Face Retrieval in Videos using Face Quality Assessment and Convolution Neural Networks

Rahma Abed, Sahbi Bahroun and Ezzeddine Zagrouba

September 10, 2020

Face Retrieval in Videos using Face Quality Assessment and Convolution Neural Networks

Rahma Abed
Laboratoire LIMTIC
Institut Supérieur d'Informatique,
Université de Tunis El Manar
Ariana, Tunisie
rahma.abed@etudiant-isi.utm.tn

Sahbi Bahroun
Laboratoire LIMTIC,
Institut Supérieur d'Informatique,
Université de Tunis El Manar
Ariana, Tunisie
sahbi.bahroun@isi.utm.tn

Ezzeddine Zagrouba
Laboratoire LIMTIC
Institut Supérieur d'Informatique,
Université de Tunis El Manar
Ariana, Tunisie
ezzeddine.zagrouba@uvt.tn

***Abstract*—With the large amount of videos produced every day, Content-Based Video Retrieval (CBVR) has become a necessity by describing each video with a compact and significant signature in order to efficiently retrieve the desired video from a large collection. In this work, we present a CBVR system applied on face recognition based on keyframes. The first step in this system consists of extracting keyframes from videos using Face Quality Assessment (FQA) and Convolution Neural Networks (CNN). Starting by generating face quality scores for each face image using three face feature descriptors (Gabor, Local Binary Pattern (LBP) and Histogram of Oriented Gradient (HoG)). Then, we train a Convolution Neural Network (CNN) in a supervised manner in order to select frames having the best face quality. Experiments on several datasets has shown that the proposed "DeepFQA" method gives promising results in terms of accuracy and precision/recall curve.**

***Index Terms*—Content Based Video Retrieval, Keyframe extraction , Face quality assessment, Convolution Neural Network.**

I. Introduction

Due to the significant growth of video data, Content-Based Video Retrieval (CBVR) systems have become an active topic research for computer vision tasks [34]. Researchers note that we can improve performance when the CBVR system describes videos based on specific objects that are also adapted to the user's request [1]. Human beings and especially faces are usually one of the most important objects in a video. Therefore, numerous studies focus on face image for several tasks such as face recognition, tracking, emotion recognition or for extracting other characteristics like age, sex, origin, family relationships, etc. Besides, most of the face images in these videos are either irrelevant or duplicated. This is due to the conditions in which these images were captured: people or head motion (head pose variation), occlusions, illumination conditions, distance from camera, facial expressions [10].

For these reasons, and for better results, we need to build a mechanism that aims to extract the best frames that describe well each identity present in the video. This mechanism is named Keyframe Extraction [33]. keyframe play an important role in a videos indexing and retrieval system, since they provide the most useful information for retrieval purpose [1]. Extracting keyframes based on faces, consists in defining each video by face images of each identity appearing in this video. These frames will be selected based on an objective criteria named Face Quality Assessment (FQA). Using FQA for keyframe extraction leads to select, for each identity, the face image having the higher face quality score. In other words, the most suitable image in which all the details of faces are visible to represent an identity.

The contribution of this work is to define a keyframe extraction module to be integrated into a CBVR system applied to face recognition. Keyframe extraction will be based on FQA and CNN. Moreover, we use several face features in order to generate face quality for a large collection of face images that will be considered as a training set. Then, and based on these face image set and the generated face score (considered also as label), we train a CNN in order to be able to automatically predict the face image quality.

The rest of this paper is organized as follows. In Section II, we introduce an overview of related works on keyframe extraction based on FQA. Section III explains the proposed keyframe extraction method. Then, the results of the experiments and the observations are discussed in Section IV. We conclude the paper in Section V.

II. Related Work

We distinguish three methods for keyframe extraction based on face image [28]. Classification based, optic flow based and quality based. In this work, we focus on the image quality based methods due to their ability to filter the low-quality face images and keep only useful frames [11].

The image quality based methods aim to describe face image based on several aspects associated with various facial conditions such as head rotation, expressions, accessories and occlusions [5]. In addition, These methods aim to choose the best face image based on quality from a set of images [5]. In real scenarios, most of the face images are useless due to several problems such as not facing the camera, motion blur, illumination and low resolution. That is why we need to choose the best face image in order to represent each identity. In the beginning, Face Quality Assessment (FQA) was calculated using the geometric face shape and face features, including pose, resolution of the face area, confidence score of the detected eyes, lighting, facial expressions, etc [21]. Among the major advantages of these methods is the low cost of calculation, simplicity and speed.

In fact, the face image with the best quality is the one, in which, all the face details are well visible. In [2], two metrics were used to estimate the brightness measurement, combined with the head pose, sharpness, presence of human skin and resolution. Nasroallahi et.al [19] use pose, sharpness, brightness and resolution to generate a quality score for each face image. Qi et al. [27] propose to use the symmetry measurement instead of estimating pose, combined with resolution, sharpness and brightness. Anantharajah et al. [3] apply face image quality for face image clustering in news videos. Face quality module was based on symmetry, sharpness, contrast and brightness. Subsequently, these metrics are combined into a single value to be considered as face quality score. Moreover, the previous researches associate predefined weights for each metric. The use of these fixed weights is neither capable nor suitable to deal with several videos having different backgrounds, illumination conditions, head positions, etc [28].

To solve this issue, researchers start to use face feature rather than using metrics for estimating face quality. And with the technology progress and the success achieved by the use of Deep Learning (DL) techniques, several works focus on the use of DL for keyframe extraction based on FQA.

Chen et.al [23] propose a learning-to-rank framework to estimate face quality in which, they use five face features: CNN, Gist, Gabor, LBP and HoG combined into one face quality score. Also, three categories of image are combined to prepare a huge learning set including high and low quality face images and non-face images. Vignesh et.al [23] use two face features extractors (HoG and LBP) to provide face quality scores. The obtained face scores are considered as label for the face image set. Then, the face image and their labels are used for training a CNN to predict automatically face image quality. Since then, most research has been interested in annotating face images for training step. Vishal et.al [25] compares each face image against a chosen template based on the Euclidean distance between the face feature vectors provided by FaceNet [8]. Two way for estimating face quality was proposed and combined in [15]. The first is a comparison of images using machine learning techniques and the second is based on a manual annotation. They concluded that the human quality is a better accurate predictor. Recently, Hernandez et.al [35] propose a Quality Assessment system based on deep learning. They use a framework to label the VGGFace2 images [36] used as a learning set. Then, the authors use the FaceNet model for feature extraction, and the face scores (used as ground truth) are obtained using the Euclidean distance between the obtained face features. For training, they fine-tune the ResNet-50 model to perform the quality prediction

The purpose of this work is to describe video using keyframes that are selected based on face quality. Instead of using face metrics, we use face features to define face quality. These labels and their corresponding face images are fed to a CNN in order to learn it how to automatically predict face quality. thanks to the use of CNN, the estimation of the quality of face images will be based on a single face image rather than using a set of images. For label generation, several points should be taken into consideration. On the one hand, the manual estimation of face image quality requires a huge human effort and a lot of time [15]. On the other hand, using the recognition accuracy between two images as a quality score depends on the used face recognition algorithm and the template image chosen. To deal with this, we use the similarity between face features in order to estimate face quality. In addition, we provide an objective criteria to select templates. We have addressed not only the label generation method, but also the learning set. Our learning set contains several face images gathered from different dataset that are well-known in the literature and are useful for face recognition and face quality assessment in unconstrained environment. More details are shown in the following sections.

III. Proposed method

The proposed method ”**DeepFQA**” for keyframe extraction has two main steps. First, the label generation for the training images set. Second, we train a deep CNN based on the face image set and their generated labels in order to automatically predict face quality. Face detection is performed in the beginning in order to eliminate useless frames. A complete flowchart of the proposed method is shown in Figure 1.

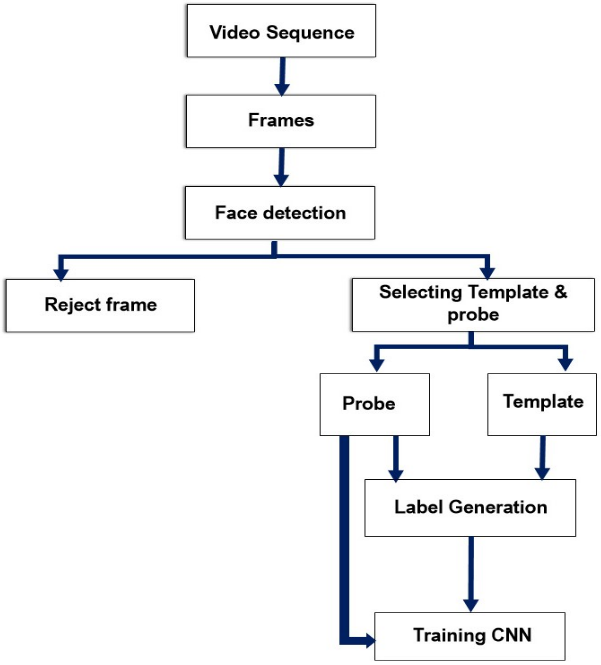
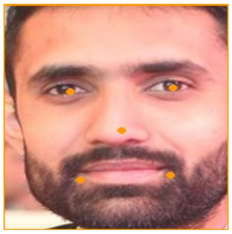


Fig. 1: Flowchart of the proposed method

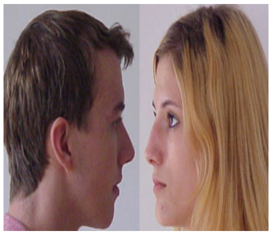
A. Face Detection

We use the Multi-task Cascaded Convolutional Networks (MTCNN) [13] as a face detector. The MTCNN provide five landmarks (the two eyes, the two corners of the mouth and the nose) and a confidence score indicating the probability that the detected object is a face. Figure 2a illustrate an example of a detected face.

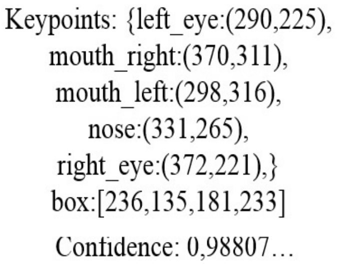


(a) Detected face

The main advantage of this detector is the ability to reject frames in which the MTCNN cannot detect the five landmarks. This condition allows us to reject rotated faces or low-resolution face images (Two examples are presented in Figure 2b). Figure 2c present the confidence score and the coordinates of the five landmarks.



(b) Rejected faces



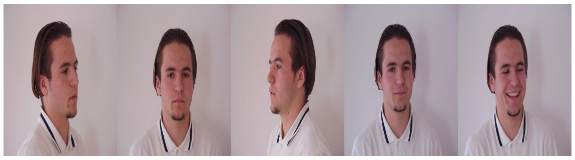
```
Keypoints: {left_eye:(290,225),
mouth_right:(370,311),
mouth_left:(298,316),
nose:(331,265),
right_eye:(372,221),}
box:[236,135,181,233]
Confidence: 0,98807…
```

(c) Output

Fig. 2: Output of the MTCNN detector

B. Selecting Template and probe Set

Each set of face images will be divided into two subsets: Template and Probe [25]. The template contains only one image per identity. This template will be selected in an objective manner, thanks to the use of confidence score provided by the MTCNN. Indeed, the face image with the highest confidence score will be considered as template.
The remaining face images are forwarded to the probe set.



(a) Face image Set

We present in Figure 3a an example of a face image set and the chosen as Template (Figure 3b).

(b) Template

Fig. 3: Selection of the Template and probe Set

In the next section, we present the label generation process in which, we estimate face quality score for each image in the probe set. Then, the probe set and the corresponding scores are fed to a CNN in order to predict automatically face quality.

C. Label generation

To generate labels, we use three face features: LBP, HoG and Gabor filters. These features are widely used to describe face images. Gabor filters are robust against unbalanced illumination conditions and noises [10]. The wavelet coefficients of different scales and orientations make this filter robust against rotation, translation, distortion, and scaling [24]. That is why these filters are widely used for facial landmark location, tracking, face classification and head pose estimation [16]. LBP is known as an invariant to monotonic illumination variations caused by slight lighting deviation.
LBP works as a filter that extracts the pixels difference in order to generate at the end either a binary code or a histogram [10]. The use of histograms as a descriptor also makes it robust against the misalignment and pose variations. The HoG is considered as the most adequate to describe facial expressions. In fact, HoG is adopted for facial expression recognition based on the use of the face muscle shapes that are modeled by a contour analysis [22].
We start by extracting the feature vectors using the three descriptors. Subsequently, we calculate the cosine distance between each pairs of vector from the template and each image from the probe set. The obtained similarity values are summed and the average is considered as the face quality score.

D. Convolution Neural Network Architecture

The CNN structure is described in Figure 4. Inspired by the architecture proposed in [28], we start by applying an inception module in which, we use PReLU [12] as activation function after each convolution layer even in the inception node.

The inception module used is composed of three convolution layers having 16 filters with three different paths:5×5, 3×3 and ×1. Each convolution layer is followed by a PReLU layer and a max-pooling layer with stride and size equal to 2. The advantage of using the inception module is the concatenation of different features extracted of several convolution kernels having different sizes [28]. After that, we add a simple convolution bloc containing a convolution layer with 128 filters with 5×5 kernels, followed by a PReLU function and a max pooling layer.

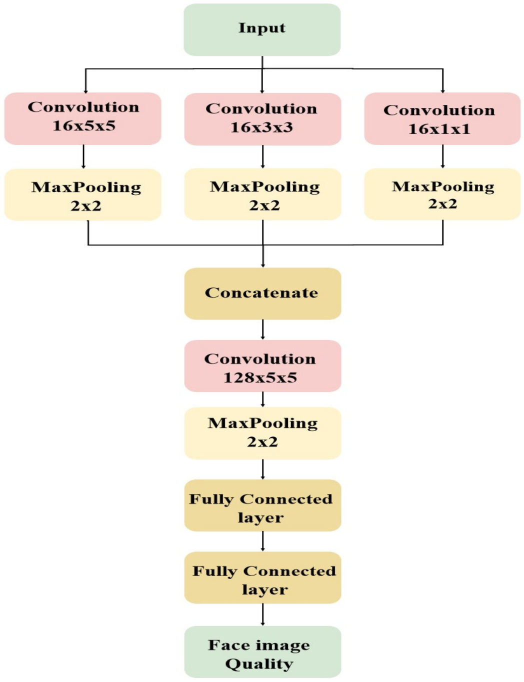


Fig. 4: CNN architecture

Then, we put in the end of the CNN two fully connected layers. We use a sigmoid function in the last node in order to generate a numeric output between 0 and 1 [28]. We use the adam optimizer and a mini-batches of 128 samples. The model is regularized using dropout applied before the two fully connected layers with a rate of 0.5, and a learning rate set to 0.001. We train our CNN for 500 epochs.

IV. Experimental Results

In this section, we will evaluate our DeepFQA method based on subjective and objective tests.

We start by giving an overview of the used datasets to train and test our system. The first test presents the ranking result of a sequence using our label generation module then DeepFQA, compared to several methods. The last test will evaluate our method into a face recognition task and in a CBVR system on real face datasets.

A. Datasets

In order to generate a large set of images that presents the difficulties that can be found in real environments (e.g pose variation , illumination conditions, occlusion,low-quality images, etc), we combine several datasets in order to collect a set of 0.22M frames. The first dataset is AT&T [4]. The images in this set were taken in a dark homogeneous environment with small pose variation and different facial expressions. The FRI CVL dataset [9] in which, the face images have different head positions and facial expressions and a fixed resolution. The third dataset is Face Recognition Data provided by the University of Essex [7]. Several accessories are presented, and people are moving toward the camera and the background is complex. We also used the cropped face images from the Yale Face dataset B [14] and the Extended Yale Face dataset B. The Yale Face dataset B containing 16434 images of 28 human subjects under different viewing and poses conditions.

To test our system, we use the YouTube Face dataset (YTF) [17]. This dataset contains 3,425 videos of 1,595 people collected from YouTube, with an average of 2 videos per identity, and it is a standard benchmark for face verification in video.

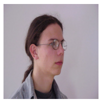
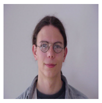
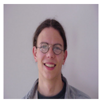
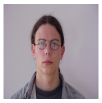
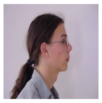
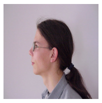
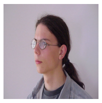
B. Subjective Evaluation

In this test, we try to sort the face images based on their quality scores (Table I). We compare our ranking score provided by our label generator, in order to prove the effectiveness of the labels used for training, and the whole method (DeepFQA) against a ground-truth provided by [19], a face quality metric based [19] and a face feature based method [23]. The ground truth presents a human perception of the quality of the face. In fact, Nasroallahi et.al [19] annotated the images of this sequence according to the visual features and the face visibility. Then, they sorted them manually according to their perception of quality. While, Chen et.al use five face features including CNN, Gist, Gabor, LBP and HoG.

Based on the ranking results presented in Table I, we aim to prove the effectiveness of our label generator module. In fact, while using only the label generator module, the face ranking is based on the similarity measurement between Template and each image of the probe set. Which mean that the face quality depends on the chosen template. In addition, we notice that the most rotated faces have the last ranks. Moreover, the use of DeepFQA offer two advantages. First, the predicted face quality is based on all the learning set and not based on a given sequence, which allows us to deal with frames in a real time. In addition, we don't need to select a template or to use a weighting systems. The reliability of such a system depends only on the given learning set. In other words, using a CNN like in our method we could estimate face quality based on a single image as input.

We note also that the use of the MTCNN face detector allows us to reduce the number of face images to deal with by rejecting rotated face images.

TABLE I: Ranking Result for a CVL sequence: the numbers from 1 to 7 rank the images based on their quality score: Notes that the ranking number is increased, while the image quality decreases.

Frames							
Ground-truth,2008 [19]	4	2	3	1	5	5	4
Nasrollahi and Thomas [19]	5	1	2	1	6	5	3
Chen et al.,2015 [23]	4	2	3	1	-	-	-
Label Generator	**3**	**1**	**Template**	**2**	**Rejected**	**Rejected**	**Rejected**
DeepFQA	**4**	**2**	**3**	**1**	**Rejected**	**Rejected**	**Rejected**

C. Objective Evaluation

To validate our proposed method, we use the keyframe extraction method for a face verification task using the YTF dataset following this scenario: We start by keyframe extraction using DeepFQA. Then, we use the extracted keyframes as input to perform the face recognition algorithm. Face recognition module used in this experiment is the FaceNet system [8]. In Table II, we summarizes the obtained results by testing different keyframe extraction methods in a facial recognition task.

TABLE II: Comparison on the YouTube Faces dataset against similar methods

Method	Accuracy
Adam and Laganiere, 2007 [2]	69,82%
Yongkang et al., 2011 [32]	79,92 %
Mikhail et al., 2014 [18]	74.46%
Anantharajah et al., 2013 [3]	89.7 %
Xuan and Chen, 2015 [27]	92,6 %
DeepFQA	**95.2 %**

From Table II, We note that the best accuracy rates are obtained using the following four metrics: pose, sharpness, resolution and brightness [27]. Indeed, 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 some metrics over others. In addition, those methods do not consider the facial expression, which be verified in our cases.

Next, we compare our result against recent deep face recognition's methods. The results in table III show that our DeepFQA achieves higher accuracy rates than some deep face methods like deep face [30] and Deep ID+ [31]. On the one hand, we achieve the same accuracy as FaceNet [8] even with the use of smaller learning set (2.6 M image for FaceNet). on the other hand, our approach is a little far from deep methods like [29] and [20] which are characterized by a wide learning ability due to the use of a large dataset (0.7 M for Center loss [29] and 4.7M for DFCL [26]) and a very deep CNN.

TABLE III: Comparison on the YouTube Faces dataset against deep methods

Method	Accuracy
Deep face,2014 [30]	91.4%
Deep Id+,2014 [31]	93.2%
FaceNet,2015 [8]	95.1%
VGG-face,2015 [20]	**97.3%**
Center loss,2016 [29]	94.9%
DFCL , 2017 [26]	**96.06%**
DeepFQA	95.2%

For the last test, we integrate our keyframe method into a CBVR system. In the first stage, the CBVR system performs DeepFQA. Then, the extracted keyframes are stored in the face image dataset. The last stage consists in comparing the face image of the user's request with all face images in the dataset (Figure 5).

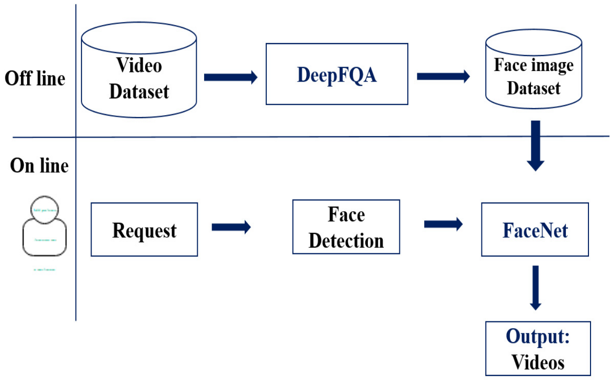


Fig. 5: The proposed Content based video retrieval system

We integrate two other methods into the same CBVR system in order to evaluate the effectiveness of the proposed keyframe extraction method. The method proposed by [27] use four features: symmetry, sharpness, brightness and resolution. Besides, the other mentioned in [3] use symmetry, sharpness, brightness and contrast.

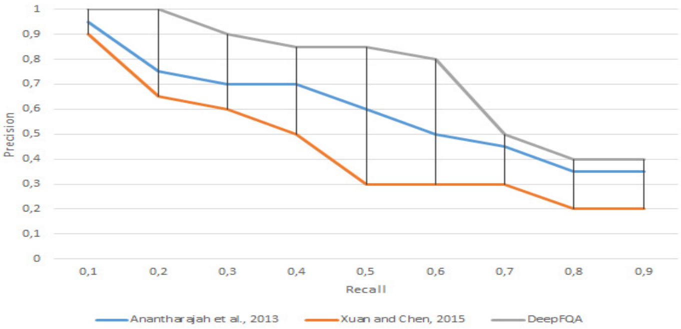


Fig. 6: Precision and recall curve using three keyframe extraction methods: Xuan and Chen, 2015, [27] represented by the blue curve, and Anantharajah et al., 2013, [3] by the orange curve and DeepFQA represented by the gray curve

Figure 6 shows the precision/recall curve obtained using our DeepFQA with two other methods that use metrics. These two methods use almost the same metrics. The curve of our method is higher rather the other curves.

In other words, we achieve better performance in terms of accuracy and recall. These results prove the effectiveness of our keyframe extraction method into a CBVR system. Selecting the best face image in video based on several factors such as brightness, face expression, pose may improve the robustness of such system.

V. Conclusion

Face video retrieval aims to search from a large database the videos containing a particular person, with the same face image as the query. This field have attracted more and more attentions in recent years.

In this paper, we present a new Content Based Video Retrieval system applied on face recognition. We integrate a keyframe extraction method in order to describe each video with a set of face image from people appearing in this videos. The keyframe extraction module is based on the use of Face Quality assessment to assign for each face image a quality score and a Convolution Neural Network in order to predict automatically face image quality. For this purpose, we start by generating quality scores for the learning set using several face features. Then, the labels and their corresponding images are fed to CNN. The experimental results prove the utility of our method for keyframe extraction. We conclude that our method does not obtain the best accuracy. It may be due to the learning set used or the CNN architecture. The latter proved its effectiveness for classification task, but improvement is always possible.

As part of future work, we plan to improve the performance of our DeepFQA method focusing on our label generator module, basically the face feature used. First, we will use other feature that aim to alleviate the impact of noise, and blur effect. Further, we plan to improve our CNN architecture using pre-trained models. Also we aim to add a facial neutralization module to be performed before forwarding face frames towards the face recognition system.

References

[1] Ansari, A. and Muzammil H.M., *Content based video retrieval systems-methods, techniques, trends and challenges*. International Journal of Computer Applications, 2015, 112(7):13–22.

[2] A. Fourney and R. Laganiere,*Constructing face image logs that are both complete and concise.* Fourth Canadian Conference on Computer and Robot Vision, 2007,pages 488–494.

[3] K. Anantharajah, S. Denman, D. Tjondronegoro, S. Sridharan, C. Fookes and X. Guo,*Quality based frame selection for face clustering in news video*. International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2013 pages 1–8.

[4] AT&T (2002). *Laboratories cambridge face database*. Available at: http://www.cl.cam.ac.uk/research/dtg/attarchive/ facedatabase.html.

[5] S. Bhattacharya and A. Routray, *Score based face quality assessment (FQA).* 14th IEEE India Council International Conference (INDICON), 2017:, pages 1–6.

[6] J. Chen, Y. Deng, G. Bai and G. Su. *Face image quality assessment based on learning to rank.* Signal Processing Letters, IEEE,(2015) 22(1):90–94.

[7] Faces96. *Face recognition data, university of essex, uk.* Available at: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.

[8] Schroff.F,, Kalenichenko,D., Philbin.J, *Facenet: A unified embedding for face recognition and clustering.* IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2015, pages 815–823.

[9] Solina, F., Peer, P., Batagelj, B., Juvan, S., Kovač, J. *Colorbased 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,2003 pages 38–47.

[10] H. Wang, J. Hu and W. Deng *Face feature extraction: A complete review.* IEEE Access, 2018 6:6001–6039.

[11] Barr, J. R., Bowyer, K. W., Flynn, P. J., Biswas, S. *Face recognition from video: a review.* International Journal of Pattern Recognition and Artificial Intelligence, 2012, 26(5).

[12] He, K., Zhang, X., Ren, S., Sun, J. *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification.* IEEE Proceedings of the IEEE International Conference on Computer Vision, 2015 1026–1034.

[13] K. Zhang, Z. Zhang, Z. Li and Y. Qiao *Joint face detection and alignment using multi-task cascaded convolutional networks.* Signal Processing Letters, 2016 23(10):1499 – 1503.

[14] Kuang-Chih Lee, J. Ho and D. J. Kriegman *Acquiring linear subspaces for face recognition under variable lighting.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5):684–698.

[15] L. Best-Rowden and A. K. Jain*Learning face image quality from human assessments.* IEEE Transactions on Information Forensics and Security, 2018.13(12):3064–3077.

[16] Shen, L., and Bai, L. *A review on gabor wavelets for face recognition.* Pattern Analysis and Application, 2006, 9(2-3):273–292.

[17] Wolf, L., Hassner, T., and Maoz, I. *Face recognition in unconstrained videos with matched background similarity.* Conference on Computer Vision and Pattern Recognition, 2011.pages 529–534.

[18] Nikitin, M., Konushin, V., Konushin, A*Face quality assessment for face verification in video.* Proceedings of GraphiCon2014, pages 111–114.

[19] Nasrollahi, K., Moeslund, T. B.*Face quality assessment system in video sequences.* Biometrics and Identity Management. Springer Berlin Heidelberg, 2008 pages 10–18.

[20] Parkhi, O. M., Vedaldi, A., Zisserman, A *Deep face recognition.* British Machine Vision Conference (BMVC),2015 1(3):1–12.

[21] Griffin, P *Understanding the face image format standards.* Computer Vision and Pattern Recognition Workshops (CVPRW), ANSI/NIST Workshop, Gaithersburg, MD. 2005

[22] Carcagnì, P., Del Coco, M., Leo, M., Distante, C.*Facial expression recognition and histograms of oriented gradients: a comprehensive study.* Springer-Plus, 2005 4(1):645.

[23] S. Vignesh, K. V. S. N. L. M. Priya and S. S. Channappayya, *Face image quality assessment for face selection in surveillance video using convolutional neural networks.* IEEE Global Conference on Signal and Information Processing (GlobalSIP),2015 pages 577–581.

[24] See, Y. C., Noor, N. M., Low, J. L., Liew, E. *Investigation of face recognition using gabor filter with random forest as learning framework.* Region 10 Conference, TENCON 2017 IEEE, pages 1153–1158.

[25] Agarwal, V. *Deep face quality assessment.* arXiv preprint arXiv:1811.04346, 2018 page 6.

[26] W. Deng, B. Chen, Y. Fang and J. Hu *Deep correlation feature learning for face verification in the wild.* IEEE Signal Processing Letters, 2017 24(2):1877–1881.

[27] X. Qi and C. Liu, *GPU-accelerated key frame analysis for face detection in video.* IEEE workshop on Delay Sensitive Video Computing in the Cloud, DSVCC.2015

[28] X. Qi, C. Liu and S. Schuckers, *Boosting face in video recognition via cnn based key frame extraction.* international Conference of Biometrics(ICB) 2018, pages 132–139.

[29] Wen, Y., Zhang, K., Li, Z., Qiao, Y *A discriminative feature learning approach for deep face recognition.* European Conference on Computer Vision, 2016 pages 499–515.

[30] Taigman, Y., Yang, M., Ranzato, M. A., Wolf, L *Deepface: Closing the gap to human-level performance in face verification.* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014 page 1701–1708.

[31] Sun, Y., Wang, X., Tang, X *Deep learning face representation from predicting 10,000 classes.* IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014 , page 1891–1898.

[32] Wong, Y., Chen, S., Mau, S., Sanderson, C., Lovell, B. C*Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition.* IEEE Biometrics Workshop, Computer Vision and Pattern Recognition (CVPR) Workshops, 2011, pages 81–88.

[33] Y. Yuan, H. Li and Q. Wang, *Spatiotemporal modeling for video summarization using convolutional recurrent neural network.* IEEE Access,2019 7:64676–64685.

[34] K. Zhang, H. Sun, W. Shi, Y. Feng, Z. Jiang and J. Zhao, *A video representation method based on multi-view structure preserving embedding for action retrieval.* IEEE Access,2019 pages 50400–50411

[35] Hernandez-Ortega, J., Galbally, J., Fierrez, J., Haraksim, R., Beslay, L*FaceQNET: quality assessment for face recognition based on deep learning*, arXiv preprint arXiv:1904.01740, 2019

[36] Q. Cao, L. Shen, W. Xie, O. M. Parkhi and A. Zisserman, *Vggface2: A dataset for recognizing faces across pose and age*. 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG) 2018 pages 67–74