The invention and development of face recognition systems

Law, Security

The beginnings of face recognition systems date back to the 60s when the technology was based on pointing the coordinates of distinguishable parts (features) of the face, such as eyes, nose or mouth. Nowadays, deep convolutional neural networks take image pixels to obtain a uniquely n-dimensional representation of the face, facial features of the person are mapped into a numerical vector representation. This vector is often called an embedding. Since these embeddings share the same vector space, we can use vector distances to determine the similarity between two embeddings, similarity between two faces. This process is called matching.

As a general rule, a deep face recognition process acquires more than one picture to generate a robust facial template. Many studies claim that FR performance gets increased when using multiple images. Therefore, a deep-based features extractor produces as many embedding as input images. However, in order to match a face template of N images against another, we need to combine the N into a single vector. Features fusion is the procedure of merging face embeddings to obtain a compact feature representation.

In this section, we explore some global face quality measures and propose to use them in three features fusion methods. Subsection A explains the global face quality references employed. Subsection B describes the implementation of the methods for features fusion. Finally, subsection C includes an evaluation of the whole impact on FR performance. Typically, face detectors give a score of how likely is that the detected region is really a face. High resolution and frontal faces usually have a higher detection

confidence when compared to blurry or extreme-pose faces. Hence, the face detection score can be indicative of the quality

Authors in 21 studied the relationship between deep convolutional neural network features and the original image. They showed that top-level features included information about the pose of a face. In addition, they proved that images in the centre of the feature space are low-quality whereas the farthest images are high-quality. In other words, face image quality increases with distance from the origin. Castillo confirmed this observing that the L2-norm of features learned using softmax loss is indicative of the face quality.

Authors of RQS introduced a method based on learning to rank a face image between three different quality-type databases. They trained a Convolutional Neural Network (CNN) extracting up to five different image features and mapped the results into a global quality score. Classical approaches confront features fusion problem by simply averaging the features extracted from each image/frame of the face template. However, this may lead to non-compact feature representations since both good and poor quality faces are weighted equally. vspace{2mm}

An intuitive approach is to average only the top N features corresponding to images whose quality is above a certain threshold. The following pseudo-code details the method. Ranjan introduced a feature-level fusion method based on weighting the vectors according to its face image quality. More precisely, they used the confidence score from a face detector. We propose

to evaluate this fusion method not only with the detection probability but with other quality references.

Although exhaustive testing data is very important for the evaluation of face recognition systems, most of the public datasets to evaluate face recognition performance are limited by restrictions in illumination, pose, expressions or occlusions. Alternatively, in 2015, IARPA and NIST developed the IARPA Janus Benchmark A (IJB-A), cite{IJB-A} a challenge characterized by unconstrained in-the-wild images to drive research in face recognition. This database contains pictures from 500 subjects with ground truth face locations using 1, 501, 267 million crowdsourced annotations. Its two main protocols are the search protocol (identification) and compare protocol (verification). In this work, we focus our attention on the effect of quality-based features fusion methods in verification accuracy.

The compare protocol defines comparisons between face templates of the same user (genuine verifications) and templates belonging to different users (impostor verifications). The comparisons are divided into 10 splits, each of them containing 333 subjects for training and 147 for testing. For a given split, there are about 10, 000 impostor comparisons and 1, 800 genuine comparisons. In order to ensure that the protocol is challenging, the two subjects to compare are always chosen to have the same gender and similar skin tone. vspace{2mm}

Furthermore, the IARPA Janus Benchmark-B (IJB-B)cite{IJB-B} extends the IJB-A dataset. It consists of 1, 845 subjects represented by 21, 798 still

images and 55, 026 frames from 7, 011 videos. Its verification protocol (emph{test1}) is far more challenging than the verification protocol from IJB-A. Unlike IJB-A, it is not structured in splits and does not provide training sets. The total number of impostor comparisons amounts to 8, 000, 000 matches while the genuine comparisons reach about 10, 000. This increase in the total number of matches allows a higher resolution in the results. vspace{2mm}

To test the features fusion methods in those benchmarks, we have implemented an evaluation pipeline in emph{Python 2. 7} based on the emph{bob} toolbox. footnote {emph{Bob} is a machine learning/signal-processing toolbox developed by the Idiap Research Institute in Switzerland. cite{bob2017}} Figure~ref{fig: evaluationPipeline} shows the evaluation scheme. First of all, the dataset is parsed in order to obtain the information about the templates and their corresponding images. Then, every image goes through the MTCNN face detector. cite{MTCNN} If no face is detected, we use the ground truth annotations provided by the benchmarks.

Afterwards, the face is aligned by a similarity transform (scale, rotation, and translation) according to the landmarks positions of a predefined emph{ideal} face. Next, the selected features extractor by Gradiant obtains the vector representation of the face, which is internally stored. Once all the emph{embeddings} have been extracted and saved, the emph{features fusion} block loads every template to apply the desired fusion method between its vectors. Finally, the emph{matching} block receives the list of comparisons to perform from the database’s protocol and calculates the

Euclidean distance between the fused emph{embeddings} of both templates. vspace{2mm}

Typically, the FAR is plotted versus the FRR in the Receiver Operating Characteristic curve (ROC). This graphical representation allows the evaluation of face recognition systems at various threshold settings. In addition, biometrics systems usually measure the error as the FRR at a certain FAR point, ie fixing the security of a system at a certain operating point and measuring the usability at such point.

Before evaluating the impact of the fusion methods and global quality measures in the IJB-A/B benchmarks, we need to set the quality thresholds for Mean Top N fusion method. Mean and emph{Quality Pooling} fusions do not require any tuning of parameters. To this purpose, we have used the training sets of every IJB-A split and calculated the mean ROC for every global quality measure and a range of thresholds $alpha$. The resulting curves are shown in Figure~ref{fig: mtnRocRqs} for the case of emph{RQS}. This plot reflects that the optimal threshold for the emph{MTN+RQS} combination is between 50 and 60. Similarly, we reproduced this procedure for the rest of combinations (emph{MTN+detector confidence}, emph{MTN+detector L2-norm}, and emph{MTN+extractor L2-norm}) and find its bests $alpha$’s. vspace{2mm}

Once we have set the parameters in a disjoint training set, we can run a comparison between all combinations of fusion methods and global quality measures. Table V lists the performance of all combinations on the

verification protocol of the IJB-A benchmark. Results are presented as mean and standard deviation of the 10 splits. At first sight, we can see that the significant differences appear at low FARs, while the EER and FRR@FAR10^{-2} show similar results regardless of the combination. The detector confidence obtains the best results across all operating points. emph{QP+detector confidence} obtains the top performance at FAR10^{-3} reducing the error up to -4. 2% with respect to emph{MEAN} fusion. emph{MTN+detector confidence} obtains the top performance at FAR10^{-4} reducing -4. 5% the emph{MEAN} fusion error. The rest of the combinations do not improve the recognition performance of the emph{MEAN} method. In fact, RQS quality measure even worsens the results. vspace{2mm}

% IJB_B RESULTS COMMENTS

In the case of IJB-B, the differences are smaller due to the increased protocol’s complexity. The total number of matches to perform in IJB-B is about emph{x}80 times the number of comparisons in IJB-A. emph{QP+detector confidence} tops performance again across the most restrictive operating points (-5. 6% @FAR10^{-5} with respect to emph{MEAN} fusion). Differences at ERR and higher FAR remain insignificant.