Face image quality for actor profile image curation

Yash Pandya, Abhinav Aggarwal, Manivel Sethu, Laxmi Ahire, and Kaustav Nandy
Amazon, India
(yaspan, aggarw, mssethu, ahilaxmi, kaustn)@amazon.com

Abstract

Selecting an ideal profile image to represent a person is a common problem with many applications. The ideal characteristics of a representative or profile image differ based on the application. In this work, we focus on selecting a representative face which is easy to recognise and aesthetically pleasing. Manually curating these images is time consuming, repetitive, and subjective. This makes the quality of curated images inconsistent. We have built a solution to automate this process and make it efficient, scalable, and consistent. In this work, we describe the various factors which affect the suitability of a face image for recognition by humans. We propose efficient solutions which can solve the problem without the use of ground truth data. We train a regression model using weak supervision provided by heuristics based on features which affect face quality. Finally, we use professional photography techniques to create standardized and aesthetically pleasing profile images.

1. Introduction

Prime Video (PV) is a streaming service by Amazon used by hundreds of millions of customers worldwide for both video on demand and live sports. X-Ray on PV enhances the viewing experience of customers by engaging their curiosity to explore and learn during playback about the actors in a scene, music, trivia, and dive into behind-the-scenes content. The actor profile headshot, name and other metadata is surfaced to customers via the X-Ray UI within playback (example in Figure 1). Providing this X-Ray experience on PV is a challenging scaling problem because the metadata needs to be generated accurately for a larger number of titles every year across multiple languages to maintain customer experience. Manual annotation process for generating this metadata takes many hours of effort for every hour of content and does not scale for an ever growing video catalog at PV scale.

Using actor headshots of the cast list available in IMDb is helpful for actor identification. However with the growing number of titles in local languages with actors who are not globally popular, this is not useful for all titles. We also discovered that profile images in IMDb vary wildly in quality and lesser known actors or titles (non english languages) have poor quality profile images. This led to the development of a solution [2] which can be used for titles with no headshots available for the actors. This solution is based on clustering all the faces present in a title and then getting a small number of faces manually annotated to predict the presence of actors across the title. This Human-in-the-loop approach reduced the effort required for metadata generation by a significant percentage. Despite this reduction in manual effort, there are still open problems of determining the quality of the extracted actor headshots and automatically extracting the best image from the entire video in such a way that it looks similar to a professional headshot for PV X-Ray customers. Manual process for such extraction is time consuming and presents additional problems like lack of standardisation across operators performing the annotations, leading to inconsistencies in headshot image parameters and quality, thereby resulting in poor customer experience. To address these open problems and drive towards full automation, in this paper we focus on selecting a representative face for a group of faces. We look at determination of the quality of a face image for recognition by both ML algorithms and human viewers. This is a challenging problem as there is no prior work which we can use directly. There is no ground truth data for this problem as well. Creating

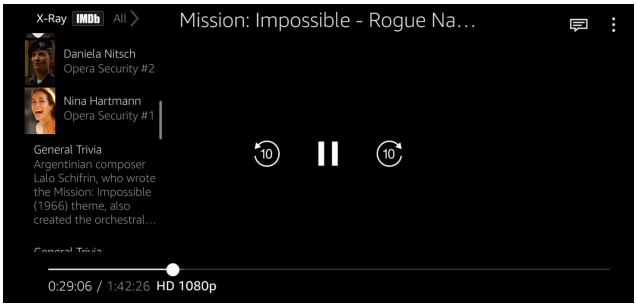

Figure 1. An example of the X-Ray user experience. Credited actors present in the scene are listed with names and profile headshots.

ground truth data is challenging and ambiguous as it is very likely that people will have very different opinions about the quality of the same image. We also adopted an additional constraint of frugality and developed a solution without using supervision. Our developed solution uses deep learning based feature extractors and heuristic based weak supervision. We leverage the vast number of faces extracted from PV video catalog to develop an effective solution without using any ground truth data.

2. Related Work

There are many different studies (e.g., [1] [6] [8] [10]) which cover impact of face quality on performance of face recognition algorithms. They look at different features like pose, illumination, expression, resolution, etc. Most of these works are with respect to suitability of a face image for algorithmic matching. There is some overlap between the problems of face image quality prediction from human perspective and matching algorithms.

Given the sensitivities of recognition performance when input faces deviate from constrained conditions, many earlier works (e.g., [16] [4] [19]) predict the quality of a face image using its similarity to a reference, or "ideal", face images. But the models we are interested in can predict quality of a face image using just the single face image as input.

There are face recognition system which train a network to learn a single face representation(embedding) from a cluster of faces images of a given person. Although these methods are trained for improving face matching between clusters, they can be used to measure face quality because the weights or coefficients learned for combining the multiple faces into a single representation reflects on the quality of a face and its' embeddings for recognition purposes. For example, Disentangled Representation learning - Generative Adversarial Network (DR-GAN) [18], learns a single representation from multiple images of a subject, gives confidence coefficients which can be used to predict the face quality. Similarly, Yang et al. [20] propose Neural Aggregation Network(NAN), a Convolutional Neural Network based model which has attention blocks which learn weights for pooling the embeddings from different face images of a person into a single representation. Both studies have examples to show that their model can be used to suggest face quality.

The most relevant work for face quality from a human perspective is a recent study [5] which works on different models to predict face quality for algorithmic matching and human assessment. They propose (and compare) two different methods to learn face image quality based on target face quality values from

1. human assessments of face image quality (matcher-independent); and
2. quality values computed from similarity scores (matcher-dependent)

They train a support vector regression model trained on face features extracted using a deep convolutional neural network(ConvNet) to predict the quality of a face image. They claim that this is the first study to utilize human assessments of face image quality in designing a predictor of unconstrained face quality that is shown to be effective in cross-database evaluation.

Their model based on human assessment should ideally be suitable for our use case. To the best of our knowledge, we aren't aware of any study which evaluates a face image quality model for manual matching. Even though there are many studies [13] [12] which compare and contrast the performance difference between humans and matching algorithm.

We couldn't inspect the efficacy of the model trained using human assessment as neither the data or model is publicly available. We test another recent model called FaceQNet [10] which is based on a convolutional neural network created by fine-tuning a pre-trained ResNet50 [9]. It takes a face image as input and gives it a score $\in \{0, 1\}$ for it's suitability for face matching using FaceNet [15]. We use this model in our work as mentioned in the features section.

3. Problem Statement

Selecting faces from a human perspective using an algorithm is fundamentally challenging, as it tries to learn features that machines don't naturally learn and prioritize when learning face recognition. Getting human supervised data can help, but we try to solve this problem using only unsupervised or pre-trained tools and resources.

We can define the problem as – Given a set of clusters of face images, $C_i = \{F_1^i, F_2^i, \ldots, F_n^i\}$ for $i \in \{1, 2, ..., m\}$, select a face for each of the clusters - $F_{k_i}^i$, which is the easiest to recognize and distinguish for a human. These face images are picked from a frame in a video, we also need to find a standardised image crop which is aesthetically appealing and suitable to be used as a profile image.

For some applications such as filtering out bad quality faces, we need an absolute score to determine the quality of the face. We define a second problem - Given a face image, predict a score for the quality of the face image (from a human perspective) in the range of $\{0, 1\}$ such that a higher quality image has a higher score.

4. Features

The performance of any face recognition system (automated or manual) is highly influenced by the variability of the samples [3]. There are many factors from the image acquisition conditions which effect its suitability for a recognition system - illumination, location, background homo-

geneity, focus, sharpness, etc. There are many factors associated to the properties of the face itself like pose, presence of occlusions, and different expressions.

We shortlist the following features for selection of faces from a human perspective:

1. **Brightness** - We calculate brightness of the face image by converting the image to the HSV format and calculating the mean of value in percentage(we divide it by 255 and multiple with 100) of all pixels.

2. **Resolution**(Size of the face image) - We capture face images from the video by using MTCNN face detection [22] on every alternate frame in the video. We calculate the area of the bounding box we get in square pixels.

3. **Sharpness** - A tool designed for quality assessment of a face to check it's suitability for automated face recognition systems called FaceQNet [10] is very effective in judging the sharpness of the image. This convolutional neural network based on ResNet50 [9] takes a face image as input and gives it a score between 0 and 1 for it's suitability for face matching. Based on our observations, we decide to use the FaceQNet score obtained by the face to judge it's sharpness. This feature helps us in selecting face images which aren't blurred or hazy.

4. **Pose** - A face with a frontal pose is easier to recognise than a side face. Pose detection is a standard problem in Computer Vision. There are many different proprietary (AWS Rekognition) and open-source solutions (Deep Gaze [11], Hopenet [14], FSANet [21]) to calculate the roll, pitch and yaw values of a face. FSANet is a model which aggregates results from different Capsule Networks. It outperforms other alternatives and is a natural choice for us.

5. Dataset

Getting useful data for training and validation of solutions for this problem is very difficult. The definition of what properties we need is itself very ambiguous, some people might feel that pose is more important; while others feel sharpness is more important. Even when comparing two images, people's opinions can be very different.

We don't have any ground truth for this problem but we do have access to large number of face images and some features which can be calculated using pre-trained models.

The distribution of face images of different qualities in standard face datasets is very different from the distribution seen in videos. Movies and TV episodes have many challenging face shots due to the direction and cinematography techniques and effects. The faces obtained from video frames are truly unconstrained and in the wild.

We use face images and cluster of faces from more than 200 titles, where each title has approximately 10-100 clusters and 10k-40k faces. The faces are captured from every alternate frame in the video using MTCNN [22] face detector and they're clustered using a solution [2] developed for actor tagging for X-Ray metadata.

6. Approach

The selection of a representative headshot comprises of two major steps. The first is to identify the frame of the representative face for each credited actor. The challenge is that we do not have influence over frame conditions (such as illumination, location, background homogeneity, focus, sharpness in the input). The faces themselves can have diverse pose, presence of occlusions, and different expressions.

The second step is to extract the optimal headshot from the selected frame. Based on in-depth study of various photography blogs and professional photography courses, we have identified some key techniques to yield automatic professional quality headshots. This is the first known work to automatically extract professional quality headshot from video frame. This also helps to ensure a standardised experience to the customers. By these steps, we have solved the problem of selecting actor headshots from an unconstraint input video using computer vision algorithms.

6.1. Frame Selection

In this section, we propose multiple solutions for the selection of the frame from which we obtain the representative face image. We use the features described in section 4 for determining the quality of a face. We start with a simple baseline model for comparison. These solutions are compared and evaluated in section 7.

1. **Baseline Model (Pose based Heuristic Cost Function)** - One of the earliest deployed solution was a simple heuristic based on the features we have obtained. It is chosen such that faces which are more frontal facing have a smaller pose cost.

   Pose Cost = $(max(|yaw|-1,0)+1.5*max(|pitch|-1,0)+0.5*max(|roll|-1,0))^3$

   Size Cost = $max(200-\sqrt{resolution},0)$

   Face Cost = Pose Cost + Size Cost

   The face with the lowest cost is selected from a cluster. This is an absolute score function but the value for it can be arbitrarily large, there is no bound on the score and thus it is difficult to interpret the values. Hand crafted heuristics are very likely to be inconsistent and sub-optimal as there is likely to be inefficient trade-off between different features.

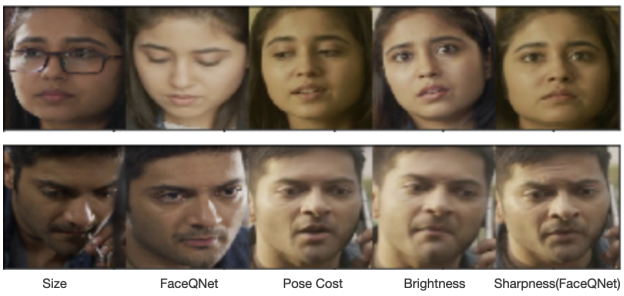

Figure 2. Examples of the best face selected according to the filtering metric in each stage.

2. **Filtering Algorithm** - There are usually many good faces in a large cluster and we can design an effective algorithm by trying to avoid selection of a bad face for any cluster. If we want our selected face to be good in all features, we should focus on selecting a face which does not rank poorly in any feature instead of selecting a face which ranks very highly for only a few features.

   In this algorithm, we filter out bad faces based on a different feature in multiple stages. We pass the cluster of faces through a series of filters based on their scores for different features. In each stage, we filter out all faces which do not match any of the following conditions:

   (a) It has a feature value better than a threshold T.

   (b) It ranks within the top M faces for the feature value.

   (c) It's percentile score based on the feature is greater than minimum percentile P.

   In each stage, we filter out faces which aren't good according to the metric, the next filtering stages operates on the faces which have already successfully passed the previous quality checks/filters. We select filtering parameters (T, M, P) by tuning it using random search(we don't have a way to calculate accuracy of the system, the tuning relies on visual inspection of the results).

   We can understand how the filtering algorithm works by looking at the best face selected according to the filtering metric in each stage in Figure 2.

3. **Regression based on weak supervision** - Since the filtering logic relies on the percentile distribution of feature scores, it is more effective for selection of faces from larger clusters. The percentile distribution of scores over a large number of faces is quite useful. We can use percentile score of each feature in a heuristic score function to obtain weak supervision labels for training our model to predict face quality score between 0 and 1.

   We combine two different heuristics to obtain more robust scores and to introduce non-linearity between the features we use and the labels. The first heuristic is a linear function which combines their percentile scores for our features. The second heuristic is inspired by the filtering logic and is similarly based on score thresholds and percentile thresholds to select a quality bucket for each face. To qualify for a bucket, the percentile score must be higher than P_i or the feature score must be higher than S_i $\forall$ features i. We get the weak supervision labels by taking the average of the normalised $Heuristic1$ score and the quality bucket score of a face. The heuristics are tuned using random search followed by visual inspection of results.

   Our training set consists of more than 4 million faces from over 200 titles. Using the percentile scores from a large set of faces makes our heuristics reliable.

   We try to train different models to fit our weak supervision labels:

   (a) **ConvNet based image models** - We fine tune different ImageNet [7] and face verification models such as ResNet [9], EfficientNet [17], and FaceNet [15] by using the face image and our noisy labels. But we realise that the ImageNet trained models learn only the pose based features of the face. The face verification models are only able to learn the sharpness based features of the face. No single ConvNet model could extract and utilise all the features we need to effectively predict face quality.

   (b) **Regression using features** - We have reliable scores for different features. We need to combine them effectively to avoid under-indexing any feature. Extracting and using different features gives us more transparency and tractability. Thus, we use our features (brightness, resolution, sharpness, roll, pitch and yaw) as inputs for a regression model. We try many different regression models based on Decision Trees, Linear Regression, Boosting, and Deep Learning. Many models can fit our weak label scores effectively. We select Random Forest based regression as our preferred solution because it performs better than other models in our visual inspection tests.

We look at how each of our different approaches perform in section 7.

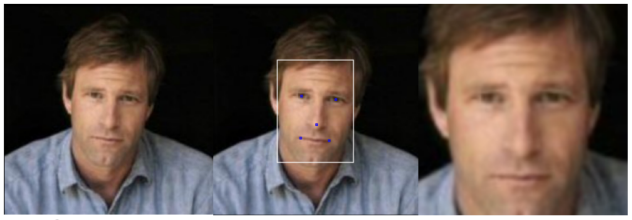

Figure 3. Example of face alignment using landmark points.

6.2. Obtaining optimal headshot from selected frame

Human viewers and ML algorithms have very different requirements for how a headshot should be obtained from an image/frame. There are many standard practices for preprocessing of faces for algorithmic matching. It includes warping of the face to align the face which artificially reduces the roll of the face. This makes the face more frontal in pose and reduces the variance. These aligned faces are also cropped very tightly and don't include the person's neck or hair. The alignment of the faces is done using the landmarks detected by face detection models such as MTCNN [22]. We can look at how the faces are aligned in Figure 3.

The face alignment and preprocessing used for ML algorithms isn't suitable to obtain images for human viewers. When selecting actor profile images, we also need to ensure that the image looks aesthetically appealing. We use the following principles based on our research of professional photography techniques.

1. Actor headshot should be in upper half or 1/3rd of the image
2. Actor headshot should be centered.
3. Full actor hair is generally not be shown. Their hair should be partially visible.
4. Actor shoulders should be included to give the idea of body type.

These guidelines can be used to ensure standardisation of the quality of actor profile images. We asked viewers to choose between the non-standardised IMDb profile images with the same image after our standarised crop. They found the standardised crop to be equivalent or better for all images in the test. The standardised crop was judged to be significantly better for 813 images from a total of 1597 images.

Combining the techniques for image/frame selection with headshot selection gives us a very robust and effective method to generate a profile image. These methods can also be used for other applications such as showing profile images in photo gallery apps such as Amazon Photos. We can look at the profile images generated from some titles in Figure 4.

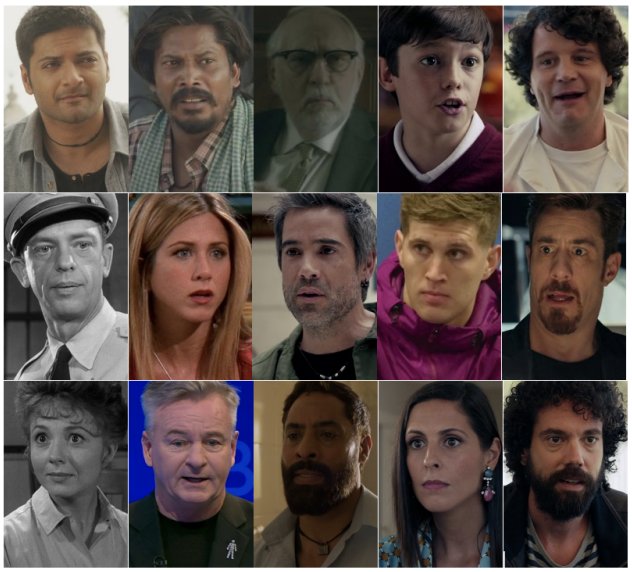

Figure 4. Example of profile images generated from some titles using Filtering Algorithm with standardised cropping of face images.

7. Evaluation

In the first deployed version of the solution[2] to cluster all faces in a title, we used the baseline heuristic to select faces to represent a cluster to human annotators. We later replaced it with the Filtering Algorithm. We compare the results between the baseline algorithm, using just FaceQNet scores and the Filtering algorithm. We manually evaluated the faces selected by the different approaches into three categories -

1. Unidentifiable Images
2. Identifiable/Usable Images
3. Excellent Images

We compare the face selected by the algorithms for more than 350 clusters based on visual inspection. We find the filtering logic to be consistently better. The result estimates are summarised in Table 1.

We realise that filtering logic performs optimally for large clusters. To get more consistent results for face selection, we need models for face quality prediction which would work effectively for small clusters as well. The regression model fulfills this requirement. We compare the face selected by both algorithms for 130 clusters and pick which selected face is better. The Regression Model significantly outperforms the Filtering Algorithm. The results are summarised in Table 2.

Result	Baseline	FaceQNet	Filtering Algorithm
Unidentifiable Faces ↓	13.4%	**8.2%**	**8.2%**
Usable Faces(including excellent) ↑	87.6%	**92.7%**	**92.7%**
Excellent Faces ↑	50.5%	56.7%	**80.4%**

Table 1. Comparing the quality distribution the faces selected for different clusters by the Baseline, FaceQNet and Filtering Algorithm. The evaluation is performed on more than 350 clusters.

Filtering Algorithm	Weakly Supervised Regression	Equal
12(9.2%)	**89(68.5%)**	29(22.3)%

Table 2. Count of the number of clusters in which mentioned algorithm selected a better face than the other algorithm. The evaluation is performed on 130 clusters.

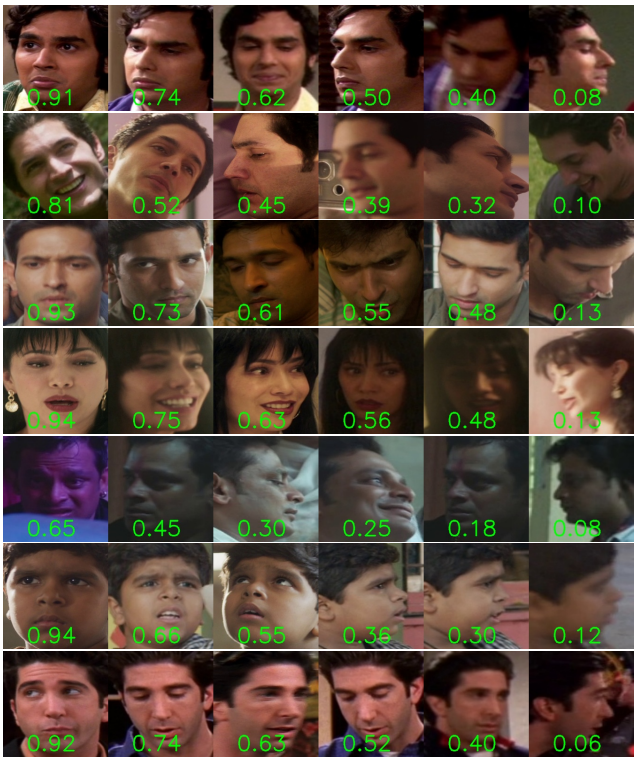

Figure 5. Few examples for the scores assigned by Regression Model. Faces are ranked in a cluster according to the score and then sampled at equal intervals.

For models which return a quality score, we are also interested in knowing if it can predict which faces are bad in quality. We can look at the distribution of scores for faces from some clusters in Figure 5. We can see that the model is able to discriminate between images based on it's quality.

Our priority is to ensure that the best faces we select are good with respect to all features. We compare the median value of features for the faces selected from 991 clusters in Table 3. We can see that our regression model is able to maintain good scores for all the features and outperforms all other algorithms.

8. Conclusion and Future Work

Selection of a high quality representative face for a group of faces is a critical component in many applications such as actor identification systems. We explore multiple solutions which give excellent results without using supervision. We develop a unique method to generate weak labels using a large number of unsupervised faces. We use standardisation techniques to provide optimal headshots for actor profile images.

For our proposed solution, the quality of features extracted is very important. There is scope for improvement in the model used for sharpness of an image. We also plan to train another weakly supervised model for quality prediction for automated matching where we can use scores based on the matches made by a face verification models. We can then use the predicted score for automated matching as another feature for our current model. The heuristics used for weak labels are largely hand-crafted and subjective. There is scope to create these labels without relying heavily on hand-crafted parameters.

References

[1] A. Abaza, M. A. Harrison, T. Bourlai, and A. Ross. Design and evaluation of photometric image quality measures for effective face recognition. *IET Biometrics*, 3(4):314–324, 2014.

[2] Abhinav Aggarwal, Yash Pandya, Lokesh A. Ravindranathan, Laxmi S. Ahire, Manivel Sethu, and Kaustav Nandy. Robust actor recognition in entertainment multimedia at scale. In *ACMMM 2022*, 2022.

[3] Fernando Alonso-Fernandez, Julian Fierrez, and Javier Ortega-Garcia. Quality measures in biometric systems. *Security Privacy, IEEE*, 10:52–62, 11 2012.

[4] L. Best-Rowden, H. Han, C. Otto, B. F. Klare, and A. K. Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. *IEEE Transactions on Information Forensics and Security*, 9(12):2144–2157, Dec 2014.

Feature	Dataset	Baseline	FaceQNet	Filtering Algorithm	Regression Model
Sharpness ↑	0.51	0.50	**0.58**	0.53	0.56
Size/Resolution ↑	20511	15458	23309	17930	**28539**
Brightness ↑	39.64	39.90	39.53	40.10	**40.20**
$\|Yaw\| \downarrow$	22.35	**3.33**	18.86	11.21	9.83
$\|Pitch\| \downarrow$	8.40	**2.32**	7.21	6.55	4.44
$\|Roll\| \downarrow$	6.19	**3.11**	5.41	4.04	3.53

Table 3. Comparing the median value of features for faces selected by Baseline, FaceQNet, Filtering Algorithm and Regression Model for 991 clusters (284k faces)

[5] Lacey Best-Rowden and Anil Jain. Learning face image quality from human assessments. *IEEE Transactions on Information Forensics and Security*, PP:1–1, 01 2018.

[6] J. Ross Beveridge, Geof H. Givens, P. Jonathon Phillips, and Bruce A. Draper. Factors that influence algorithm performance in the face recognition grand challenge. *Computer Vision and Image Understanding*, 113(6):750 – 762, 2009.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[8] Ralph Gross, Jianbo Shi, and Jeff Cohn. Quo vadis face recognition. In *In Third Workshop on Empirical Evaluation Methods in Computer Vision*, pages 119–132, 2001.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[10] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality assessment for face recognition based on deep learning, 2019.

[11] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet, 2014.

[12] P. Jonathon Phillips and Alice O'Toole. Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing*, 32, 01 2013.

[13] P. Jonathon Phillips, Amy N. Yates, Ying Hu, Carina A. Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G. Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, Jun-Cheng Chen, Carlos D. Castillo, Rama Chellappa, David White, and Alice J. O'Toole. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018.

[14] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015.

[16] Harin Sellahewa and Sabah Jassim. Image-quality-based adaptive face recognition. *Instrumentation and Measurement, IEEE Transactions on*, 59:805 – 813, 05 2010.

[17] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2019.

[18] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1283–1292, July 2017.

[19] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *CVPR 2011 WORKSHOPS*, pages 74–81, June 2011.

[20] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition, 2016.

[21] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1087–1096, 2019.

[22] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.