrishnan, S.: Face recognition: semisupervised classification, subspace projection and evaluation methods. In: BoD: Books on Demand. InTech, Croatia (2016)
41. Pan, Le, et al.: A key frame extraction algorithm based on clustering and compressive sensing. Int. J. Multimedia Ubiquitous Eng. 10(11), 385–396 (2015)