# Correlation-based Face Detection for Recognizing Faces in Videos

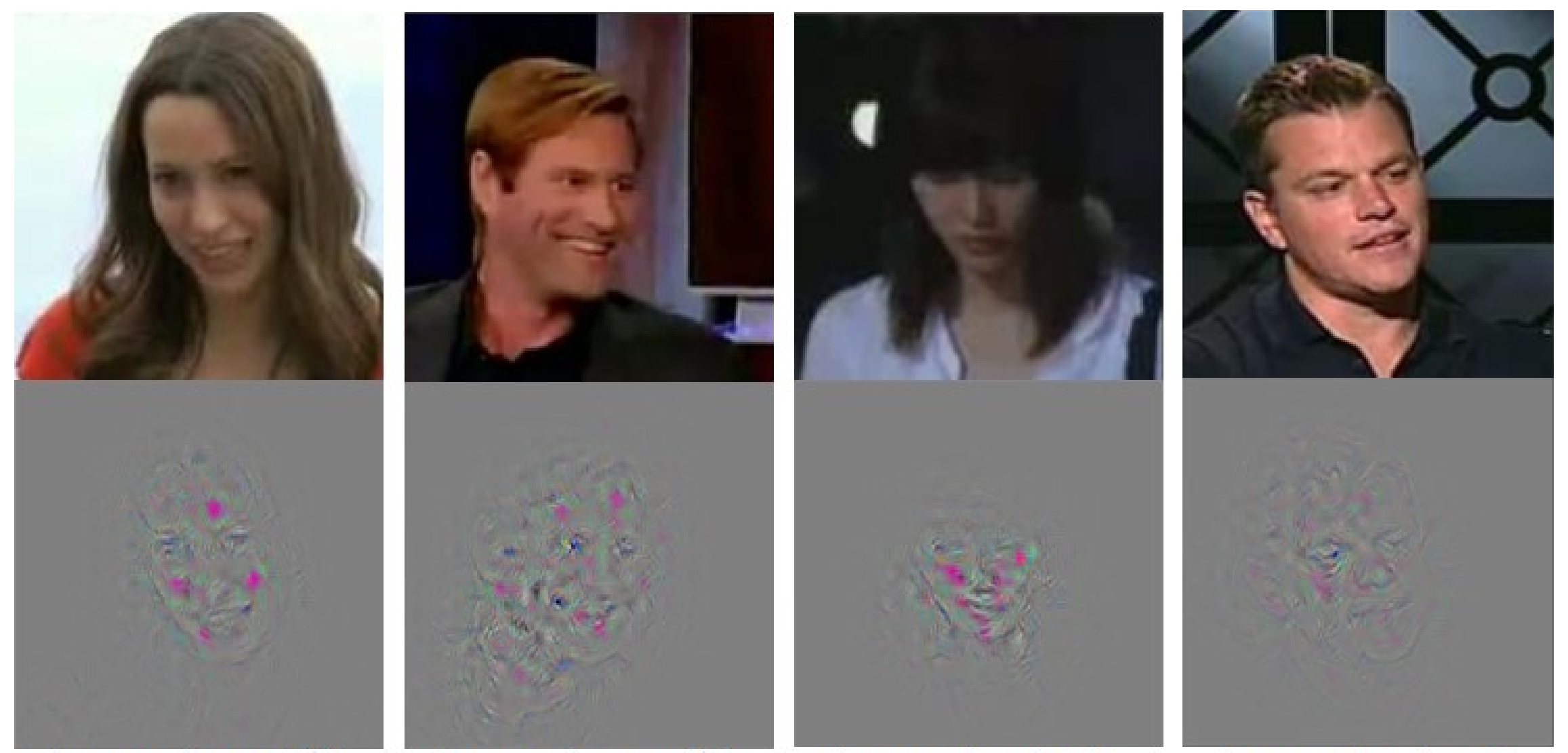
Heng-Wei Hsu[1], Tung-Yu Wu[2], Wing-Hung Wong[2,3] and Chen-Yi Lee[1]

[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

## Introduction

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. We leverage the concept of [1] to analyze network predictions to automatically identify important neurons within each video sequence and translate them to input space to generate interpretable response maps (clear silhouette of the faces). The correlation between these response maps are utilized to find the optimal face locations.

We compare our result on the YouTube Faces (YTF) dataset with recent face detection algorithms to demonstrate the superiority of our approach. Deformable part models (DPM) based method [2] is one of popular and widely adopted approaches. We demonstrate that compared with such an accurate face detector, faces cropped by our approach generate more consistent embeddings, resulting in better face recognition performance.



($\mathbf{conv5}, \mathbf{237^{th}}$) ($\mathbf{conv5}, \mathbf{237^{th}}$) ($\mathbf{conv5}, \mathbf{151^{st}}$) ($\mathbf{conv5}, \mathbf{151^{st}}$)

Figure 1: The automatically selected neurons and corresponding response maps of images from four different videos.

## Proposed Framework

**Automatic Neuron Selection** The neuron is selected by maximizing the objective function $I$,

$$I(l,i) = m_{l,i} - \lambda v_{l,i} \tag{4}$$

where $\lambda$ balances the effect of activation magnitude $m_{l,i}$ and variance $v_{l,i}$. Let $x_n \in \{x_1, ..., x_N\}$ denote the frame image, and $z_n^{l,i}$ denote the activations of the $i^{th}$ neuron in the $l^{th}$ layer of a CNN model. The overall variance $v_{l,i}$ is calculated by averaging variance of each $(r,c)$ element in $z_n^{l,i}$.

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

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

To enhance the stability, we introduce a target specific prior, a Gaussian kernel $g^{l,i}$ with peak centered at the max element of the neuron.

$$m_{l,i} = \frac{1}{RC} \sum_{r=1}^{R} \sum_{c=1}^{C} \mu_{r,c}^{l,i} \odot g^{l,i} \tag{3}$$

**Response Map Generation** The selected neuron is attached to a deconvnet for deconvolving back to the image space to generate response maps that show clear silhouette of the faces.

**Correlation Calculation** We apply circular shift to each response map and average the correlation over all combinations to find the optimal face location $(a,b)$ between two response maps $h$ and $h'$,

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

where $^*$ is the complex-conjugate, and $\mathcal{F}$ is the DFT operation. The faces in each pair of response map are cropped by centering a bounding box at $(a,b)$.

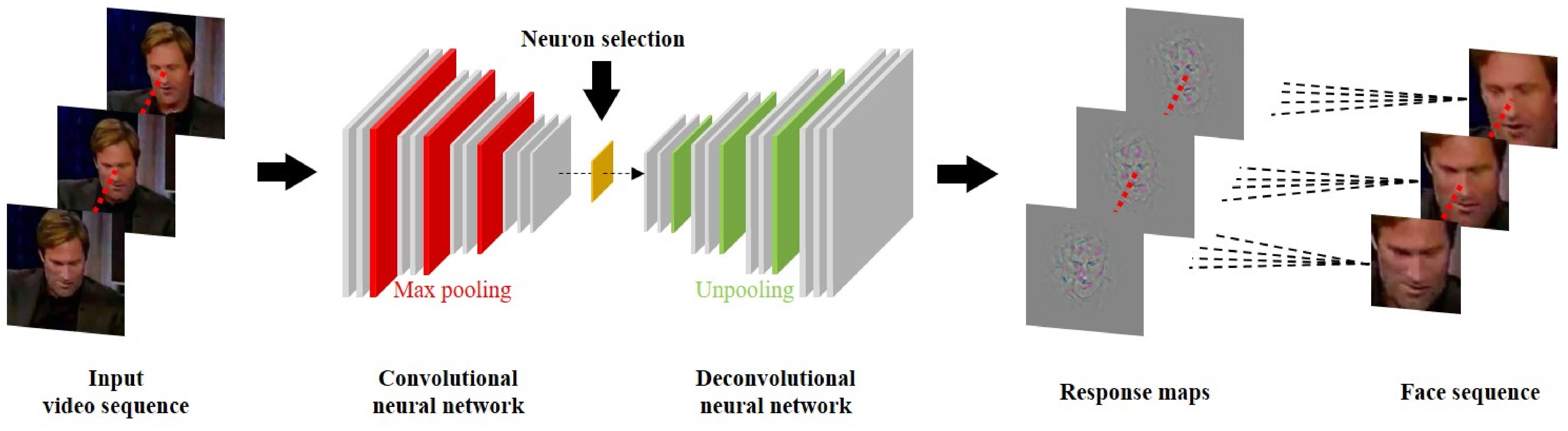


Figure 2: The overall framework of our approach.

## Experiment

We evaluate the feature consistency within each video sequence by calculating the absolute value difference of three different indexes between our 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. 3, where the blue bins represent the number of videos that our approach outperforms DPM, and the red bins represent the number of videos DPM outperforms ours. Embeddings extracted by our 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 approach outperforms, it improves the outcomes by a great margin.

Extracting embeddings from faces cropped by our approach consistently leads to better performance on public models such as VGG [3] and LCNN [4]. If faces are cropped by our approach and the embeddings are extracted from the same network structure (LCNN), the performance can be improved to 93.0% compared to the result 91.6% in [4] and achieve comparable result to the state-of-the-arts.

## 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. Experiments show that embeddings generated from faces cropped by our approach are more consistent and representative which significantly improve the baseline accuracy of the YTF dataset.

## References (Partial)

[1] 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.

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

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

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

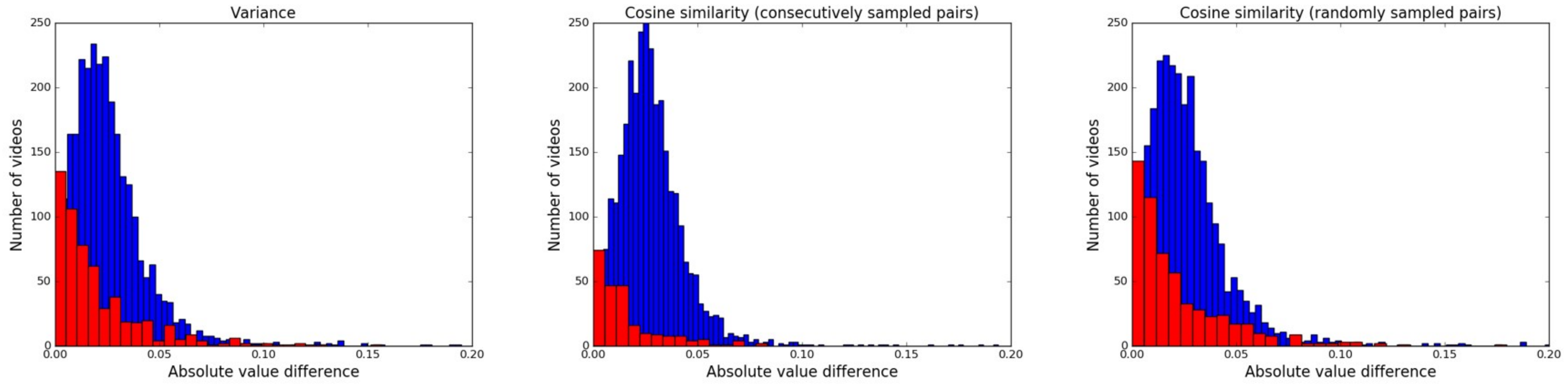


Figure 3: Histogram plots for the absolute value difference of the three indexes between our approach and DPM.

## Acknowledgements

Work supported by MOST of Taiwan under 105-2218-E-009 -001, 106-2218-E-009 -009 and NSF of USA under Grant DMS-1407557 and DMS-1721550.