NOVEL ANALYTICAL MODELS OF FACE RECOGNITION ACCURACY IN TERMS OF VIDEO CAPTURING AND ENCODING PARAMETERS

Hayder R. Hamandi and Nabil J. Sarhan

Electrical and Computer Engineering
Wayne State University
Detroit, Michigan 48202
hayder@wayne.edu, nabil.sarhan@wayne.edu

ABSTRACT

To fit the tight resource constraints, including network bandwidth, the video streams in Computer Vision systems are adapted dynamically by changing the video capturing and encoding parameters. We propose two novel analytical models that characterize the face recognition accuracy in terms of these parameters, specifically resolution, quantization, and actual bitrate. We find that the accuracy is a logistic function of the video quantization parameter, with the value of the Sigmoid's midpoint being a function of the resolution. Alternatively, we find that the accuracy is equal to the sum of two exponentials of the actual video bitrate, with the resolution as a multiplicative factor with one exponential. We develop an evaluation framework to validate the models using two distinct video datasets with 99 videos and the widely used Labeled Faces in the Wild (LFW) dataset with $13,233$ images. We conduct $1,668$ experiments that involve varying combinations of encoding parameters. We show that both models hold true for the deep-learning and statistical-based face recognition. The developed models achieve an average coefficient of determination (R^2) of 98.7% to 99.8%.

Index Terms— Analytical Modeling, Computer Vision Accuracy, Face Recognition, Video Rate Adaptation.

1. INTRODUCTION

The video streams in Computer Vision (CV) systems are adapted dynamically by changing the video capturing and encoding parameters to fit the tight resource constraints, including network bandwidth, energy, and storage. Therefore, these adaptations lead to various tradeoffs involving the accuracy and the aforementioned constraints. The overwhelming majority of studies on CV focused on the development of robust algorithms to improve the accuracy in primarily image datasets, through statistical and deep learning approaches.

We mathematically model and analyze CV accuracy, focusing primarily on face recognition, which is used in many applications, including authentication systems, personal photo enhancement, video surveillance, and photo search engines. We make **a fundamental contribution** by developing two novel analytical models that help in assessing the effect of combining adaptation strategies for the same video stream. The first model, called *Quantization Parameter and Resolution based MODel for Accuracy* (QRMODA), characterizes CV accuracy in terms of the spatial resolution and Quantization Parameter (Q_p) as a logistic function of Q_p, with the x_0 value of the Sigmoid's midpoint being a function of the resolution. In contrast, the second model, called *Bitrate and Resolution based MODel for Accuracy* (BRMODA), shows the accuracy in terms of the spatial resolution and actual bitrate, as the sum of two exponentials of the *actual* bitrate, with the resolution as a multiplicative factor with one exponential.

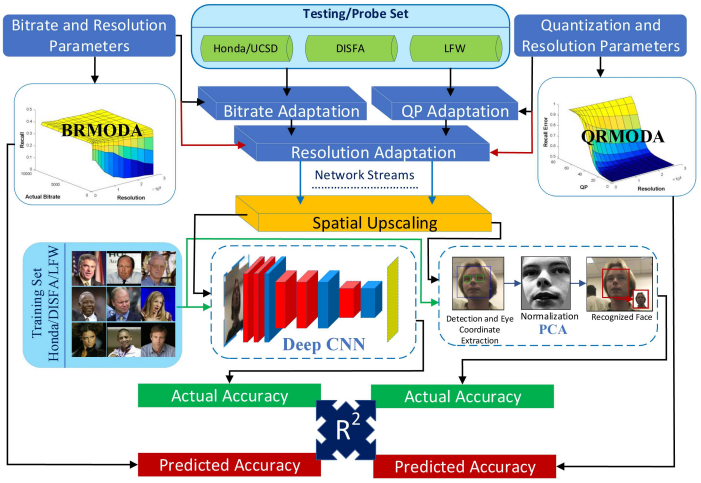


Fig. 1. Evaluation Framework for Model Validation

Moreover, we develop an *evaluation framework*, as illustrated in Fig. 1 and detailed in Section 4, to validate each model against deep-learning and statistical-based approaches of face recognition, utilizing two greatly distinct video datasets (Honda/UCSD [1] and DISFA [2]), and a large image dataset (Labeled Faces in the Wild (LFW) [3]). We conduct $1,668$ actual experiments on 99 videos and $13,233$

978-1-7281-1331-9/20/$31.00

images, with 47 and 5,749 subjects, respectively. Subjects have different gender, ethnicity, and pose variations. The results indicate that both proposed models hold true for both face recognition approaches and using different datasets. The results also show that the models can characterize face detection. Moreover, we assess the goodness of fit using the coefficient of determination (R^2). Finally, we discuss the factors impacting the constants of each proposed model and how to compute them.

The rest of this paper is organized as follows. Section 2 discusses the related work. Section 3 shows the model development. Subsequently, Section 4 explains the evaluation framework and experimental setup, and Section 5 presents and analyzes the main results. Section 6 provides additional discussion and analysis of the factors impacting the constants of both proposed models. Finally, conclusions are drawn.

2. RELATED WORK

Due to its non-intrusive nature, face recognition has gained wider acceptance than other biometric modalities [4]. Face recognition approaches can be categorized into two main categories: *holistic* and *deep learning*. The first relies on statistical analysis, whereas the latter employs Convolutional Neural Networks (CNNs). Prominent examples include Principal Component Analysis (PCA) for the first approach and ArcFace [5], SphereFace [6], and FaceNet [7] for the latter. We validate our proposed models using both approaches.

The overwhelming majority of research on face recognition considered the development of robust algorithms to improve accuracy in image datasets [5] [6] [7]. Some studies explored the impact of neural network design parameters on accuracy [8], while others [9] addressed the effects of facial expression, gender, illumination, and/or occlusion.

3. DEVELOPMENT OF THE PROPOSED MODELS

We develop two novel analytical models of the CV (primarily face recognition) accuracy in terms of the video capturing and encoding parameters. The first model characterizes face recognition accuracy with respect to variations in Q_p and resolution, and thus we refer to it as QRMODA (Q_p and Resolution based MODel for Accuracy). Similarly, we develop another model for accuracy with respect to the actual bitrate and resolution, and we refer to it as BRMODA (Bitrate and Resolution based MODel for Accuracy). Subsequently, we show that both models apply to face detection as well, but with different constant values.

The models help in analyzing the effectiveness of combining video adaptation strategies in terms of the CV accuracy. We consider adapting the video streams by changing both the spatial resolution and the Signal-to-Noise Ratio (SNR). We utilize a super-resolution algorithm to upscale the videos to their original resolutions before the analysis at the destinations in order to boost the accuracy. For the SNR adaptation, we consider both changing the target bitrate and Q_p. We do not consider temporal adaptation as missing frames will trivially lead to zero detection and therefore no recognition.

Table 1. Summary of Used Notations

Notation	Explanation
$\mathcal{E}$	Recall Error
T_P	True Positive
F_N	False Negative
Q_p	Quantization Parameter
r	Actual Video Bitrate
$N \times M$	Video Resolution
$f_{logistic}(x)$	The Logistic Function
c_1, c_2, c_3, c_4, and c_5	QRMODA Constants
c'_1, c'_2, c'_3, c'_4, and c'_5	BRMODA Constants
R^2	Coefficient of Determination

The recall captures the sensitivity of the system. Given a video of k frames, the recall error $\mathcal{E}$ for the entire video can be given by $\mathcal{E}=1-(\sum_{i=1}^{k} T_P(i)/\sum_{i=1}^{k}(T_P(i)+F_N(i)))$, where $T_P(i)$ and $F_N(i)$ are the numbers of correctly and erroneously identified faces in frame i. Table 1 summarizes the notations used in the paper. We focus on the recall because a positively classified (False Positive) face is not a catastrophic issue, whereas an overlooked face (false negative) that should have been flagged as positive may compromise security.

3.1. QRMODA

Since video adaptation is imposed due to network and resource limitations, we expect the CV accuracy to suffer starvation beyond a certain Q_p. However, due to a simultaneous independent adaptation in video resolution, a compensation for the accuracy loss will be granted if the video adapts to a higher resolution. Our empirical data, discussed in Section 5, indicate that $\mathcal{E}$ follows an exponential trend with respect to changes in resolution and a *bounded exponential* bias towards Q_p variations. Hence, we determine that $\mathcal{E}$ is a function that combines the characteristics of both (exponential and bounded exponential) functions, namely the *logistic function*. We find that $\mathcal{E}$ is a logistic function of Q_p with the x-axis of the Sigmoid's midpoint (x_0) being a function of spatial resolution. Specifically, given a video with a resolution of $N \times M$, quantized at Q_p, $\mathcal{E}$ can be characterized as

$$\mathcal{E}_{QRMODA}(Q_p, N, M) = f_{logistic}(x = Q_p, x_0 = c_1(N \times M)^{c_2}) + c_3, \quad (1)$$

where

$$f_{logistic}(x) = \frac{c_4}{1 + e^{c_5(x - x_0)}}.$$

We introduce c_3 as a bias to the model. This value defines the lowest achievable recall error (i.e. at the original resolution with no quantization). Constants c_1 through c_5 vary with

different factors, as discussed in Section 6. c_1 and c_2 define the sharpness in the change of the Sigmoid slope and impact the model's trend with respect to variations in only the spatial resolution. Small values (less than 1) result in a smooth transition in recall (with a slope of around $80°$, depending on the value of c_5) that is slightly affected by spatial adaptation. Contrarily, values greater than 1 result in a sharp transition in recall as more quantization is imposed (especially with low resolution). As the resolution is increased, the recall transition flattens. Constant c_4 determines the maximum value of the logistic function without the bias. Specifically, $(c_3 + c_4)$ determines the lowest recall rate, regardless of adaptation variations. Lastly, c_5 determines the logistic growth rate (steepness of the curve) and is always negative as recall error increases with quantization.

3.2. BRMODA

Videos with low resolutions tend to produce low bitrates when high target bitrates are imposed. Likewise, videos with high resolutions tend to produce higher bitrates than the imposed target values. Low bitrate videos have lower recall rates due to reduction in video quality. As higher bitrates are granted, the video quality increases, thereby causing the recall error to drop drastically. Our empirical results show an exponential relationship. We determine that $\mathcal{E}$ is a function of two exponentials of the *actual* bitrate, with the number of pixels in the frame being a multiplicative factor with one of the exponentials. Given an $N \times M$ resolution video with an actual bitrate r, $\mathcal{E}$ can be given by

$$\mathcal{E}_{BRMODA}(r, N, M) = c_1'(N \times M)^{c_2'} e^{c_3' r} + c_4' e^{c_5' r}, \quad (2)$$

where c_1' through c_5' are constants. This model uses the value of the actual bitrate because the target bitrate may not be achieved precisely by the encoder. Constants c_1' and c_2' are similar in purpose to their counterparts in QRMODA. They define the steepness of the exponential drop with respect to spatial resolution variation. c_3' is always negative because $\mathcal{E}$ is inversely proportional to the actual achieved bitrate. In other words, high-resolution videos require high bitrates, and thus produce high recall errors when low bitrates are imposed. c_4' and c_5' control the bias exponential.

4. EXPERIMENTAL SETUP

Fig. 1 shows the setup and evaluation framework.

4.1. Used Datasets

We utilize two greatly distinct video datasets: Honda/UCSD [1], and DISFA [2]. The former is a standard dataset for the evaluation of face detection, tracking, and recognition algorithms. The latter is used to study *Facial Action Coding Systems* (FACS). Honda/UCSD has lower quality videos, thereby serving as an example of limited-bandwidth network systems. In contrast, DISFA has High Definition (HD) quality videos. In addition, the subjects in Honda/UCSD make different combinations of 2-D and 3-D head rotations and have different facial expressions with varying speed. Conversely, subjects in DISFA have limited pose variations, but great variations in facial action expressions. Furthermore, we utilize a large image dataset: LFW [3]. This dataset aims at studying the problem of unconstrained face recognition. The main properties of the used datasets are summarized in Table 2. We divide each dataset into three main sets: *Training*, *Validation*, and *Testing*. In Honda/UCSD, we use the first included dataset, which is already categorized into 3 groups: Training, Testing, and Testing with Partial Occlusion. We use the latter for validation. Contrarily in DISFA, we split the right camera videos to training and validation sets and use the left camera videos for testing. For LFW, we use the split method suggested by [3]. The adaptation is performed only on the testing sets to avoid overfitting and selection bias towards adapted frames.

Table 2. Characteristics of the Used Datasets

	Honda/UCSD	DISFA	LFW
Camera	SONY EVI-D30	PtGrey stereo	Varies
Resolution	640×480	1024×768	250×250
Frame Rate	15 frame/sec	20 frame/sec	N/A
Format	Uncompressed AVI	Uncompressed AVI	JPEG
Subjects	20 (2 females and 18 males)	27 (12 females and 15 males)	$5,749$
Size	45 videos	54 videos	$13,233$ images

4.2. Video Adaptation Generation

We perform H.264 encoding on the videos/images of all the testing sets using FFmpeg. We generate two sets of adapted videos to analyze the adaptation impact on CV accuracy. The first includes videos with combined resolution and target bitrate adaptations, whereas the second includes videos that have a combination of Q_p and resolution adaptations. We use the Lanczos algorithm to upscale the videos, as it provides the best tradeoff in performance and execution time [10]. For LFW, only a combination of Q_p and resolution is used because bitrate adaptation is inapplicable to images.

4.3. Face Detection and Recognition Implementations

We use **CNN-based** face detection and recognition, utilizing a modified FaceNet [7] implementation using the Inception-ResNet architecture [11] as the deep learning platform. We develop an interface that interacts with FaceNet to perform normalization, CNN training, face detection, and face recognition on the entire video frames rather than images. The experiments start by organizing all training frames in a tree-like

fashion such that each subject maintains its own directory of the respective frames. These frames are then aligned, maintaining the same directory structure. The aligned frames are then used to train the fully-connected layers of the deep CNN, generating a classifier model for use by the recognition module. Subsequently, we fine tune this model through validation.

We use the adapted testing set videos as an input to the CNN and detect the faces in every frame of those videos using FaceNet. We employ the aforementioned classifier model to classify each frame. The result of this step is a list of probabilities for each probe with respective classes. We pick the class with the highest probability and consider it as the best candidate identifying the probe (Top-1 Class). Finally, we collect a *confusion matrix* to compute the overall recall.

In the **statistical-based** approach, we develop a face detector using the Viola-Jones algorithm. We develop a platform for extracting eye coordinates from all faces. Since eye classifiers are not accurate and may return falsely detected eyes, we develop a mechanism to filter true eyes based on their sagittal coordinates. The eye coordinates are vitally important for recognition because they represent input parameters for the preprocessing steps, including geometric normalization, histogram equalization, and masking. We utilize the *CSU Face Identification Evaluation System* [12] to perform training and face recognition. We employ PCA because of its effectiveness in generating simpler representations of the huge video dataset with all adaptations.

5. MODEL VALIDATION AND ANALYSIS

5.1. Baselines and Evaluation Metrics

We use FaceNet and PCA as the baselines for face recognition to benchmark the validity of BRMODA and QRMODA. We also employ the Viola-Jones algorithm as the baseline for face detection. In contrast with the methods used by [7] and [12] to perform image analysis, we develop interfaces to work with adapted video frames from the Honda/UCSD and DISFA datasets. We utilize R^2 to assess the goodness of fit of the proposed models and use recall as the accuracy metrics. R^2 values are shown with each figure caption.

5.2. Result Presentation and Analysis

For each analytical model, we show the results for the experiments performed using the deep-learning and statistical methods for face detection and recognition, utilizing three different datasets for QRMODA and two video datasets for BRMODA (since bitrate adaptation is not applicable to the images). We generally present subfigures that demonstrate the extremes of the resolution adaptations considered. Table 3 lists some of the constants used in this study.

We validate QRMODA in terms of detection and recognition sensitivity at different spatial resolutions. Fig. 2 shows the recall error with respect to Q_p variations for the Honda/UCSD dataset. In Subfigures. 2(a) and 2(b), FaceNet is used for both face detection and recognition tasks, whereas Viola-Jones and PCA are used in Subfigures. 2(c) and 2(d). These results demonstrate QRMODA's robustness to a change in the detection/recognition method. Fig. 4 further validates QRMODA using the DISFA dataset.

Table 3. List of Constants for QRMODA/BRMODA [Deep Learning]

Const.	Honda/UCSD [Detect., Recogn.]	DISFA [Detect., Recogn.]
c_1	$17.98, 24.03$	$0.7, 1.54$
c_2	$0.08493, 0.05211$	$1.255, 1.121$
c_3	$0.5, 0.61$	$0.003, 0.003$
c_4	$0.5, 0.3838$	$0.039, 0.5913$
c_5	$-0.2, -0.2864$	$-0.4, -0.517$
c_1'	$0.414, 0.0363$	$2.64 \times 10^{-4}, 1.867 \times 10^{-6}$
c_2'	$0.175, 0.292$	$0.65, 1.02$
c_3'	$-0.126, -0.054$	$-0.2, -0.117$
c_4'	$0.174, 0.273$	$0.0229, 0.06102$
c_5'	$-7.97 \times 10^{-6}, -4.718 \times 10^{-6}$	$-4.8 \times 10^{-6}, -3.03 \times 10^{-6}$

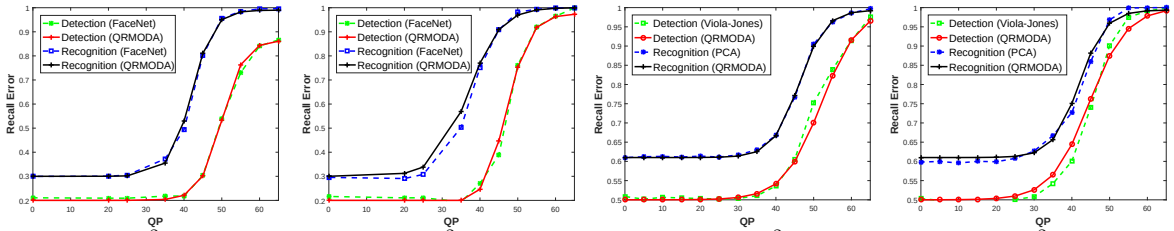
We also validate QRMODA using the LFW image dataset. The results, reported in Table 4, compares QRMODA's predicted accuracy to the actual accuracy achieved by FaceNet at different combinations of resolution and Q_p.

Table 4. Validation on LFW

Resolution (Pixels)	Q_p	Avg. Image Size (KB)	Actual (FaceNet)	Predicted (QRMODA)
$40,000$	2	9.33	0.987	0.988
$40,000$	10	3.82	0.98	0.9787
$40,000$	25	2.64	0.954	0.952
$40,000$	31	2.41	0.939	0.941
$22,500$	2	6.73	0.939	0.954
$22,500$	10	2.87	0.938	0.953
$22,500$	25	1.91	0.927	0.928
$22,500$	31	1.73	0.9	0.896
$10,000$	2	4.31	0.961	0.959
$10,000$	10	1.94	0.938	0.937
$10,000$	25	1.34	0.839	0.83
$10,000$	31	1.25	0.779	0.788
$2,500$	2	2.05	0.759	0.76
$2,500$	10	1.13	0.682	0.681
$2,500$	25	0.87	0.57	0.574
$2,500$	31	0.84	0.57	0.566

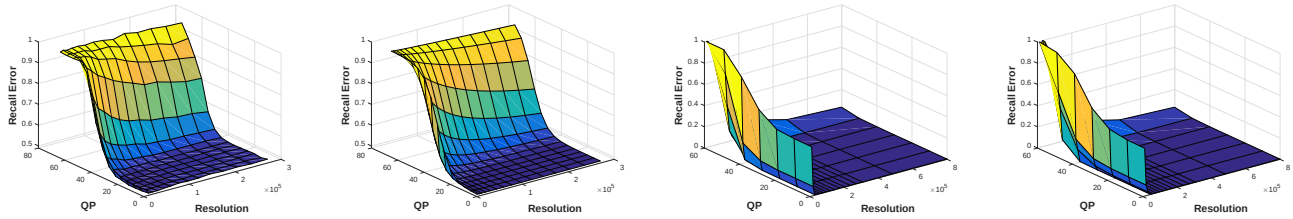
Additionally, we demonstrate QRMODA's behavior with respect to simultaneous variations in both Q_p and resolution, represented by x and y axes, respectively in Fig. 3. The recall error is color-coded over the z-axis. The results demonstrate that QRMODA is highly accurate regardless of the used dataset.

The recall error increases slowly with Q_p up to a critical point, represented by the Sigmoid's midpoint of the Logistic function. After that point, the error increases sharply with Q_p until it becomes 100%. The lower bound for face detection



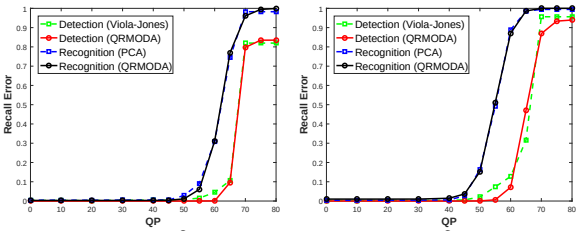
(a) 600×450 (R^2: 0.992, 0.995) (b) 280×210 (R^2: 0.982, 0.998) (c) 600×450 (R^2: 0.9962, 0.999) (d) 240×180 (R^2: 0.973, 0.962)

Fig. 2. Validation of QRMODA using Honda/UCSD Dataset



(a) FaceNet (HONDA/UCSD) (b) QRMODA (HONDA/UCSD) (c) FaceNet (DISFA) (d) QRMODA (DISFA)

Fig. 3. Validation and Analysis of QRMODA



(a) 480×360 (R^2: 0.9968, 0.998) (b) 280×210 (R^2: 0.895, 0.995)

Fig. 4. Validation of QRMODA using DISFA Dataset

error is about 0.2 for the Honda/UCSD and approximately 0 for the DISFA. Additionally, both the detection and recognition recall rates in DISFA are much higher than those in Honda/UCSD. This difference in threshold levels is because of the variation in video contents, which contain fewer frontal face poses in the Honda/UCSD than those in DISFA.

Fig. 5 validates BRMODA using deep learning and statistical methods. Each subfigure shows the normalized recall error versus the actual bitrate for selected resolutions. As the target bitrates may not be achieved precisely by the encoder, we report the actual bitrates, which are depicted in the figures on a logarithmic scale because of the wide range of considered bitrates in our experiments. The results demonstrate that the model is highly accurate in terms of R^2.

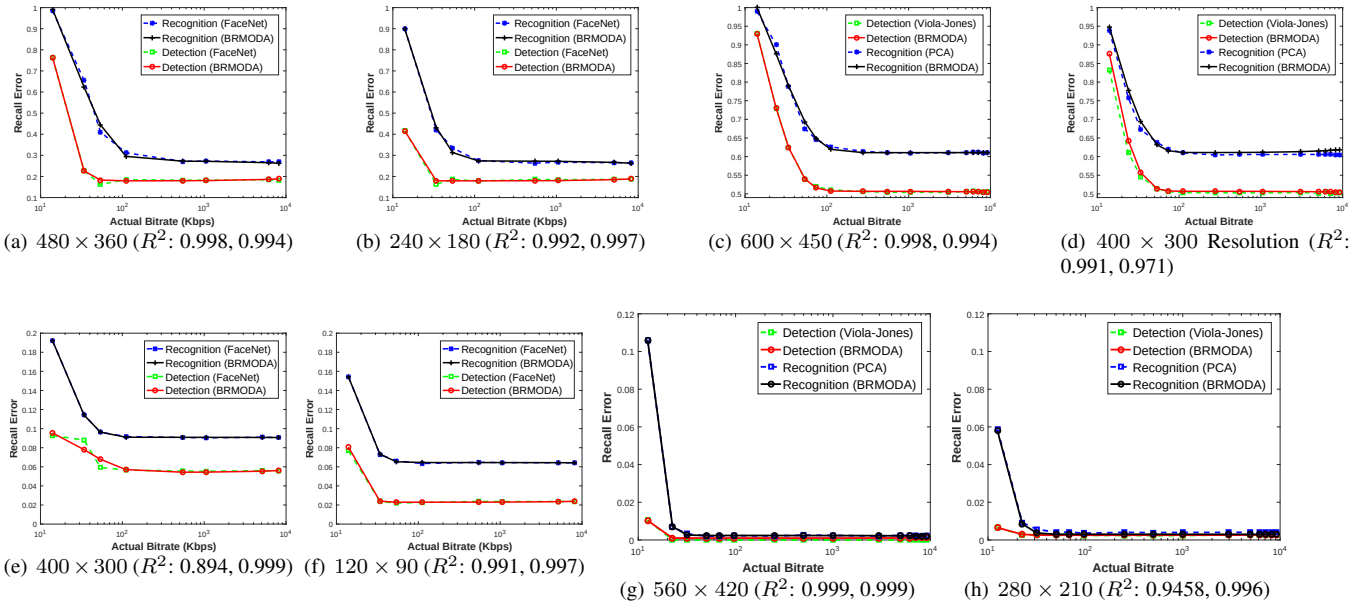
The recall is inversely proportional to the actual bitrate achieved due to the negative value of c_3'. This behavior varies with spatial resolution variation because high resolution videos require high bitrates, thereby producing high errors when low bitrates are imposed.

6. DISCUSSION

The model constant values can be determined upon system calibration, preferably based on the videos captured by the cameras in the actual deployment. The system can generate different adaptations and then determine the constants that best fit the model(s) using for example *monotone regression splines*. As indicated by the widely varying datasets, the constants capture different factors, including the subject's pose angle, inter-ocular distance in pixels, facial expressions, lighting, and potential occlusion.

7. CONCLUSIONS

We have proposed two novel analytical models that characterize the CV accuracy and have validated them using a developed evaluation framework. Both models demonstrate an average R^2 of 98.7% on the video datasets. On LFW, our QRMODA model shows an average R^2 of 99.8%. We find it remarkable that the two models apply to distinct video/image datasets and to both face recognition and detection. The models also apply to both deep-learning and statistical-based methods.



(a) 480×360 (R^2: 0.998, 0.994) (b) 240×180 (R^2: 0.992, 0.997) (c) 600×450 (R^2: 0.998, 0.994) (d) 400×300 Resolution (R^2: 0.991, 0.971)

(e) 400×300 (R^2: 0.894, 0.999) (f) 120×90 (R^2: 0.991, 0.997) (g) 560×420 (R^2: 0.999, 0.999) (h) 280×210 (R^2: 0.9458, 0.996)

Fig. 5. Validation and Analysis of BRMODA [5(a)-5(d): using Honda/UCSD, 5(e)-5(h): using DISFA]

8. REFERENCES

[1] Kuang-Chih Lee, Jeffrey Ho, Ming-Hsuan Yang, and David Kriegman, "Visual tracking and recognition using probabilistic appearance manifolds," *Journal of Computer Vision and Image Understanding*, vol. 99, no. 3, pp. 303–331, 2005.

[2] Mohammad Mavadati, Mohammad Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Trans. on Affective Computing*, vol. 4, no. 2, pp. 151–160, April 2013.

[3] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Proc. of Work. on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[4] Daniel Sáez Trigueros, Li Meng, and Margaret Hartnett, "Face recognition: From traditional to deep learning methods," *arXiv preprint arXiv:1811.00116*, 2018.

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699.

[6] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 1, p. 1.

[7] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[8] Soheil Hashemi, Nicholas Anthony, Hokchhay Tann, R Iris Bahar, and Sherief Reda, "Understanding the impact of precision quantization on the accuracy and energy of neural networks," in *Proc. of IEEE Design, Autom. & Test in Europe Conf. & Ex. (DATE)*, 2017, pp. 1474–1479.

[9] Milos Pavlovic, Ranko Petrovic, Branka Stojanovic, and Srdan Stankovic, "Facial expression and lighting conditions influence on face recognition performance," in *Proc. of IcETRAN conf.*, 2018.

[10] Yousef Sharrab and Nabil Sarhan, "Accuracy and power consumption tradeoffs in video rate adaptation for cv applications," in *Proc. of IEEE Int'l Conf. on Multimedia and Expo (ICME)*, July 2012, pp. 410–415.

[11] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning.," in *Proc. of AAAI Conf. on Artificial Intelligence*, 2017, pp. 4278–4284.

[12] Ross Beveridge, David Bolme, Bruce A. Draper, and Marcio Teixeira, "The CSU face identification evaluation system," *Machine Vision and Applications*, vol. 16, no. 2, pp. 128–138, February 2005.