Surveillance Video Quality Assessment Based on Face Recognition

Wen Heng, Tingting Jiang
National Engineering Laboratory for Video Technology, Cooperative Medianet Innovation Center, School of EECS,
Peking University, Beijing 100871, China
{wenheng,ttjiang}@pku.edu.cn

ABSTRACT

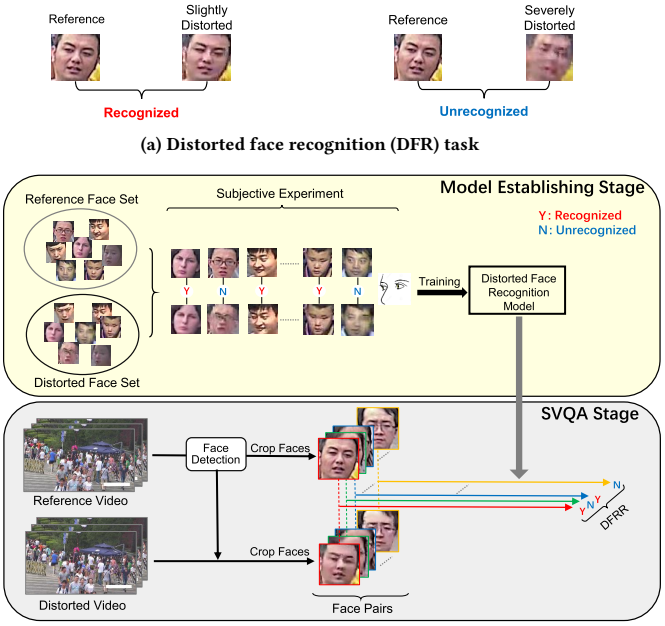
Nowadays video surveillance is widely used for public safety. In practice, multiple factors e.g. video compression, would cause the quality degradation and weaken the value of surveillance videos. So, proper quality assessment methods are needed for guiding the deployment and configuration of surveillance video system. In general, surveillance video quality assessment (SVQA) is different from conventional video quality assessment, because surveillance videos are usually used for one specific task e.g. recognition. We propose a face recognition (FR) task driven SVQA framework. In this paper, we mainly focus on one newly defined FR task: distorted face recognition (DFR) task, which is illustrated in Fig. 1 (a). Our goal is to establish an objective DFR model which can be used to measure the quality of distorted videos when compared to reference videos. To do that, first, we construct a face dataset collected from the real-world surveillance videos considering multiple factors e.g. light intensity, compression level, and conduct subjective experiments to collect subjective labels for the DFR task. Based on subjective labels we learn an objective distorted face recognition model and take it to assess the quality of distorted surveillance videos. In objective experiments, we analyze how different factors affect the quality of surveillance videos. In addition, the comparisons to PSNR and SSIM are made to show the advantages of the proposed method. At last, we give some suggestions for the practical applications of our proposed SVQA framework.

CCS CONCEPTS

• **General and reference** → **Measurement**; *Metrics*; • **Social and professional topics** → *Quality assurance*;

KEYWORDS

surveillance videos; quality assessment; deep learning; face recognition



(a) Distorted face recognition (DFR) task

(b) SVQA framework based on the DFR task

Figure 1: (a) Illustration of distorted face recognition task. (b) Workflow chart of our proposed full reference surveillance video quality assessment framework.

1 INTRODUCTION

In recent years, public safety problems have caused widespread concern. To monitor public incidents, more surveillance cameras have been deployed on the public places e.g. airports, stations, urban roads . However, many factors such as network transmission, video compression, cause the quality degradation of surveillance videos, which weaken the value of video data in some degree. Therefore, proper quality assessment methods are needed for guiding the deployment and configuration of surveillance videos systems .

Strictly speaking, surveillance video quality assessment (SVQA) is one special issue in video quality assessment (VQA) research filed. Conventional VQA methods pay more attention on the users' holistic perception of the videos for entertainment purpose. However, surveillance videos are mostly used for recognition task, e.g. pedestrian recognition. Taking conventional VQA methods e.g. SSIM [20] to measure the quality of surveillance videos, the results may not reflect the expected quality indeed. Instead, the proper SVQA method should mainly measure the usefulness of the distorted videos for recognition tasks rather than their entertainment value.

ThematicWorkshops'17, October 23–27, 2017, Mountain View, CA, USA

ACM ISBN 978-1-4503-5416-5/17/10…$15.00
https://doi.org/10.1145/3126686.3130239

Quality of recognition (QoR) is such a research topic that focuses on how to measure video quality from the perspective of recognition. Recently, the project named "Quality Assessment for Recognition and Task-based multimedia applications" (QART) was created by Video Quality Expert Group (VQEG) , which aims to "study the quality of video used for recognition tasks and task-based multimedia applications". Following the guidance, some studies [5, 8, 17] have been made to find the dependency between recognition rate and the reduced video quality based on the car license plate recognition (LPR) task. Usually, only a few factors are taken into account in their experiments such as video compression and resolution. However, considering the reliability of real-world surveillance videos quality assessment method, we believe more factors should be covered, e.g. target size, light intensity.

Besides the car license plate, human face is another important semantic carrier in surveillance videos. Korshunov and Ooi [6, 7] measured the quality of surveillance videos from the perspective of face detection, recognition and tracking tasks, and they emphatically analyzed how the bit rate of compressed video affects the performance of face detection, recognition and tracking algorithms. However, these objective algorithms are usually not perfect and can't achieve as good performance as human on the benchmarks, which probably affects the precision of quality assessment result. In this paper, we propose a full reference SVQA framework based on the distorted face recognition (DFR) task. The proposed framework can provide more authentic results than existing works because we learn the objective models to act consistently with human. Different from the conventional FR (cFR) [26] task, the DFR task is making a judgement on whether the distorted face can be still recognized when given its undistorted face as reference, which is illustrated in Fig. 1 (a). The reason why we define the DFR task for SVQA is that this practice can directly capture how much semantic information lost of faces in distorted surveillance videos when compared to reference videos. The core of the proposed SVQA framework is to learn an objective DFR model, which can be used for SVQA as shown in the bottom chart in Fig. 1 (b). To do that, we first construct a faces dataset from surveillance videos.

The construction of dataset should be based on real-world scenarios. Both the reference faces and distorted faces are included in the dataset, and the factors that affect the quality of surveillance videos are also introduced into the dataset for quantitatively analysis. In the dataset, for simplicity, we take video compression as the distortion maker. This is because in practice video compression is widely used to reduce the data volume of surveillance videos and make them more easily to store or transmit. So the distortion caused by compression can't be neglected. In addition, four factors are taken into account in the dataset which may affect SVQA: video codec selection, compression level, face resolution and light intensity. Based on the dataset, we conduct subjective experiments to collect subjective labels for the DFR task. They are taken as ground truth to learn the objective DFR model.

The methods for training the DFR model in the proposed SVQA framework can be freely chosen, and we pick two face recognition methods as examples in the experiments. In practice, users can take any reliable face recognition methods for their application purpose. The experiments show that the deep learning based face recognition method [21] achieves better performance on the DFR task than the traditional one [14].

Based on the learned DFR model, the quality of distorted surveillance videos can be assessed as shown in the bottom chart of Fig. 1 (b). We propose using the distorted face recognition rate (DFRR) as the quality metric for surveillance videos. Based on this metric, we analyze how different factors affect the quality measurement of surveillance videos, and further make comparisons with PSNR and SSIM.

In summary, our contributions are as follows. (1) We proposed a new full-reference SVQA framework based on the distorted face recognition task. (2) We construct a face dataset collected from real-world surveillance videos, which takes account of four important factors. And we also collect subjective labels for the DFR model learning. (3) Based on the experimental analysis, we give some suggestions for the practical applications of our proposed SVQA framework.

2 RELATED WORKS

2.1 Surveillance Video Quality Assessment

The widely used methods for SVQA are mostly introduced from traditional VQA field, such as PSNR, SSIM [20]. These methods were proposed from Quality of Experience (QoE) perspective, which are not suitable for task-based surveillance video. Video Quality in Public Safety (VQiPS) Working Group developed a guide for public safety that defines video quality requirements [4] in 2010. Under the framework proposed by VQiPS, Leszczuk et al. [9] presented a summary of definitions, research experiments and trends for quality assessment in surveillance video. The LPR task has been addressed in several works as the example case for SVQA. Leszczuk et al. [8] used a logistic model to show the dependency between human recognition rate and video bit rates. Janowski et al. [5] studied both human recognition and automatic LPR (ALPR) and found that human outperformed ALPR because ALPR software had lower recognition rate compared to human. Ukhanova et al. [17] conducted objective experiments on two video codecs: H.264 and H.265, and took logarithmic/logistic model to find the dependency between the ALPR recognition rates and bit rates/compression ratios respectively. Besides the LPR task, Korshunov and Ooi [6, 7] firstly used face images as the semantic carrier for surveillance video quality assessment. They proposed the concept "critical video quality" that can be used to reduce the bit rate without decreasing the performance of objective face detection, recognition and tracking algorithms.

2.2 Face Quality Assessment, Detection and Recognition

One similar research topic to face recognition based SVQA is face quality assessment, which measures the quality of a sequence of faces in videos with the aim of filtering the faces with bad quality to achieve good recognition performance in videos. One common way is extracting multiple features e.g. brightness first and combining them using weights to get the quality score [12, 24]. What's more, Wong et al. [22] proposed a patch-based method which quantifies the similarity of a face image to a probabilistic face model.

Figure 2: Subjective experiment interface with one example face group compressed by H264, where faces are arranged from left to right with ascending compression levels. The subjects need click the two threshold faces, and the backend will record the corresponding levels. The two faces with red box are label examples.

Face detection is often an essential step before the face recognition task. Traditional methods [11, 18] usually perform badly on the real-world surveillance videos. Recently, the cascaded CNN architecture was proposed in [10, 25], which gains an improved performance for surveillance videos .

Similarly, deep learning has remarkably pushed forward the progress of face recognition task too. Multiple methods have achieved better performance than human on the LFW [3] benchmark such as Face++ [1], DeepID3 [15] and Facenet [13]. However, to achieve the performance reported in the papers, it usually requires a large amount of data for model training. Wen et al. [21] proposed a new regularization term called center loss which improves the performance of sofmax classifier in some degree. The authors show that a simple CNN structure without much refinement can achieve over 99% accuracy on LFW benchmark with much less training data.

Besides deep-learning-based methods, traditional face recognition methods e.g. EigenFace [16], FV-Face [14] commonly extract hand-crafted features and train a non-linear classifier. These methods are usually time-consuming, but compared to deep learning methods they need much less training data.

3 DATASET CONSTRUCTION AND SUBJECTIVE EXPERIMENTS

In this section, we firstly describe the process of the SURveillance FACE (SURFACE) dataset construction. Then we introduce the subjective experiment which is conducted to collect subjective labels for the DFR task with different face resolutions, light conditions, video codecs and compression levels. Based on the experimental results, we quantitatively analyze the influence of these factors on the DFR task. In the next section, we will demonstrate how to use these subjective labels as the ground truth for learning objective distorted face recognition models.

3.1 Source Videos and Dataset Construction

A total of 6 source surveillance video sequences were captured from the real scenes of main road intersections and pavements in one big city, which are termed reference videos (RVs) here. The RVs are mainly captured at four periods with different light conditions: midday, afternoon, sunset and night. For daylight there are front lighting and back lighting conditions. For night the street lamps are turned on. The RVs record pedestrians walking toward or walking away from the cameras, and cameras are set at different distances from the pedestrians so that they can capture different resolutions of faces. In addition, all RVs were directly recorded from surveillance cameras without any post-processing. All the videos are in YUV420 format and have a resolution of 1920×1080 pixels with frame rate of 25 fps. The frame numbers of RVs range from 3650 to 5530.

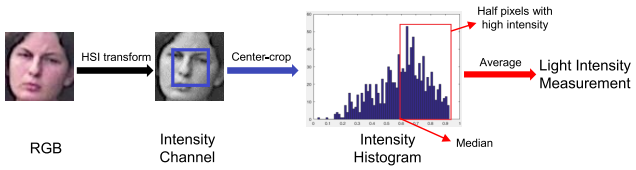


Figure 3: Method for measuring the light intensity of one face image.

Next, we conduct face detection to collect the frontal faces in the videos. The frontal faces are defined as faces with angle deviation within 45 degrees to the left, the right, the up and the down. We perform face detection on the Baidu Crowd Source Platform[1] which is similar to Amazon Turk to acquire more accurate face location and size than the automatic face detection results. Finally, we get 16452 frontal faces in total with the window size ranging from 30 to 104 pixels.

Light Intensity: Based on the detected faces, we calculate the light intensity for each RV. First, we transform the raw face images into HSI color space, and only keep the intensity channel. Then we crop center region of each intensity map with half window size, which guarantees that the cropped region always covers the face skin. Next find the median value of pixel intensity in the cropped region and take the average intensity of pixels whose intensity larger than the median as the measured light intensity for this face image. This process is shown in Fig. 3. At last, we average the light intensity of face images belonging to one RV as its measured light intensity. According to the ascending light intensity, we mark these 6 RVs as RV1 to RV6 and their measured light intensities are 0.381, 0.490, 0.513, 0.525, 0.620 and 0.642 respectively. The light intensity is in the range of 0 to 1.

Codec and Bit-rate: In this paper, we take video compression as the only causer of distortion. To explore how video compression affects the video quality, we generate distorted videos (DVs) based on RVs with multiple video codecs and a series of compression levels. We representatively picked three video codecs: H264, H265 and AVS2 [2], which are widely used for surveillance video compression. All three codecs we adopted in our experiments are their real-time versions[2]. Next, we aligned the compression levels across different video codecs. We predefine 17 levels of bit rates for each RV, then compress each RV with three video codecs based on these bit rates. The deviation of bit rates of three compressed videos at each level is restricted within the range of ±5%.

[1] http://zhongbao.baidu.com/
[2] H264 and H265 codecs are taken from FFmpeg (http://ffmpeg.org/).

Face Resolution: To quantitatively analyze how different face resolutions affect the DFR task, we representatively select the faces with the resolutions of 32×32, 48×48 and 64×64 pixels as examples, the numbers of which are 1340, 1282 and 1164 respectively. Then, we crop the corresponding compressed faces from DVs for each video codec. For convenience, we here define a concept: face group (FG), where each FG consists of one reference face from the RV and its 17 corresponding compressed faces from the DVs for one specific codec. That means one reference face corresponds to three FGs which are compressed by H264, H265 and AVS2 respectively . In addition, if the reference face in one FG is collected from RVx (x = 1, . . . , 6), we say this FG belongs to RVx for simplicity.

3.2 Subjective Experiments

According to the definition of the DFR task, we let the subjects make a judgment on whether one compressed face can still be recognized when given its corresponding uncompressed face as reference. The subjective experiment aims to test subjective performance on the DFR task.

Fig. 2 shows the interface for subjective experiments, where we arrange 18 faces in each FG according to their ascending compression levels and the reference face is treated as level 0. The subjects were asked to make comparison between the reference face and the compressed faces from left to right and label out two faces: the first is the rightmost face which can be recognized confidently, the second is the leftmost face which can not be recognized confidently. The reason for labeling two thresholds rather than one is that we found it hard for subjects to decide the only one threshold face with high confidence in practice. In other words, there is ambiguity for human to find the threshold face which distinguishes the recognizable faces and non-recognizable faces. We recorded the levels of the two faces for each FG, which are marked as L_1 ($Level1$)and L_2 ($Level2$) respectively.

In the DFR subjective tests, we recruited 7 subjects consisting of 5 males and 2 females, whose ages range from 20 to 25. All subjects have good corrected eyesight and have the research background of image processing. All tests were conducted on the same Matlab based interface as shown in Fig. 2 and each subject was shown all the selected FGs with three face resolutions and three video codecs.

3.3 Subjective Experimental Results Analysis

The outlier elimination is necessary and crucial in subjective experiments. We took the following actions to eliminate outliers in our experiments. For each FG,

(1) First, we deleted both the maximum and minimum values among 7 subjects' labels for L_1 and L_2.
(2) Next, we computed the variance of the rest labels for L_1 and L_2 respectively.
(3) If its sum of variance of L_1 and L_2 larger than 5, we deleted this FG and its corresponding two FGs which share the same reference face with it. This is to guarantee a fair comparison among three codecs in the following experiments.

The numbers of FGs left with resolution of 32×32, 48×48 and 64×64 pixels are 1031, 999 and 1000 respectively for each codec. Finally, we average the rest 5 subjects' labels for each FG with rounding down for L_1 and rounding up for L_2 as the final labels. Table 2 shows the quantity distribution of FGs belonging to different RVs.

Next, we try to quantitatively analyze how different factors, including light intensity, face resolution and codec selection, affect subjective performance on the DFR task.

We propose using the median bit-rate under the labeled compression level as the quantitative index. Assume that the RV, face resolution and codec are given, the median bit-rate of $L_1$is computed as follows,

$$M(BR_1) = median(\{BR(L_1^i)\}),\ i = 1, 2 \ldots, \|RV\| \tag{1}$$

where i is the index of FG belonging to one specific RV, $BR(L_1^i)$ denotes the corresponding bit-rate of L_1^i and $\|RV\|$ denotes the number of resolution-specific FGs belonging to RV. $M(BR_2)$ can be computed in the same way.

With the results shown in Table 1, we have the following observations. (1) Face resolution has an effect on subjective performance on the DFR task indeed. If considering on specific RV and codec, the higher resolution always corresponds to a lower median bit-rate, which means human vision has a higher tolerance for compression distortions when watching faces with higher resolution. (2) Light intensity also has an effect on the DFR task. Since the increasing light intensity of RV1 to RV6, the median bit-rate shows a roughly decreasing trend for one specific codec and resolution.

4 LEARNING DISTORTED FACE RECOGNITION MODEL

The core of the proposed framework is learning an objective model that can act like human on the DFR task. The objective face recognition methods we adopt in our experiments are initially proposed for the conventional face recognition task, but they still perform well for DFR task because of the good generalization ability. We use the collected labels from human in Sect. 3 as ground truth to learn the objective models, and evaluate its performance on the DFR task.

4.1 Selection of Face Recognition Methods

We adopt two face recognition methods for our SVQA framework representatively: Fisher Vector Face (FVF) [14] and Center-Loss Face (CLF) [21].

FVF is one state-of-the-art traditional face recognition method without using deep learning technology. Compared to deep learning based methods, FVF needs less training data, thus we can learn a new model from scratch using SURFACE dataset only. FVF method extracts fisher-vector features based on dense SIFT descriptors, then measures the similarity between two faces using the Mahalanobis distance. The source codes we adopt in experiments are taken from FVF project page[3]. CLF is one state-of-the-art face recognition method using Convolutional Neural Network (CNN). The authors release the pretrained model[4] which achieves a nearby 99% accuracy on LFW benchmark with CAISA-WebFace [23] database as training data only.

[3]http://www.robots.ox.ac.uk/~vgg/publications/2013/Simonyan13/
[4]https://github.com/ydwen/caffe-face

Table 1: The top and bottom table show the results of $M(BR_1)$ and $M(BR_2)$ respectively. There is no 32×32 FGs belonging to RV1 and RV2. The lower value indicates the better subjective quality.

$M(BR_1)$ $[Mb/s]$	H264			H265			AVS2		
	32×32	48×48	64×64	32×32	48×48	64×64	32×32	48×48	64×64
RV1	-	6.012	4.314	-	5.973	3.116	-	8.660	3.189
RV2	-	3.430	2.526	-	2.522	1.866	-	3.520	1.917
RV3	2.155	2.155	2.155	2.149	1.564	1.564	2.128	2.128	2.128
RV4	3.435	2.512	2.170	3.414	1.817	1.302	2.520	1.832	1.312
RV5	2.829	1.900	1.900	2.820	1.323	0.948	2.746	1.284	1.284
RV6	1.288	1.288	0.938	1.294	0.943	0.682	1.289	1.289	1.289
$M(BR_2)$ $[Mb/s]$	H264			H265			AVS2		
	32×32	48×48	64×64	32×32	48×48	64×64	32×32	48×48	64×64
RV1	-	1.630	1.184	-	1.174	0.844	-	2.279	1.195
RV2	-	0.975	0.975	-	0.590	0.590	-	1.000	0.494
RV3	0.579	0.579	0.579	0.574	0.403	0.403	0.801	0.566	0.566
RV4	0.932	0.932	0.656	0.924	0.647	0.506	0.930	0.652	0.452
RV5	0.690	0.499	0.499	0.490	0.336	0.336	0.660	0.343	0.343
RV6	0.481	0.481	0.335	0.486	0.486	0.486	0.485	0.485	0.339

Table 2: Quantity distribution of FGs belonging to different RVs.

Resolution	RV1	RV2	RV3	RV4	RV5	RV6
32×32	0	0	869	99	28	35
48×48	70	18	123	708	38	42
64×64	408	143	43	330	11	65

In our experiments, the fine-tuning and non-fine-tuning modes are both taken for CLF method. For fine-tuning mode, we fine tune the pretrained model released by the author with our own training data from SURFACE dataset. Based on the network architecture proposed in [21], we withdraw the last pooling layer and decrease the units number from 512 to 128 at the last fully-connected layer, and this makes it more adapted to our dataset. For non-fine-tuning mode, we directly test its performance on our dataset without making any modifications on the pretrained model. For simplicity, we let CLF-NFT denote the non-fine-tuned CLF model and CLF-FT denote the fine-tuned CLF model.

4.2 Training And Testing Details

For learning face recognition models it usually needs to construct positive pairs and negative pairs. According to the definition of DFR task, it should combine the reference face with the compressed faces whose levels are not larger than L_1 as positive pairs and with compressed faces whose levels are not smaller than L_2 as negative pairs in each FG. In practice this way can't generate enough data for training, so we take an augmentation approach. For one FG, we define the subset of faces whose levels are not larger than L_1 as $\mathbf{S_1}$, and the subset of faces whose levels are not smaller than L_2 as $\mathbf{S_2}$. Then we take all pairs of faces in $\mathbf{S_1}$ as positive pairs, which means any two faces in $\mathbf{S_1}$ can be recognized, and take all pairs, each of which contains one face of $\mathbf{S_1}$ and one face of $\mathbf{S_2}$, as negative pairs which means any face in $\mathbf{S_2}$ when compared to any face in $\mathbf{S_1}$ can't be recognized. It's noteworthy this augmentation approach only works for training, not for testing.

To test the generalization ability of trained models, we take the FGs belonging to RV5 and RV6 exclusively as testing data. And the rest are taken as training/fine-tuning data. This split is mainly due to the unbalanced quantity distribution of FGs belonging to different RVs as shown in Table 2. During training, we conduct 10-folds cross-validation to avoid overfitting.

4.3 Model Comparison and Performance Analysis

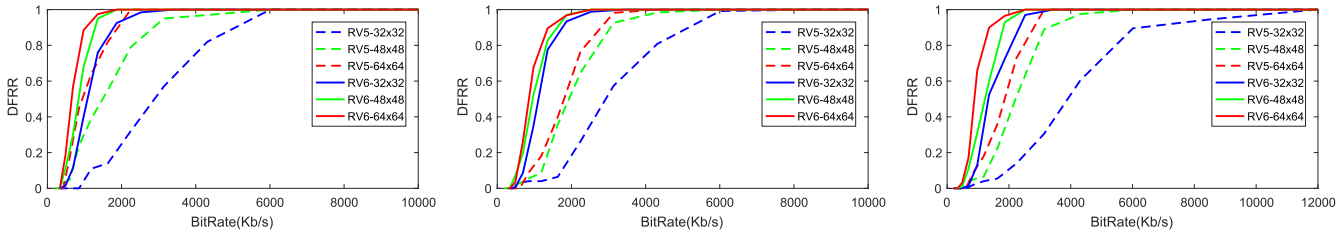
As emphasized above, we want the objective model to perform like human on the DFR task. So it's necessary to test the performance of objective models on the DFR task before using it for surveillance video quality assessment. A better performance on testing data means higher consistency between subjective and objective results. The performance is measured by the prediction accuracy defined as $N_{correct}/N_{test}$, where $N_{correct}$ and N_{test} represent the number of pairs predicted correctly and all test pairs respectively.

4.3.1 Single-Scale Models. For most conventional face recognition methods, the face images are resized into one certain size to learn the model. However, in DFR task, we assume that that face resolution would have an influence on the DFR task, and the subjective experiments have proven this assumption too, as shown in Sect. 3.3. So, firstly we train/fine-tune one model for each face resolution. Table 3 shows the prediction accuracies of different models, where FVF and CLF-FT-ind denote the models trained on one specific resolution, and CLF-NFT denotes the pretrained model without fine-tuning. It is easy to see that even without fine-tuning CLF method still performs better than FVF, which indicates the deep learning based face recognition method has a better generalization ability than the traditional one. Furthermore, the CLF-FT-ind achieves a slightly better performance than CLF-NFT.

4.3.2 Unified Model. Secondly, we try another way that interpolating all faces into a same size to learn a unified model. In Table 3 , we only report the result of fine-tuned CLF model (CLF-FT-one). Because of the bad performance of single-scale models of FVF compared to CLF, we didn't make further experiments on FVF method.

Table 3: The prediction accuracies of different objective methods and settings. The column termed "Overall" denotes the overall performance of three face resolutions. The bold types denote the best three results among different methods and settings for each tested codec. FVF: individual FVF models for different resolutions. CLF-NFT: CLF model without fine-tuning. CLF-FT-ind: individual fine-tuned CLF models for different resolutions. CLF-FT-one: one unified CLF model for different resolutions.

Training/ Fine-tuning Data	Method	H264 test				H265 test				AVS2 test			
		32 × 32	48 × 48	64 × 64	Overall	32 × 32	48 × 48	64 × 64	Overall	32 × 32	48 × 48	64 × 64	Overall
-	CLF-NFT	0.963	0.975	0.974	0.971	0.971	0.983	0.983	0.980	0.960	0.957	0.966	0.961
H264	FVF	0.805	0.975	0.929	0.911	-	-	-	-	-	-	-	-
	CLF-FT-ind	0.990	0.993	0.981	**0.988**	0.992	0.989	0.974	0.984	0.955	0.985	0.979	0.974
	CLF-FT-one	0.983	0.993	0.992	**0.990**	0.987	0.983	0.983	0.984	0.986	0.980	0.991	**0.986**
H265	FVF	-	-	-	-	0.872	0.911	0.944	0.912	-	-	-	-
	CLF-FT-ind	0.976	0.986	0.986	0.983	0.992	0.997	0.997	**0.996**	0.954	0.985	0.982	0.975
	CLF-FT-one	0.992	0.985	0.982	0.986	0.983	0.992	0.997	**0.991**	0.986	0.991	0.976	**0.984**
AVS2	FVF	-	-	-	-	-	-	-	-	0.851	0.929	0.887	0.892
	CLF-FT-ind	0.986	0.991	0.984	0.987	0.993	0.998	0.996	**0.996**	0.970	0.986	0.970	0.976
	CLF-FT-one	0.991	0.992	0.992	**0.991**	0.970	0.993	0.996	0.989	0.968	0.978	0.984	**0.978**



(a) CLF-FT-one (trained on H264) **(b) CLF-FT-one (trained on H265)** **(c) CLF-FT-one (trained on AVS2)**

Figure 4: The variation trend of DFRR with the increasing bit rates. The test data is in H264 format. (a), (b) and (c) show the testing results with CLF-FT-one model trained on the data in H264, H265 and AVS2 format respectively.

CLF-FT-one performs as well as CLF-FT-ind, which indicates combining different resolutions to learn one unified model will make no harm to its performance.

4.3.3 Single-Codec Models. In addition, one critical problem is how to train the FR models with different video codecs. We train individual models for different video codecs. But for each model, we test its performance on data compressed by all three video codecs. For example, the model is trained on the data compressed by H265, and we test its performance not only on H265 codec but also on the other two codecs. The first reason for this strategy is that we hope to compare the quality of videos compressed by different codecs with the same objective model. Another reason is that we want to explore whether different models trained on different video codecs showing different performance. As shown in Table 3, the model (for CLF-FT-ind and CLF-FT-one) trained on the data compressed by any one codec shows equally good performances on all three tested codecs. For example, the CLF-FT-one model trained on H265 not only perform well on H265 (0.991) but also achieve a good performance on H264 and AVS2 (0.986 and 0.984). This shows the fine-tuned model based on one specific codec will not constraint its good performance on this codec self, but gain a generalization ability for cross-codec testing.

In summary, CLF-FT-one model is the best choice for SVQA framework among the considered methods because of its convincing performance and convenience for dealing with multiple face resolutions with one unified model. So we take CLF-FT-one as the objective DFR model in our SVQA framework. Considering the similar performance, we take all three models trained on three codecs for further experiments.

5 OBJECTIVE SVQA

In this section, we take the objective DFR models for assessing the quality of surveillance videos. Meanwhile, we analyze how different factors e.g. light intensity, codec selection, affect the measured quality of compressed videos. A comparison to PSNR and SSIM is made to show the advantage of our method. Finally, we give some suggestions for the practical applications of our framework.

5.1 Surveillance Video Quality Assessment

First, we define a metric termed distorted face recognition rate (DFRR), which is taken as the quality measurement of compressed videos in our SVQA framework.

The DFRR is computed as follows. Given the pairs generated between one DV and its reference video RV, we conduct the DFR test on them using the objective model. If $pair_j$ can be recognized, its label l_j would be 1, otherwise be 0, where j is the index. Then,

$$DFRR = \frac{\sum_{j=1}^{\|pair\|} l_j}{\|pair\|} \tag{2}$$

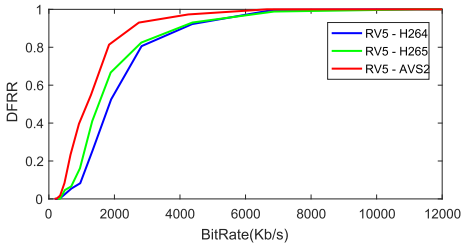
where $\|pair\|$ is the number of given face pairs. It's obvious that the higher DFRR score denotes the higher video quality.

Next, we conduct experiments that using the objective model for SVQA. We mainly test on the DVs corresponding to RV5 and RV6, which are taken exclusively as testing data as mentioned in Sect. 4.2.

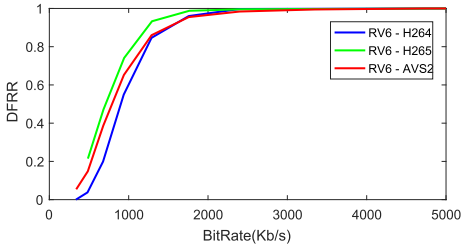
Firstly, we try to explore the relationship between DFRR and compression level, face resolution and light intensity respectively. In Fig. 4, we draw the BitRate-DFRR curve for each resolution and each testing RV, where, for example, RV5-32x32 can be seen as one separate video that share the same light intensity with RV5 but only have 32x32 faces. We mainly have the following observations. (1) In a certain range, a lower bit-rate corresponds to a lower DFRR score. This is an intuitive observation that one distorted video with large compression level often has a low quality. (2) For a low bit rate, the video with a higher face resolution corresponds to a higher DFRR score. (3) For one specific resolution, the video with high light intensity owns a high DFRR score, because RV6 is brighter than RV5. In addition, Fig. 4 (a), (b) and (c) show the test results of CLF-FT-one model trained on H264, H265 and AVS2 respectively, and they show the roughly similar trend. This indicates the CLF-FT-one model trained on one specific codec shows the good generalization ability for other codecs. Here we only show the results of testing data in H264 format.

Secondly, we explore the relationship between DFRR and codec selection. Since the faces in one video are usually with multiple resolutions, so we will calculate the DFRR score of one video with combining all three face resolutions. The testing results are shown in Fig. 5. It's obvious that for one specific low bit-rate ($< 2Mb/s$) there is a disparity of DFRR among different codecs. But there is an interesting phenomenon that the compressed videos of RV5 in AVS2 format have a higher DFRR than the videos in H265 and H264 format. But for RV6 the DFRR of compressed videos in H265 format is higher than the other two codecs. Because RV5 is darker than RV6, we conjecture that H265 works better (compressed videos have higher quality) on the videos with strong light intensity, and AVS2 works oppositely.

Since the similar results of CLF-FT-one models trained on three different codecs have been observed from Fig. 4, we show the test results based on the model trained on H265 in Fig. 5 for an example. And the models trained on the other two codecs show similar results.



(a) Test on DVs of RV5



(b) Test on DVs of RV6

Figure 5: The quality comparison of three video codecs. The two figures show the results for DVs of RV5 and RV6 respectively. All test results are based the CLF-FT-one model trained on H265.

5.2 Comparison to PSNR and SSIM

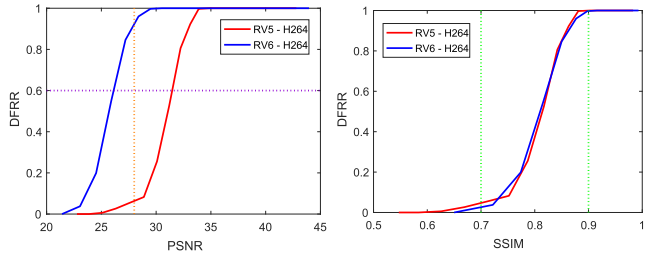
To show the advantages of our proposed method, we make a comparison between DFRR and the traditional VQA metrics: PSNR and SSIM.

Firstly, we conduct tests for all three methods on the DVs of RV5 and RV6, and draw the PSNR-DFRR curve in Fig. 6 (a), SSIM-DFRR curve in Fig. 6 (b). In Fig. 6 (a), there is a huge gap between the PSNR-DFRR curves of RV5 and RV6. Considering the orange dotted line, it's obvious that two DVs with the same PSNR measurement but correspond to vastly different DFRRs, which means taking PSNR as the metric to assess the quality of surveillance video could not reflect its true quality from the perspective of recognition. In Fig. 6 (b), the two SSIM-DFRR curves of RV5 and RV6 almost coincide, this indicates that SSIM is more robust to light intensity variations compared to PSNR. This can be explained by the involvement of light intensity in the computation of SSIM metric. However, it's easy to see while SSIM increasing from 0.7 to 0.9, the corresponding DFRR almost increase from 0 to 1. And in the ranges of SSIM<0.7 or SSIM>0.9, no matter how SSIM varies, the DFRR is almost saturated. This means SSIM metric is consistent with DFRR only within a certain range, and outside this range SSIM is not useful for SVQA from the perspective of face recognition.

In addition, we show some example faces in Fig. 7 through which we can find the differences between PSNR(SSIM) and objective recognition results intuitively. Fig. 7(a) shows two face pairs picked from RV5 and RV6 respectively. It's shown that the two distorted faces with similar PSNR values but the recognition results are opposite. In Fig. 7(b), it's shown that the distorted faces with a SSIM lower than a certain value would never be recognized. And oppositely, the distorted faces with a SSIM higher than a certain value would always be recognized.

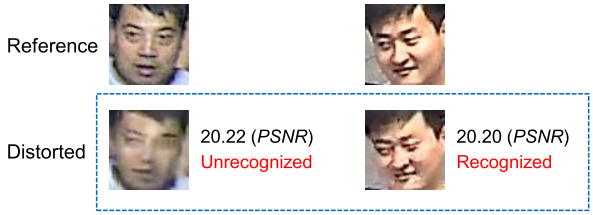
5.3 The Guide for Practical Applications

5.3.1 Model Selection. One important issue is which model we should use in practice. The experiments show that the fine-tuned CLF model performs better than the non-fine-tuned one. So, if possible, it's a better choice to use the fine-tuned model. Since we have verified that the fine-tuned model based on one specific codec shows good generalization ability for other codecs, there is no restriction for the selection of fine-tuning data. In addition, if the users have no additional data for fine-tuning, one compromising practice is directly taking the non-fine-tuned model for practical applications. Although the performance is slightly worse, we believe it's feasible for SVQA.

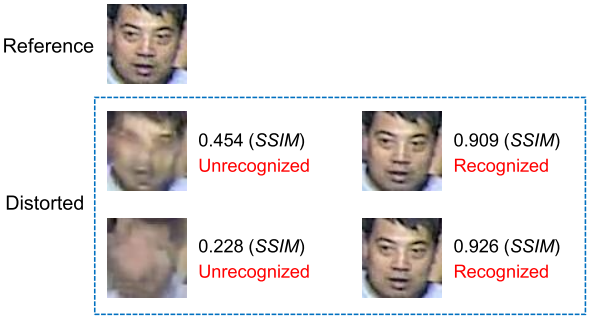


(a) PSNR vs. DFRR (b) SSIM vs. DFRR

Figure 6: (a) PSNR vs. DFRR (b) SSIM vs. DFRR. The test videos are in H264 format and the model we take is CLF-FT-one trained on H265. (Best viewed in color.)



(a) Comparison between PSNR and objective recognition result



(b) Comparison between SSIM and objective recognition result

Figure 7: Face examples for comparison between PSNR or SSIM and objective recognition results.

5.3.2 Deal with Multiple Face Resolutions. Another important issue is how to deal with different face resolutions in one video. In fact, the users can take all the faces in one video with any resolutions for its quality measurement. We have proved that one video with higher face resolutions always has a better quality. Meanwhile, the distribution of face resolutions in one video can be seen as its intrinsic property. So, in the same conditions, higher percentage of faces with large resolutions one video has, the better measured quality it gets.

5.3.3 An Application Scenario for QP Selection. In addition, based on our proposed framework, we propose an application scenario for QP selection in surveillance video compression, which is similar to the idea in [19]. Once the surveillance cameras are installed, the user can capture some short sequences at different time periods (noon, night etc. with different light conditions) or with different zoom factors, and compress them with a series QP values. Then, detect faces on the reference and compressed sequences with the state-of-the-art face detection algorithm e.g. MTCNN [25]. Next, calculate DFRR scores for all compressed videos as described in Subsection 5.1. Finally, the users can select the suitable QP value for each time period or zoom factor according to their acceptable DFRR score.

6 CONCLUSION

In this paper, we propose one objective SVQA framework from the perspective of face recognition. We try to keep the consistency between the objective methods and subjective labels. To do that, we first construct a face dataset named SURFACE which is collected from real-world surveillance videos considering multiple factors: video codec selection, compression level, light intensity and face resolution. With subjective experiments, we reveal how these factors affect subjective performance on the DFR task. Based on human labels on this dataset, we further learn one objective model as the core of the proposed SVQA framework. The experiments show that the framework has a better performance than PSNR and SSIM when measuring the quality of surveillance videos from the perspective of QoR. Finally, we give some suggestions for practical applications of the SVQA framework.

ACKNOWLEDGMENTS

This work was partially supported by National Basic Research Program of China (973 Program) under contract 2015CB351803 and the Natural Science Foundation of China under contracts 61572042, 61390514, 61421062, 61210005, 61527084.

REFERENCES

[1] Haoqiang Fan, Zhimin Cao, Yuning Jiang, Qi Yin, and Chinchilla Doudou. 2014. Learning deep face representation. *arXiv preprint arXiv:1403.2802* (2014).
[2] Zhichu He, Lu Yu, Xiaozhen Zheng, Siwei Ma, and Yun He. 2013. Framework of AVS2-video coding. In *Image Processing, 2013 20th IEEE International Conference on*. IEEE, 1515–1519.
[3] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* Technical Report 07-49. University of Massachusetts, Amherst.
[4] Video Quality in Public Safety Working Group et al. 2010. Defining video quality requirements: a guide for public safety. (2010).
[5] Lucjan Janowski, Piotr Kozłowski, Remigiusz Baran, Piotr Romaniak, Andrzej Glowacz, and Tomasz Rusc. 2014. Quality assessment for a visual and automatic license plate recognition. *Multimedia Tools and Applications* 68, 1 (2014), 23–40.
[6] Pavel Korshunov and Wei Tsang Ooi. 2005. Critical video quality for distributed automated video surveillance. In *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 151–160.
[7] Pavel Korshunov and Wei Tsang Ooi. 2011. Video quality for face detection, recognition, and tracking. *ACM Transactions on Multimedia Computing, Communications, and Applications* 7, 3 (2011), 14.
[8] Mikołaj Leszczuk, Lucjan Janowski, Piotr Romaniak, Andrzej Głowacz, and Ryszard Mirek. 2011. Quality assessment for a licence plate recognition task based on a video streamed in limited networking conditions. In *International Conference on Multimedia Communications, Services and Security*. Springer, 10–18.
[9] Mikolaj I Leszczuk, Irena Stange, and Carolyn Ford. 2011. Determining image quality requirements for recognition tasks in generalized public safety video applications: Definitions, testing, standardization, and current trends. In *Broadband Multimedia Systems and Broadcasting, IEEE International Symposium on*. IEEE, 1–5.
[10] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. 2015. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5325–5334.
[11] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. 2014. Face detection without bells and whistles. In *European Conference on Computer Vision*. Springer, 720–735.
[12] Ramachandra Raghavendra, Kiran B Raja, Bian Yang, and Christoph Busch. 2014. Automatic face quality assessment from video using gray level co-occurrence

matrix: an empirical study on automatic border control system. In *Pattern Recognition, 2014 22nd International Conference on*. IEEE, 438–443.

[13] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 815–823.

[14] Karen Simonyan, Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2013. Fisher Vector Faces in the Wild. In *British Machine Vision Conference*, Vol. 2. 4.

[15] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. 2015. DeepID3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873* (2015).

[16] Matthew A Turk and Alex P Pentland. 1991. Face recognition using eigenfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 586–591.

[17] Anna Ukhanova, Jesper Støttrup-Andersen, Søren Forchhammer, and John Madsen. 2014. Quality assessment of compressed video for automatic license plate recognition. In *Computer Vision Theory and Applications, 2014 International Conference on*, Vol. 3. IEEE, 306–313.

[18] Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1. IEEE, I–511.

[19] Dandan Wang, Dong Liu, and Fangdong Chen. 2015. Image semantic quality assessment for compression of car-plate images. In *Visual Communications and Image Processing*. IEEE, 1–4.

[20] Zhou Wang, Ligang Lu, and Alan C Bovik. 2004. Video quality assessment based on structural distortion measurement. *Signal processing: Image communication* 19, 2 (2004), 121–132.

[21] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A Discriminative Feature Learning Approach for Deep Face Recognition. In *European Conference on Computer Vision*. Springer, 499–515.

[22] Yongkang Wong, Shaokang Chen, Sandra Mau, Conrad Sanderson, and Brian C Lovell. 2011. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 74–81.

[23] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923* (2014).

[24] Guangpeng Zhang and Yunhong Wang. 2009. Asymmetry-based quality assessment of face images. In *International Symposium on Visual Computing*. Springer, 499–508.

[25] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

[26] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. 2003. Face recognition: A literature survey. *ACM computing surveys* 35, 4 (2003), 399–458.