A Comparative Study of Face Re-identification Systems under Real-World Conditions


Published in:

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
© 2018 The Authors and The Irish Pattern Recognition & Classification Society.
This work is made available online in accordance with the publisher’s policies. Please refer to any applicable terms of use of the publisher.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Download date: 26. May. 2019
A Comparative Study of Face Re-identification Systems under Real-World Conditions.

Authors

Institution

May 2018

Abstract

Face re-identification is largely thought of as a solved problem in the research community, with State-of-the-art systems attaining human-level performance on unconstrained image datasets. However, these results do not seem to translate to the real world. In systems matching people in surveillance-like footage to high-quality images, reported performance is much lower than what the literature would suggest. In this work, we contribute a multi-modal dataset for evaluating real-world performance of facial re-identification systems. We then perform a verification and re-identification evaluation for state-of-art systems on both this dataset and the popular benchmarking dataset Labelled Faces in the Wild (LFW).

Keywords: Deep Learning, Person Re-identification, Face Recognition, Image-to-Video

1 Introduction

Face verification is a well-known problem with interesting applications such as biometrics, access control and user authentication. Facial re-identification is an evolution of the previous problem in a much more complex setting, where the identity of a user in the probe (which has never been seen before) should be paired against a full gallery of users.

The emergence of Deep Learning has revolutionised the computer vision field by bringing unprecedented performance. Deep Neural Networks (DNNs) may have different network architectures that vary in size, efficiency, and performance, but the weights learned through training are defined by the dataset the network is trained on. As a requirement, deep neural networks require large and complex datasets to be trained on, which are representative of the real world complexity. Face Re-identification systems are no different, and a substantial part of the field’s recent success can be attributed to the availability of large-scale datasets to train on, such as VGGFace [Parkhi et al., 2015], CelebFaces+ [Sun et al., 2014], CASIA-WebFace [Yi et al., 2014], Ms-Celeb-1M [Guo et al., 2016], and others.

Among the available datasets with facial imagery, Labelled Faces in the Wild (LFW) [Huang et al., 2007] is of particular interest due to its highly unconstrained nature, and the large number of images and people it contains. While this dataset leads in complexity among the available standard facial datasets, it can be argued if its complexity is enough to develop and evaluate systems truly ready for real-life problems. For instance, due to the acquiring process, that dataset consists mostly of static images from high-quality sources. Thus, the classes within these datasets tend to be unimodal: Every image in a class has roughly similar properties to every other image in the same class, making the reidentification easier by not requiring domain adaptation.

This is not always desirable though since it does not always reflect real-world acquisition conditions. For example: when matching still-images of people facing towards the camera to frames from surveillance video, the two sets of images have different intrinsic properties, i.e. belong to different domains, besides the actual subject
in the image. This bias will severely impact the performance of models trained on standard datasets when applied to realistic and complex scenarios. Dataset bias has been noted before in the literature [Klare et al., 2015] leading to researchers performing cross-dataset experiments [Cao et al., 2017, Bansal et al., 2017].

In order to prove the severity of this effect, in this work, we perform a comparative study of state-of-the-art deep learning models for re-identification under a complex cross-dataset setting. All models are first trained on unconstrained still image datasets and then tested on a still-to-video dataset that we created and released to the community. Doing this, we aim to show that using a one-size-fits-all dataset for evaluating systems may lead to less than expected performance in the real world.

2 State of the Art

Still-to-Still (S2S). Many approaches have successfully been applied for facial verification on unconstrained images. In [Schroff et al., 2015], the authors introduced the triplet-loss, a metric learning loss function that aims to embed images into Euclidian space with similar images pulled towards each other and dissimilar images pushed apart. Similarly, VGG16 architecture [Parkhi et al., 2015] combines the triplet-loss with a CNN and a large unconstrained dataset. This was further extended [Cao et al., 2017] with an even larger dataset, and a network based on Resnet50 [He et al., 2016] with Squeeze-and-Excitation blocks [Hu et al., 2017].

In spite of their good performance in standard sets, authors in [Klare et al., 2015] found that the performance of re-identification systems (specifically an unnamed Government-Off-The-Self system and OpenBR [Klontz et al., 2013]) was heavily dependant on the subject’s pose. Unconstrained frontal images can be recognised with near human-level performance, but images with full pose variation such as profile shots cannot reliably be detected or recognised.

Still to Video (S2V). Recently, research approaches have emerged to tackled Still-to-Video (S2V) re-identification, where still images are enrolled in a system and then video sequences are matched against them. These approaches have screening as an intended application, aiming to match mugshots in a watchlist with CCTV footage of suspects. This task introduces extra domain adaptation challenges between cameras that increases its complexity. A common approach is to use application-specific models that mitigate the multi-modality of the gallery and probe images. One such S2V model is the HaarNet [Parchami et al., 2017a], which consists of a trunk network which extracts features, with three branch networks based on Haar-like features. Another model is the Canonical Face Representation CNN (CFR-CNN) [Parchami et al., 2017b], which uses a supervised auto-encoder network to generate canonical face representations. The performance of the CFR-CNN does not surpass that of the HaarNet, but it is comparable and is much less complex: requiring 3 orders of magnitude fewer operations, with only a tenth of the parameters as the HaarNet.

3 MMF Dataset

In this section, we present an independent dataset called Muti-Modal Faces (MMF) that has been recorded for evaluating applications such as watchlist screening or e-gates in airports. The dataset also aims to provide multiple modalities to better reflect the complexity of the real world.

Our dataset captures 77 subjects, consisting of 2 sets A and B. Set A consists of frontal images whereas Set B consists of images mined from surveillance-like videos. The cameras chosen at the time to record the video and stills were simple off-the-shelf hardware and were not high quality.

3.1 Data Collection

The original data was collected in many sessions, over several days. Inside a rectangular room, a temporary division was put up along the centre. Three checkpoints were set up. Each checkpoint had a face camera
recording still images, and a body camera recording video overlooking it. Figure 1 shows the layout with face cameras labelled $A$ and full body cameras labelled $B$.

Each subject would enter from the door and proceed to Checkpoint $C_1$. Once there, they were to pause and look straight into the appropriate face camera. After a few moments, they were to proceed to the next checkpoint and repeat the process. Once all three checkpoints were visited, the subject left the room.

### 3.2 Data Processing

To extract faces from the raw data, we adopted an approach similar to LFW, by using a Viola-Jones detector [Viola and Jones, 2001]. Both the still images and videos had face detection and extraction applied to them. No alignment was done at this stage. The videos were fed into the detector frame-by-frame, with the output forming set $B$. The Viola-Jones detector was used to make the resulting dataset more comparable to LFW. The images were not resized after extraction. Images were excluded from the sets based on the following criteria:

1) If an image does not include a face.
2) If the image was cropped in such a way that the face is not at the centre of the image.
3) If the image was in set $A$ and the subject was not facing the camera.

Sample images from each camera are shown in Figure 2.

The final dataset contains 77 identities, with 60,924 images: 15,202 in Set $A$ and 45,722 in Set $B$. Figure 3 shows both the number of images in each class per set and the number of images from each camera.

**Dataset Variance:** MMF differs from other unconstrained datasets in that its intra-class similarity is multi-modal: that is, each class contains images with intrinsically different properties, such that the images form multiple natural clusters around these properties. These are shown in Figure 4. Figure 4 (left) shows that the average intra-class similarity for LFW over all identities is largely unimodal, whereas Figure 4 (centre) shows that the MMF average intra-class similarity has distinctive peaks and modalities, with clear separation between Set $A$ and Set $B$. Moreover, when observing individual classes distribution, the multi-modalities due to the three checkpoints can be more easily observed.

### 4 Methods

Two main state-of-art systems were tested for face re-identification and face verification: a ResNet50 [He et al., 2016] based DNN and the VGG16 [Parkhi et al., 2015] DNN.

The ResNet50 based network was pre-trained on the VGGFace2 dataset [Cao et al., 2017] using the softmax loss function. As the network was pre-trained on a classification problem, we remove the final layer in order to get image embeddings as output. We chose this network because it is State-of-the-Art with regards to unconstrained facial re-identification.
We also compare with the successful VGG16 architecture trained on the original VGGFace dataset [Parkhi et al., 2015] using the triplet-loss function [Schroff et al., 2015]. Like with the ResNet50 network above, we remove the final layer to get out feature vectors. We included this network in our comparison because it is a shorter, simpler network while still attaining close to State-of-the-Art performance.

All models are executed in Keras [Chollet et al., 2015] with the TensorFlow backend [Abadi et al., 2015]. All pre-trained weights were downloaded from VGG’s website and converted into Keras/Tensorflow models.

Face alignment (FA) has been proved [Bansal et al., 2017] as crucial for a correct identification, particularly in unconstrained environments. With the aim to validate this in the context of our scenario, all tested methodologies were evaluated with and without it. It is provided by DLIB [King, 2009] and OpenCV [Bradski, 2000]. DLIB is used to detect the positions of facial landmarks, namely the eyes. Knowing their location and the relative angle between them, OpenCV’s warpAffine function was used to place the eyes into standard positions, rotating the image as necessary so the eyes are horizontal relative to each other. Since we are explicitly testing performance without FA, we are using the original un-aligned LFW dataset.

5 Experiments and Results

Two types of experiments were performed in this evaluation, Face Verification and Face Re-Identification. We aim for a comparison between LFW and our own dataset, as well as the methods themselves in both scenarios.
As specified in Section 4, all training was done on either VGGFace or VGGFace2 depending on the network. LFW and MMF are used for evaluation only, being this validation a cross-dataset setting.

5.1 Experimental Setup

5.1.1 Verification

Face verification consists of taking two images, and determining if both are of the same person. Standard practice for evaluation is to create a set of labelled pairs; pass them into face verification system which produces some distance measurement -Euclidean distance in our case-; which is then thresholded by a value $\tau$ to produce positive or negative labels. In our work, $\tau$ is found per fold per experiment by performing a grid-search, optimising for verification accuracy. Since $\tau$ is obtained from our test datasets, we do not report the best value for $\tau$, nor the derived verification accuracy.

LFW. For LFW the standard "Unrestricted with Outside Data" protocol will be used [Learned-Miller, 2014]. This consists of 10 sets of 600 test pairs, provided by the authors of LFW. The pairs of images in each fold (a fold consisting of 9 of the pair sets, with a different holdout set each time) are then passed through the system which predicts if both images contain the same person. The ROC Curve is then generated for each of the 10 folds. The final result is the average ROC curve and averaged accuracy over the folds.

MMF. For MMF, we mirror the protocol used by LFW as much as is practical. In our case, we generate 4620 pairs with 50% positive matches and 50% negative matches (same as LFW). Each image in Set A is matched uniformly to two randomly selected images in Set B: one of the same class, and the other of a different class. The rest of the protocol follows LFW, with the final results averaged over 10 folds.

5.1.2 Re-identification

Re-identification is where one matches a single probe image to a gallery of images where neither the probe nor gallery images were in the original training set. We will evaluate performance with the Cumulative Matching Characteristic (CMC) Curve. Re-identification naturally gets more difficult the larger the gallery: a system using purely random chance has probability 0.5 matching a probe to a gallery of 2, and probability 0.001 matching to a gallery of 1000. To mitigate this, we make the gallery size of both LFW and MMF identical.

LFW. To construct galleries and probes for LFW, only the classes with 2 or more images are selected (of which there are 1680). Then, we divide the classes into groups so each group contains 77 classes (with the last group taking the remainder). This results in 22 class groups. We then construct a probe and gallery for each group. Galleries are constructed by randomly selecting an image from each class, with probes then containing the images not used in the gallery. The CMC scores are then computed for each gallery-probe pair, with the final result being the mean of all 22 CMC curves.
**MMF.** Given the size of the dataset and that our probe and gallery will be made from disjoint images *Set A* and *Set B*, we construct the MMF probe and gallery differently to *LFW*. Iterating over each class, we randomly select 15 images from the class in *Set A*, and 30 images from the class in *Set B*. We then construct 15 galleries with the images from *Set A* (every class represented in each gallery), and combine the all of the images from *Set B* into a single probe.

### 5.2 Results

**Verification.** The resulting ROC curve comparing the difficulty of the verification problem on *LFW* and *MMF* can be seen in Figure 5: the Area Under Curve in square brackets in the legend). The best results were all achieved on *LFW*, regardless of model or whether FA was used. This shows the more complex conditions encompassed in our dataset, which is more obvious if no FA is applied. Regarding methods, ResNet50 performs better and seems better generalisation properties than VGG16 in this cross-dataset setting.

![Figure 5: ROC Curve comparing LFW and MMF Datasets](image)

**Re-identification.** Figure 6 shows the resulting CMC curves for the re-identification experiments, with Table 1 showing the CMC scores for ranks 1, 5, and 20. In this more complex setting, *MMF* still exhibits more complex conditions than *LFW*, showing even bigger performance drops, in like-for-like examples (same model and using FA or not). The only exception is for *VGG16* without FA. However, since FA is not applied, which appears to be a requirement for good re-identification, the results are inherently unreliable. Regarding methods, although *ResNet50* is clearly better in *LFW*, a dataset closer to the VGG training sets, *VGG16* obtains better results on *MMF* than *ResNet50*. In spite of *VGG16* containing more parameters (138,357,544 vs 25,636,712 for *ResNet50*) due to the bottleneck structure of *ResNet50*, *VGG16* is a shorter network, which means it may exhibit smaller levels of overfitting for this complex setting.

### 6 Conclusions

In this work, we contributed a dataset to evaluate facial re-identification systems. By comparing our dataset *MMF* to a widely used, unconstrained, facial dataset *LFW*, we show that State-of-the-Art systems are still not fully robust, and give a much poorer performance when used on real-world still-to-video data. We also show
that face alignment gives a large boost to current systems. Furthermore, MMF complexity and the use of cross-dataset experimentation may unravel relevant conclusions such as overfitting of current architectures.

Regarding possible future work, for real-world performance of facial re-identification systems to improve, application specific datasets need to be created which are large enough to facilitate training: either entirely, or with fine-tuning. Until such datasets are available, via either harvesting or data augmentation, the field will struggle to implement robust solutions that reliably work in real-world applications.

References


