Multiple Instance Learning with Deep Instance Selection for Video-based Face Recognition

Ning Liu[1, a]

[1]IRIP Lab, School of Computer Science and Engineering, Beihang University, Beijing, China

[a]lnfw01@163.com

Keywords: Multiple instance learning, deep learning, video-based face recognition.

Abstract. In many real-world video-based face recognition scenarios, videos are usually captured under unconstrained conditions. It is very challenging because of low face resolutions, varying head pose and complex lighting. To address this issue, we present a new method by formulating the video-based face recognition issue as a multi-instance learning (MIL) problem. Specifically, given a pair of videos, we generate a bag composed of all the frame pairs from the two videos. The bag is positive if the given pair of videos is from the same person, otherwise it is negative. In this way, the recognition task is formulated as a binary classification problem in MIL. Then we propose a novel MIL algorithm with deep instance selection (MILDIS), which maps each bag into a feature space defined by the selected instances via an instance similarity measure. Our work achieves the state-of-the-art performances on the real-world datasets YouTube Faces (YTF) according to the restricted protocol.

Introduction

Video-based face recognition (whether two face videos belong to the same identity) has drawn increasing attention in recent years. This issue becomes more difficult when faces are captured in the wild. The key challenge is the dramatic intra-person variations caused by poses, illuminations, expressions, and occlusions. Many approaches have been proposed. Most of them choose to strip off temporal dynamics that is inherent in video sequences. Basically because it is hard to extract useful person specific facial dynamics while the video is captured under unconstrained conditions. Without temporal information, video-based face recognition can be converted to the matching of two image sets.

The image-set based classification approaches generally fall into two categories [1]: parametric model methods and nonparametric sample methods. The former [2, 3] exploit some parametric distribution to represent each image set and then measure the between-distribution similarity. One limitation of the parametric methods is that they have to assume some distribution and handle the parameter estimation problem. If the data set does not follow the predefined statistic distribution, the estimated model will not consist with the data set. Some non-parametric methods attempt to represent an image set as a linear subspace [4, 5] or a nonlinear manifold [6, 7]. Then solve the problem by developing some algorithms to measure the similarity or distance between two subspaces or manifolds. Such approaches do not impose any assumption on data distribution, and have shown many merits compared to parametric models. The main disadvantage is that classification performance is dependent on the effectiveness of clustering training data into meaningful subspaces or manifolds.

To improve the robustness, we present a new framework to deal with the video-based face recognition using deep learning approach. Our main idea is to formulate the video-based face recognition as a multi-instance learning problem. Specifically, given a pair of videos, we generate a bag composed of all the possible frame pairs of the two videos. The bag is positive if the given pair of videos is from the same person, otherwise it is negative. For every pair of frames, we call it an instance (in the multi-instance learning terminology). In this way, the recognition task is formulated as a binary classification problem in multi-instance learning. Compared with the previous methods, our method has two merits. One is that we impose no assumption on the data distribution or the feature space. The other is that we use all the frames from the videos to maintain the information as much as possible. Then, for representation, we extract features to represent every face pair by utilizing a deep convolutional neural network which is trained to verification face image pairs. Finally, for classification, since we use all the frames, the huge amount of data makes many multi-instance learning algorithms getting bad performance. So we propose a novel multi-instance learning algorithm highly connected to the extracted features, which is not only useful but also efficient. Our work achieves the state-of-the-art performances on the real-world datasets YouTube Faces (YTF) according to the restricted protocol.

Related work

Deep learning Recently the performance of face recognition is extremely improved with the success of deep convolutional neural networks (e.g. Deep Face[8] and DeepID[9]), becoming comparableto human-level performance. The deep feature shows great advantage in handling the intra-personal variations than hand-crafted feature.

Convolutional Neural Network (CNN) [10, 11] is a type of feed-forward artificial neural network which is inspired from biology. The individual neurons are designed to simulate cells within visual cortex, which are sensitive to small sub-regions of input space, named receptive fields. Thus the connections among neurons are tied in such a way that each output neuron only responds to a local region of input neurons. This mechanism is better suited to exploit the strong spatially local correlations presented in natural images.

In our method we use a practiced 9-layers deep model [12, 13] and train our CNN on a binary classification task, namely to predict whether two faces in comparison belong to the same person. The CNN takes a pair of gray face regions of size 2×63×55 pixels as input. Its four convolutional layers (followed by max-pooling layers except the last one) extract the relational features hierarchically. The top two layers (F8 and F9) are fully connected: each output unit is connected to all inputs. The output of the last fully-connected layer is fed to a 2-way soft-max which indicate the probability distribution over the two classes; that is, whether they are the same person. The goal of training is to maximize the probability of the correct class. We achieve this by minimizing the cross-entropy loss for each training sample. If k is the index of the true label for a given input, the loss is :$L = -\log p_k$. We use mini-batch stochastic gradient descent with momentum which is shown to be an effective method for training CNN. The gradients are computed by standard back propagation of the error. We use ReLU as an activation function which is a popular choice especially for deep networks.

MILIS The MIL with instance selection (MILIS) [14] is efficient in training and well suited for large-scale MIL problems. In the training phase, it first performs instance selection on the training instances. This is achieved by modeling the distributions of negative instances and picking the least negative one from each positive bag. After doing this, we obtain a set of instance prototypes (IPs). Each of these is chosen from the corresponding positive bag. Then the MIL problem is converted into a single instance problem via a similarity-based feature mapping using the selected IPs. We will give a detailed description of bag-level feature representation later. Given the bag-level feature vectors, a standard linear SVM classifier is trained on the bag features. Based on the classification results on the training data, it updates and reselects the IPs. This step sequence is interleaved until convergence. In the testing phase, it commence by extracting the feature vector for the test bag using the feature mapping defined over the IPs obtained in the training phase. The trained SVM classifier is then applied so as to obtain the classification result.

To put it simply, its main idea is to select a single instance representation for each bag, and use the chosen subset of instances for bag-level feature computation. So the key problem is instance selection. In MILIS, the principle of instance selection is to picking the most positive one from each positive bag and the most negative one from each negative bag. Since the positive bags maybe contain negative instance, it is hard to find the most positive one directly. So the main idea of MILIS is to build a favorable adaptive instance selection algorithm.

Different from the MILIS, we use a deep CNN to extract deep feature, and the soft-max layer (F9) can indicate the probability distribution over the two classes. For CNN, the training stage is also getting optimal solution on the training set. When the training data is sufficient, the result of CNN is very dependable. Compared with MILIS, CNN can better handle the large-scale data. So we do not have to calculate other distribution.

Our method: MILDIS

In this paper, we present a new method to video-based face verification by formulating the video-based face recognition as a multi-instance learning problem. Given a pair of videos, we generate a bag composed of all the possible frame pairs of the two videos. Formally, two video sequences, are denoted respectively by $S_1 = \{a_i | i = 1,2, \cdots m\}$ and $S_2 = \{b_j | j = 1,2, \cdots n\}$ where a_i and b_j represent the face images cropped from the video frames, m and n are the image numbers in S_1 and S_2 respectively. Then the set $P = \{\langle a_i, b_j\rangle, \langle b_j, a_i\rangle | a_i \in S_1 \wedge b_j \in S_2\}$ contains all the possible frame pairs of the two video sequences. Use the DNN to extract the deep feature of every element (a face pair) in P. Then combine these feature vectors into a new set B which is exactly the bag we need in Multi-instance learning. The bag is positive if the given videos are from the same person, otherwise it is negative. Each instance in B is a feature vector which represents the similarity of a face pair. In this way, the recognition task is formulated as a binary classification problem in multi-instance learning. Then we select a single instance as the IP of each bag, and employ a similarity-based feature mapping to extract a single bag-level feature per bag. Finally, we convert the MIL problem into a single instance problem which can be solved by a SVM classifier.

Instance selection by DNN To construct bag-level feature vectors, a single instance is selected from each training bag to form the subset of IPs for feature mapping. Intuitively, we want to include true positive and negative instances in the subset to compute a discriminative feature map. So the principle of instance selection is to picking the most positive one from each positive bag and the most negative one from each negative bag. So we input the face pairs into the trained 9-layer DNN. Then we use the output of fully connected layer (F8) as the feature of each face pair and the output of soft-max layer (F9) as the probability distribution over the two classes.

Denote $B^{tr} = \{B_1, \cdots, B_n\}$ as the training set of bags and $Y = \{y_1, \cdots, y_n | y_i \in (-1,1)\}$ as the labels associated with each bag. For each bag $B_i = \{x_{i,j} | j = 1, \cdots, m_i\}$, given by the output of DNN, denote $P_i = \{p_{i,j} | j = 1, \cdots, m_i\}$ where $p_{i,j}$ the probability of positive of $x_{i,j}$. To a positive bag B_i^+, we chose the instance $x_{i,l}$ as the IP where $p_{i,l}$ is maximal. To a negative bag B_i^-, we chose the instance $x_{i,l}$ as the IP where $p_{i,l}$ is minimum.

Bag-level feature representation To effectively employ the similarity-based feature mapping, a distance metric needs to be defined first between bags and instances. The Hausdorff distance is a natural distance metric for this purpose. Specifically, the distance between bag B_i and instance x is given by the distance between x and its nearest neighbor in B_i.

$$d(Bi, x) = \min_{x_{i,j} \subset B_i} \left\| x_{i,j} - x \right\|^2. \tag{1}$$

Given the distance metric above, we can then derive the following similarity measure using an exponential function

$$s(B_i, x) = \exp(-\lambda d(B_i, x)) = \max_{x_{i,j} \subset B_i} \exp(-\gamma \left\| x_{i,j} - x \right\|^2). \tag{2}$$

After instance selection, we have obtained a subset of instance prototypes (IPs), $x_1, \cdots, x_n$, where x_i is the prototype selected from the ith bag in the training set. The bag-level feature is then given by

$$z_i = [s(B_i, x_1), s(B_i, x_2), \cdots, s(B_i, x_n)]. \tag{3}$$

A single feature vector is formed per bag.

Classification We then train a classifier that can be applied to the bag features. To this end, we use linear SVM, which employs the L2 norm for both the regularization term on feature weights w and the data term as follows:

$$f(w) = \frac{1}{2}\|w\|^2 + C\sum_i(1 - y_i w^T z_i)^2, \qquad (4)$$

where y_i is the label value for bag i, C is the regularization parameter that controls the influence of the second term on the right-hand side of the above equation. The resulting classifier is given by $w^T z$, where w are the linear weights for the features. Here, we have also absorbed the bias term into the weight vector for the sake of simplicity.

Experiments and results

We test our method on the recent video-level face verification dataset. YTF database contains more than 3425 videos of 1595 subjects obtained from YouTube, with significant variations on expression, illumination, pose, resolution and background. An important number of existing methods have been tested on this database [15,16,17]. Specifically, they randomly collect 5,000 video pairs from the database, half of which are pairs of videos of the same person, and half of different people. These pairs were divided into 10 splits. Each split containing 250 ’same’ and 250 ’not-same’ pairs.

The DNN architecture is shown in table 1. First we use about 200 thousand face image pairs, which we collected from LFW, Honda/UCSD and YouTube celebrity database, to pre-train our DNN. Then we use 9 splits of YTF to fine-tuning our DNN and the left one split to test. We random 2000 ’same’ and 2000 ’not-same’ pairs from the 9 splits, and use these 4000 bags as the training set of the MIL. After the feature mapping, a 4000-d feature vector is formed per training bags. Then we use them to train the SVM classifier. Given a test pair, we first extract the bag-level feature and then use the trained SVM to classify.

Table 1: Architecture of DNN

Layer	Layer Type	Size	Output Shape
C1	Convolution + ReLU	32×44 filters	(32, 60, 52)
M2	Max Pooling	22, stride 2	(32, 30, 26)
C3	Convolution + ReLU	64×33 filters	(64, 28, 24)
M4	Max Pooling	22, stride 2	(64,14,12)
C5	Convolution + ReLU	128×33 filters	(128,12,10)
M6	Max Pooling	22, stride 2	(128,6,5)
C7	Convolution + ReLU	64×22 filters	(64,5,4)
F8	Fully Connected+ReLU	400 hidden units	400
F9	Softmax	2 way	2

For video-based face verification on the YTF dataset, we compare our method MILDIS with the several existing methods, including LM3L [18], DDML(LBP) [19], DDML(combined) [19], EigenPEP [20], DeepFace-single [3] and DeepID2+ [21]. The recognition accuracies and standard deviations of different methods are reported in Table 2. Note that our method MILDIS is better than most other methods except DeepID2+. Possibly because DeepID2+ has a much more complex architecture and its developers also have much more data and computing resource to tuning it. From this table, out method achieves the state-of-the-art performances on the YTF dataset.

Table 2: The comparisons on the YouTube Faces dataset
(Mean Accuracy ± Deviation in %).

Method	Result
LM3L [18]	81.3±1.2
DDML(LBP) [19]	81.3±1.6
DDML(combined) [19]	82.3±1.5
EigenPEP [20]	84.8±1.4
DeepFace-single[8]	91.4±1.1
DeepID2+[21]	93.2±0.2
MILDIS	92.8±1.4

Conclusions

In this paper, we have proposed a new method for video-based face recognition. To improve the robustness, we formulate the issue as a multi-instance learning (MIL) problem and then apply a new MIL algorithm (MILDIS) to solve it. MILDIS first extract the deep feature of face pairs. Then though instance selection and feature mapping, MILDIS reduce the large amount of instances and obtain the bag-level feature. Because of this, our method is both robust and efficient. Our work achieves the state-of-the-art performances on the real-world datasets YouTube Faces (YTF) according to the restricted protocol.

References

[1] Zhen Cui, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen, “Image sets alignment for videobased face recognition,” in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 2626–2633.

[2] Ognjen Arandjelovi´c, Gregory Shakhnarovich, John Fisher, Roberto Cipolla, and Trevor Darrell, “Face recognition with image sets using manifold density divergence,” in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005, vol. 1, pp. 581–588.

[3] Gregory Shakhnarovich, John W Fisher, and Trevor Darrell, “Face recognition from long-term observations,” in Computer VisionłECCV 2002, pp. 851–865. Springer, 2002.

[4] Masashi Nishiyama, Mayumi Yuasa, Tomoyuki Shibata, TomokazuWakasugi, Tomokazu Kawahara, and Osamu Yamaguchi, “Recognizing faces of moving people by hierarchical image-set matching,” in Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on. IEEE, 2007, pp. 1–8.

[5] Kazuhiro Fukui and Osamu Yamaguchi, “The kernel orthogonal mutual subspace method and its application to 3d object recognition,” in Computer Vision–ACCV 2007, pp. 467–476. Springer, 2007.

[6] Ruiping Wang, Shiguang Shan, Xilin Chen, and Wen Gao, “Manifold-manifold distance with application to face recognition based on image set,” in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008, pp. 1–8.

[7] Ruiping Wang and Xilin Chen, “Manifold discriminant analysis,” in Computer Vision and pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009, pp. 429–436.

[8] Yaniv Taigman, Ming Yang, Marc’Aurelio Ranzato, and Lars Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, 2014, pp. 1701–1708.

[9] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang, “Deep learning face representation by joint identification-verification,” in Advances in Neural Information Processing Systems, 2014, pp. 1988–1996.

[10] Yann LeCun, L´eon Bottou, Yoshua Bengio, and Patrick Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.

[12] Zhen-Zhong Lan, Lu Jiang, Shoou-I Yu, Shourabh Rawat, Yang Cai, Chenqiang Gao, Shicheng Xu, Haoquan Shen, Xuanchong Li, Yipei Wang, et al., “Cmuinformedia at trecvid 2013 multimedia event detection,” in TRECVID 2013 Workshop, 2013, vol. 1, p. 5.

[13] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann, “Support vector machines for multipleinstance learning,” in Advances in neural information processing systems, 2002, pp. 561–568.

[14] Yixin Chen, Jinbo Bi, and James Z Wang, “Miles: Multiple-instance learning via embedded instance selection,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 28, no. 12, pp. 1931–1947, 2006.

[15] Qi Zhang and Sally A Goldman, “Em-dd: An improved multiple-instance learning technique,” in Advances in neural information processing systems, 2001, pp. 1073–1080.

[16] Junlin Hu, Jiwen Lu, Junsong Yuan, and Yap-Peng Tan, “Large margin multi-metric learning for face and kinship verification in the wild,” in Computer Vision–ACCV 2014, pp. 252–267. Springer, 2015.

[17] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid, “Is that you? metric learning approaches for face identification,” in Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009, pp. 498–505.

[18] Hieu V Nguyen and Li Bai, “Cosine similarity metric learning for face verification,” in Computer Vision–ACCV 2010, pp. 709–720. Springer, 2011.

[19] Junlin Hu, Jiwen Lu, and Yap-Peng Tan, “Discriminative deep metric learning for face verification in the wild,” in Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, 2014, pp. 1875–1882.

[20] Haoxiang Li, Gang Hua, Xiaohui Shen, Zhe Lin, and Jonathan Brandt, “Eigen-pep for video face recognition,” in Computer Vision–ACCV 2014, pp. 17–33. Springer, 2015.

[21] Yi Sun, Xiaogang Wang, and Xiaoou Tang, “Deeply learned face representations are sparse, selective, and robust,” arXiv preprint arXiv:1412.1265, 2014.