Deep Attributes for One-Shot Face Recognition

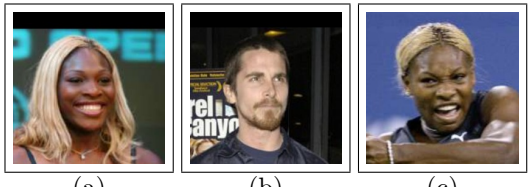
Aishwarya Jadhav[1,3], Vinay P. Namboodiri[2], and K. S. Venkatesh [3]
[1]Xerox Research Center India, [2]Department of Computer Science,
[3]Department of Electrical Engineering, IIT Kanpur
aishwaryauj@gmail.com, vinaypn@iitk.ac.in, venkats@iitk.ac.in

Abstract. We address the problem of one-shot unconstrained face recognition. This is addressed by using a deep attribute representation of faces. While face recognition has considered the use of attribute based representations, for one-shot face recognition, the methods proposed so far have been using different features that represent the limited example available. We postulate that by using an intermediate attribute representation, it is possible to outperform purely face based feature representation for one-shot recognition. We use two one-shot face recognition techniques based on exemplar SVM and one-shot similarity kernel to compare face based deep feature representations against deep attribute based representation. The evaluation on standard dataset of 'Labeled faces in the wild' suggests that deep attribute based representations can outperform deep feature based face representations for this problem of one-shot face recognition.

Keywords: Face Recognition, Attributes, One-shot classification

1 Introduction

Consider that we have seen a face. How would we go about recognizing the face that we have seen only once? The problem of recognizing examples from a single training example is termed one shot recognition. The task becomes especially challenging for unconstrained pose with variable illumination setting ("in the wild"). While, the problem of face recognition has been widely studied in computer vision [2], that of one shot face recognition has not yet been as well studied. In this paper, we focus on this scenario.



(a) (b) (c)

Fig. 1. Given a query image (a), the task is to determine if it resembles (b) or (c)

Our approach towards solving this problem is based on a deep attribute based description of a face. While, this approach was prevalent some time back [4], lately this approach has been overshadowed. Current computer vision related research has shown great progress in describing faces using deep neural network based approaches [11, 16]. These have shown remarkable performance for unconstrained face recognition in real world settings. However, these make use of a large number of training data for training for face recognition. One commonly used option is that of pre-training these deep neural networks using large amount of training data, and then using them as a means to obtain high-level features, which are then matched for face recognition.

The task of one shot recognition differs from the general datasets in that we have at least one sample of the test class. We can make use of the limited information in training in order to obtain a better representation of the class. To solve the problem, we evaluate two classes of methods, one that is based on the deep learned face feature and the other that is based on attribute based features. Our evaluation suggests that for one shot recognition, attribute based one-shot methods outperforms the deep learned face features. We further analyse this performance in different settings.

Table 1 shows importance of attribute based representation in face recognition where the list of attribute scores predicted by CNN suggest that person (a) must resemble person (c) which is evident in fig. 1.

Through this paper we make the following contributions:

(1) We show that attribute based deep feature representation outperforms deep learned face features in one-shot face recognition.

(2) We observe that a one-shot recognition system that uses the attribute based deep representation from the pre-final layer output of a convolutional neural network is more suited for various one-shot face recognition settings.

2 Related Work

There have been a number of techniques that address one-shot recognition. One such set of methods make use of Bayesian formulation to categorise objects [7, 13].

Table 1. L:Labels of groundtruth '0': absence '1': presence of attribute. P: probability score of attribute predicted by our CNN. Score clearly shows that face (a) is more closer to (c) than (b). Thus concatenating these attribute scores gives a good representation of face. Our attribute representation is of higher dimension and these outputs are shown for illustration of the concept

Attribute	L(a)	P(a)	L(b)	P(b)	L(c)	P(c)
Male	0	0.0333	1	1.0000	0	0.1097
Blond Hair	1	0.9990	0	0.0000	1	0.8589
Mustache	0	0.0083	1	0.9998	0	0.0035
Black Hair	0	0.0000	0	0.4984	0	0.0000
Oval Face	0	0.0722	0	0.0014	0	0.0439
Cheekbones	1	1.0000	0	0.0000	1	0.9913
Pointy Nose	0	0.0005	1	0.5039	0	0.0007
Chubby	0	0.0019	0	0.0146	0	0.0727

Another stream in one-shot learning focuses on building generative models to build extra examples [5]. These methods rely on elaborate feature vector representations. There have also been a number of interesting discriminative methods like [18] and [9] that explicitly make use of the one-shot recognition setting. In this paper, we evaluate both these methods for one-shot recognition.

All the models described above do not explore ways to generalise concepts learnt from one or few examples. Generalization of semantic concepts based on attributes has been widely studied in the problem of zero shot learning [14], [10], [6] . However, these ideas have not been explored in the context of one-shot learning. There has been several works that address the task

of obtaining attribute based representations for faces [4, 19, 8]. We compare our attribute prediction results with the results in [8].

3 Method

3.1 Deep Attribute Representation

We first obtain attribute vectors of face images using convolutional neural networks (CNN). The architecture of our CNN (fig. 2) is similar to VGG-Face CNN [11]. The filters in CNN are initialised with pre-trained parameters from VGG-Face CNN. For each attribute a separate CNN is trained in binary classification setting.

3.2 Exemplar-SVM

In Exemplar-SVM [9] a separate linear SVM is trained for each face in positive training set using negative faces. A set of negative faces does not contain any face from positive training identities. The final identity of the query face is then predicted by comparing calibrated scores of all SVMs. If x is input and f(x) is decision value given by SVM, the calibrated score is given by

$$p(x) = \frac{1}{e^{Af(x)+B}} \tag{1}$$

where A and B are estimated independently for each Exemplar-SVM. Calculating calibrated scores generalises output of all SVM and makes them comparable. Higher score indicates the query face is closer to the positive face on which corresponding SVM is trained.

3.3 One-Shot Similarity

In this approach we use one-shot similarity (OSS) kernel to train SVM. In this, similarity between two faces is calculated by first learning a model for each face with a set of negatives and then these models are used to predict similarity between the two faces [17]. Wolf et al. [18] show that for free-scale Linear Discriminant Analysis (LDA), one shot similarity score and its exponent can be used as kernels in one versus all SVM.

Let A be set of negatives of size n_A containing feature vectors a_i. m_A and S are mean and covariance of vectors in A. S is given by

$$S = \frac{1}{n_A}\sum_{i=1}^{n_A}(a_i - m_A)(a_i - m_A)^T \tag{2}$$

In case of binary classification, consider two positive faces are represented by feature vectors x and y. Their one-shot similarity with free-scale LDA (Linear Discriminant Analysis) is given by

$$OSS(x, y) = (x - m_A)^T S^+ (y - \frac{x + m_A}{2})(y - m_A)^T S^+ (x - \frac{y + m_A}{2}) \tag{3}$$

where S^+ is pseudo-inverse of S. Using above formula, similarity score between two training faces is calculated which is then used to train SVM classifier.

4 Experiment

4.1 Dataset

We use Large-scale CelebFaces Attributes (CelebA) dataset [15] (202599 face images and 40 binary attributes) to train CNN for attribute classification. The test dataset is LFW dataset [3].

4.2 Deep Attribute Representation

For all the attributes, we fine tuned CNN (fig. 2) using randomly chosen 10000 images from CelebA dataset. Randomly cropped and horizontally flipped with probability 0.5 patches of 224x224 size of rescaled images are fed to the network. We do not apply any alignment to input images. Learning rate is varied from 10e-4 to 10e-6.

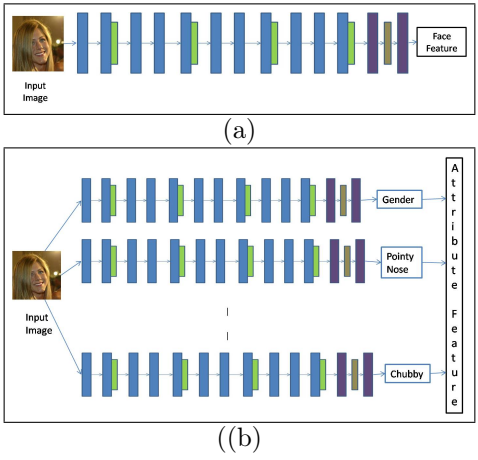


Fig. 2. Architecture of VGG-CNN shown in figure (a). Each colour block represents a specific layer given by Blue: Convolutional Layer +ReLu , Green: Max pooling , purple:Fully Connected and Brown: ReLu. Figure (b) shows collection of CNN from figure (a)

The performance of fine tuned CNNs is evaluated on LFW dataset for attribute prediction. The results are given in the table 2. It can be seen that even though our CNNs are fine tuned with limited number of training images, they predict attributes with good accuracy compared to LNets+ANet.

While choosing 8 attributes to represent face for one-shot recognition, we have considered several points [4]:

(1) Attributes which are related to accessories or facial expression are not considered for selection. (2) From equation 3, complexity to calculate one-shot similarity kernel per pair is O(d^2)where d=4096*n and n is number of attributes. (3) More accurate classifier is more likely to extract true attribute features. (4) An attribute which is specific to a region of face is more likely to help in recognition. We use five local attributes Pointy Nose, High Cheekbone, Black Hair, Blond Hair, Mustache and remaining three are global.

The output of the pre-final fully connected layer is used as descriptor of that attribute while binary output of CNN is used for attribute prediction. A single vector of size 8*4096 is then formed by concatenating descriptors of 8 attributes for each face image and used as its deep attribute based representation.

4.3 One-Shot Face Recognition

To evaluate performance of one-shot face recognition, we follow experiments given in [18]. We use positive set as subset of 6733 images of 610 identities from LFW dataset such that each has at least 4 images. Negative examples are images of identities having only one image each from LFW. Negative set is formed by randomly choosing 1000 negative examples.

To compare performance of one-shot face recognition with [18], we vary number of identities by 5, 10, 20 and 50. For each identity we randomly select two probe images and two gallery images from positive set. Then we compare performance of two one-shot methods by training Exemplar-SVMs and OSS-SVM

with deep attribute descriptors of gallery images and negative images. We use Libsvm [1] to train all SVMs. Calibration score is calculated in Libsvm by using improved Platt method [12].

For each number of identities, 20 repetitions are performed by randomly choosing different identities in each iteration. The result of test accuracy in terms of mean and standard deviation is shown in table 3. As per our knowledge, there are no experimental results available for one-shot face recognition on LFW dataset more recent than [18]. So accuracies shown in table 3 can be used to compare any future work on this task. Also it can be seen that as number of classes increase to 50, accuracy of recognition is decreasing as there is more chance of misclassification. Also as expected, deep attribute based features perform far better than bag of features taken from [18].

Table 2. Percentage accuracy of attribute classification on LFW. LNets+Anet uses aligned faces while our method does not apply any alignment

Attribute	our method	LNets+ANet [8]
Black hair	86	90
Blond hair	94	97
Cheekbones	81	88
Chubby	90	73
Mustache	93	92
Gender	96	94
Oval Face	69	74
Pointy Nose	73	80

Table 3. Results of one-shot face recognition for different classes using Bag of features (BoF) representation (first row) and deep attribute representation (last two rows) with Exemplar-SVM (E-SVM) and OSS-SVM

Class	5	10	20	50
OSS (BoF) [18]	0.7550 ±0.1432	0.7300 ±0.0768	0.7000 ±0.0782	0.5855 ±0.0365
E-SVM	0.9600 ±0.0730	0.9375 ±0.0521	0.8887 ±0.0539	0.8407 ±0.0366
OSS	0.9600 ±0.0940	0.9450 ±0.0473	0.8887 ±0.0573	0.8613 ±0.0413

4.4 Face Representation Vs Deep Attribute Representation

In this experiment, we select 10 identities randomly from positive set and repeat exactly same steps as above experiment in section 4.3 using attribute based representation of faces. These experiments are further repeated when images are represented as VGG-Face descriptors. The results are shown in table 4. It can be seen that, deep attribute features give more accurate recognition results than using just deep learned face features for each of the three methods. Also exponential OSS-SVM trained with deep attribute features gives most accurate performance.

Table 4. Comparisons of accuracy and standard deviation for 10 identities represented by VGG-face and attribute descriptors with Exemplar-SVM (E-SVM), SVM with free-scale LDA OSS kernel and exponential of OSS kernel. Attribute features perform better than face features

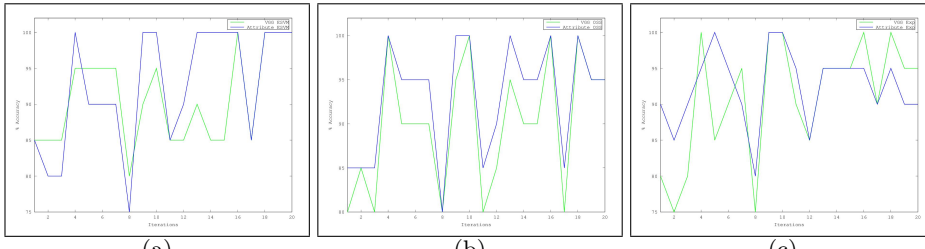
Input	ESVM	OSS	Exponential OSS
VGG-Face	0.9075 ±0.0638	0.9000 ±0.0725	0.9100 ±0.0815
Attribute	0.9250 ±0.0829	0.9250 ±0.0512	0.9300 ±0.068

Fig. 3 shows comparison of accuracies during testing with VGG-Face and deep attribute descriptors when the experiment is repeated 20 times. For most of the experiments, attribute features perform better than face features.

In OSS based SVM, positive faces are first compared with each other using similarity scores. These scores are then used determine decision boundaries in the similarity space. In Exemplar-SVM, one compares a positive sample with a fixed set of negatives and the scores of other positive samples are not considered. As a result of these differences, it can be seen that Exponential OSS-SVM performs better than Exemplar-SVM.

In these experiments the deep attribute based feature vectors are observed to perform better. These encode both the characteristics of faces as well as specific attribute characteristics. As explained earlier, in one-shot recognition knowledge from negative examples is used to generalise concepts learnt from one or few positive examples. Attributes provide better generalisation than face features over negative and positive identities. Attribute feature space has higher dimension than face feature space. Also, since each attribute is represented by 4096 vector, we believe that it contains much higher level description of that attribute for a person.

Due to all these advantages of attribute features over face features, attribute space enables one-shot methods to characterise entire positive identity from one example using knowledge acquired from other identities. Hence the attributes aid in face recognition and therefore as expected, we observe the performance of attribute based feature vectors to be better.



(a) (b) (c)

Fig. 3. Accuracy for 20 repetitions of experiment using VGG-Face and attribute descriptors with classifiers (a) Exemplar-SVM (b) OSS kernel (c) Exponential of OSS

5 Conclusion

In this paper we have proposed the use of deep attribute based representation for one-shot face recognition. The deep attribute representations are obtained by fine-tuning a deep CNN for face recognition on data for specific attributes such as gender and shape of face. While, specific face information is challenging, it is far more easier to obtain attribute related information. We observed that the face features when further adapted by various attributes yield consistent improvement in accuracy for one-shot recognition. This was observed for two different methods, one-shot recognition using Exemplar-SVM based and one-shot similarity kernel based techniques. In future we would be interested in exploring the kind of attributes that are useful for improving face recognition.

References

1. Chang, C., Lin, C.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3), 27 (2011)
2. Ding, C., Tao, D.: A comprehensive survey on pose-invariant face recognition. ACM Trans. Intell. Syst. Technol. 7(3), 37:1–37:42 (Feb 2016)
3. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., Technical Report 07-49, University of Massachusetts, Amherst (2007)
4. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Describable visual attributes for face verification and image search. Pattern Analysis and Machine Intelligence, IEEE Transactions on 33(10), 1962–1977 (2011)
5. Lake, B.M., Salakhutdinov, R., Gross, J., Tenenbaum, J.B.: One shot learning of simple visual concepts. In: Proceedings of the 33rd Annual Conference of the Cognitive Science Society. vol. 172, p. 2 (2011)
6. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. Pattern Analysis and Machine Intelligence, IEEE Transactions on 36(3), 453–465 (2014)
7. Li, F.F., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(4), 594–611 (2006)
8. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3730–3738 (2015)
9. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-svms for object detection and beyond. In: Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 89–96. IEEE (2011)
10. Palatucci, M., Pomerleau, D., Hinton, G.E., Mitchell, T.M.: Zero-shot learning with semantic output codes. In: Advances in neural information processing systems. pp. 1410–1418 (2009)
11. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. Proceedings of the British Machine Vision 1(3), 6 (2015)
12. Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers 10(3), 61–74 (1999)
13. Salakhutdinov, R., Tenenbaum, J., Torralba, A.: One-shot learning with a hierarchical nonparametric bayesian model (2010)
14. Socher, R., Ganjoo, M., Manning, C.D., Ng, A.: Zero-shot learning through cross-modal transfer. In: Advances in neural information processing systems. pp. 935–943 (2013)
15. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in Neural Information Processing Systems. pp. 1988–1996 (2014)
16. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1701–1708 (2014)
17. Wolf, L., Hassner, T., Taigman, Y.: Descriptor based methods in the wild. In: Workshop on faces in'real-life'images: Detection, alignment, and recognition (2008)
18. Wolf, L., Hassner, T., Taigman, Y.: The one-shot similarity kernel. In: Computer Vision, 2009 IEEE 12th International Conference on. pp. 897–902. IEEE (2009)

19. Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.: Panda: Pose aligned networks for deep attribute modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1637–1644 (2014)