Unsupervised Learning of Generative Topic Saliency for Person Re-identification

Hanxiao Wang
hanxiao.wang@qmul.ac.uk
Shaogang Gong
s.gong@qmul.ac.uk
Tao Xiang
t.xiang@qmul.ac.uk

School of Electronic Engineering and Computer Science,
Queen Mary, University of London,
London E1 4NS, UK

Existing approaches to person re-identification (re-id) are dominated by supervised learning based methods, which requires a large number of manually labelled pairs of person images across every pair of camera views. This thus limits their ability to scale to large camera networks. To overcome this problem, a novel unsupervised re-id model, Generative Topic Saliency (GTS), is proposed in this paper for localised human appearance saliency selection in re-id by exploiting unsupervised generative topic modelling. It yields state-of-the-art re-id performance against existing unsupervised learning based re-id methods. For supervised methods, it also retains comparable re-id accuracy but without any need for pairwise labelled training data.

We are motivated by a very intuitive principle – humans often identify people by their salient appearances and ignore the more common traits in people's appearance. Compared to the pioneering work of [2] which is also based on learning appearance saliency for re-id, our model has two advantages: (1) Interpretability - our work explicitly models human appearances and backgrounds through learning a set of latent topics corresponding to localised human appearance components and also image backgrounds, so that the background cannot be mistaken as distractions to true foreground local salient region discovery. In addition, through associating saliency with *atypical* human appearances, the learned saliency is also more interpretable by human sense. (2) Complexity - only a *single* model is needed for computing saliency for all the images in a camera view, instead of learning a different discriminative saliency model (k-NN or one-class SVM) for every patches of every image.

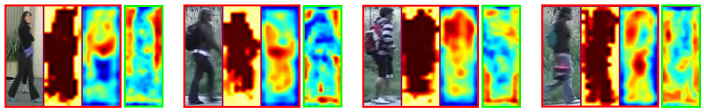

Figure 1: Saliency maps comparison (left to right): A person image in detected bounding box, GTS-computed background map, GTS-computed saliency map, saliency map computed by the model of [2] (green bounding box).

Our model is a generalisation of the Latent Dirichlet Allocation (LDA) model [1] with an added spatial variable to make the learned topics spatially coherent. Given a dataset of M images, each image will be factorised (clustered) into a unique combination of K shared topics, with each topic generating its own proportion of words on that image. Conceptually, one topic encodes a certain distribution of visual words (patches), whose vocabulary and spatial location revealing certain patterns, in our case the visual characteristics of human appearances and backgrounds. We thus learn two types of latent topics in our model corresponding to foreground and background respectively. Since foreground appearance are in general more 'compact' than background, we choose a Gaussian distribution to encode foreground human appearance topics and a Uniform distribution to encode more spread-out background topics.

A key objective of our model is to discover salient local foreground patches in a person's image that make the person stand out from other people, i.e. the model seeks not only visually distinctive but also *atypical* localised appearance characteristics of a person. In specific, we define a patch P_A's saliency according to three factors: The first one is how *unlikely* this patch will appear in a training set $\mathcal{I}^R$ of J images at the proximity of a particular spatial location in the images (i.e. its prevalence level). The less likely P_A repeatedly appears, the higher saliency score it should possess. Second, a patch with high probability of belonging to background topics should have low saliency scores. Third, even if a patch belongs to a human appearance topic, but if this topic is very dominant/popular in the training dataset (e.g. many people wearing jeans), the patch also should have low saliency score. With $Prevalence(P_A)$ measuring the prevalence level of P_A, Z_A denoting P_A's topic, T^{cb} the set of camera background topics, T^{pop} the set of *popular* human appearance topics, L and H the learned latent variables set and hyper-parameter set, patch P_A's saliency score is computed by:

$$\begin{aligned} Saliency(P_A) = h(Prevalence(P_A)) - \eta_1 \cdot \sum_{t_k \in T^{cb}} Pr(z_A = t_k | L, H) \\ - \eta_2 \cdot \sum_{t_k \in T^{pop}} Pr(z_A = t_k | L, H), \qquad 0 < \eta_1, \eta_2 < 1 \end{aligned} \tag{1}$$

where $h(x)$ is a inverse function defined as taking the additive inverse and normalising the result into the $[0,1]$ interval. The prevalence of P_A and the probability for P_A's topic Z_A falling into background topics and dominant/popular human appearance topics can all be computed from our model parameters inferred from training set. η_1, η_2 are the latter two factors' weights to affect the saliency score, determined by cross-validation. If one considers that $Prevalence(P_A)$ simply measures how likely the exact same patch appears repeatedly across images, its topic's *popularity* (the third component) takes much larger amounts of patches into consideration. These patches may even be visually different from P_A, but they are inherently related by the same topic. This model avoids the topic being simply data-driven; it also considers more inherent structure of the large-scaled data. The comparison between computed saliency are shown in Fig. 1.

Given the patch level saliency score, we adopt the same patch-based image matching scheme in [2]. In this patch-matching scheme, patches with higher saliency scores will contribute more to the distance between a pair of probe/gallery images. We conduct 10-trail experiments on both VIPeR and iLIDS dataset, compared with existing unsupervised learning methods, the GTS model improves re-id accuracy significantly, especially on Rank-1. The GTS model is also competitive against the state-of-the-art supervised learning based methods, but without requiring manual labelling of data, resulting in greater scalability to large scale re-id problems in many practical applications.

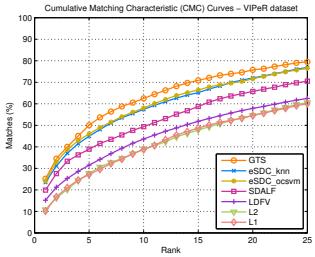

Figure 2: VIPeR test: CMC comparison of unsupervised learning based re-id models.

Method	r=1	r=5	r=10	r=20
ELF	12.00	31.50	44.00	61.00
PRDC	15.66	38.42	53.86	70.09
PCCA	19.27	48.89	64.91	80.28
LMNN-R	20.00	49.00	66.00	79.00
KISSME	19.46	48.10	62.50	78.32
RPLM	27.00	-	69.00	83.00
LF	24.18	-	67.12	-
GTS	25.15	50.03	62.50	75.76

Table 1: VIPeR test: Comparing the GTS model to supervised learning based models.

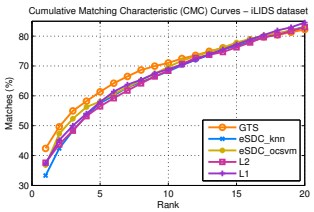

Figure 3: iLIDS test: CMC comparison of unsupervised learning based re-id models.

Method	r=1	r=5	r=10	r=20
SDC_knn	33.31	57.55	68.22	83.13
SDC_ocsvm	36.81	58.10	69.69	82.94
PRDC	37.83	63.70	75.09	88.35
LMNN	27.97	53.75	66.14	82.33
PLS	22.10	46.04	59.95	78.68
ITM	28.96	53.99	70.50	86.67
GTS	42.39	61.35	71.04	82.21

Table 2: iLIDS test: Comparing the GTS model against other unsupervised (top) and supervised (bottom) learning based models.

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, pages 993–1022, March 2003.

[2] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.