GenLR-Net: Deep framework for very low resolution face and object recognition with generalization to unseen categories

Sivaram Prasad Mudunuri* [1], Soubhik Sanyal* [2] and Soma Biswas[1]
[1]Dept. of Electrical Engineering, Indian Institute of Science, Bangalore, India.
[2]Perceiving Systems Dept., Max-Planck Institute for Intelligent Systems, Tuebingen, Germany.
sivaramm@iisc.ac.in, soubhik.sanyal@tuebingen.mpg.de, somabiswas@iisc.ac.in

Abstract

Matching very low resolution images of faces and objects with high resolution images in the database has important applications in surveillance scenarios, street-to-shop matching for general objects, etc. Matching across huge resolution difference along with variations in illumination, view-point, etc. makes the problem quite challenging. The problem becomes even more difficult if the testing objects have not been seen during training. In this work, we propose a novel deep convolutional neural network architecture to address these problems. We systematically introduce different kinds of constraints at different stages of the architecture so that the approach can recognize low resolution images as well as generalize well to images of unseen categories. The reason behind each additional step along with its effect on the overall performance is thoroughly analyzed. Extensive experiments are conducted on two face and object datasets which justifies the effectiveness of the proposed approach for handling these real-life challenging scenarios.

1. Introduction

Recognizing faces and objects from very low resolution (LR) images is important in long distance surveillance applications [33], street-to-shop matching, etc. When images are taken from a distance, the region of interest in the image is usually very small and thus lack discriminatory information usually present in images which helps to distinguish one object from another. But, the images in the database are usually of high resolution (HR) and thus the images need to be matched across significant difference in resolution along with variations in view-point, illumination, etc.

Also due to the ever increasing number of object categories, it is not practical to assume that all the categories are available during training [34], making the task even more challenging. Though recognition of LR facial images has been reasonably studied [39][20], recognizing LR images of general objects is relatively unexplored. For general objects, classifying data from unseen categories is being extensively studied and is referred to as zero shot learning (ZSL) [34]. Here, the attributes of the seen as well as the unseen categories are provided and during testing, the unseen query image is classified by matching it with the available attributes. But many times, we may want to retrieve similar items instead of classifying them, eg., we take a picture of a dress that we like in the street and we want to retrieve similar clothing items from online shops [10][8][18]. In our work, the goal is to retrieve similar images when the query belongs to seen/unseen object, and so we do not require attribute information. Thus the problem addressed in this work is very different from ZSL.

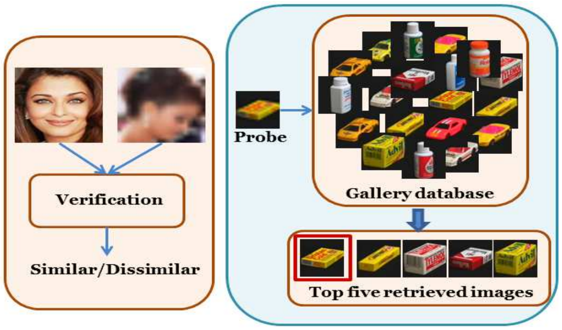


Figure 1. Applications of low resolution face verification [25] (left) and low-resolution object recognition [21] (right).

In our work, using the same framework, we address two problems, namely, (1) uncontrolled face verification between LR query and HR gallery images captured during enrolment (Figure 1 left) and (2) object recognition when the input is LR and the database consists of HR images (Figure 1 right). In the second application, given the query, the similar items are retrieved and the one which has the highest similarity with the query is taken as the correct match. For face verification, the training and testing subjects are completely disjoint. For general objects, during testing, we

*Equal Contribution. [2]This work was done while the author was a research student at IISc, Bangalore.

first evaluate the performance of the proposed framework when the input object classes are seen, and then we evaluate how well it generalizes to unseen categories. For training, we require image pairs belonging to same and different classes, but the class label is not required. Since the goal is not object/person classification, the object categories or the subject identities are not required during training.

In this work, we propose a novel deep CNN architecture which can handle large difference in resolution as well as generalize to unseen categories. We build upon an existing architecture by incorporating different losses at different stages with appropriate analysis and evaluation. We use the VGG face network [23] for the face verification experiment and the VGG-object network [27] for the object retrieval task, though any other relevant architectures can also be seamlessly used. First, we analyze how the performance of a standard architecture is effected when the HR images are replaced by LR images. In order to compensate for the drastic drop in performance due to very low-resolution, we introduce three different losses, namely, contrastive loss at the high-level features, inter-intra classification loss at the mid-level features and super-resolution loss at the low-level features. We analyze why and how each of the losses is beneficial for improving the network performance. Extensive experiments are performed on modified LFW [7] and CFP in wild [25] databases for the face verification task; and COIL-100 [21] and Toy Cars [22] databases for the object retrieval task. The experiments for unseen object categories shows the generalization ability of the proposed network. The contributions of the proposed work are as follows:

- Analyze the effect of very low-resolution for the applications of face verification and object recognition.
- A novel deep architecture which can handle large difference in resolution as well as generalize to unseen categories.
- Extensive experiments on different datasets show the effectiveness of the proposed network for matching very LR images as well as its generalization ability.

The rest of the paper is organized as follows. Section 2 gives pointers to the related work. The proposed approach and the experimental results are presented in Section 3 and Section 4 respectively. Finally, the paper ends with a conclusion.

2. Related Work

In this section, we provide pointers to some of the works related to low resolution face and object recognition.

Low-resolution Face Recognition: It is only recently that the researchers have started looking into the low resolution problem. Wang *et al.* [32] demonstrated the problem of recognition under very low resolution cases through a systematic deep learning based architecture. Zeng *et al.* [37] propose a resolution-invariant deep network to learn the resolution invariant features across the domains. Yang *et al.* [35] propose a discriminative multi-dimensional scaling approach by adding an extra inter-class constraint that enlarges the distances among different subjects in the learnt space. Farrugia *et al.* [4] propose a face hallucination technique using linear model of coupled sparse support framework that constructs linear models based on the local geometrical structure of the high resolution manifold. Zhu *et al.* [38] propose a deep bi-network architecture that solves face hallucination and dense correspondence field estimation problems together for low resolution facial images. Mudunuri *et al.* [19] propose an automatic low resolution face recognition approach based on MDS at fiducial points. These methods are designed to learn projection mappings from LR and HR images so that the same subjects show similar characteristics in the common space. Synthesis based approaches have also been developed to reconstruct the HR image from a given LR image. A synthesis based approach that learns class and resolution specific dictionaries is desribed in [26]. Kolouri and Rohde [13] propose a method that synthesizes corresponding HR face image from a given LR image by learning a nonlinear Lagrangian model on HR images. Zou *et al.* [39] propose a relationship learning based super-resolution method by enforcing the discriminative constraints between HR and LR image spaces. The authors demonstrated the performance for LR images of 7×6 to analyze very low resolution problem. A transformative and discriminative auto-encoder model is designed in [36] to hallucinate the LR faces that are not aligned and nosiy. A more generalized similarity measure is developed and unified into deep neural network architecture for discriminative feature representation learning in [17].

Low-resolution Object Recognition is relatively less explored. Peng *et al.* [24] propose a knowledge transfer framework for distinguishing fine-grained objects in LR images. Cai *et al.* [2] propose an end-to-end resolution-aware CNN architecture to classify LR objects by modeling the super-resolution and classification together. A detailed investigation on the effect of performance for different applications under optical blur is presented in [31] . Su and Maji [28] modeled a cross-quality model adaptation by addressing cross-domain variations including image degradation. Most of the object recognition algorithms address the problem of classifying the LR image. In contrast, in this work, we address the task of retrieving similar HR images for a given LR probe.

3. Proposed Method

Here, we provide details of the proposed architecture by taking the example of face verification. We gradually build our architecture with different losses at different parts of the network and analyze and evaluate each of them.

3.1. Problem Statement

In this work, we address two problems, face verification and object recognition, both under low resolution settings. We explain the proposed network architecture by taking the example of face verification in the subsequent sections. During training, we assume that we have access to both LR and HR images of the training subjects, i.e. the training data is given in the form $\{\mathbf{x}_i^h, \mathbf{x}_i^l, l_i\}, i = \{1, 2, 3, ...N\}$, where N is the total number of training images. Here l_i is the binary label, which is 1 if the images belong to the same subject (similar pair) and 0 otherwise (dissimilar pair). The subject id is not used, since the goal is not classification, rather verification, and the testing subjects are completely different than the training subjects. Though $\mathbf{x}_i^h$ and $\mathbf{x}_i^l$ belong to the same class, they differ in resolution, view-point, illumination, etc. (Figure 2 and 3). During testing, given an image pair, where one is LR and the other is HR, the goal is to verify whether they belong to the same subject or not.

3.2. Motivation

First, we analyze the effect of LR input images on the verification performance of a standard deep architecture developed for faces. For this work, we have taken the VGG face network [23] as the baseline architecture, though the proposed approach is general and can be applied to other base networks as well. For this experiment, we use the LFW database [7] and the standard experimental protocol [7]. We conduct the experiment on fold 1 of the database using the LFW-deepfunneled images. Since the VGG face network is trained on HR images, we first evaluate the performance by providing HR images for both views, and we obtain a verification rate of 93.83%. In all our experiments, we take the activations of the corresponding $fc7$ layer (Figure 2) as the feature for a given image. To analyze the effect of low resolution, we retain the images from view 1 at its original resolution (224 × 224) and downsample the view 2 images to 20×20. These images are then resized to 224 × 224 using bilinear interpolation (Figure 3) so that features from the same network can be computed and compared. With these view 2 images, the face verification performance drops drastically to 69.16%, signifying the importance of resolution for this application. Now, we describe the proposed deep architecture which is built on top of the VGG framework, and is termed as GenLR-Net, since it works for LR images and also generalizes to unseen categories as explained later.

3.3. Proposed GenLR-Net

The VGG face architecture [23] is shown by the shaded portion in Figure 2. This architecture is developed for classifying the facial images and has 16 trainable layers including convolutional and fully connected layers. The average prediction log-loss after the softmax layer is employed during the training to minimize the classification error. The network is trained on 2.6M images of 2622 identities.

In this work, our goal is to verify whether a pair of LR and HR image belongs to the same subject or not. Thus we want the HR and LR images belonging to the same subjects to come closer to one another and those from different subjects to move apart. Since we have two images of different resolutions, the proposed network (Figure 2) has two channels, where one channel takes the input as HR image and the other channel takes the input as the LR image. The HR channel is kept fixed (shaded part indicates that the weights are locked and not updated during training) since the network has already been trained for computing discriminative features from HR facial images. Our first modification is that we replace the final classification layer of the VGG network with a contrastive layer for the higher level feature, i.e. the final fully connected layer ($fc7$).

Contrastive Loss at the higher level features: First, contrastive loss [30][6] is applied between $fc7$ layers of both the channels of the network and the network is trained. This will help to bring the HR and LR positive samples closer and move the negative samples far apart. The contrastive loss between any two features are computed as

$$L_{Cont} = \frac{1}{2N} \sum_{i=1}^{N} l_i D_i^2 + (1 - l_i) max(\delta - D_i, 0)^2 \qquad (1)$$

Here, D_i is the L_2 distance between the corresponding features and δ is the margin by which the features of dissimilar subjects are to be separated. N is the total number of training samples and l_i is the corresponding binary label of each training pair. As already mentioned, the loss will only be propagated through the channel which takes the LR input data, since the channel weights corresponding to the HR image is fixed and not updated. This is different from the standard metric learning where the weights corresponding to both the channels are updated simultaneously. With this modification, we obtain a verification rate of 84.00%.

In some recent literature [24][16], it has been observed that enforcing the losses between intermediate layers can boost the performance. We investigated the impact of introducing the contrastive loss between the previous layers (like $fc6$, $pool5$ and $pool4$) and so on. We observe that the results improves to 86.00% when this loss is incorporated in $fc6$ layer. But when we enforce the same loss in the previous layers, the performance started to decrease.

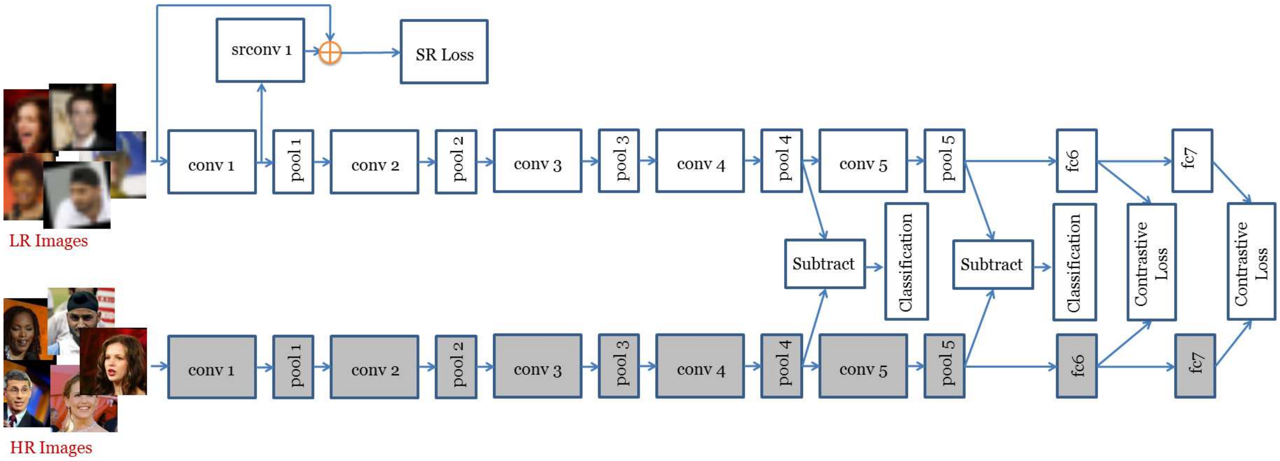


Figure 2. Simplified block diagram of the proposed GenLR-Net framework. The two channel CNN takes HR images in one channel and the LR images in the other channel. The shaded blocks indicate that the weights are initialized and fixed to those of the baseline network (VGG face network in this example). All the layers and the non linear activations etc. are not shown for ease of visualization.

This is probably because of the fact that this loss is not suitable for the mid-level or low-level features, which needs more flexibility to adapt to the different characteristics of the input data. This leads to a question, *what kind of supervision is suitable for the initial layers for improving the performance?*

Inter-intra Classification Loss at the mid-level features: Based on the previous analysis, we incorporate a constraint which is *softer* than the contrastive loss, and thus may be more suitable for the mid-level features. Inspired by [16] , we propose an inter-intra classification loss for the mid-level features. Here the difference between the two images (HR and LR) is computed and classified as 1 if they belong to the same class (subject) and 0, if they belong to different classes. Thus a N-class problem is converted to a 2-class problem. This loss also tries to bring samples from the same class closer and push samples from different classes apart. But it is less constrained than the contrastive loss since it does not enforce a strict margin between same and different classes and thus we feel it suits our purpose. We take the difference between the activations of $pool5$ of both the channels and and apply classification loss on these features, and the verification performance with this modification improves to 87.24%. By also including this loss on the difference vectors at $pool4$, we observe that the performance further improves to 89%. As with the contrastive loss, the performance starts decreasing if the loss is enforced in the initial low-level features like $pool3$, $pool2$, etc. So in the final network, we apply this loss only on the $pool5$ and $pool4$ layers.

Let f_1 and f_2 be the activations of the $pool5$ layers of the LR and HR channel respectively. In the face verification experiment, both f_1 and $f_2 \in R^{7\times7\times512}$ since there are 512 filters at the last convolution layer of $conv5$. The difference feature is denoted as (f_1 - f_2) and the resultant tensor is converted into a column vector (denoted by f). This vector is connected to a softmax layer with two nodes, such that f is classified as 1 if the input image pair belongs to the same class and 0 otherwise. Let θ_f be the weights connecting f to the softmax layer, the activations can be computed as

$$g = \phi\left(\theta_f^T f\right) \tag{2}$$

Here, ϕ is the non linear function *reLU*. The softmax probabilities are computed as follows

$$P_j = \frac{exp(g_j)}{\sum_k exp(g_k)} \tag{3}$$

The cross-entropy classification loss [29] is employed on the softmax probabilities for classifying the input pair

$$L_{Cls} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{1} -\mathbb{1}[l_i = k] log P_k \tag{4}$$

Here, $\mathbb{1}[l_i = k]$ = 1 if the input pair belongs to class k and zero otherwise.

Super-Resolution Loss at the low-level features: One way of matching a LR image with a HR image is to first apply super-resolution on the LR image to make it HR and then perform matching. Though super-resolution (SR) approaches are very useful for enhancing the image resolution, they are not designed to perform well for recognition applications [1][39]. We also observe that applying SR algorithms on LR images separately and using the enhanced images does not result in significant improvement in the recognition performance. So inspired by [2], we include the super-resolution objective along with the verification task. This should result in a boost in the performance [2] since the weights of the network are now responsible for simultaneously improving the resolution and the verification performance. We can use this loss only if the HR images corresponding to the LR images are available. This is different from the HR image pair available for training, since this along with the resolution difference also has difference in pose, illumination, expression, etc. and may confuse the network. To this end, we take the output of *conv1* layer and give it as an input to a *srconv1* layer. The *srconv1* layer predicts the residual images and so this output is added with the corresponding LR images to give the final super-resolution output [12]. Next, the reconstruction loss between the super-resolution output and the original HR image is computed. Let $\mathbf{s}_i^l$ be the output of the *srconv1* layer, then the reconstruction loss is given below as:

$$L_{SR} = \left\| (\mathbf{s}_i^l + \mathbf{x}_i^l) - (\mathbf{x}_i^l)_{hr} \right\|_2^2 \tag{5}$$

Here, $(\mathbf{x}_i^l)_{hr}$ is the HR image of the corresponding $\mathbf{x}_i^l$. Here, we assume that the HR version of the LR images are available for training which is usually the case with super-resolution tasks. With this, we observe that the verification rate improved to 90.00%.

In summary, we train the entire network by jointly minimizing all the losses. Specifically, there is a super-resolution loss after *conv1* layer, two classification losses each at *pool4* and *pool5* layers and two contrastive losses each at $fc6$ and $fc7$ layers. The final loss can be formulated as below:

$$\begin{aligned} L = \lambda_1 L_{SR} + \lambda_2 L_{Cls}^{pool4} + \lambda_3 L_{Cls}^{pool5} + \\ \lambda_4 L_{Con}^{fc6} + \lambda_5 L_{Con}^{fc7} \end{aligned} \tag{6}$$

The summary of results of the above losses applied step by step is given in Table 1.

Table 1. Illustrating the step-wise motivation of the proposed approach.

Method	Verification rate (%)
L_{Con}^{fc7}	84.00
$L_{Con}^{fc6} + L_{Con}^{fc7}$	86.00
$L_{Cls}^{pool5} + L_{Cont}^{fc6} + L_{Cont}^{fc7}$	87.24
$L_{Cls}^{pool4} + L_{Cls}^{pool5} + L_{Con}^{fc6} + L_{Con}^{fc7}$	89.00
$L_{SR} + L_{Cls}^{pool4} + L_{Cls}^{pool5} + L_{Con}^{fc6} + L_{Con}^{fc7}$	90.00

In our experiments, we set λ_1 = 10^{-3} and λ_m =1 for *m* =2,3,4,5. We optimize the proposed deep network using Caffe [9] deep learning framework. For the final network with five losses, we initialize the network with pretrained VGG face (in case of face verification experiments) weights and trained the architecture with a learning rate of 10^{-8} and weight decay of 0.0005. The momentum is fixed to 0.9. We drop the learning rate in steps by a factor of 10 after every 20000 iterations and the network is learnt for 80000 iterations. The super-resolution layer *srconv1* is the 3 channel layer (RGB) with kernel size 3. The weights of *srconv1* are randomly initialized from *xavier* distribution with a standard deviation of 0.01. For the experiments on object recognition, we use the same parameters with VGG object network but we drop the learning rate in steps by a factor of 10 after every 20000 iterations and the network is learnt for 60000 iterations.

4. Experimental Results

Here, we describe in detail the experiments performed to evaluate the performance of the proposed network. Specifically, we want to address the following questions:

- How effective is the proposed framework for matching/verifying across large variations in resolutions.
- How does the proposed approach compare with state-of-the-art super-resolution approaches?
- How does the proposed approach generalize to images from unseen categories?

In this work, we focus on two applications: (1) cross-resolution face verification, where one image is LR and the other is HR in addition to variation in pose, illumination, etc. and (2) cross-resolution object recognition, where the probe is LR and the images in the database are HR.

4.1. Cross-resolution face verification

For the application on cross-resolution face verification, we evaluate our approach on a modified version of LFW database [7] and CFP in wild database [25].

(A) Experiments on LFW database [7]: The LFW database [7] has labeled facial images captured under real unconstrained environments. The face images have wide range of pose, expression, race, clothing, hairstyles, lighting and other parameters and thus is one of the most widely used databases for face verification applications. So we have chosen a modification of this database along with the standard experimental protocol for our application. But our application is much more challenging as compared to the standard setting, since in our case, one of the images have very low-resolution.

We conduct the experiment on fold 1 of the database and using the LFW-deepfunneled images. We train our network on 2700 similar pairs and 2700 dissimilar pairs and test on 300 similar and 300 dissimilar pairs as per the standard protocol [7] but with modified resolutions as explained in Section 3.2. The LR images are obtained by downsampling the

Table 2. Performance (%) of the proposed approach on modified LFW database [7]. Comparisons with super-resolution, metric learning and domain adaptation techniques with deep features are also provided.

Method	Verification rate (%)
HR-HR (original VGG)	93.83
HR-LR (original VGG)	69.16
SSR [11] + $fc7$ features	72.10
SRCNN [3] + $fc7$ features	73.16
LapSRN [15] + $fc7$ features	76.16
$fc7$ features + SA [5]	72.50
$fc7$ features + LSML [14]	71.00
Proposed GenLR-Net	**90.00**

original images to 20×20 and then upsampled to the original resolution using bi-linear interpolation. Few sample images as per our protocol are shown in Figure 3.

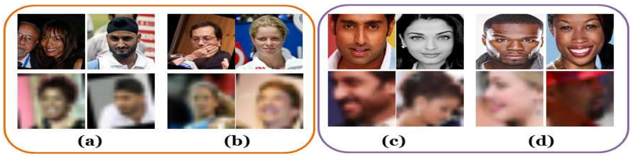
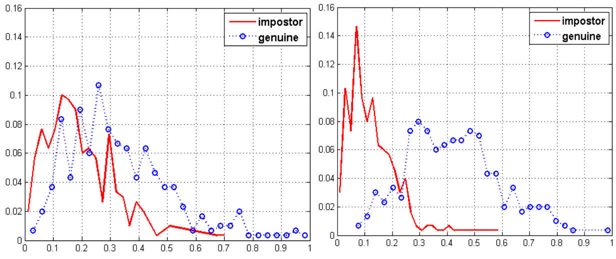


Figure 3. Sample facial images from modified LFW database [7] (left) and CFP in wild database [25] (right) used in our experiments. Each column of (a,c) and (b,d) shows example of similar and dissimilar pairs of the respective datasets. As per our protocol, the images of view 1 are HR images and view 2 are LR images.

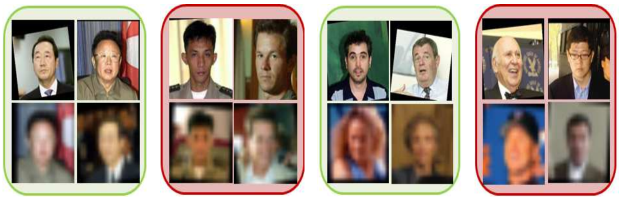
The results are reported in Table 2. As already mentioned in the motivation section, when both HR images are given to the original VGG face network, we obtain a verification rate of 93.83%. The performance drops to 69.16% when view 2 images are changed to LR as per our protocol signifying the importance of resolution for this application. The proposed network is able to improve the verification performance to 90.00%.

One of the standard techniques to improve the resolution of a LR image is to use super-resolution (SR) techniques. Recently, research in this area has advanced significantly, and several SR techniques have been proposed which give impressive outputs. Here, we evaluate the verification performance when SR techniques are applied to the LR images to enhance its resolution and then the features are extracted using the original VGG network. We have evaluated three state-of-the-art SR techniques [11][15][3] to understand the effect of super-resolution on the verification performance. Deep Laplacian Pyramid Super-Resolution Network (LapSRN) [15] is a deep architecture that systematically reconstructs the residual content from the HR images. SRCNN [3] is a lightweight deep architecture formulated based on the conventional sparse coding based SR techniques. SSR [11] is a Kernel ridge regression based sparse encoding approach that learns the mapping between LR and HR images. We observe from Table 2 that the SR approaches are successful in improving the performance, but the proposed GenLR-Net performs significantly better. The proposed approach also performs significantly better compared to standard metric learning [14] and domain adaptation [5] techniques applied on $fc7$ features.



(a) VGG face network (b) Proposed Approach

Figure 4. Similar (termed as genuine) and dissimilar (termed as imposter) score distribution of the HR - LR images using VGG net and the proposed GenLR-Net. The distributions are better separated using GenLR-Net. For each plot, X and Y axes refers to similarity score and fraction of test pairs respectively.



(a) Similar (✓) (b) Similar (×) (c) Dissimilar (✓) (d) Dissimilar (×)

Figure 5. Verification results of the proposed GenLR-Net on the modified LFW database [7]. Similar (dissimilar) indicates the columns which belong to same (different) subjects. The checkmarks (✓) and (×) denotes that the network classified the input pair correctly and incorrectly respectively. i.e. Similar (✓) indicates that, the input pair (column-wise) is similar and GenLR-Net classifies it correctly.

We further analyze the effectiveness of proposed network using the distribution of distances of similar and dissimilar pairs. We observe from Figure 4 that using the proposed GenLR-Net, the distributions are better separated which is essential for good performance. This results in improved face verification performance of the proposed network compared to the base network as evident in Table 2 under the challenging low resolution settings. Figure 5 shows few image pairs which are correctly and incorrectly classified by the proposed GenLR-Net.

Effect of resolution variations: Here we study the performance of the proposed approach under different resolutions of the view 2 images. For this study, we

Table 3. Verification performance (%) using the proposed GenLR-Net for different resolutions of the LR images.

Models	20×20	10×10	5×5
Original VGG-Face	69.16	56.17	54.50
Proposed GenLR-Net (Fine-tuned with 20×20 images)	90.00	67.70	62.30
Proposed GenLR-Net (Fine-tuned with 10×10 images)	-	72.17	65.00
Proposed GenLR-Net (Fine-tuned with 5×5 images)	-	-	65.80

experiment with LR images of size 10×10 and 5×5. The original images are downsampled to the required size and then upsampled to the same resolution using bilinear interpolation. For these even lower resolution images, the training is done in a stage-wise manner, inspired by [24]. i.e. for handling LR images of size 10×10, the network is first trained on 20×20 LR images, and then fine-tuned for 10×10 LR images, instead of directly training for the lower resolution. This approach helps the network to learn the variations systematically [24]. The results are reported in Table 3. The second row indicates that if we directly take the network that has been trained for resolution of 20×20 for images of even lower resolution, the performance starts decreasing. We obtain verification rates of 67.7% and 62.3% for images of size 10×10 and 5×5 respectively. But if we fine-tune the network for images of size 10×10, the performance improves to 72.17%. We also observe from the table, that the performance of the proposed GenLR-Net degrades gradually with decrease in resolution, and performs reasonably well even for very poor resolution.

(B) Experiments on CFP in wild database [25]: Here, we evaluate the proposed framework on frontal to profile face verification under low resolution settings. Few sample images of the database as per our protocol are shown in Figure 3 (right). This scenario is even more challenging compared to the earlier case since in this database, there exists significant variations in pose and expressions between the frontal (view 1) and profile (view 2) faces. This dataset presents a very challenging and realistic scenario involving cross-resolution and cross-pose matching. Similar to LFW, the database has 10 splits and each split has 350 matched pairs and 350 non-matched pairs. In this experiment also, the probe faces are downsampled to 20×20 and upsampled to the original resolution using bilinear interpolation. We evaluate the proposed approach on the first fold in which we train our network on 9 splits and test it on the remaining split. We follow the frontal-to-profile matching protocol of the database (fold 1) and the results are shown in Table 4. We observe that the results are significantly better compared to the baseline performance of 71.71% using the original VGG $fc7$ features.

Table 4. Verification performance (%) of the proposed approach on CFP in wild database [25].

Method	Verification rate (%)
HR-HR (original VGG)	88.57
HR-LR (original VGG)	71.71
Proposed GenLR-Net	**77.28**

4.2. Cross-resolution object recognition

We conduct experiments on COIL 100 [21] and Toy Cars Database [22] for the second application and also evaluate the generalization capability of the proposed approach. We initialized our network with VGG object network [27] weights for all the experiments related to object recognition.

(A) Experiments on COIL-100 Database [21]: Majority of the approaches which address the object classification task assume that the object classes for the training and testing are the same. i.e. the model is trained on a set of **N** categories and during testing, it is required to classify the query image which belongs to one of these **N** categories. But because new object categories are continuously being discovered, this is quite a restrictive assumption. In many realistic scenarios, the query image may come from a class which the model has not seen during the training stage.

For example, in online shopping, the user wants to search for a dress that is similar to the picture that he/she has captured. Firstly, the images displayed in online shops are usually professionally photographed, with cleaner backgrounds, good lighting whereas the consumer captured photos may be captured using a low resolution camera. Also, new clothing items are continuously added and so it is unlikely that the model will be trained on all the possible clothing items that is there in the database. Thus the final goal is to match the uncontrolled, unseen query with the relatively controlled data items and retrieve the similar ones.

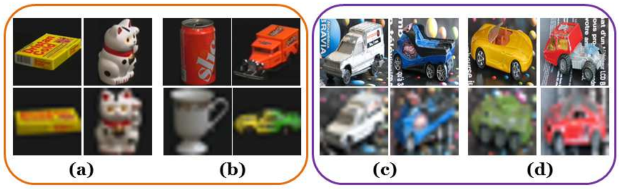


Figure 6. Sample object images from COIL-100 database [21] (left) and Toy Cars database [22] (right) used in our experiments. Each column of (a,c) and (b,d) shows example of similar and dissimilar pairs of the respective datasets used for training.

We evaluate the generalizability of the proposed GenLR-Net on the COIL-100 database [21] (Figure 6 left). The database has 100 categories and each category has 72 images with different pose, with a total of 7200 images. Each image is of size 128×128. The objects in the database have

Table 5. Rank-1 accuracy (%) on COIL-100 database [21] under different protocols.

Method	Seen in seen	Unseen in unseen	Seen in all	Unseen in all
HR-HR (original VGG object)	97.57	99.66	97.57	88.66
HR-LR (original VGG object)	68.99	90.67	67.17	69.67
Fine-tuned VGG-Object on LR data	78.28	92.33	77.17	75.67
LapSRN [15] + $fc7$ features	88.18	94.00	87.47	76.67
Proposed GenLR-Net	**93.13**	**98.00**	**91.21**	**81.00**

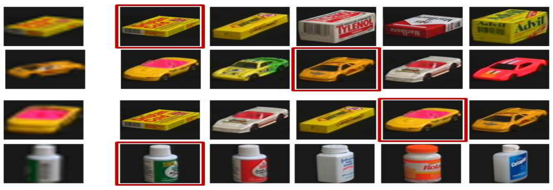


Figure 7. Cross-resolution object retrieval results of GenLR-Net on COIL-100 [21]. Each row shows top five retrieved results (column 2-6) corresponding to the LR query (first column). The first two rows are from **seen in all** protocol, and last two rows are from **unseen in all** protocol. Correct match is denoted by the red box.

wide variety of complex geometric and reflectance characteristics. We divide the dataset into two sets, where 90 categories are selected as seen categories and the remaining 10 are treated as unseen categories. We randomly select 60 images (out of 72) from each of those 90 objects to generate the matched and non-matched pairs for training and validation. In each pair, one image is of high resolution and the second one is of low resolution. The LR image is obtained by downsampling the original image to 20 × 20 and then upsampling it to 224 × 224 (resolution required for VGG object network input). The remaining 12 images from each of the 90 categories are used for testing. After training, the testing stage is divided into four different protocols.

Seen in Seen: One image from the 12 testing images of each of the 90 training categories is kept as a gallery image and the remaining 11 images are used as the probe images. In all the four protocols, the gallery is a HR image and the probes are LR images.

Unseen in Unseen: Here we use the remaining 10 categories which were not used for training. One randomly chosen image from each of the 10 categories is used as the HR gallery. We randomly choose 30 images from the remaining images for these categories and use them as the LR probe.

Seen in All: In this case, we keep the probe images same as the settings of the protocol ”seen in seen”. The gallery consists of both seen and unseen categories, and is formed using the gallery of ”unseen in unseen” along with that of ”seen in seen”.

Unseen in All: We keep the gallery same as in the protocol of ”seen in all” and the probe same as ”unseen in unseen”. The last two protocols are more realistic since we will usually have no apriori knowledge whether the query image belongs to a seen or unseen class. So the query will have to be compared with both the seen and unseen categories.

The performance of GenLR-Net for all the protocols is presented in Table 5. We observe that the results for the unseen categories are higher compared to seen categories. This is because the number of samples are much lesser for the unseen categories. However, we see that the proposed framework performs significantly better compared to the base network and also compared to SR approach in all the scenarios. Each row in Figure 7 shows top five retrieval results (column 2-6) for a given LR probe image (first column) using the proposed framework. The correct match is indicated by the red box.

(B) Experiments on Toy Cars database [22]: Here, we explore the problem of object verification in uncontrolled settings as we did for the face verification application. The database has images of 14 different toy trucks and cars. The training set contains 7 object instances and has 1185 similar and 7330 dissimilar image pairs. The remaining 7 object instances are used for testing. The images in the dataset have wide range of pose and lighting variations (Figure 6 right). The task is to verify whether an image pair belongs to the same object or not. The performance of the proposed approach is reported in Table 6. We observe that GenLR-Net is successful in improving the performance of the base network even in the challenging LR setting.

Table 6. Verification performance (%) of the proposed approach on Toy Cars database [22].

Method	Verification rate (%)
HR-HR (original VGG)	94.54
HR-LR (original VGG)	86.15
Proposed GenLR-Net	**88.09**

5. Conclusion

In this work, we propose a novel deep learning framework to address the challenging problem of matching low resolution probe images (faces/objects) against high resolution images in the database. We also addressed the very challenging and practical problem of unseen object recognition, which is a relatively unexplored area. We believe that this study will serve as a motivation as well as benchmark for researchers to address this challenging problem in addition to the more traditional seen objects classification task.

References

[1] S. Biswas, G. Aggarwal, P. J. Flynn, and K. W. Bowyer. Pose-robust recognition of low-resolution face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):3037–3049, 2013.

[2] D. Cai, K. Chen, Y. Qian, and J. K. Kamarainen. Convolutional low-resolution fine-grained classification. *arXiv preprint, arXiv:1703.05393*, pages 1–7, 2017.

[3] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016.

[4] R. A. Farrugia and C. Guillemot. Face hallucination using linear models of coupled sparse support. *IEEE Transactions on Image Processing*, 26(9):4562–4577, 2017.

[5] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. *IEEE International Conference on Computer Vision*, pages 2960–2967, 2013.

[6] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1735–1742, 2006.

[7] B. G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *University of Massachusetts, Amherst, Technical Report 07-49*, 2007.

[8] J. Huang, R. S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. *IEEE International Conference on Computer Vision*, pages 1062–1070, 2015.

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[10] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. *IEEE International Conference on Computer Vision*, pages 3343–3351, 2015.

[11] I. K. Kim and Y. Kwon. Single-image super-resolution using sparse regression and natural image prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1127–1133, 2010.

[12] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016.

[13] S. Kolouri and G. K. Rohde. Transport-based single frame super resolution of very low resolution face images. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4876–4884, 2015.

[14] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2288–2295, 2012.

[15] W. S. Lai, J. B. Huang, N. Ahuja, and M. H. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 624–632, 2017.

[16] C. Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. *Artificial Intelligence and Statistics*, pages 562–570, 2015.

[17] L. Lin, G. Wang, W. Zuo, X. Feng, and L. Zhang. Cross-domain visual matching via generalized similarity measure and feature learning. *IEEE transactions on pattern analysis and machine intelligence*, pages 1089–1102, 2017.

[18] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016.

[19] S. P. Mudunuri and S. Biswas. Low resolution face recognition across variations in pose and illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):1034–1040, 2016.

[20] S. P. Mudunuri and S. Biswas. Dictionary alignment for low-resolution and heterogeneous face recognition. *IEEE Winter Conference on Applications of Computer Vision*, pages 1115–1123, 2017.

[21] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (coil-100). *Technical Report CUCS-006-96*, 1996.

[22] E. Nowak and F. Jurie. Learning visual similarity measures for comparing never seen objects. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[23] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. *British Machine Vision Conference*, pages 1–6, 2015.

[24] X. Peng, J. Hoffman, X. Y. Stella, and K. Saenko. Fine-to-coarse knowledge transfer for low-res image classification. *IEEE International Conference on Image Processing*, pages 3683–3687, 2016.

[25] S. Sengupta, J. C. Cheng, C. D. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. *IEEE Winter Conference on Applications of Computer Vision*, pages 1–9, 2016.

[26] S. Shekhar, V. M. Patel, and R. Chellappa. Synthesis-based robust low resolution face recognition. *IEEE Transactions on Information Forensics and Security*, pages 1–10, 2017.

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, pages 1–15, 2015.

[28] J. Su and S. Maji. Adapting models to signal degradation using distillation. *British Machine Vision Conference*, pages 1–14, 2017.

[29] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. *IEEE International Conference on Computer Vision*, pages 4068–4076, 2015.

[30] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. *European Conference on Computer Vision*, pages 135–153, 2016.

[31] I. Vasiljevic, A. Chakrabarti, and G. Shakhnarovich. Examining the impact of blur on recognition by convolutional networks. *arXiv preprint arXiv:1611.05760*, pages 1–10, 2016.

[32] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang. Studying very low resolution recognition using deep networks. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4792–4800, 2016.

[33] Z. Wang, Z. Miao, Q. J. Wu, Y. Wan, and Z. Tang. Low-resolution face recognition: a review. *Springer: The Visual Computer*, 30(4):359–386, 2014.

[34] Y. Xian, B. Schiele, and Z. Akata. Zero-shot learning-the good, the bad and the ugly. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4582–4591, 2017.

[35] F. Yang, W. Yang, R. Gao, and Q. Liao. Discriminative multidimensional scaling for low-resolution face recognition. *IEEE Signal Processing Letters*, 25(3):388–392, 2017.

[36] X. Yu and F. Porikli. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3760–3768, 2017.

[37] D. Zeng, H. Chen, and Q. Zhao. Towards resolution invariant face recognition in uncontrolled scenarios. *International Conference on Biometrics*, pages 1–8, 2016.

[38] S. Zhu, S. Liu, C. C. Loy, and X. Tang. Deep cascaded bi-network for face hallucination. *European Conference on Computer Vision*, pages 614–630, 2016.

[39] W. W. Zou and P. C. Yuen. Very low resolution face recognition problem. *IEEE Transactions on Image Processing*, 21(1):327–340, 2012.