Low-Resolution Face Recognition with Deep Convolutional Features in the Dissimilarity Space

Mairelys Hernández-Durán, Yenisel Plasencia-Calaña,
and Heydi Méndez-Vázquez(✉)

Advanced Technologies Application Center, 7ma A # 21406, Playa, Havana, Cuba
{mhduran,yplasencia,hmendez}@cenatav.co.cu

Abstract. In video surveillance and others real-life applications, it is usually needed to match low resolution (LR) face images against high-resolution (HR) gallery images. Although extensive efforts have been made, it is still difficult to find effective representations for low-resolution face recognition due to the degradation in resolution together with facial variations. This paper makes use of alternative representations based on dissimilarities between objects. Unlike previous works, we construct the dissimilarity space on top of deep convolutional features. We obtain a more compact representation by using prototype selection methods. Besides, metric learning methods are used to replace the standard Euclidean distance in the dissimilarity space. Experiments conducted on two data sets particularly designed for low-resolution face recognition showed that the proposal outperforms state-of-the-art methods, including some neural networks designed for this problem.

Keywords: Face recognition · Low-resolution
Dissimilarity representation · Convolutional networks

1 Introduction

Face recognition systems based on good quality and high-resolution images have obtained good results in practical applications [1]. However, high-resolution (HR) images are often not available, especially in real scenarios with uncontrolled conditions. On the contrary, low-resolution (LR) images are usually captured on these environments, which generates the so-called dimensional mismatch problem (different resolutions between gallery/probe images).

Different approaches have been proposed for low-resolution face recognition [2–4]. However, the performance of traditional methods suggests that current feature representation approaches are not enough to cope with the low resolution problem [5]. Recently, alternative representations based on dissimilarities between objects have been explored. It was shown in [6] that discriminative information for classification can be obtained if the LR images are analyzed in

Y. Hernández Heredia et al. (Eds.): IWAIPR 2018, LNCS 11047, pp. 95–103, 2018.
https://doi.org/10.1007/978-3-030-01132-1_11

the context of dissimilarities with other images and good results can be achieved in low-resolution face recognition.

The dissimilarity space (DS) representation brings a proximity information between prototypes and the rest of the training set, instead of representing the characteristics of each object individually [7]. It is necessary a base representation to build a dissimilarity space, unless an expert directly defines it. Intuitively, if more discriminative features are used as basis, the DS will be more effective. Recently, convolutional neural networks (CNNs) have been successfully applied in many domains and also in this context [3]. Taking this into account, we believe that features obtained from a deep architecture could replace traditional features as basis to construct a DS and thus, a more robust representation can be obtained.

In this paper, we propose a method for low-resolution face recognition that builds a DS on top of features learned by a convolutional neural network. We take advantage of a pre-trained network to avoid costs in terms of computer resources, and time in adjusting the parameters of the training stage. We also propose a reduced DS by using prototype selection methods and a learned metric is used in order to improve the classification accuracy in this space. Extensive experimental evaluations are conducted on two complex databases, where our method achieves state-of-the-art results, outperforming previous methods based on dissimilarity representations as well as some others that use deep convolutional features directly.

2 Dissimilarity Space from Deep Convolutional Features

The general scheme of the proposal can be found in Fig. 1. First, the low-high strategy proposed in [6] is used to deal with the dimensional mismatch. Deep convolutional features are obtained by means of the pre-trained VGG-Face [8] network and the DS is constructed on top of these features. Finally, a metric learning is applied to replace the standard Euclidean distance in the DS.

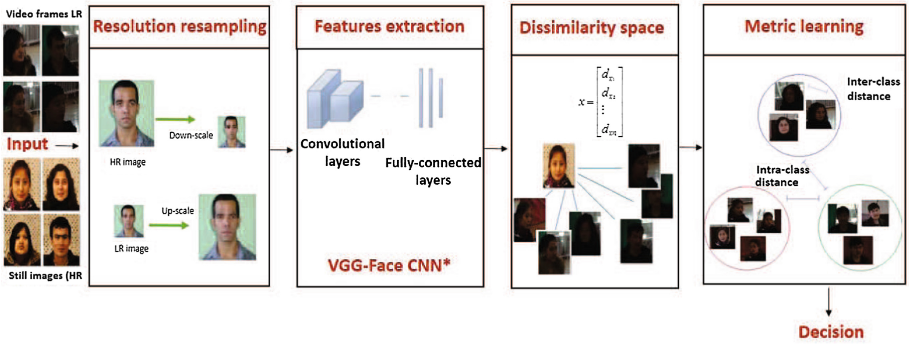

Fig. 1. General flow of the process.

2.1 Deep Convolutional Features Extraction

Recently, CNNs have become the state-of-the-art for many tasks [3]. In deep architectures, feature layers are not manually designed but they have learning procedures focused on general purpose, while specialization is given by the training data of the network. Despite their effectiveness, CNNs are difficult to train because they have many hyper-parameters (learning rate, momentum, etc.). Achieving the best combination requires complex calculations and powerful equipments. However, once the model is learned, it can be used to solve other similar problems without any additional training. Since it is convenient for users who do not have the resources needed for training, the use of intermediate layers from pre-trained networks have been growing recently [9].

Different already trained CNNs models are publicly available. We have selected VGG-Face [8] that currently reports one of the best performances for face recognition. We follow the idea in [9] which suggests that when using a pre-trained network for a given task, intermediate representations may achieve better results for a similar task. Considering this, we get not only the original net-descriptor from the last fully conected layer (dimension 4096), but also an average pooling descriptor obtained from the third convolutional layer of the 8th block (dimension 512).

2.2 Dissimilarity Space and Prototype Selection

Once the deep convolutional features are extracted from face images, cosine similarity is used as the distance measure to create the DS. The DS was first introduced by Pękalska and Duin [10]. Let X be the space of objects, let $R = \{r_1, r_2, ..., r_k\}$ be the set of prototypes such that $R \in X$, and let $d : X \times X \rightarrow \mathbb{R}^+$ be a suitable dissimilarity measure for the problem. For a finite training set $T = \{x_1, x_2, ..., x_l\}$ such that $T \in X$, a mapping $\phi_R^d : X \rightarrow \mathbb{R}^k$ defines the embedding of training and test objects in the DS by the dissimilarities with the prototypes:

$$\phi_R^d(x_i) = [d(x_i, r_1)\ d(x_i, r_2)\ ...\ d(x_i, r_k)]. \tag{1}$$

The prototypes may be chosen based on some criterion or even at random; but they should have good representation capabilities [11]. These methods allow to generate a new space to represent the whole set while keeping or even improving its discriminative power. We will focus on selective schemes since we are interested in exploiting a given dissimilarity matrix computed directly from the initial features. Some methods from this group were evaluated in this work:

Random. The selection of a representation set defines a dissimilarity space in which the entire training set is used to train the classifier. For this reason, even a randomly selected representation can be useful as a basic procedure [7].

Farthest-First Traversal (FFT) [12]**.** It selects an initial prototype at random from a sample of objects S, being $S = \{X_1, X_2, ..., X_k\} \subset X$; and then, each new prototype is defined as the farthest element of S from all previously chosen prototypes. It stops when it reaches the desired cardinality without any refinement.

kCentres [7]**.** It groups k-centroid objects from a symmetric distance matrix. These centroids are chosen so that the maximum distance of the objects with their closest centroid is minimized. Initialization can be randomly done. The value of k can alter the representation of the new space, therefore some adjustment would be necessary during the experiments.

Center [7]**.** As its name indicates, it selects prototypes situated in the center of a given set. Due to their central position all prototypes are structurally similar which may origin redundant prototypes. However, samples at the border are not considered, and thus, the set of prototypes is not negatively influenced by outliers.

Genetic Algorithm (GA) [11]**.** It is a search method based on heuristics that mimic natural mechanisms, by evolving individuals created after each generation by the best fitted ones. The basic idea is to maintain a population of chromosomes, which represent plausible solutions to a particular problem, and the evolution of this population through a process of competition and controlled variation. We use a variant proposed in [11] (GAsup), in which a stage of clustering is added before the random initialization which guarantees a faster convergence of the method.

2.3 Metric Learning in the Dissimilarity Space

Metric learning algorithms take advantage of prior information in form of labels over standard similarity measures. The effectiveness of using a learned metric to improve the DS representation was shown in [6]. In the present study, we selected two metric learning methods to replace the standard Euclidean distance in the dissimilarity space: LDA (Linear Discriminant Analysis) [13] and LMNN (Large Margin Nearest Neighbor) [14]. LDA computes a linear projection L that maximizes the amount of between-class variance relative to the amount of within-class variance. The linear transformation L is chosen to maximize the ratio of between-class to within-class variance, subject to the constraint that L defines a projection matrix, in a way that better separation between objects of different classes is achieved. In LMNN, the metric is trained following the criterion that the k-nearest neighbors belong to the same class, while samples from different classes are separated by a large margin. These metrics are used in this work to compute distances in the nearest neighbor (1-NN) classification.

3 Experimental Evaluation and Discussion

Two public face datasets designed for low-resolution face recognition evaluation were selected for the experiments. The **COX database** [15] includes 1000 still HR images and 3,000 videos corresponding to 1000 subjects. The still images were captured with a high-resolution camera. Videos were taken simulating a video-protection environment, with three cameras at different locations. Video-to-Still (V2S) and Still-to-Videos (S2V) evaluation protocols are evaluated. On the other hand, **SCFace database** [1] is composed by images captured simulating surveillance scenarios, making this database one of the most suitable set to evaluate LR case. It contains images of 130 subjects including high quality frontal images (mugshot). To capture the LR ones, three distances were used; each one with five video cameras with different qualities and resolutions.

Table 1. Results in Cox with some prototype selection methods.

	Average pooling descriptor			Original net-descriptor		
Method	Camera 1	Camera 2	Camera 3	Camera 1	Camera 2	Camera 3
random	81.56	80.96	90.75	75.45	76.22	86.41
FFT [12]	80.22	82.80	89.43	78.33	81.45	87.65
kcentres [7]	82.36	85.98	90.26	81.34	83.27	90.18
center [7]	86.85	88.58	**95.86**	84.49	86.18	**92.90**
GAsup [11]	84.74	88.36	94.29	82.56	84.22	90.36

3.1 Experiments on COX Database

Considering that COX database has a larger number of subjects, we first conducted an experiment on it, in order to compare the performance of the two network descriptors (the original net-descriptor and the average pooling descriptor) and some prototype selection methods under the proposed scheme. Both networks descriptors are obtained from the pre-trained VGG-Face model. For the prototype selection methods, similar parameters to those used in [11] were selected, with a cardinality equal to 120. For down-scaling and up-scaling the images in the low-high strategy, bicubic interpolation was used. Every video contains a large number of frames, thus we selected 20 frames distributed in a spaced manner to represent the whole video, which are averaged for obtaining the final face descriptor.

We report in Table 1 a comparison between some prototype selection methods to obtain a reduced DS, using LDA metric learning. It shows the recognition rates obtained for 10 random iterations of the V2S protocol, in which three videos from 300 subjects are used for training (900 in total), and the rest 700 videos from each camera are used for testing. We found that a small set of prototypes is sufficient to obtain a good representation. From this comparative study we can see that the *center* method reported the best performance followed by the *GAsup* method with a small difference. We consider *center* is the best in this setup because automatically learned representation are sensitive to outliers, i.e., those objects that are highly deviated from its representation compared with the rest of the dataset. From the results, we can see that the average pooling descriptor reports the best results in comparison with those obtained with the original net-descriptor. This implies that also in the DS, intermediate layers are able to obtain more discriminative representations when a pre-trained network is used. This is an advantage since it brings a much simple and effective representation than the original representation. On the other hand, it can be seen that in general the best results are obtained for Camera 3, the less affected by low resolution.

Table 2. Recognition rates on COX database following the S2V and V2S protocols.

Method	V2S-cam1	V2S-cam2	V2S-cam3	S2V-cam1	S2V-cam2	S2V-cam3
PSCL [15]	38.60	33.20	53.26	36.39	30.87	50.96
VGG-Face [8]	79.10	77.53	79.03	59.31	65.21	74.29
CERML-EG [16]	85.71	82.51	87.23	88.80	85.69	90.99
CERML-EA [16]	86.40	83.13	86.76	**88.97**	85.84	90.26
CERML-ES [16]	86.21	82.66	86.64	88.93	85.37	89.64
Our **center LMNN**	85.33	86.18	88.27	81.66	84.77	90.10
Our GAsup LMNN	80.54	80.96	87.44	81.20	85.44	89.96
Our **center LDA**	**87.54**	**89.04**	**91.91**	86.85	**88.58**	**95.86**
Our GAsup LDA	82.34	81.19	91.59	84.74	88.36	94.29

Considering the above results, we compare the proposal, using the average pooling descriptor and the best prototype selection methods (*center* and *GAsup*), with different state-of-the-art algorithms. We follow the comparison protocol in [15] to also evaluate the performance of the two metric learning methods (LMNN and LDA). Table 2 shows the recognition rates obtained not only in V2S, but also in S2V scenario. It can be seen that in general our proposal achieves higher recognition rates that previous methods evaluated on COX database, including those specially designed for dealing with the dimensional mismatch problem such as Point-to-Set Correlation Learning (PSCL) [15] and the three variants of the Cross Euclidean-to-Riemannian Metric (CERML) [16]. Moreover, when comparing with the results obtained by the original VGG-Face network the obtained improvement is significant. This shows the influence of the proposed pipeline on the classification accuracy, i.e. the use of intermediate layers, the construction of the DS and the use of a metric learning. From the table we can also observe that in general, *center* method performs better than *GAsup* with both metric learning methods. Our result follows the statement in [17] because LDA has a simple closed-form solution that is useful to handle large-scale learning.

3.2 Experiments on SCFace Database

For evaluating the proposal in the SCFace database we follow the protocol used in [3], in which the HR images (mugshot) are used as gallery and the images of distance 1 (the most affected by LR) are used for test. We randomly selected 80 subjects for training and the remaining 50 subjects for testing. For this database, a single video representation was obtained taking the average of 5 images per subject (the database only has 5 images per subject in each distance). In this case only the average pooling descriptor with the *center* method of prototype selection was used, since it was the best performing combination in previous experiments. The results in terms of average recognition rates (10 random iterations) are presented in Table 3. They are compared with state-of-the-art methods reported on this database, also including the original VGG-Face network.

Table 3. Recognition rates in SCFace database.

Method	Recognition rates(%)
MDS [2]	61.14
Proposal in [3]	74.00
VGG-Face [8]	68.75
Our center LDA	92.23
Our center LMNN	**94.96**

In contrast with the results in COX dataset, we found that LMNN metric learning shows better results in comparison with LDA metric learning in SCFace database. It is important to mention that the dimension of the vectors with LDA is always smaller than the number of classes, therefore, for problems with a small number of classes it does not offer good results. We consider this is the reason why in the case of Cox database that contains images from 1000 subjects LDA performs better, while for the SCFace database that only has 130 subjects, the LMNN method exhibits better results than LDA. As shown, the proposed scheme achieves a significantly higher recognition rate (94.96%) than other state-of-the-art methods. When comparing our results with those obtained by the VGG-Face, it is corroborated that the strategy allows us to obtain more robust descriptors from the network and also, the importance of using metric learning to emphasize discriminative information from the descriptors.

4 Conclusions

One important contribution of our work is the proposal of obtaining a dissimilarity representation from deep convolutional features to address low-resolution face recognition. We found that dissimilarity representations constructed from convolutional features are more effective than representations directly obtained from convolutional network, as in the case of the VGG-Face. It was shown that since automatically learned representations are sensitive to outliers, it is convenient the use of prototype selection methods that take into account central objects. On the other hand, it was corroborated that the low-high strategy and the metric learning methods used in previous works are effective for this problem. As future work, we aim at improving our results by using neural networks in two main phases of our approach. First, to address the dimensional mismatch using a super-resolution neural network. Second, trying to find more discriminative descriptors from neural network to construct the dissimilarity space.

References

1. Grgic, M., Delac, K., Grgic, S.: SCface-surveillance cameras face database. Multimed. Tools Appl. **51**(3), 863–879 (2011)
2. Biswas, S., Aggarwal, G., Flynn, P.J., Bowyer, K.W.: Pose-robust recognition of low-resolution face images. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 3037–3049 (2013)
3. Zeng, D., Chen, H., Zhao, Q.: Towards resolution invariant face recognition in uncontrolled scenarios. In: 2016 International Conference on Biometrics (ICB), pp. 1–8. IEEE (2016)
4. Lu, T., Xiong, Z., Zhang, Y., Wang, B., Lu, T.: Robust face super-resolution via locality-constrained low-rank representation. IEEE Access **5**, 13103–13117 (2017)
5. Rajawat, A., Pandey, M.K., Rajput, S.S.: Low resolution face recognition techniques: a survey. In: 2017 3rd International Conference on Computational Intelligence and Communication Technology (CICT), pp. 1–4. IEEE (2017)
6. Hernández-Durán, M., Plasencia-Calaña, Y., Méndez-Vázquez, H.: Metric learning in the dissimilarity space to improve low-resolution face recognition. In: Beltrán-Castañón, C., Nyström, I., Famili, F. (eds.) CIARP 2016. LNCS, vol. 10125, pp. 217–224. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-52277-7_27
7. Pękalska, E., Duin, R.P., Paclík, P.: Prototype selection for dissimilarity-based classifiers. Pattern Recognit. **39**(2), 189–208 (2006)
8. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. BMVC **1**, 6 (2015)
9. López-Avila, L., Plasencia-Calaña, Y., Martínez-Díaz, Y., Méndez-Vázquez, H.: On the use of pre-trained neural networks for different face recognition tasks. In: Mendoza, M., Velastín, S. (eds.) CIARP 2017. LNCS, vol. 10657, pp. 356–364. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75193-1_43
10. Pękalska, E., Duin, R.: Dissimilarity Representation For Pattern Recognition: Foundations and Applications, vol. 64. World Scientific, Singapore (2005)
11. Plasencia-Calaña, Y., Orozco-Alzate, M., Méndez-Vázquez, H., García-Reyes, E., Duin, R.P.W.: Towards scalable prototype selection by genetic algorithms with fast criteria. In: Fränti, P., Brown, G., Loog, M., Escolano, F., Pelillo, M. (eds.) S+SSPR 2014. LNCS, vol. 8621, pp. 343–352. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-44415-3_35
12. Olivetti, E., Nguyen, T.B., Garyfallidis, E.: The approximation of the dissimilarity projection. In: 2012 International Workshop on Pattern Recognition in NeuroImaging (PRNI), pp. 85–88. IEEE (2012)
13. Hastie, T., Tibshirani, R.: Discriminant adaptive nearest neighbor classification and regression. In: Advances in Neural Information Processing Systems, pp. 409–415 (1996)

14. Weinberger, K.Q., Blitzer, J., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. In: Advances in neural information processing systems, pp. 1473–1480 (2006)
15. Huang, Z., et al.: A Benchmark and comparative study of video-based face recognition on COX face database. IEEE Trans. Image Process. **24**(12), 5967–5981 (2015)
16. Huang, Z., Wang, R., Van Gool, L., Chen, X., et al.: Cross euclidean-to-riemannian metric learning with application to face recognition from video. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
17. Liao, S., Lei, Z., Yi, D., Li, S.Z.: A Benchmark study of large-scale unconstrained face recognition. In: 2014 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–8. IEEE (2014)