Summaries for "DeepFace:Closing the Gap to Human-Level Performance in Face Verification"

Weituo Hao 109241801
Stony Brook University
Computer Science Department
Advanced Computer Vision

Contribution

The author proposes to combine nine-layer neural network without weights sharing with accurate model-based alignment to complete the face recognition. This method can reach an accuracy of 97.35% on the Labeled Faces in the Wild (LFW) as claimed by the author.

Main idea

First 2D alignment is adopted by detecting 6 fiducial points distributing inside the detection crop, center of the eyes and tip of the nose and mouth locations. They are used to scale, rotate and translate the image into six anchor locations.

Then for 3D alignment, the author uses a generic 3D model and registers a 3D affine camera that are used to warp the 2D-aligned crop to the image plane of the 3D shape.

For the representation, the author proposes a architecture as follows. A 3D-aligned 3-channels (RGB) face image of size 152 by 152 pixels is given to a convolutional layer with 32 filters of size 11 by 11 by 3 and lead these to a max-pooling layer which takes the max over 3 by 3 spatial neighborhoods with a stride of 2 followed by another convolutional layer that has 16 filters. Since several levels of pooling would cause the network to lose information about the precise position of detailed facial structure. So max-pooling structure is only applied to the first convolutional layer.

On the other hand, the author uses three large locally connected layers without weights sharing because of the fact that each output unit of locally connected layer is affected by a very large patch of the input. And ReLU is applied to every convolution, locally connected and fully connected layer, making the whole representation sparse. The final problem is about the identity of two images. For the similarity metric, the author uses the inner product between the two normalized feature vectors.

Experimental results

The whole method is applied on three different datasets: Social Face Classification (SFC), Labeled Faces in the Wild (LFW) and YouTube Faces (YTF). On the SFC dataset, the error result grows very modestly as the dataset grows larger.

Network	Error	Network	Error	Network	Error
DF-1.5K	7.00%	*DF-10%*	20.7%	*DF-sub1*	11.2%
DF-3.3K	7.22%	*DF-20%*	15.1%	*DF-sub2*	12.6%
DF-4.4K	8.74%	*DF-50%*	10.9%	*DF-sub3*	13.5%

Table 1. Comparison of the classification errors on the SFC w.r.t. training dataset size and network depth.

And comparison compared to different methods can be seen as the following two tables.

Method	Accuracy ± SE	Protocol
Joint Bayesian [6]	0.9242 ±0.0108	restricted
Tom-vs-Pete [4]	0.9330 ±0.0128	restricted
High-dim LBP [7]	0.9517 ±0.0113	restricted
TL Joint Bayesian [5]	0.9633 ±0.0108	restricted
DeepFace-single	**0.9592** ±0.0029	unsupervised
DeepFace-single	**0.9700** ±0.0028	restricted
DeepFace-ensemble	**0.9715** ±0.0027	restricted
DeepFace-ensemble	**0.9735** ±0.0025	unrestricted
Human, cropped	0.9753	

Table 2. Comparison with the state-of –the-art on the LFW dataset

Method	Accuracy (%)	AUC	EER
MBGS+SVM- [31]	78.9 ±1.9	86.9	21.2
APEM+FUSION [22]	79.1 ±1.5	86.6	21.4
STFRD+PMML [9]	79.5 ±2.5	88.6	19.9
VSOF+OSS [23]	79.7 ±1.8	89.4	20.0
DeepFace-single	**91.4** ±1.1	96.3	8.6

Table 3. Comparison with the state-of-the art on the YTF dataset

Conclusion

The author claims that coupling a 3D model-based alignment with large capacity feedforward models can effectively learn from many examples to overcome the drawbacks and limitations like generalizing issues exist in previous methods, and can be a potential method to other vision domains as well.

Summaries for "Recurrent Convolutional Neural Networks for Scene Labeling"

Weituo Hao 109241801
Stony Brook University
Computer Science Department
Advanced Computer Vision

Contribution

The author proposes a method for scene labeling based on recurrent convolutional neural network, and it does not require any engineered features, and does not rely on any label space searching. The approach yields state-of-art performance on Stanford Back-ground Dataset and the SIFT Flow Dataset.

Main idea

To solve the label dependencies, the author proposes recurrent network. This architecture consists of the composition of P instances of the "plain" convolutional network f(.). And f(.) means a standard CNN out put for a given input patch $I_{i,j,k}$, where (i,j) means the location of pixel and k means the kth image. Each instance has the same trainable parameters (**W**, **b**). The pth instance of the network ($1 \leq p \leq P$) is fed with an input "image" $\boldsymbol{F}^p$ of N+3 features maps

$$\boldsymbol{F}^p = [f(\boldsymbol{F}^{p\text{-}1}), I^p_{i,j,k}], \boldsymbol{F}^1 = [\mathbf{0}, I_{i,j,k}]$$

And the whole system is trained by maximizing the likelihood

$$L(f) + L(fof) + \cdots + L(fo^p f)$$

where L(f) is the likelihood in the case of plain CNN, and o^p denotes the composition operation performed p times.

The following step is proposed by the writer is scene inference that will be solved by such equation. Given a test image I_k, for each pixel at location (i,j) the network predicts a label as:

$$\widehat{\iota_{i,j,k}} = argmax_{c \in classes} p(c|I_{i,j,k}; (\boldsymbol{W}, \boldsymbol{b}))$$

Experimental results

The author test the method on two different fully labeled datasets. The first one is the Stanford Background and the SIFT Flow Dataset. All networks are trained by sampling patches surrounding a randomly chosen pixel from a randomly chosen image from the training set. Two accuracy measures are used to compare the proposed method with the previous work. The first one is the accuracy per pixel of test images and the second one is the averaged per class accuracy.

METHOD	PIXEL/CLASS ACCURACY (%)	COMPUTING TIME (S)
(GOULD ET AL., 2009)	76.4 / -	10 TO 600
(TIGHE & LAZEBNIK, 2010)	77.5 / -	10 TO 300
(MUNOZ ET AL., 2010)‡	76.9 / 66.2	12
(KUMAR & KOLLER, 2010)	79.4 / -	< 600
(SOCHER ET AL., 2011)	78.1 / -	?
(LEMPITSKY ET AL., 2011)	81.9 / 72.4	> 60
(FARABET ET AL., 2013)⋆	78.8 / 72.4	0.6
(FARABET ET AL., 2013)†	81.4 / 76.0	60.5
PLAIN CNN_1	79.4 / 69.5	15
CNN_2 ($\circ^1$)	67.9 / 58.0	0.2
$RCNN_2$ ($\circ^2$)	79.5 / 69.5	2.6
CNN_3 ($\circ^1$)	15.3 / 14.7	0.06
$RCNN_3$ ($\circ^2$)	76.2 / 67.2	1.1
$RCNN_3$ 1/2 RESOLUTION ($\circ^3$)	79.8 / 69.3	2.15
$RCNN_3$ 1/1 RESOLUTION ($\circ^3$)	80.2 / 69.9	10.7

Table 1. Comparison with previous work on Stanford Background Dataset

METHOD	PIXEL/CLASS ACCURACY (%)
(LIU ET AL., 2011)	76.67 / -
(TIGHE & LAZEBNIK, 2013)	77.0 / 30.1
(FARABET ET AL., 2013)	78.5 / 29.6
PLAIN CNN_1	76.5 / 30.0
CNN_2 ($\circ^1$)	51.8 / 17.4
$RCNN_2$ ($\circ^2$)	76.2 / 29.2
$RCNN_3$ ($\circ^2$)	65.5 / 20.8
$RCNN_3$ ($\circ^3$)	77.7 / 29.8

Table 2. Comparison with previous work on SIFT Flow Dataset

From the above two tables the author claims that their method achieves state-of-the-art results and runs very fast.

Conclusion

The author proposes a novel feed-forward approach for full scene labeling based on recurrent architecture to overcome the long range label dependencies and claims no expensive graphical model or segmentation technique is needed and the whole method can run in a relatively low computing cost.