End-to-End Localization and Ranking for Relative Attributes

UC DAVIS UNIVERSITY OF CALIFORNIA

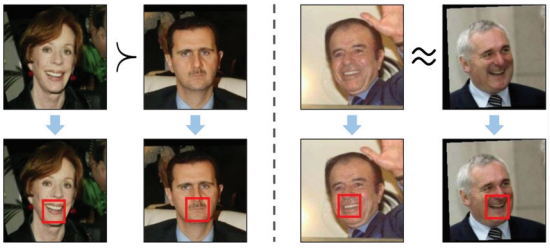
Krishna Kumar Singh Yong Jae Lee

ECCV'16 EUROPEAN CONFERENCE ON COMPUTER VISION

Motivation and Idea:

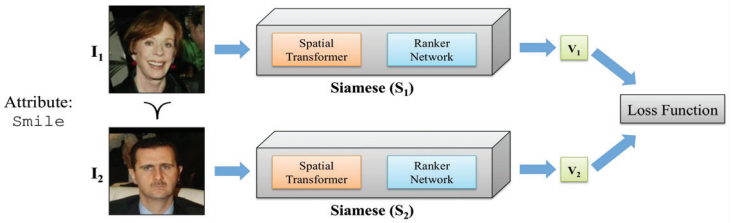
End-to-end deep network to simultaneously rank and localize relative attributes.

Attribute: Smile



- Local representation of visual attributes often leads to better performance [Bourdev ICCV 11, Duan CVPR 12, Sandeep CVPR 14].
- Existing localization methods have two limitations:
  - ➢ Require human supervision [Bourdev ICCV 11, Duan CVPR 12, Sandeep CVPR 14].
  - ➢ Not optimized jointly to rank and localize attribute [Xiao & Lee ICCV 15].
- **Key Idea**: Jointly optimize relative attribute ranking and localization with an end-to-end deep network.

Approach:



For training, pair of images (I_1, I_2) and label denoting their relative ordering is given as input to Siamese network which consists of:

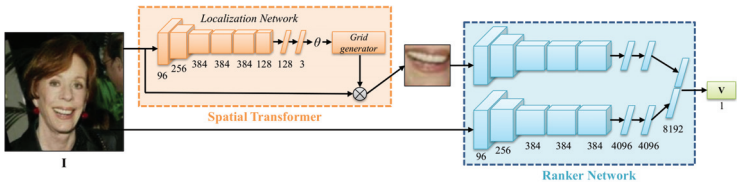
Spatial Transformer Network (STN): Learn image transformation parameters (t_x, t_y, s) to choose most relevant image region for ranking.

$$\begin{pmatrix} x_i^{in} \\ y_i^{in} \end{pmatrix} = \begin{bmatrix} s & 0 & t_x \\ 0 & s & t_y \end{bmatrix} \begin{pmatrix} x_i^{out} \\ y_i^{out} \\ 1 \end{pmatrix}$$

Ranker Network (RN): Takes STN output and original image (for context) as input, and combine their features to generate attribute strength (v).

For testing, one branch of the Siamese network is used to predict attribute strength.

Network Architecture and Loss:



Ranking Loss:

- $v_1 > v_2$ when I_1 has higher attribute strength than I_2, $L = 1$.
- $v_1 = v_2$ when I_1 and I_2 have similar attribute strengths, $L = 0.5$.

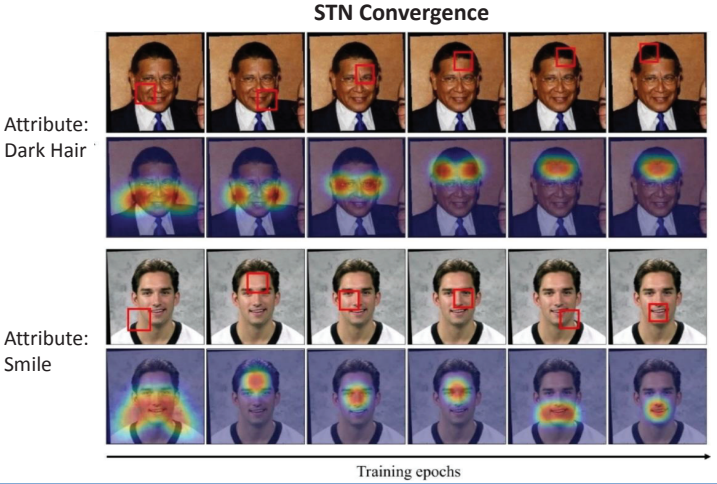
$$Rank_{loss}(I_1, I_2) = -L \cdot \log(P) - (1 - L) \cdot \log(1 - P), where$$
$$P = e^{v_1 - v_2} / 1 + e^{v_1 - v_2}$$

Spatial Transformer Loss: If output of STN goes outside of image boundaries:

$$ST_{loss}(I) = (C_x - s \cdot t_x)^2 + (C_y - s \cdot t_y)^2, where\ (C_x, C_y)\ is\ image\ center$$

Stochastic gradient descent is used to optimize the loss.

STN Convergence



Quantitative Results:

Attribute Ranking Accuracy:

LFW-10	BH	DH	EO	GL	ML	MO	S	VT	VF	Y	Avg
P & G* + CNN	78.10	83.09	71.43	68.73	95.40	65.77	63.84	66.46	81.25	72.07	74.61
Sandeep et al.	82.04	80.56	83.52	68.98	90.94	82.04	85.01	82.63	83.52	71.36	81.06
Xiao & Lee	83.21	88.13	82.71	**72.76**	93.68	88.26	**86.16**	86.46	**90.23**	75.05	84.66
Ours	**83.94**	**92.58**	**90.23**	71.21	**96.55**	**91.28**	84.75	**89.85**	87.89	**80.81**	**86.91**

Shoe-Attribute	Open	Pointy	Sporty	Avg
P & G* + CNN	77.1	72.5	71.6	73.7
Xiao & Lee	80.2	**82.5**	88.1	83.6
Ours	**89.3**	**82.5**	**93.6**	**88.5**

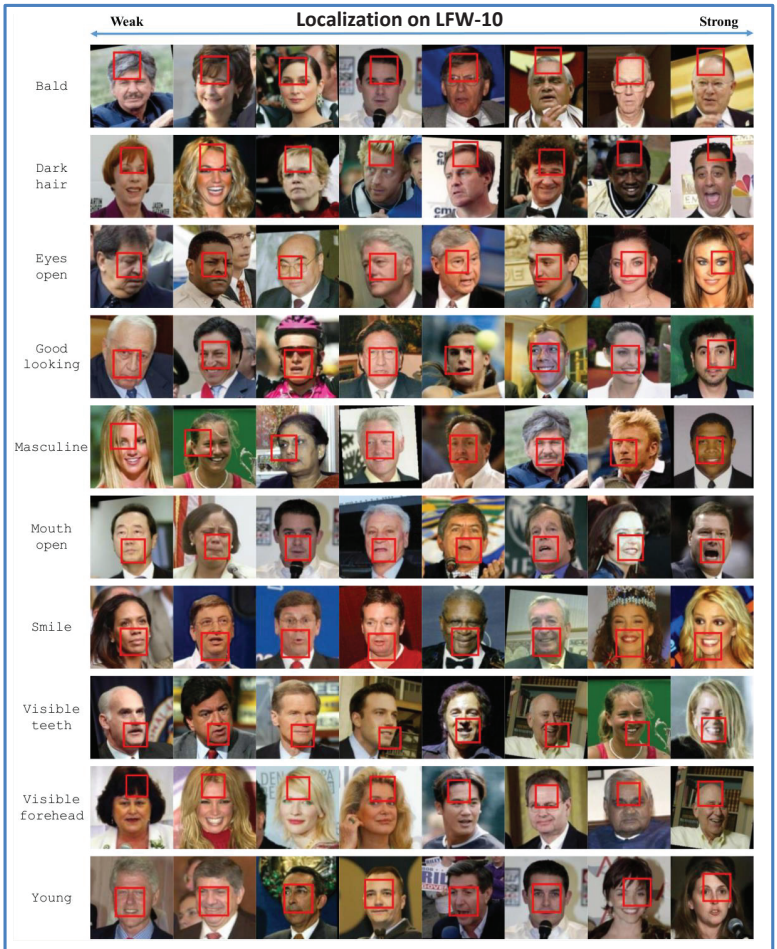
OSR	Nat	Opn	Per	S-L	Dia	Dep	Avg
Parikh+CNN	98.0	94.5	93.0	94.0	95.0	95.3	95.0
Li et al.	95.2	92.4	87.6	88.3	89.3	89.5	90.4
Yu et al.	95.7	94.1	90.4	91.1	92.4	90.5	92.4
Ours	**98.9**	**97.2**	**96.3**	**96.0**	**97.6**	**96.1**	**97.0**

Ablation Study (LFW-10)	BH	DH	EO	GL	ML	MO	S	VT	VF	Y	Avg
Global Image	**84.3**	90.2	82.7	70.0	94.8	80.2	80.8	79.4	85.6	77.2	82.5
STN output	78.1	89.3	**93.2**	66.6	95.4	**91.3**	84.5	88.6	85.6	73.8	84.6
Combined	83.9	**92.6**	90.2	**71.2**	**96.6**	**91.3**	**84.8**	**89.9**	**87.9**	**80.8**	**86.9**

- State-of-the-art results for attribute ranking accuracy.
- Global image and STN output contain complementary information.
- Our approach is 100 times faster than [Xiao & Lee].

P & G : Parikh and Grauman*

Qualitative Results:



- STN localizes relevant image regions.
- Attribute ranking is accurate overall.
- Localization to small and large regions for local and global attributes respectively.

Acknowledgement: This work was supported in part by an Amazon Web Services Education Research Grant and GPUs donated by NVIDIA.