CORRELATION-BASED FACE DETECTION FOR RECOGNIZING FACES IN VIDEOS

Heng-Wei Hsu[1]*, Tung-Yu Wu*[2]*, Wing Hung Wong*[2,3] *and Chen-Yi Lee*[1]

[1]Institute of Electronics, National Chiao Tung University, Hsinchu, Taiwan
[2]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA
[3]Department of Statistics, Stanford University, Stanford, CA, USA

ABSTRACT

Finding the locations and identities of faces in videos is a very important task in numerous applications. In this paper, we propose a correlation-based face detection approach to improve the performance of face recognition tasks for videos. We apply correlation measures to pairs of response maps which are generated from automatically selected neurons in deep convolutional neural network (CNN) models to detect faces in each video frame. The embeddings extracted from faces cropped by our proposed approach are more consistent across each video sequence and more suitable for face recognition and clustering tasks. Experimental results from the YouTube Faces (YTF) dataset demonstrate that our proposed approach is more robust and achieves better recognition accuracy compared to state-of-the-art face detection approaches.

Index Terms— Convolutional neural network, deep learning, neuron selection, face detection, face recognition

1. INTRODUCTION

Recognizing faces in videos has gained much interest recently due to the fast growth of social media. In this context, each person is represented by a sequence of faces in video frames rather than one single image. Therefore, to improve the recognition performance, it is important to utilize the correlation between consecutive video frames to detect faces. If correlation information is not utilized and faces are detected independently for each frame, the discontinuity of the face images within each video sequence will introduce large variations in the embedding space leading to inferior performance.

In our work, we propose a correlation-based approach that utilizes response maps from CNN models to detect faces in video sequences such that the face features of each identity are better aligned in the embedding space. CNN models have significantly advanced the state-of-the-art in many computer vision tasks, such as image classification [1, 2], face detection [3, 4, 5], and face recognition [5, 6, 7, 8, 9, 10], but the understanding of what they learn has been far behind. Zeiler *et al.* [11, 12] proposed Deconvolutional Network (deconvnet) to visualize features learned by different neurons by reversing the operations of normal forward computations into backward computations. Several works [13, 14, 15] propose explanatory frameworks that analyze network predictions to automatically identify important neurons. These neurons can be translated to input space and generate interpretable response maps for understanding CNN models. Our framework leverages this concept to generate response maps that show clear silhouette of the faces in each video sequence, and the correlation between these response maps are utilized to detect faces. We then input these face sequences to recognition models to generate face embeddings for each identity. These recognition models [6, 7, 8, 9, 10, 16] learn compact face embeddings which discriminate faces of different identities.

We evaluate the performance of our proposed approach on the YouTube Faces (YTF) [17] dataset. Each identity in the YTF dataset contains several video sequences, and only one embedding is needed to represent each video sequence. Most methods average embeddings extracted from the K frames sampled from each video sequence either randomly or under certain conditions. If the faces are detected independently for each frame, the features in the embedding space may not be consistent. We compare our result with recent face detection algorithms to demonstrate the superiority of our proposed approach. Deformable part models (DPM) based method [18] is one of popular and widely adopted approaches. It applies a trained latent SVM as classifier to find geometric relations within parts converted from faces. Despite its high computational complexity, DPM achieves state-of-the-art on several benchmark datasets and even outperforms some CNN based methods [3, 4]. We demonstrate that compared with such an accurate face detector, faces cropped by our proposed approach generate more consistent embeddings, resulting in better face recognition performance.

2. PROPOSED APPROACH

In this section, we introduce our proposed approach which utilizes the correlation between response maps generated from CNN models to detect faces in video sequences. First, we extract response maps for all the images in each video sequence. This is achieved by a forward pass through the

Work supported by MOST of Taiwan under Grant 105-2218-E-009-001 and NSF of USA under Grant DMS-1407557 and DMS-1721550.

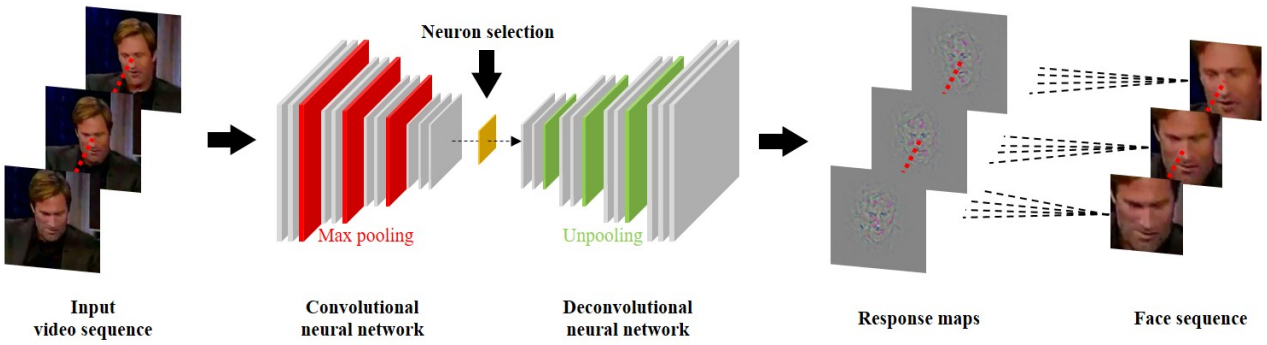

Fig. 1: The overall framework of our proposed approach.

CNN model, stopping at an automatically selected neuron. A deconvnet is then attached to this neuron to generate the response maps. Finally, the correlation between these response maps are calculated to find the face locations and the regions with higher response will be cropped as face images. The overview of our proposed approach is shown in Fig. 1.

2.1. Automatic Neuron Selection

CNN models encode semantic features in neurons of different layers. Neurons in deeper layers capture complex features, while neurons in shallow layers only activate on simple features such as edges or color. Inspired by [15], we target to automatically select a neuron that activates on the faces while having small variations within each video sequence. Let $x_n \in \{x_1, ..., x_N\}$ denote the frame image from the input video sequence, and $z_n^{l,i}$ denote the activations of the i^{th} neuron in the l^{th} layer of a CNN model given x_n. The overall variance $v_{l,i}$ is calculated by averaging variance of each (r, c) element in $z_n^{l,i}$ which is of size $R \times C$.

$$v_{l,i} = \frac{1}{RCN} \sum_{r=1}^{R} \sum_{c=1}^{C} \sum_{n=1}^{N} (z_n^{l,i}[r][c] - \mu_{r,c}^{l,i})^2 \quad (1)$$

$$\mu_{r,c}^{l,i} = \frac{\sum_n z_n^{l,i}[r][c]}{N} \quad (2)$$

Neurons that have small variation are considered important, since they consistently contribute similar activation magnitude despite the variations within the video sequences. However, one case to be excluded is when some neurons are not activated throughout the video sequence which results in zero variance. Hence we also require the important neurons to have large activation magnitude. Let $m_{l,i}$ denote the activation magnitude of $z_n^{l,i}$ which is obtained by averaging responses of the neuron within each video sequence.

$$m_{l,i} = \frac{1}{RC} \sum_{r=1}^{R} \sum_{c=1}^{C} \mu_{r,c}^{l,i} \quad (3)$$

To enhance the stability of neuron selection, we introduce a target specific prior which emphasizes the important elements in the neuron. We multiply each neuron by a Gaussian kernel $g^{l,i}$ with the peak centered at the max element of the neuron. The standard deviation σ of the Gaussian kernel controls the concentration level of the target with respect to our prior knowledge. This Gaussian kernel penalizes elements that have large activations but are far from the max element of the neuron to avoid selecting neurons that activate on certain texture of the background. Therefore, the activation magnitude $m_{l,i}$ is reformulated as,

$$m_{l,i} = \frac{1}{RC} \sum_{r=1}^{R} \sum_{c=1}^{C} \mu_{r,c}^{l,i} \odot g^{l,i} \quad (4)$$

The ability of each neuron to encode important features across the video sequence is approximated by maximizing the objective function I,

$$I(l, i) = m_{l,i} - \lambda v_{l,i} \quad (5)$$

where λ balances the effect of activation magnitude and variance. Since faces in the videos have various angles, we set λ to 0.1 to tolerate the variations induced by the faces. Neurons that maximize the objective function are chosen to generate the response maps.

2.2. Response Map Generation

All neuron responses in the selected layer are set to zero, except the chosen one which is attached to a deconvnet [12] for deconvolving back to the image space such that the important patches are emphasized. This process includes unpooling, rectification and transpose filtering, and is repeated until it reaches the input image space. The generated response maps have the same dimension as the input images and show clear silhouette of the faces which provide information of the face locations. Response maps of consecutive frames are slightly different due to the noise of the background and motion of the faces. Therefore, cropping regions with large response directly may lead to crooked face images which introduce large variations in the face embedding space even in consecutive frames. This issue can be mitigated by the method introduced in the next section.

Fig. 2: An example of vertical circular shift. The original image is in the middle. The images from left to right correspond to pixels shifted by -30, -15, 0, 15, 30 pixels.

2.3. Correlation Calculation

Given the response maps of the input video sequence, we exploit the correlation between pairs to find the optimal face locations. Instead of considering the correlation only between two response maps, we apply circular shift to each response map and average the correlation over all possible combinations to find the optimal face location. The operation of circular shift is to shift the image patch in one direction, and the pixels out of boundary are then warped back to fill the pixels on the other side. Fig. 2 shows the concept of circular shift. Calculating the correlation for all circular shifts of two response maps demands high computational complexity. Henriques *et al.* [19] proposed that matrices which contain circular shifts of data in each row can be made diagonal by applying Discrete Fourier Transform (DFT). With this property, all operations in frequency domain are done element-wise on the diagonal elements which greatly simplifies the computational complexity. Thus the optimal face location (a, b) between two response maps h and h' is obtained by,

$$\arg\max_{a,b} \sum_{d} \mathcal{F}^{-1}(\hat{h}_d^* \odot \hat{h}'_d) \tag{5}$$

where * denotes complex-conjugate, $\hat{h}$ denotes the DFT of h, and $\mathcal{F}$ is the DFT operation such that $\hat{h} = \mathcal{F}(h)$. An inverse DFT is applied to the correlation of each channel d of the response maps to transform them back to spatial domain. These correlation response maps from different channels are summed up and we target to find the coordinate (a, b) with maximum correlation response. The faces in each pair of response map are cropped by centering a bounding box at (a, b) and the corresponding size is determined by the region that has response outside three standard deviations of the mean. To fully utilize the correlation among the response maps within video sequences, the face location of each frame is obtained by averaging the coordinates (a, b) from all possible pairs within a window of length L.

We take advantage of the natural characteristics of video to search for the position of faces that have the maximum correlated response among the response maps. Since there are limited differences between consecutive frames, faces appear in the previous frames tend to remain in the local neighborhood, as a result only minor shifts are needed to cope with the minor differences. By cropping strongly correlated faces, consistent embeddings can be obtained to improve face recognition performance.

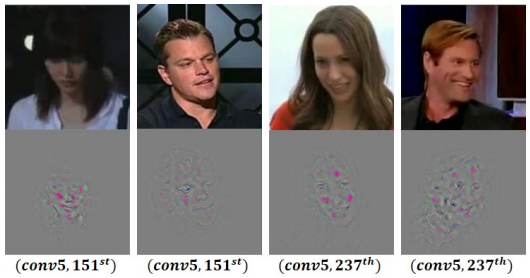

Fig. 3: The automatically selected neurons and corresponding response maps of frame images from four video sequences.

Fig. 4: Faces cropped by different approaches from the first video of Aaron Eckhart. The top row is cropped by directly finding the maximum response from independent frames. The middle row is cropped by our proposed approach. The bottom row is cropped by DPM. Faces are resized to the same scale.

3. EXPERIMENTS

3.1. Experimental Setup

We use the caffenet model proposed in [20] as the response map generation model, and σ is set to 3 for the Gaussian kernel. When calculating the correlations among response maps, we set the overlapping window to length $L = 4$ and stride $L/2$. Two public face recognition models, VGG deep face (VGG) [6] and Lightened_CNN_B (LCNN) [7] are used to extract the face embeddings. The embedding vectors are of size 4,096 and 256 respectively for these two models. We input the same video sequences to our proposed approach and DPM to detect the faces and extract the embeddings for all cropped faces. The performance is evaluated by reporting the accuracy which is defined as $100\%-$EER (Equal Error Rate) on the YTF dataset following conventional settings and checking the feature consistency within each video sequence. The feature consistency is measured by comparing the variance and cosine similarities of consecutively/randomly sampled pairs within the embeddings of each video sequence.

3.2. Results and Discussion

Qualitative results: Fig. 3 illustrates the response maps of the automatically chosen neurons from four different video sequences. The chosen neurons clearly reveal the silhouette of the faces. Fig. 4 shows face images cropped by different approaches. Our proposed approach crops more consistent faces across the video sequence.

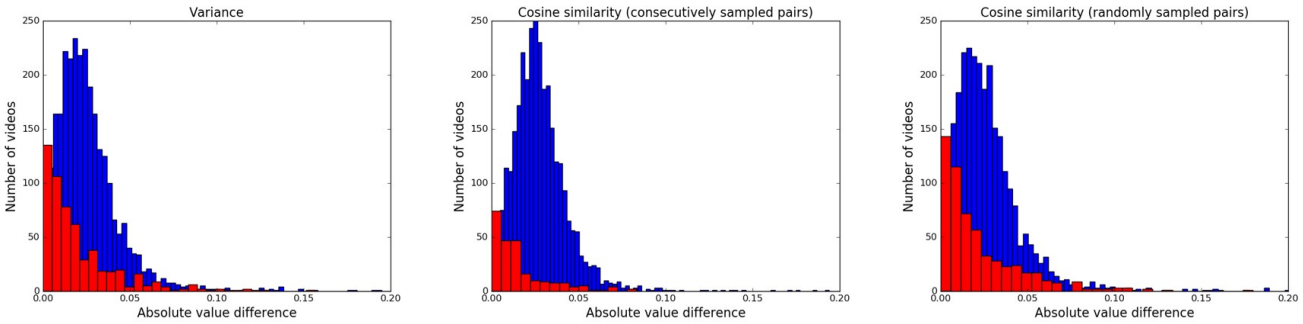

Fig. 5: Histogram plots for the absolute value difference of the three indexes between our proposed approach and DPM. The number of videos our proposed approach outperforms DPM are shown as the blue bins, and otherwise shown as the red bins.

Quantitative results: We evaluate the feature consistency within each video sequence by calculating the absolute value difference of three different indexes between our proposed approach and DPM, (i) variance, (ii) average cosine similarity of consecutive pairs, and (iii) average cosine similarity of random pairs. The distributions of the absolute value difference are shown in Fig. 5, where the blue bins represent the number of videos that our proposed approach outperforms DPM, and the red bins represent the number of videos DPM outperforms ours. Embeddings extracted by our proposed approach have lower variance and higher cosine similarity in most of the videos which reveals the consistency of our embeddings within each video sequence. The peak value of the blue histograms are much larger than the red histograms, which shows that when our proposed approach outperforms, it improves the outcomes by a great margin.

Table 1: Comparison of our proposed approach and DPM when using same face recognition models on YTF.

	DPM	Our proposed approach		
		Original	Max	Avg
VGG	87.8	88.8	**88.9**	88.9
LCNN	90.4	93.0	**93.0**	92.9

We apply two different regularization methods to the response maps to preserve important information and compare the results of the three approaches. (i) Original: The response map generated from the deconvnet. (ii) Max: Only the pixel with the maximum value among the depth of the original response map is preserved. (iii) Avg: The value of each image pixel is averaged across the depth of the original response map. In Table 1, we compare the result of faces cropped by our proposed approach and DPM on YTF dataset. Although the original response maps show clear silhouette of the faces, the activation on the three channels does not represent the RGB values, it only reflects the intensity of the back propagating gradients from the chosen neuron. Thus the Max method preserves larger activations and leads to better results. Extracting embeddings from faces cropped by our proposed approach consistently leads to better performance on both public models which demonstrates the effectiveness of our proposed approach.

Table 2: Comparison of state-of-the-art approaches on YTF.

Models	#Net	Accuracy
WebFace [16]	1	90.6
DeepFace [9]	1	91.4
LCNN [7]	1	**91.6**
VGG [6]	1	92.8
LCNN*	1	**93.0**
DeepID2+ [8]	25	93.2
FaceNet [10]	1	95.1

Table 2 demonstrates that if faces are cropped by our proposed approach and the embeddings are extracted from the same network structure (LCNN*), the performance can be greatly improved compared to the result in [7] and achieve comparable result to state-of-the-art. Note that no preprocessing methods are applied to the face images before extracting the embeddings, thus the performance can be further improved if preprocessing methods such as 2D/3D face alignment are used.

4. CONCLUSION

In this paper, we target to enhance the face recognition performance in videos by exploiting the correlation within response maps generated from automatically selected neurons to find the optimal face locations throughout video sequences. Experiments show that embeddings generated from faces cropped by our proposed approach are more consistent and representative which significantly improve the baseline accuracy of the YTF dataset.

5. REFERENCES

[1] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[2] Z. Li, Y. Song, I. Mcloughlin, and L. Dai, "Compact convolutional neural network transfer learning for small-scale image classification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 2737–2741.

[3] Sachin Sudhakar Farfade, Mohammad J Saberian, and Li-Jia Li, "Multi-view face detection using deep convolutional neural networks," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 2015, pp. 643–650.

[4] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua, "A convolutional neural network cascade for face detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5325–5334.

[5] W. Jiang and W. Wang, "Face detection and recognition for home service robots with end-to-end deep neural networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 2232–2236.

[6] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[7] Xiang Wu, Ran He, and Zhenan Sun, "A lightened cnn for deep face representation," *arXiv preprint arXiv:1511.02683*, 2015.

[8] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2892–2900.

[9] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1701–1708.

[10] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[11] Matthew D Zeiler, Graham W Taylor, and Rob Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 2018–2025.

[12] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[13] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje, "Not just a black box: Learning important features through propagating activation differences," *arXiv preprint arXiv:1605.01713*, 2016.

[14] Pang Wei Koh and Percy Liang, "Understanding black-box predictions via influence functions," *arXiv preprint arXiv:1703.04730*, 2017.

[15] Benjamin J. Lengerich, Sandeep Konam, Eric P. Xing, Stephanie Rosenthal, and Manuela M. Veloso, "Visual explanations for convolutional neural networks via input resampling," *CoRR*, vol. abs/1707.09641, 2017.

[16] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[17] Lior Wolf, Tal Hassner, and Itay Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 529–534.

[18] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool, "Face detection without bells and whistles," in *European Conference on Computer Vision*. Springer, 2014, pp. 720–735.

[19] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.

[20] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson, "Understanding neural networks through deep visualization," in *Deep Learning Workshop, International Conference on Machine Learning (ICML)*, 2015.