Cross-Resolution Face Recognition using Quintuplet Metric Learning

Anonymous SBS submission

Paper ID 1

Abstract

Face recognition systems, trained in controlled environment, often fail to match low resolution images with high resolution images. State-of-art systems for face identification often fail when used on surveillance images, because of lack of usage low resolution images when training models.

In this paper, we propose quintuplet metric learning method, which aims to learn useful metric by distance comparisons. Idea is pretty similar to triplet metric learning approach, but uses high resolution and low resolution images for training. Triplet metric uses anchor image and compares it to positive and negative classes of the image. All three images are of the same quality. In our proposed method, we compare anchor image to positive and negative class of the same quality image and downsampled positive and negative images. This means that five images are used for each quintuplet, where three are used for previously mentioned triplet.

In this paper we compare Open Source Biometric Recognition framework, VGG neural network, triplet metric and quintuplet metric on two datasets. We show that probability of correct face recognition and/or identification is improved when using triplet metric learning, and furtherly improved when using quintuplet.

1. Introduction

With advancements in technology, surveillance cameras now have a profound presence and are widely used in security and law enforcement applications. There are sevedal instances where surveillance videos have helped agencies in apprehending individuals who have commited crime or identifiy individuals with the intent to commit crime. For example, in 2005 subway bomb blasts in London [1], CCTV footage helped law enforcement officers in identifying the bombers. In 2008 Mumbai terrorist attacks [2], surveillance cameras helped the agencies to track the activities of terrorists and later identifying them. In both presented cases, surveillance cameras could not foil the terrorist attacks, however they served as the primary evidence in leading the investigation and also recognizing the individuals at the end.

Fig. 1 illustrates the idea of proposed quintuplet metric learning. Beside positive (P) and negative (N) class image, for loss minimization we also use downsampled image of both positive (Pd) and negative (Nd) class. We feed images to VGG network which returns feature vectors (X_A,X_P,X_N,X_{Pd},X_{Nd}) for each of the images. We then use these feature vectors to learn quintuplet metric.

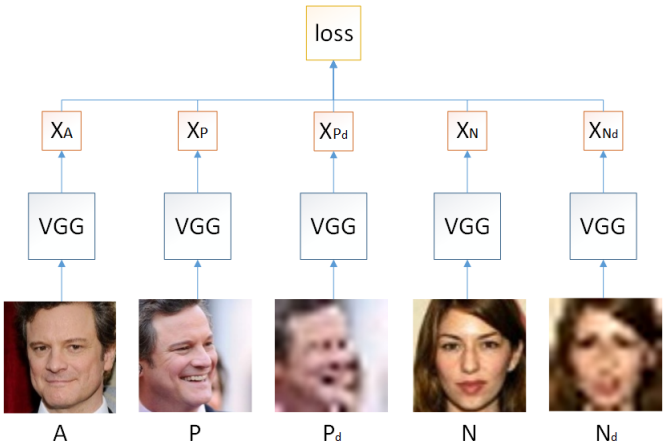

Figure 1. We feed five images (Anchor (A), Positive (P), Negative (N), Positive downsampled (Pd), Negative downsampled (Nd)) to VGG neural network which returns feature vectors (X_A,X_P,X_N,X_{Pd},X_{Nd}). Feature vectors are then used to calculate quintuplet loss.

With the increasing use of video surveillance systems for applications in security and forensics, the demand for face recognition has been growing. However, recognizing faces using such systems in real-world scenarios not only requires one to deal with the facial image variations of pose, illumination and expression, but also those with insufficient resolution due to long distances between the subjects of interest and the camera sensors. For example, query images with low resolution (LR) need to be verified using gallery ones with high resolution (HR). How to match images across different resolution turns out to be a practical yet challeng-

ing task. [54] One approach to match cross resolution images, i.e. low resolution probe with high resolution gallery, is to downsample high resolution images to the level of low resolution images before matching. However, information useful for face recognition such as texture, edges and other high frequency information is compromised while downsampling the images. Another widely used approach is to enhance the low resolution face images using super-resolution techniques [3, 4] and then match with high resolution images. Super-resolution techniques are intended for reconstructing a high resolution view from low resolution image(s) and are not optimized for face recognition applications. Though there are few techniques that incorporate face recognition with super-resolution [5], they remain susceptible to environmental variations and introduce distortions.

Recently, FaceNet [6] introduced a triplet loss to minimize the difference between an anchor image and a positive one (i.e. with the same identity), while the distance between it and its negative one is to be maximized. In addition, Wen et al. [7] introduced a center loss function into existing CNN models, which also resulted in better recognition performance.

Although the above methods have reported promising results on challenging and large-scale benchmark datasets, these approaches typically assume that both the query and gallery images are with the same or similar resolution. In other words, as we verify later in the experiments, these frameworks cannot be easily extended to cross-resolution recognition.

In this paper, we propose a novel metric learning approach called quintuplet. The goal is to learn a suitable and efficient metric function to effectively measure the similarity of face samples, under which the similarity of positive pairs is enlarged and that of negative pairs is reduced as much as possible. The proposed metric learning approach quintuplet is similar to triplet network [6, 8, 9], with five inputs used instead of three. More detailed description of proposed approach is given in section 3.

In summary, we make the following contributes:

- Presemt novel metric learning procedure that is useful for matching cross-resolution images
- Evaluate proposed method on two datasets
- Compare method with state-of-art face recognition and identification systems for matching cross-resolution images

2. Related work

The related work review is divided into three parts: (1) review of cross-resolution face recognition related work, (2) metric learning and (3) deep learning.

2.1. Cross-resolution Face Recognition

In the literature, several approaches have been presented to match cross-resolution face images. These approaches can be classified into two categories: super-resolution and transformation based approaches. Super-resolution based approaches for cross-resolution matching enhance the low quality probe image before recognition. On the other hand, transformation based approaches extract features that are resilient to resolution changes and matching cross-resolution face images.

Super-resolution based approaches: In recent years, many super-resolution face recognition approaches have been presented [32] - [37]. Due to environmental variations and distortions, the proposed approaches failed to significantly improve the recognition performance. However, there are some approaches that simultaneously optimize both super-resolution and face recognition [39, 5, 40]. Proposed super-resolution technique from [40] improved face recognition performance for the very low resolution problem.

Transformation based approaches: Unlike super-resolution, another method to match cross-resolution images is to downsample high resolution images to the level of low resolution images before matching. To address this problem, several approaches were proposed [41] - [38]. Li *et al.* [41] proposed to project both high resolution and low resolution images to a feature space using coupled mappings. Biswas *et al.* [42] proposed a multidimensional scaling approach to simultaneously transform the features from high resolution gallery and low resolution probe images. Lei *et al.* [43] proposed a local frequency descriptor based on the magnitude and phase information to match cross-resolution face images in the frequency domain. Shekhar *et al.* [24] proposed a generative approach using the information from high resolution gallery to match low resolution probe images with illumination variations. Lei *et al.* [38] proposed a coupled discriminant analysis for heterogeneous face recognition (matching high vs. low resolution images). To maintain the discriminative power and generalizability of their approach, they utilized multiple samples from different resolutions along with locality information in the kernel space.

2.2. Metric Learning

Many metric learning algorithms have been proposed over the past decade, and some of them have been successfully applied to address the problem of face verification in the wild [10] - [14]. Metric learning aims to learn a similarity (distance) function. Traditional metric learning [15] - [18] usually learns a matrix $\mathbf{A}$ for a distance metric $||x_1 - x_2||_{\mathbf{A}} = \sqrt{(x_1 - x_2)^T \mathbf{A}(x_1 - x_2)}$ upon the given features x_1, x_2.

Recently, prevailing deep metric learning [19, 20, 21,

22, 23, 6] usually uses neural networks to automatically learn discriminative features x_1, x_2 followed by a simple distance metric such as Euclidean distance. Most widely used loss functions for deep metric learning are contrastive loss [26, 27] and triplet loss [25, 6, 8], and both impose Euclidean margin to features. Our proposed approach is similar to triplet approach, but instead of three factors uses five.

Quadruplet metric learning: After triplet loss, quadruplet loss was introduced in [53] where authors used quadruplet network to investigate the presence of smile on face images. Their approach aims at efficiently modeling similarity from rich or complex semantic label relationships. Proposed method is shown on Fig. 2.

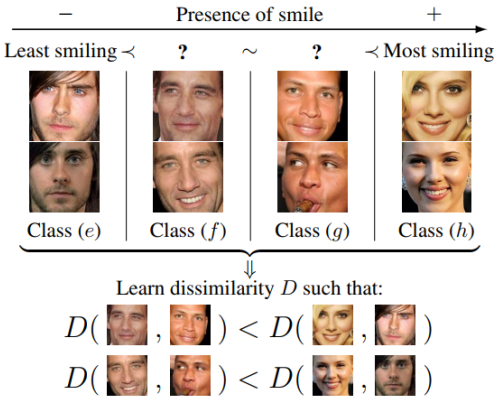

Figure 2. Quadruplet-wise (Qwise) strategy on 4 face classes ranked according to the degree of presence of smile. Instead of working on pairwise relations that present some flaws, Qwise strategy defines quadruplet-wise constraints to express that dissimilarities between examples from (f) and (g) should be smaller than dissimilarities between examples from (e) and (h). [53]

2.3. Deep Learning

Deep learning is arguably one of the most active research area in the past few years. Generally, deep learning aims to learn hierarchical feature representations by building high-level features from low-level ones. Existing deep learning methods can be categorized in three classes: unsupervised, supervised and semi-supervised, and they have been successfully applied to many visual analysis applications such as object recognition [14], human action recognition [28, 29] and face verification [30]. While many attempts have been made on deep learning in feature engineering such as deep belief network [31], stacked autoencoder [29], and convolutional neural networks [28], little progress has been made in metric learning with a deep architecture. More recently, Cai et al. [10] proposed a nonlinear metric learning method by combining the logistic regression and stacked independent subspace analysis. Differently, our proposed quintuplet method employs a network to learn the distance metric where the back propagation algorithm can be used to train the model.

3. Metric learning

Metric learning is the task of learning distance function over . A metric or distance function has to obey four axioms: non-negativity, identity of indiscernibles, symmetry and subadditivity / triangle inequality. In practice, metric learning algorithms ignore the condition of identity of indiscernibles and learn a pseudo-metric. [55]

In next two subsections, we describe triplet metric learning and present our proposed approach quintuplet metric learning. We will briefly discuss the similarities between them and how they differ from each other. In chapter 4, we present results both for triplet metric learning and for quintuplet metric learning.

3.1. Triplet metric learning

Triplet network aims to learn feature embedding by optimizing the relative distance between the samples from the same class and dissimilar classes. Fig. 3 illustrates the idea of metric learning using triplet network.

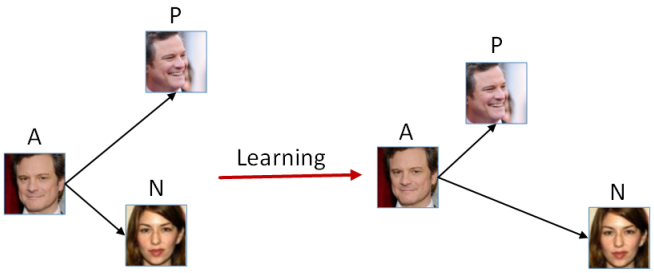

Figure 3. Triplet example before and after learning. The essence of metric learning is to make Anchor (A) and Positive (P) image as close as possible and at the same time distance between Anchor (A) and Negative (N) image as far as possible. [6]

Triplet loss is learned on a series of triplets, which consist of images A (anchor image), P (positive class) and N (negative class). The goal of the triplet loss is to keep A closer to P than N. The triplet loss is formulated as:

$$L_{trp} = \sum_{a,p,n}^{i} [||f(x_a) - f(x_p)||_2^2 - ||f(x_a) - f(x_n)||_2^2 + \alpha], \tag{1}$$

where $f(x_a)$ is feature vector extracted from Anchor (A) image, $f(x_p)$ feature vector from Positive (P) image and $f(x_n)$ feature vector from Negative (N) image.

Triplet network structure is illustrated on Fig. 4.

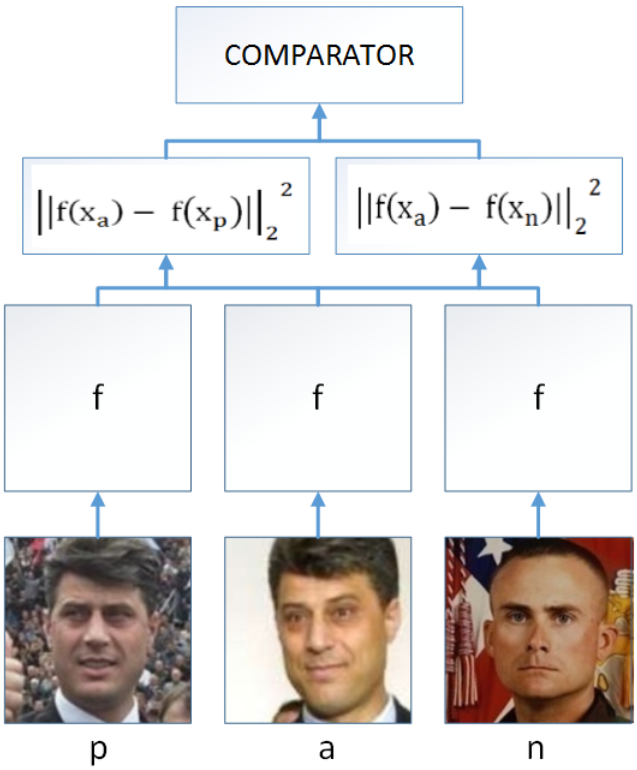

Figure 4. Firstly, features are extracted from Anchor (a), Positive (p) and Negative (n) images. Aim of triplet loss is to minimize distance between Anchor (a) and Positive (p) images and to maximize distance between Anchor (a) and Negative (n) images.

Training is performed by feeding the network with samples where A and P are of the same class, and N is of different class. The network architecture allows the task to be expressed as a 2-class classification problem, where the objective is to correctly classify which of P and N is of the same class A. 4

3.2. Quintuplet metric learning

Quintuplet metric learning is quite similar to triplet metric learning, but instead of just positive and negative classes, we add downsampled images for positive and for negative class. This gives as quintuples x_A, x_P, x_N, x_{Pd} and x_{Nd}. We use them to define quintuplet loss.

For quintuplet metric learning, we had to downsample images for positive and negative classes. For downsampling, we resized images from 224x224 to 16x16 and then resized images back to 224x224. Using bicubic interpolation, as can be seen from Fig. 5, a lot of information is lost.

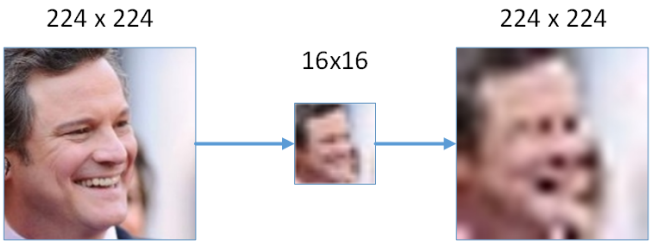

Figure 5. We take image of size 224x224 and resize it to 16x16 and then we resize it back to 224x224.

Quintuplet loss is learned on a series of quintuplets, which consist of images A (anchor image), P (positive class), N (negative class), Pd (positive downsampled class), Nd (negative downsampled class). The goal of the quintuplet loss is to keep A closer to P and Pd than to N and Nd. The quintuplet loss is formulated as:

$$L_{quin} = \sum_{A,P,N,Pd,Nd}^{i} [||f(x_A) - f(x_P)||_2^2 - ||f(x_A) - f(x_N)||_2^2 + ||f(x_A) - f(x_{Pd})||_2^2 - ||f(x_A) - f(x_{Nd})||_2^2 + \alpha], \quad (2)$$

where $f(x_A)$, $f(x_P)$, $f(x_N)$, $f(x_{Pd})$, $f(x_{Nd})$ are feature vectors extracted from Anchor (A), Positive (P), Negative (N), Negative downsampled (Pd), Positive downsampled (Nd) images, respectively.

Quintuplet network structure is illustrated in Fig. 6.

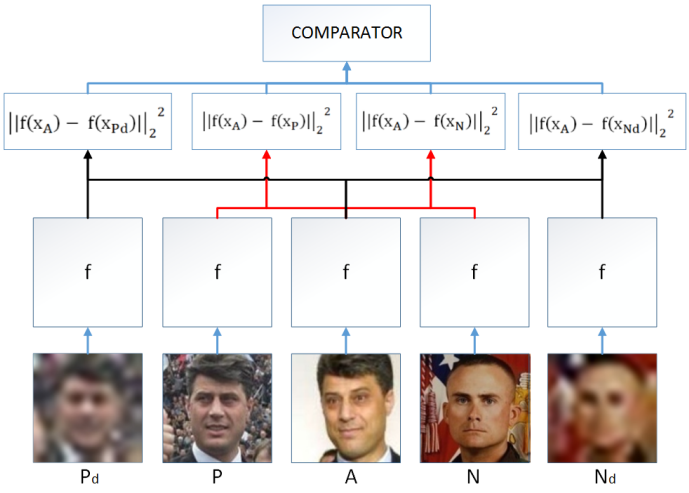

Figure 6. Firstly, features are extracted from Anchor (A), Positive (P), Negative (N), Positive downsampled (Pd) and Negative downsampled (Nd) images. Aim of quintuplet loss is to minimize distance between pairs Anchor (A) - Positive (P) and Anchor (A) - Positive downsampled (Pd) images and to maximize distance between pairs Anchor (A) - Negative (N) and Anchor (A) - Negative downsampled (Nd) images.

Training is performed by feeding the network with samples called quintuples, where A, P and Pd are of the same class, and N and Nd are of different class. Since we still have only two classes, task is still expresses as a 2-class classification problem. Similar as for learning metric distance using triplets, we used whole CASIA-WebFace dataset. Every image was used once as an Anchor (A), Positive (P), Positive downsampled (Pd), Negative (N) and Negative downsampled (Nd) were choosen randomly.

4. Experiments

4.1. Datasets

For training and testing three datasets were used. For training we used CASIA-WebFace dataset [44] and for testing Labeled Faces in the Wild [45] and SCface - surveillance cameras face database [46].

CASIA-WebFace: The CASIA-WebFace dataset [44] contains 494,414 images of 10,575 subjects and it is free for research and educational purposes. Taken into account that face recognition dataset only needs face image and identitiy, images can be obtained relatively simply by just crawling different web pages. This dataset was obtained by scrapping images and their identities from IMDb website. Each celebrity has an independent page on the website. This makes in pretty simple to scrap the data from it. Images were isolated so that only one face is on each of the images.

We cropped the faces from the images and resized them to 224x224 pixels. Whole dataset was downsampled to 16x16 pixels and then resized to 224x224 pixels. Downsampled dataset was also used to learn quintuplet metric. Downsampling process is illustrated on Fig. 5.

Labeled Faces in the Wild: The Labeled Faces in the Wild [45], often reffered to as LFW is a dataset of face photographs designed for studying problem of unconstrained face recognition. It consists of 13233 images fo 5749 people. All of the images are 250 x 250 pixels in size and were gathered using the Viola-Jones face detector. We used LFW dataset for evaluation of our proposed approach. As a comparison benchmark, 10-fold cross validation using randomly generated splits was used. We additionally cropped faces from the images and resized them to 224 x 224 pixels. Results are presented in the form of ROC curves with inforamtion about area under the curve (AUC) and equal error rate (EER).

SCface: The SCface dataset [46] consists of 4160 static face images of 130 subjects. Images were taken in uncontrolled indoor environment using five video surveillance cameras of various qualities. Images from different quality cameras should mimic real-world conditions and enabled robust face recognition algorithms testing, emphasizing, different law enforcement and surveillance use case scenarios. In datasert paper, four evauluation protocols were proposed: DayTime tests, NightTime test, Performance metrics and Training scenario.

For evaluation of our proposed approach, we used the DayTime tests. This test is pretty much straightforward, where we compare the mug shot image to all the other images (from cam1 to cam5) using three different distances from where images were taken (4.20m, 2.60m and 1m). Five different cameras were used which gives us 15 possible different probes sets, verying both in distances from camera and in camera qualities. Comparing the probe image to one gallery image is the most logical real-world (law enforcement) scenario.

List of surveillance cameras used to obtain images that we used for evaluation:

- cam1 - Bosch LTC0495/51
- cam2 - Shany WTC-8342
- cam3 - J&S JCC-915D
- cam4 - Alarmcom VFD400 - 12B
- cam5 - Shany MTC-L1438

As for both previous datasets, we cropped faces from images and resized images to 224 x 224 pixels. Results are presented as rank n identification rate, where we present results for first 20 ranks.

4.2. Performance metrics

Performance of metric learned is demonstrated on both face recognition problem (LFW dataset) and face identification problem (SCface dataset).

Graphically, the results on LFW dataset are presented as Receiver Operating Characteristic (ROC) [52], which shows relation between true positive rate (TPR) and false positive rate (FPR) for different thresholds.

As quantitative measurements, results for LFW dataset are shown in a table as Area Under the ROC Curve (AUC) and Equal Error Rate (EER) measurements. EER is the value where 1-TPR equals FPR. Value of AUC is always between 0 and 1, but since random guessing when facing 2-class classification problems gives us result 0.5, we expect results to be better than 0.5. AUC value can be interpreted as probability, with which classifier correctly classifies random sample.

Visually, the results on SCface dataset are presented as Cumulative Match Characteristic (CMC). CMC curve is rank based metric. Each probe sample is compared against all gallery samples. The resulting scores are sorted and ranked. We then determine the rank at which a true match occurs. We measure True Positive Identification Rate (TPIR), which is the probability of observing the correct identitiy within the top K ranks.

As quantitative measurements, we calculated how many times correct identity was observed within top 20 ranks. Results are given in Table 2.

4.3. Used Algorithms

Two algorithms were used for metric learning and evaluation. VGG neural network [47] was used for extractomg feature vectors from images for metric learning on CASIA-WebFace dataset [44] and evaluation on LFW [45] and SCface [46] and Open Source Biometric Recognition (OpenBR) [48] was used for evaluation on LFW and SCface datasets.

OpenBR: Open Source Biometric Recognition (OpenBR) [48] is a framework for investigating new modalities, improving existing algorithms, interfacing with commercial systems, measuring recognition performance, and deploying automated biometric systems. The project is designed to facilitate rapid algorithm prototyping, and features a mature core framework, flexible plugin system, and support for open and closed source development. [48]

While algorithms implemented within the OpenBR project are applicable to many biometric disciplines, particular effort has been devoted to the scenario of facial recognition. The default face recognition algorithm in OpenBR is based on the Spectrally Sampled Structural Subspaces Features (4SF) algorithm [49]. 4SF is a statistical learning based algorithm used previously to study the impact of demographics [50] and aging [51] on face recognition performance. [48]

The 4SF algorithm is not claimed to be superior to other techniques in the literature, instead it is representative of modern face recognition algorithms in its use of face representations and feature extraction. The 4SF algorithm demonstrates strong accuracy improvements through statistical learning, allowing OpenBR to differentiate itself from commercial systems in its ability to be trained on specific matching problems like heterogeneous face recognition. [48]

VGG: The VGG (Visual Geometry Group) neural network [47] is a 16 weight layer convolutional neural network used for image recognition. Authors were investigating the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. Main contribution of paper [47], where VGG is presented, is a thorough evaluation of networks of increasing depth using an architecture with very small (3x3) convolutional filters, which shows that a significant improvement on the prior-art configurations can be achieved by pushing the depth to 16-19 weight layers, which is substantially deeper than what has been used in the prior art.

VGG neural network can be used in many frameworks: Caffe, The Microsoft Cognitive Toolkit, TensorFlow, theano, Torch, MXnet, Chainer and Keras. We used Keras which is build on TensorFlow.

We used a slightly different VGG model which was trained in our laboratory, on the same dataset that authors used to train VGG-16 model. We then used this model to extract feature vectors from images that were used for metric learning and for algorithm evaluation. Features that we got from CASIA-WebFace dataset were then used for metric learning and features that we got from SCface dataset were used for algorithm evaluation.

4.4. Face Recognition

We used VGG neural network and OpenBR framework in a face recognition scanario. Firstly, we evaluated both algorithms on LFW and SCface datasets, before metric learning was done above VGG neural network. Then we used triplet and quintuplet metric learned to repeat evaluation and to prove that metric learning improves recognition rate. Furtherly, we proved that proposed quintuplet method outperforms triplet method.

For training we used 494414 triplets from CASIA-WebFace dataset [44]. Every image from dataset was used once for an anchor, positive and negative classes were chosen randomly from the set. Results are presented in chapter

Table 1 shows AUC and EER values for evalution done on LFW dataset. VGG in Table 1 is the original neural network model with 16 layers. VGG+Triplet and VGG+Quintuplet are results that we got after we learned triplet and quintuplet metric above feature vectors that we got using original VGG network, respectively.

Table 1. UAC and EER values for tests on LFW dataset.

Algorithm	**AUC**	**EER**
OpenBR	0.8460	0.233
VGG	0.9277	0.1477
VGG + Triplet	0.9692	0.09
VGG + Quintuplet	0.9741	0.0826

Fig. 7 shows ROC curves for evaluation on LFW dataset. As we can see from the Fig. 7, the best recognition improvement is achieved using quintuplet network. Using triplet network is slightly worse, because triplet metric wasn't trained for that kind of circumstances. Quintuplet metric was learned for face recognition using cross resolution images, especially those with very low quality. Better and more representative results are visible on Fig. 8 which shows CMC curve for evaluation on SCface dataset.

Table 2 shows results for Rank-N identification done on SCface dataset. All results are in percentages. Each number presents percentage of correctly identified people within the top N ranks.

Figure 8 shows CMC curves for evaluation on SCface dataset. As we can see from the Fig. 8, the best recognition is achieved using OpenBR. Beside that, quintuplet met-

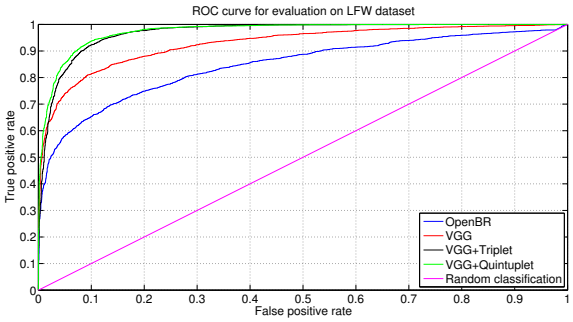

Figure 7. ROC curves for evaluation on LFW dataset. Blue line belongs to OpenBR, red line to original VGG, black line to VGG+Triplet and green line belongs to VGG+Quintuplet. Pink line presents random classification rate. We see that best result is achieved using quintuplet metric learning.

Table 2. Rank-1 recognition rate values in percentages for tests on SCface dataset.

Rank	OpenBR	VGG	VGG+Triplet	VGG+Quintuplet
1	21	4	3	5

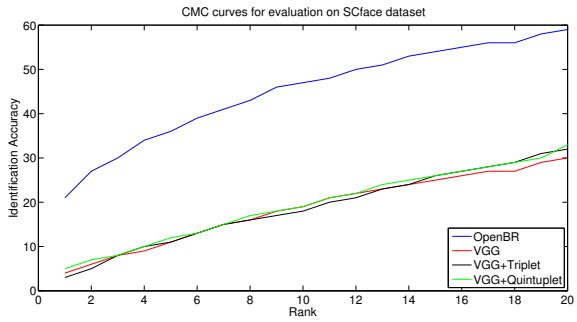

Figure 8. CMC curves for evaluation on SCface dataset. Blue line belongs to OpenBR and shows much better performance than VGG or our learned metrics. Green line belongs to VGG+Quintuplet, black to VGG+Triplet and red belongs to original VGG net. We can see that quintuplet metric learned improves results when comparing with triplet metric or original VGG net.

ric leared outperforms original VGG neural network, which means that proposed method is useful and needs to be further investigated. Problem for such poor results using VGG is in the network itself. It wasn't trained for usage on low-resolution images.

When origial VGG neural network from the VGG paper was used, we got far better results than with the net that was trained in our lab. Just for Rank-1, we got 70 % identification accuracy. When we tried learned metric, this percentages dropped, because features used for learning metric weren't extracted using original VGG network, but our lab's one.

Percentages on SCface prove that our proposed method improves face recognition, but a lot of work needs to be done in that area. For start, we will use original VGG neural network to extract features from CASIA web-face dataset, which will then be used to learn triplet and quintuplet metrics again. We expect results to improve a lot, because original network is returns much better results on cross-resolution images. We will also test metrics used on more datasets than just mentioned two, because they don't give enough transparency. Both LFW and SCface have, although used for other purposes, images of relatively high quality. Even though SCface dataset contains surveillance video images, their quality is not that low.

Most time consuming task was feature extraction using VGG for CASIA-webface dataset where we had to extract 1 million features, because we used original dataset and downsampled one, which sums up to about 28 GB of data. Additionally, learning triplet and quintuplet metric distance took a lot of time (12 hours for each attempt for each method), because we needed about 50 GB of RAM, but we didn't have GPU's on the same machine, so we had to learn using only CPU. This took about 2 hours for each model with batch size of 256 triplets/quintuplets, 2048 steps per epoch and 10 epochs. Another problem we encountered during learning was convergence of the model. We needed to stop learning a few times because model over-fitted and became useless.

5. Conclusion and futher work

In this paper, we proposed method called Quintuplet metric learning. The method is not used instead of existing neural network for feature extraction, but above neural network. This means that it is used once we already have our features, to better calculate matching score between images. The idea of quintuplet metric learning is similar to triplet metric learning described in [8], but also uses downsampled images for learning. This means that beside positive and negative class image of high-quality, positive and negative class images of low-quality are also used. To downsample images, we just resized images from 224x224 pixels to 16x16 pixels and then back again to 224x224. We managed to learn network to behave and classify images with greater performance than without proposed metric learning method, but can be improved furtherly. For start, we should use more appropriate neural network to extract features for learning. Once we have done extraction step, we can try to learn more efficient metric. We will not only use downsampled images of positive and negative class, but various combinations where downsampled image of anchor class will also be included.

References

[1] "http://www.cbc.ca/news/background/london_bombing/investigation_timeline.html", (last accessed: January, 5, 2013).

[2] "http://www.hindustantimes.com/india-news/newdelhi/who-s-keeping-watch/article1-908391.aspx", (last accessed: January, 5, 2013).

[3] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," IEEE TPAMI, vol. 32, no. 6, pp. 1127-1133, 2010.

[4] J. Yang, J. Wright, T. S. Huang and Y. Ma, "Image super-resolution via sparse representation," IEEE TIP, vol. 19, no. 11, pp. 2861-2873, 2010.

[5] P. H. Hennings-Yeomans, S. Baker and V. Bhagavatula, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in CVPR, 2008, pp. 1-8.

[6] F. Schroff, D. Kalenichenko and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[7] Y. Wen, K. Zhang, Z. li and Y. Qiao, "A discriminative feature learning approach for deep face recognition," European Conference on Computer Vision, Springer, 2016.

[8] E. Hoffer and N. Ailon, "Deep Metric Learning using Triplet Nework," ICLR 2015.

[9] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition", BMVC, 2015.

[10] X. Cai, C. Wang, B. Xiao, X. Chen, and J. Zhou, "Deep nonlinear metric learning with independent subspace analysis for face verification," ACM MM, pp. 749752, 2012.

[11] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen, "Fusing robust face region descriptors via multiple metric learning for face recognition in the wild," CVPR, pp. 35543561, 2013.

[12] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Informationtheoretic metric learning," In ICML, pp. 209216, 2007.

[13] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," ICCV, pp. 498505, 2009.

[14] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. Lecun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," CVPR, pp. 18, 2007.

[15] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning with application to clustering with sideinformation," NIPS, 2003.

[16] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," Journal of Machine Learning Research, pp. 207244, Feb, 10, 2009.

[17] M. Kstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," CVPR, 2012.

[18] Y. Ying and P. Li, "Distance metric learning with eigenvalue optimization,"JMLR, pp. 126, Jan, 13, 2012.

[19] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," CVPR, 2014.

[20] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou, "Multimanifold deep metric learning for image set classification," CVPR, 2015.

[21] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," CVPR, 2016

[22] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," CVPR, 2014.

[23] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," NIPS, 2014.

[24] S. Shekhar, V. M. Patel, and R. Chellappa, "Synthesis-based recognition of low resolution faces," in International Joint Conference on Biometrics, pp. 16, 2011.

[25] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking, "CVPR, 2014.

[26] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification,"CVPR, 2005.

[27] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping, "CVPR, 2006.

[28] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," PAMI, pp. 221231, 2013.

[29] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," CVPR, pp. 33613368, 2011.

[30] G. B. Huang, H. Lee, and E. G. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," CVPR, pp. 25182525, 2012.

[31] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets,"Neural Computation, pp. 15271554, 2006.

[32] H. Huang and H. He, "Super-resolution method for face recognitionusing nonlinear mappings on coherent features," IEEE Transactions on Neural Networks, vol. 22, no. 1, pp. 121130, 2011.

[33] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 9, pp. 11671183, 2002.

[34] A. Chakrabarti, A.N. Rajagopalan, and R. Chellappa, "Super-resolution of face images using kernel pca-based prior," IEEE Transactions on Multimedia, vol. 9, no. 4, pp. 888892, 2007

[35] H. Chang, D. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in Proceedings of Computer Vision and Pattern Recognition, 2004, pp. 275282.

[36] B. Li, H. Chang, S. Shan, and X. Chen, "Locality preserving constraints for super-resolution with neighbor embedding," in Proceedings of International Conference on Image Processing, pp. 11891192, 2009.

[37] C. Liu, H. Shum, and W. Freeman, "Face hallucination: Theory and practice," International Journal of Computer Vision, vol. 75, no. 1, pp. 115134, 2007

[38] Z. Lei, S. Liao, A. K. Jain, and S. Z. Li, "Coupled discriminant analysis for heterogeneous face recognition," IEEE Transactions on Information Forensics and Security, vol. 7, no. 6, pp. 17071716, 2012.

[39] K. Jia and S. Gong, "Multi-modal tensor face for simultaneous superresolution and recognition," in Proceedings of International Conference on Computer Vision, pp. 16831690, 2005.

[40] W. W. W. Zou and P. C. Yuen, "Very low resolution face recognition problem," IEEE Transactions on Image Processing, vol. 21, no. 1, pp. 327340, 2012.

[41] B. Li, H. Chang, S. Shan, and X. Chen, Low-resolution face recognition via coupled locality preserving mappings, IEEE Signal Processing Letters, vol. 17, no. 1, pp. 2023, 2010.

[42] S. Biswas, G. Aggarwal, and P. J. Flynn, "Pose-robust recognition of low-resolution face images," in Proceedings of Computer Vision and Pattern Recognition, pp. 601608, 2011.

[43] Z. Lei, T. Ahonen, M. Pietikainen, and S. Li, "Local frequency descriptor for low-resolution face recognition," in Proceedings of International Conference on Automatic Face Gesture Recognition and Workshops, pp. 161166, 2011.

[44] D. Yi, S. Liao and S. Z. Li, "Learning Face Representation from Scratch," Computer Vision and Pattern Recognition, 2014.

[45] G. B. Huang, M. Remesh and others. "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," University of Massachusetts, Amherst, Technical Report 07-49, 2007.

[46] M. Grgic, K. Delac and S. Grgic. "SCface - surveillance cameras face database," Multimed Tools Appl, pp. 863-879, 2011.

[47] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition," International Conference on Learning Representations, 2015.

[48] J. C. Klontz, B. F. Klare and others. "Open Source Biometric Recognition," IEEE Biometrics: Theory, Applications and Systems, 2013.

[49] B. Klare. "Spectrally sampled structural subspace features (4SF)," In Michigan State University Technical Report, MSUCSE-11-16, 2011.

[50] B. Klare, M. Burge, J. Klontz, R. Vorder Bruegge and A. Jain. "Face recognition performance: Role of demographic information," IEEE Trans. on Information Forensics and Security, 7(6):17891801, 2012

[51] B. Klare and A. K. Jain. "Face recognition across time lapse: On learning feature subspaces," In IEEE International Joint Conf. on Biometrics (IJCB), 2011, pages 18.

[52] T. Fawcett. "An introduction to ROC analysis," Pattern Recognition Letters, 27, pages 861-874, 2006.

[53] M. T. Law, N. Thome and M. Cord. "Quadruplet-wise Image Similarity Learning," Computer Vision (ICCV), 2013.

[54] T. Fu, W. Chiu and Y. F. Wang. "Learning Guided Convolutional NuuralNetworks for Cross-Resolution Face Recognition," International Workshop on Machine Learning for Signal Processing, 2017.

[55] "https://en.wikipedia.org/wiki/Similarity_learning", (last accessed: January, 16, 2018)