Attribute Recognition by Joint Recurrent Learning of Context and Correlation

Jingya Wang jingya.wang@qmul.ac.uk
Xiatian Zhu eddy@visionsemantics.com
Shaogang Gong s.gong@qmul.ac.uk
Wei Li wei.li @qmul.ac.uk

Queen Mary University of London
ICCV17 International Conference on Computer Vision 2017

1 Introduction

➢ **Problem:**

Pedestrian attributes recognition in surveillance

➢ **Challenges:**

- Poor image quality
- Complex background clutter
- Uncontrolled viewing conditions
- Small number of labelled training

➢ **Main idea:**

- Discover the interdependency and correlation among attributes
- Explore visual context as an extra information source to assist attribute recognition

➢ **Contributions:**

- A novel end-to-end encoder-decoder architecture capable of jointly learning image level context and attribute level sequential correlation
- Exploit more latent and richer higher-order dependency among attributes

2 Joint Recurrent Learning of Context and Correlation

(a) Inter-Person Similarity Context

Search top-k exemplars $\{I_i^c\}_{i=1}^k$ that are visually similar to the image and compute its context vector z_i^a using encoding process.

$$z^* = \max(z, z_1^a, \cdots, z_k^a)$$

(b) Intra-Person Attribute Context

Horizontal strip regions:

$$S = (s_1, \ldots, s_m)$$

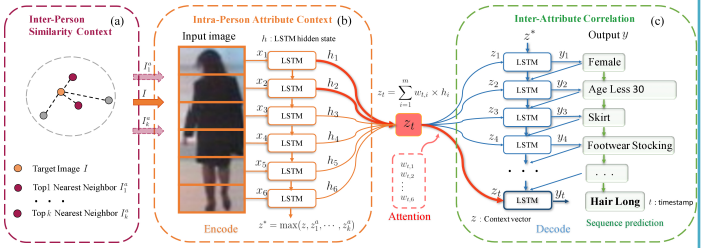
Encoder LSTM

Context representation:

$$H^{\text{en}} = (h_1^{\text{en}}, \ldots, h_1^{\text{en}}, \ldots, h_m^{\text{en}})$$

Encoder LSTM:

$$f_t = \text{sigmoid}(W_{fx}x_t + W_{fh}h_{t-1} + b_f)$$
$$i_t = \text{sigmoid}(W_{ix}x_t + W_{ih}h_{t-1} + b_i)$$
$$o_t = \text{sigmoid}(W_{ox}x_t + W_{oh}h_{t-1} + b_o)$$
$$g_t = \tanh(W_{gx}x_t + W_{gh}h_{t-1} + b_g)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$
$$h_t = o_t \odot \tanh(c_t)$$



(c) Inter-Attribute Correlation

$$h_t^{\text{de}} = \phi_\theta(h_{t-1}^{\text{de}}, y_{t-1}, z)$$

Decoder LSTM:

$$f_t = \text{sigmoid}(W_{fz}z + W_{fh}h_{t-1}^{\text{de}} + W_{fy}y_{t-1} + b_f)$$
$$i_t = \text{sigmoid}(W_{iz}z + W_{ih}h_{t-1}^{\text{de}} + W_{iy}y_{t-1} + b_i)$$
$$o_t = \text{sigmoid}(W_{oz}z + W_{oh}h_{t-1}^{\text{de}} + W_{oy}y_{t-1} + b_o)$$
$$g_t = \tanh(W_{gz}z + W_{gh}h_{t-1}^{\text{de}} + W_{gy}y_{t-1} + b_g)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$
$$h_t = o_t \odot \tanh(c_t)$$

Recurrent Attribute Attention

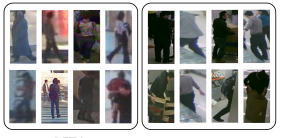
Step-wise context representation: $z_t = \sum_{i=1}^{m} w_{t,i} \times h_i^{\text{en}}$

$$w_{t,i} = \frac{\exp(\alpha_{t,i})}{\sum_{i=1}^{m}\exp(\alpha_{t,i})}, \text{with } \alpha_{t,i} = \phi_{\text{att}}(h_{t-1}^{\text{de}}, h_i^{\text{en}})$$

Attributes probability

$$p(\{y_{t,i}=1\}_{i=1}^{n_{\text{attr}}+1}) = \phi_y(h_{t-1}^{\text{de}}, y_{t-1}, z)$$
$$= W_y o_t + b_y$$

3 Evaluations



PETA RAP

Method \ Metric	mAP$^{\text{cls}}$	mPrc$^{\text{ins}}$	mRcl$^{\text{ins}}$	F1$^{\text{ins}}$
MRFr2[1]	75.60	-	-	-
ELF+SVM [2]	75.21	49.45	74.24	59.36
CNN+SVM [4]	76.65	51.33	75.14	61.00
ACN [7]	81.15	84.06	81.26	82.64
DeepSAR [3]	81.30	-	-	-
DeepMAR [3]	82.60	83.68	83.14	83.41
CTX CNN-RNN [5]	80.13	79.68	80.24	79.68
SR CNN-RNN [6]	82.83	82.54	82.76	82.65
JRL	85.67	86.03	85.34	85.42

Comparing with state of art methods on PETA

Dataset	TDS (%)	Model \ Metric	mAP$^{\text{cls}}$	mPrc$^{\text{ins}}$	mRcl$^{\text{ins}}$	F1$^{\text{ins}}$
PETA [1]	100	DeepMAR [3]	82.60	83.68	83.14	83.41
		SR CNN-RNN[6]	82.83	82.54	82.76	82.65
		JRL	**85.67**	**86.03**	**85.34**	**85.42**
	75	DeepMAR [3]	80.83	81.02	81.73	81.37
		SR CNN-RNN[6]	81.06	81.11	81.66	81.21
		JRL	**84.45**	**84.86**	**84.23**	**84.07**
	50	DeepMAR [3]	79.16	80.66	80.39	80.52
		SR CNN-RNN[6]	79.09	80.40	80.13	80.06
		JRL	**83.42**	**84.16**	**82.39**	**82.46**
	25	DeepMAR [3]	76.37	79.12	77.93	78.52
		SR CNN-RNN[6]	76.59	79.23	78.12	78.39
		JRL	**82.03**	**83.16**	**81.01**	**81.51**

Model robustness vs. training data size

Dataset	Method \ Metric	mAP$^{\text{cls}}$	mPrc$^{\text{ins}}$	mRcl$^{\text{ins}}$	F1$^{\text{ins}}$
PETA [1]	JRL (No AC)	83.45	83.96	83.97	83.89
	JRL	**85.67**	**86.03**	**85.34**	**85.42**
RAP [4]	JRL (No AC)	75.19	75.55	76.93	75.97
	JRL	**77.81**	**78.11**	**78.98**	**78.58**

Benefit of intra-person Attribute Context

Qualitative analysis of latent attribute correlation, with wrong predictions in red, true in green and missed predictions in blue.

Dataset	Method \ Metric	mAP$^{\text{cls}}$	mPrc$^{\text{ins}}$	mRcl$^{\text{ins}}$	F1$^{\text{ins}}$
PETA [1]	JRL(No Attention)	84.03	84.92	84.19	84.24
	JRL	**85.67**	**86.03**	**85.34**	**85.42**
RAP [4]	JRL(No Attention)	75.96	76.89	77.49	77.13
	JRL	**77.81**	**78.11**	**78.98**	**78.58**

Effects of recurrent attribute attention

4 Conclusion

- Joint Recurrent Learning (JRL) model for exploring attribute context and correlation
- Joint learning high-order (sequential) inter-attribute correlation
- More robust than state-of-the-art deep models

More information:
http://www.eecs.qmul.ac.uk/~jw306/
http://vision.eecs.qmul.ac.uk/

[1] Y. Deng et al. Learning to recognize pedestrian attribute. arXiv, 2015
[2] R. Layne et al, and S. Gong. Attributes-based re-identification. In Person Re-Identification, pages 93–117. Springer, 2014
[3] D. Li et al. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. Proc. ACPR, 2015
[4] D. Li et al. A richly annotated dataset for pedestrian attribute recognition. arXiv, 2016
[5] Y. Li et al. Sequential person recognition in photo albums with a recurrent network. In CVPR, 2017
[6] F. Liu et al. Semantic regularisation for recurrent image annotation. In CVPR, 2017
[7] P. Sudowe et al. Person attribute recognition with a jointly-trained holistic cnn model. CVPR Workshops, 2015