# FACE RECOGNITION WITH LOW FALSE POSITIVE ERROR RATE

N. U. Bagrov[1,2], A. S. Konushin[1], V. S. Konushin[2]

[1] Lomonosov Moscow State University
[2] Video Analysis Technologies LLC

**Commission II, WG II/5**

**KEY WORDS:** Face Recognition, Distributed systems

**ABSTRACT:**

Nowadays face recognition systems are widely used in the world. In China these systems are used in safe cities projects in production, in Russia they are used mostly in closed-loop systems like factories, business centers with biometric access control or stadiums. Closed loop means that we need to identify people from a fixed dataset: in factory it’s a list of employees, in stadium it’s a list of ticket owners. The most challenging task is to identify people from some large city with an open dataset: we don’t have a fixed set of people in the city, it’s rapidly changing due to migration. Another limit is the accuracy of the system: we can’t make a lot of false positive errors (when a person is incorrectly recognized as another person) because number of human operators is limited and they are expensive. We propose an approach to maximize face recognition accuracy for a fixed false positive error rate using limited amount of hardware.

## 1. INTRODUCTION

The main problem is to eliminate the number of false matches in face recognition system while keeping high the true match rate. This can be achieved by various ways, it the article the data pre-processing method is proposed. Another problem is system quality estimation: the real world data is also different from what is used in scientific benchmarks like CAS-Peal (Gao et al., 2008) which was taken in laboratory conditions or MS-Celeb-1M (Guo et al., 2016) which contains only celebrities photos due privacy or difficulties in data collection. Thus a private dataset from the real world data was collected and face recognition performance of the proposed system was estimated.

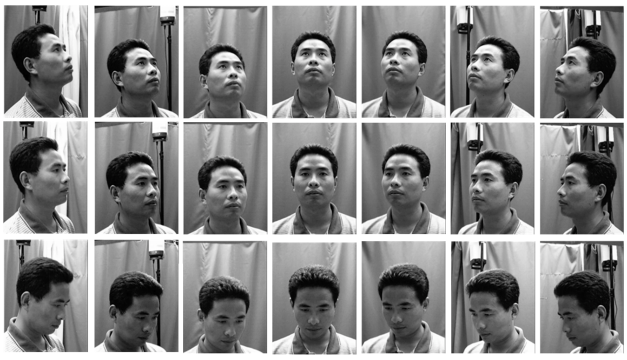

Figure 1. CAS-Peal Sample

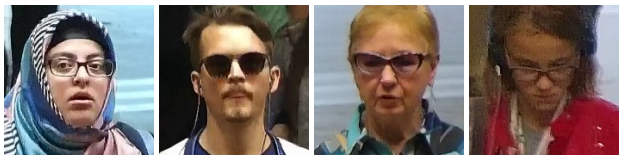

Figure 2. Prepared dataset examples

So having a private dataset is necessary for precise evaluation of face recognition systems in production. Experiments have shown that a lot of face recognition algorithms which were evaluated only on research datasets performs much worse than industrial solutions which are tested on real data in the wild conditions (where illumination, quality and face size is not fixed and can change during day due to artificial illumination or sun shades).

### 1.1 Face recognition system

Systems for face recognition usually consist of the following components:

1. Face detection module. It detects faces from camera stream and sends these faces to the data center. Usually these modules are located near the cameras to lower the required network bandwidth. This module can’t be scaled up like datacenter-based solutions because of limited resources on a site near cameras. Having a lot of high performance face detection modules require complicated air conditioning system and it’s price would be too high. Network bandwidth is usually also limited since cellular or low speed connection could be used to connect face detection module to load balancer.

2. Data storage and load balancer. This system collects faces from face detection modules and sends them to face recognition module. After recognition it stores the result and sends it to user interface. This module collects and stores data for a long time - in some countries it’s required to store face recognition logs with images from one to three months.

3. Face recognition module. This module performs face identification in a specified dataset and sends back the result. It can be easily scaled up during peak hours because of stateless design and datacenter location. This design is also cost effective when using cloud based solutions with dynamic pricing.

## 2. PROPOSED METHOD

One of the ideas is to filter some low quality faces from face detection module to improve overall system accuracy and reduce network traffic between detection and load balancer nodes. As it was confirmed by algorithm evaluation, low quality faces make a lot of false positive matches which greatly decrease overall system accuracy. Another problem are face detection false positives - sometimes faces are wrongly detected on images and generate false matches in face recognition algorithm. In terms of face recognition, low quality face can be rotated, blurred or it's resolution could be too small for a reliable identification. One of the most complicated approaches is to describe face quality in accordance to (ISO/IEC 19794-5:2011, 2011), but in practice more simple model is used. In some papers like (Sang et al., 2009) or (Ferrara et al., 2012) static model for estimating face quality is proposed. In this article an automatic approach was presented to estimate face quality distribution for the selected cameras (by detecting faces and analyzing their characteristics during some time) and to apply some filtration which improves overall system accuracy (measured on another dataset with the same face characteristics distribution). The following characteristics are estimated: yaw/pitch of face in degrees, face size, face duration in video.

### 2.1 Face tracking system

Another key point is face tracking system: faces are detected in live stream and matched into tracklets. Face detection couldn't be run for all frames because of the hardware limit so face detection is performed on so-called detection frames and face tracking with optical tracker on others. In the article an algorithm to estimate the period between face detection frames is proposed using dynamic FPS. The idea is taken from TCP congestion control (Allman et al., 2009): when the processing server is not overloaded a face detection interval can be decreased additive. But when a congestion occurs the limit is increased multiply. This approach is good for production systems because it easily copes with overloads during peak hours (Figure 3) (for example, in transport security applications it usually happens at 18-00 in working days) and effectively loads equipment during another time.

The minimal and maximal interval between detection frames were set to 100ms and 750ms. It was experimentally evaluated that FPS more than 10 frames per second is excessive and in some situations would lead to tracklet splitting. This can occur when person rotates his head and face detector didn't detect the face but optical tracker tracks it. This situation is common for transport security applications when people form a crowd during peak hours.

### 2.2 Pitch/yaw estimation

Face pitch and yaw is estimated using specially trained neural network, architecture is shown in (Figure 4). This neural network was pretrained on face recognition task. The training dataset for pitch/yaw estimation task consists of chinese CAS-PEAL dataset (30900 images) and private dataset (20000 images).

The performance of pitch/yaw estimation model is shown on (Table 1), the processing speed is lesser than 2ms on modern CPU (Intel(R) Core(TM) i5-7440HQ). Results for MLD-wJ were taken from (Geng , Xia, 2014) and for human performance from (Gourier et al., 2006).

Optimal face filtering angles (yaw, pitch) and face size are estimated during two phases:

1. Parameter fixing phase: maximum pitch and yaw angles are selected by general considerations like both of eyes of persons should be visible and inter-eye distance is more than 60 pixels. These are required values for most of face recognition systems.

2. Parameters learning phase: the face detector is applied during some period of time, in the experiments is was one day. Then the image is divided into equal blocks and for each block individual pitch, yaw face parameters are collected. Then these lists are sorted, 5% values from both sides of the sorted lists are truncated. This gives an ability to apply scene context-aware pitch/yaw angle filtration - for example, for some parts of the scene the majority of people are frontal, but in some areas like metal detectors person faces are rotated. The same procedure is applied for face size.

3. Filtration phase: using learned parameters from previous phase the new faces are filtered and sometimes are checked by an expert to cope with conditions where something was changed in the scene and people are going by different ways. An expert is required because sometimes a poster with human face can be located on the wall so such detection should be filtered to achieve true angle estimation.

This procedure helps to eliminate false positives of face detector and improve face recognition rate by removing low quality images like turned persons faces. An example of such image is shown on figure 5. This estimation can also improve face detection algorithm speed by running it on multiple scales, but it wasn't estimated in this paper.

### 2.3 Overall quality function

Overall image quality for face recognition is evaluated by the formula 1, where $pitch$, $yaw$ and $size$ are current face parameters, $pitch_{camera}$, $yaw_{camera}$, $size_{camera}$ are estimated camera parameters. Face size 32 is minimal detected size which is limited by currently used face detector.

$$\begin{aligned} Quality = & max(0, 1 - \frac{|pitch|}{pitch_{camera}}) \\ & + max(0, 1 - \frac{|yaw|}{yaw_{camera}}) \\ & + min(1, max(0, \frac{size - 32}{size_{camera} - 32})) \end{aligned} \tag{1}$$

where $pitch$ = current face pitch
$yaw$ = current face yaw
$size$ = current face size
$pitch_{camera}$ = estimated optimal face pitch
$yaw_{camera}$ = estimated optimal face yaw
$size_{camera}$ = estimated optimal face size

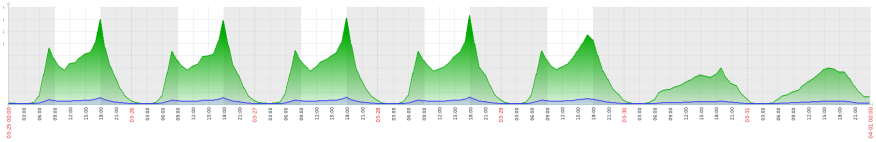

Figure 3. Network traffic during peak hours

Figure 4. Neural network architecture for face pitch,yaw estimation

Figure 5. Bad pitch angle

### 2.4 Face recognition algorithm

Face recognition model is the same for baseline and proposed methods and it is based on Resnet-32(He et al., 2016) architecture with AM-Softmax(Wang et al., 2018) loss function. Training was performed on custom dataset using stochastic gradient descent. 26M images of more than 200k different persons were used during neural network training. The images were collected from social networks and some open datasets like MS-Celeb-1M (Guo et al., 2016).

Processing speed is 140ms on modern CPU (Intel(R) Core(TM) i5-7440HQ) using custom designed neural network library. The outputs from last fully connected layer are used as face descriptors and matched using L2 metric. The descriptor size is 2048 bytes.

## 3. EXPERIMENTAL EVALUATION

In the experiment the servers with 2xE5-2680v2 were used for face detection and 2xE5-2690v2 server was used for face recognition. The dataset consisted of 35000 face samples and the required false positive rate was $10-8$ (in most public face recognition benchmarks like (US NIST FRVT 1:1 Verification, 2019) such low false positive rate values are not estimated). The scenario was security in high loaded transport hubs in metropolis: the cameras were located in the entrance to the hub and the traffic flow was very high during peak hours. Face recognition quality is measured as true acceptance rate (when a person is correctly identified as a person in dataset) with fixed false acceptance rate 1e-8. Another metric is the overall number of correctly detected faces. Dynamic detection FPS improved the detection accuracy from 79% to 84% in crowd. This was measured in peak hours (from 8-00 to 9-30 in the morning and 19-00 to 21-00 in the evening). In the other conditions without tight crowd the overall face detection accuracy is more than 94%.

Baseline algorithm was implemented to select large face without motion blur. The amout of blur is estimated by Laplacian of Gaussian method (Hua et al., 2012). Face recognition algorithm is the same for baseline and proposed methods.

### 3.1 Wild dataset

Low quality image filtering lowered network traffic more than 5 times and improved overall face recognition to more than 8% than a reference implementation without such filters (Figure 6). This was measured on the special dataset with a real-world distribution because face recognition with low FAR rate can’t be effectively measured on the real data because the actual dataset consists from the criminals and the true-match probability is too low to make statistically correct measures. Some tests were performed with modified dataset (some other people were added and recognized) and it showed comparable result with synthetic benchmark. The overall dataset size is more than 100000 images and about 50000 unique persons. Wild dataset was collected from video and photo samples from a CCTV system. Photo samples were collected from social networks and other internet resources. 50% of video sequences have a true match as a photo sample. Best shot estimation for true matches (from video sequences) was different for baseline and proposed algorithms.

### 3.2 Mugshot dataset

Another benchmark was performed on mugshot (Figure 7) dataset which contained mugshot-like photos with variable quality. Some of them were taken in good conditions with accordance to (ISO/IEC 19794-5:2011, 2011), other images were taken in low light conditions from a CCTV system in a transport hub. Best shot selection using proposed pitch/yaw angle selection have shown better quality than baseline solution without adaptive algorithm. Best shot algorithm selection is essential because false positives from detector or low quality images usually make a lot of false matches in common face recognition algorithms. So a set of low quality images can produce a Cartesian product of high score matches which greatly lowers precision in low FAR area of ROC curve.

| Algorithm | Dataset | | | |
|---|---|---|---|---|
| | Customer yaw | Pointing-04 yaw | Customer pitch | Pointing-04 pitch |
| Proposed | 4.13 | 4.26 | 3.28 | 3.17 |
| Human | - | 11.8 | - | 9.4 |
| MLD-wJ | - | 4.24 | - | 2.69 |

Table 1. MAE for pitch,yaw estimation

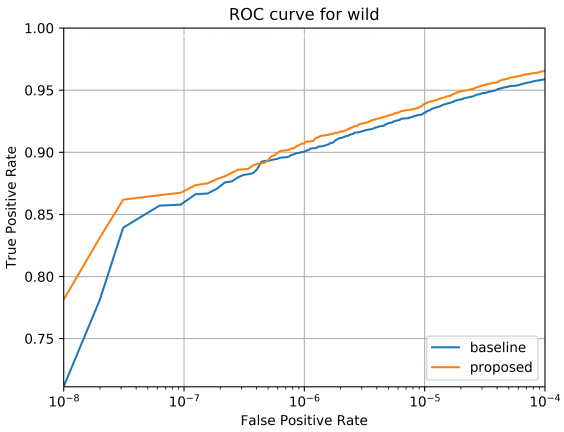


Figure 6. Face recognition in wild conditions

CCTV part of a mugshot dataset was manually labeled by experts who marked different tracklets in a video. Then a proposed algorithm was used to select best shots from each tracklet in a video. Total number of images was the same for baseline and proposed algorithm. The dataset consists of 20000 different tracklets from the video which are used as distractors for face recognition algorithm like in Megaface (Kemelmacher-Shlizerman et al., 2016) benchmark. True match pairs are formed by specially added images: one is a mugshot taken in good light condition, another is a CCTV tracklet of this person. This approach allows comparison of algorithms in low FAR area without large manually annotated dataset. Tracklets were uniformly selected from the weekly video for data in various lighting conditions. Duplicate tracklets of the same people were manually deleted after processing with face recognition algorith. An additional feature of this dataset is a large number of face samples with headwear and glasses.

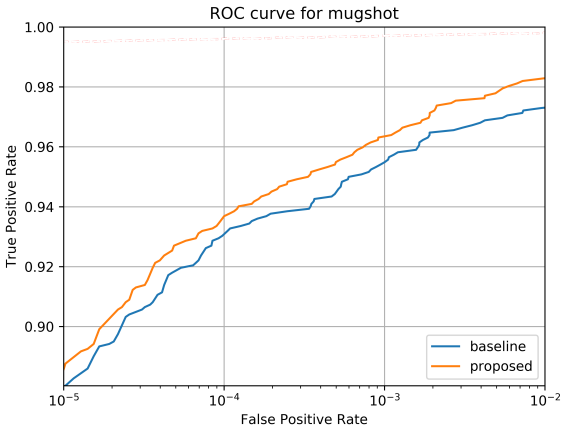


Figure 7. Face recognition on mugshot dataset

Proposed method has shown slight accuracy increase (lesser than 2%) on CAS-PEAL (Gao et al., 2008) dataset in low far area (Figure 8), but in other area it’s accuracy was worse than baseline method. Images with yaw more than 55 degrees and pitch greater than 40 degrees were filtered out: these values were selected by cross-validation to optimize accuracy for FAR 1e-7. Another cause for low accuracy on CAS-PEAL dataset are the contents of this dataset: it consists of Asian faces while the training set for neural network consists of Caucasian faces.

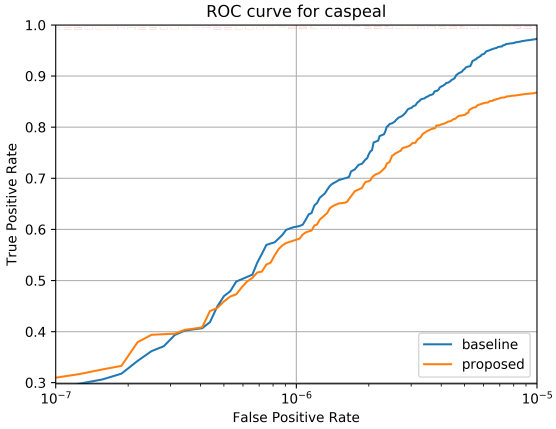


Figure 8. Face recognition on CAS-PEAL

## 4. CONCLUSIONS

Face recognition systems can be improved by applying some filtering to the input data. On the scientific datasets like CAS-PEAL which were collected in laboratory conditions, overall accuracy does not improve from this filtration but on the real world data it increases significantly. Dynamic detection FPS solution can apply some congestion control to limit the performance during system overload and utilize more resources in other hours. It is important that the volume of network traffic to the load balancer also has decreased which can improve overall system stability and reduce expluatation costs in real systems.

Future research can improve face detection speed by dividing detection area into segments and running face detector on multiple scales. Another type of research could be a machine learning model on face parameters like pitch/yaw and size using detection and face attributes estimation output. This model could be updated in real time to additional improve quality in time period or when a scene is changed: some doors near the camera could be closed or opened.

An open question is about face quality estimation: other characteristics like motion blur, camera defocus, weather conditions can be included into model and estimated for the input frame. This can improve best face estimation with long focused camera lens.

## REFERENCES

Allman, Mark, Paxson, Vern, Blanton, Ethan, 2009. Tcp congestion control. Technical report.

Ferrara, Matteo, Franco, Annalisa, Maio, Dario, Maltoni, Davide, 2012. Face image conformance to iso/icao standards in machine readable travel documents. *IEEE Transactions on Information Forensics and Security*, 7, 1204–1213.

Gao, Wen, Cao, Bo, Shan, Shiguang, Chen, Xilin, Zhou, Delong, Zhang, Xiaohua, Zhao, Debin, 2008. The CAS-PEAL large-scale Chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 38, 149–161.

Geng, Xin, Xia, Yu, 2014. Head pose estimation based on multivariate label distribution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1837–1842.

Gourier, Nicolas, Maisonnasse, Jérôme, Hall, Daniela, Crowley, James L, 2006. Head pose estimation on low resolution images. *International Evaluation Workshop on Classification of Events, Activities and Relationships*, Springer, 270–280.

Guo, Yandong, Zhang, Lei, Hu, Yuxiao, He, Xiaodong, Gao, Jianfeng, 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. *European Conference on Computer Vision*, Springer, 87–102.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, Sun, Jian, 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hua, Fang, Johnson, Peter, Sazonova, Nadezhda, Lopez-Meyer, Paulo, Schuckers, Stephanie, 2012. Impact of out-of-focus blur on face recognition performance based on modular transfer function. *2012 5th IAPR International Conference on Biometrics (ICB)*, IEEE, 85–90.

ISO/IEC 19794-5:2011, 2011. Information technology – biometric data interchange formats – part 5: Face image data. Standard ISO/IEC 19794-5:2011, International Organization for Standardization, Geneva, CH.

Kemelmacher-Shlizerman, Ira, Seitz, Steven M, Miller, Daniel, Brossard, Evan, 2016. The megaface benchmark: 1 million faces for recognition at scale. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4873–4882.

Sang, Jitao, Lei, Zhen, Li, Stan Z, 2009. Face image quality evaluation for iso/iec standards 19794-5 and 29794-5. *International Conference on Biometrics*, Springer, 229–238.

US NIST FRVT 1:1 Verification, 2019. US NIST FRVT 1:1 Verification. https://www.nist.gov/programs-projects/frvt-11-verification.

Wang, Feng, Cheng, Jian, Liu, Weiyang, Liu, Haijun, 2018. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25, 926–930.