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

Abstract

Existing approaches to person re-identification (re-id) are dominated by supervised learning based methods which focus on learning optimal similarity distance metrics. However, supervised learning based models require 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, this paper proposes a novel unsupervised re-id modelling approach by exploring generative probabilistic topic modelling. Given abundant unlabelled data, our topic model learns to simultaneously both (1) discover localised person foreground appearance saliency (salient image patches) that are more informative for re-id matching, and (2) remove busy background clutters surrounding a person. Extensive experiments are carried out to demonstrate that the proposed model outperforms existing unsupervised learning re-id methods with significantly simplified model complexity. In the meantime, it still retains comparable re-id accuracy when compared to the state-of-the-art supervised re-id methods but without any need for pair-wise labelled training data.

1 Introduction

Person re-identification (re-id) is a challenging problem for computer vision [7]. Recent efforts on solving the re-id problem are dominated by supervised learning based methods that aim to learn an optimal matching function or distance metric [5, 10, 11, 22, 29, 31]. More specifically, for each pair of camera views, a *labelled* training set is constructed. It consists of a set of people for which images of each individual must be annotated manually with an identity label across both views. A matching function is learned from the training set subject to a set of constraints, that is, a pair of images of the same person should have larger matching score/smaller similarity distance compared to that of two different people given the labelling information, regardless their visual appearance dissimilarity/similarity. By satisfying these constraints the learned model can implicitly discover visual features that are more stable against intra-class appearance variations. These variations are typically caused by viewing condition changes across a particular pair of camera views. However, there is a significant limitation of these supervised learning based methods – a large set of people must be labelled manually across every pair of camera views. Moreover, even for the same pair of camera views, once the conditions change (e.g. different time of the day), new labelling may be needed again to update the matching function. Therefore, such approaches are inherently

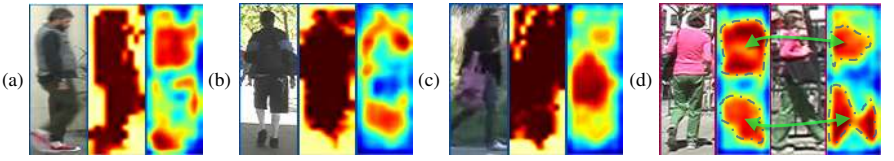

Figure 1: Each of (a)-(c) shows (left to right): person image, topic model detected background map and foreground saliency map. The saliency maps capture localised appearance features (e.g. brown jacket, red shoes, blue sleeve pattern, pink handbag, green bottom, pink shirt). (d) show that the distributions of the foreground saliency maps from two different camera views of the same person are stable and useful for re-id. Best viewed in colour.

limited in their scalability to different camera pairs at different times without the need for exhaustive and repeated manual labelling. This is impractical for large camera networks of hundreds of cameras.

An attractive method to overcome this limitation of supervised learning based re-id models is to explore unsupervised learning, which aims to make re-id models more scalable to new camera views even though it may sacrifice re-id matching accuracy. A key question is what can be learned in images for re-id without person's identity labels explicitly annotated to images across camera views. Among the few reported unsupervised learning based re-id methods, most are focused on learning view or illumination change invariant (stable) feature descriptors of human appearance [6, 17, 18]. More recently, an unsupervised learning model has also been proposed to discover localised visual appearance saliency for re-id [27]. This is based on very intuitive principles – humans often identify people by their salient appearances (local/small area) such as wearing a rare-coloured coat or a strange-shaped hat, and ignore the more common traits in people's appearance. However, this saliency model is exhaustively data-driven therefore computationally complex. This is due to the fact that the model is based on constructing a different saliency model for every local image patch in every image against a reference set whilst each image is decomposed into hundreds of patches. That is, if there are M images to be matched across two camera views and each image is decomposed to N patches, there are $M \times N$ different saliency models required to be constructed against the reference set. This data-driven approach to unsupervised saliency learning also makes it potentially unstable to large scale problems. For these problems, many images of people (from hundreds to thousands) need be matched across camera views and people's appearance necessarily exhibits greater variety.

In this paper, a novel unsupervised modelling approach to saliency detection for person re-id is proposed based on probabilistic generative topic modelling. This is significantly different from previous attempts, which are data-driven and discriminative. More specifically, given abundant unlabelled data, our model aims to learn simultaneously what people look like (background removal in a bounding box) and how their typical appearance can be represented by a collection of local and visually coherent parts. This is achieved by learning a set of latent topics that correspond to both typical and localised human appearance components, e.g. blue jeans and dark suit. This component-based typical appearance representation is then deployed for identifying *atypical* appearance by discovering local saliency. This generative topic model based representation is also inherently capable of differentiating background clutters from typical human appearance in a detected person bounding box (Figure 1), beneficial to person re-id in cluttered scenes [6, 15].

Our proposed Generative Topic Saliency (GTS) model is based on unsupervised topic modelling designed specifically to discover re-id relevant saliency that corresponds to atypi-

cal appearance of individual people (foreground). It also simultaneously removes surrounding background clutter in a person detection bounding box. It has two advantages over the existing saliency model for person re-id [27]: (1) Interpretability - each learned topic has clear semantic meaning. (2) Complexity - only a single model is needed for computing saliency for all the images in a camera view, in contrast to having to construct a different saliency model for every image patch of every image. Comparative evaluations on the VIPeR [9] and iLIDS [28] datasets demonstrate that the proposed GTS model not only outperforms existing unsupervised learning based saliency model, but also is competitive to the state-of-the-art supervised learning models without the need for expensive data labelling.

2 Related Work

Supervised learning – Most existing learning based re-id methods rely on supervised learning of discriminative distance metrics given pairwise labelled data from different cameras [5, 11, 12, 22, 31]. The scalability limitation of these supervised methods has motivated a number of transfer learning-based methods [1, 13, 16, 30] to utilise previously labelled data elsewhere to minimise the need for labelling images of every new camera pair. However, a small set of labelled images are still required at each new pair and model updating (re-learning) remains necessary when the lighting and view conditions change.

Unsupervised learning – Earlier unsupervised learning re-id methods are focused on feature design [6, 17, 18], rather than learning saliency. This is because without labelling, any saliency measure has to rely on general principles without knowledge of person-specific appearance characteristics. This is a much harder problem than supervised learning from labelled information on person-specific appearance. Recently, Liu *et al*. [15] proposed a feature importance mining scheme, aiming to optimise the weights for global feature types. Alternatively, Zhao *et al*. [27] proposed a patch-based representation to learn local saliency in a person's appearance. Their method relies on exhaustively learning a very large number of data-driven discriminative models (k-NN or one-class SVM) through constructing different saliency models for every patch of every image. In contrast, our approach learns a *single* generative model for computing saliency map for all the images in a camera view, significantly reducing model complexity. Moreover, our model segments simultaneously foreground and background, giving more accurate saliency detection compared to [27] as the latter is sensitive to false saliency detection caused by confusing background as salient foreground.

Topic modelling – Probabilistic topic models [2] have been used for image analysis which can be considered as a dimensionality reduction technique that represents image content in a low-dimensional latent topic space. Topic models have been employed to perform various tasks such as scene understanding, object classification and annotation [3, 14, 20, 25]. However, to the best of our knowledge this is the first time that probabilistic topic modelling is explored for unsupervised learning of human appearance saliency for re-id. Our model is related to the work of Shi *et al*. [24]. However, their method is a weakly-supervised model designed to localise different categories of objects in an image. In contrast, our model is fully unsupervised and designed to optimise the selection of human appearance saliency by learning localised topics for a component-based human appearance representation.

Contributions – Our contributions are: (1) A novel re-id model, Generative Topic Saliency (GTS), for localised human appearance saliency selection in re-id by exploiting unsupervised generative topic modelling. (2) The GTS model is capable of simultaneous foreground saliency detection and background clutter removal. (3) The GTS model yeilds state-of-the-art re-id performance against existing unsupervised learning based re-id methods.

3 Methodology

3.1 Image Representation

Similar to [27], we adopt an over-sampled local patch based representation for each person image. More precisely, each image is represented by 50% overlapped uniform-sized square patches on a dense grid. From each patch, a 32-bin color histogram is computed in the LAB color space with 3 levels down-sampled. SIFT features are also computed in the 3 color channels, with each patch divided into 4×4 cells and 8-bin orientations of local gradients. The final patch descriptor is computed by $L2$ normalisation and concatenation of the colour histogram and SIFT, giving a 672 dimensional feature vector ($32\times3\times3+8\times4\times4\times3$). Patch size and grid step length are 10 and 4 pixels respectively. Our overall image representation builds on the patch descriptors and differs from that of [27]. Specifically, a topic model treats each document (image) as a certain combination of visual words and requires a bag-of-words representation. Given the patch feature vectors from each image, we cluster all the patch feature vectors from an unlabelled training set into a $N_v = 2000$ words codebook by K-means clustering. Given this codebook, each patch is assigned with a word label by its cluster index. An image I_m is then represented by N_m words together with their image positions, denoted as $\{w_{nm}, l_{x_{nm}}, l_{y_{nm}}\}_{n=1}^{N_m}$, with w_{nm} the word label of a patch, $l_{x_{ij}}$ and $l_{y_{ij}}$ the image coordinates of that patch.

3.2 Joint Modelling Human Appearance and Camera Background

Given a set of M images of people in bounding boxes, typically extracted from a person detector, we wish to learn a joint topic model capable of capturing the typical appearance of people in foreground patches and simultaneously separating the background patches within each bounding box, without any labelling information. The topic model essentially factorises the image patches and attempts to find localised coherent patches (not necessarily connected) that correspond to common appearance traits of people such as grey top and blue jeans, without any supervised learning. However, the bounding boxes inevitably contain backgrounds, which are often also spatially and visually coherent. To differentiate them, background patches are also modelled explicitly by the generative topic modelling. 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, similar to [24] we choose a Gaussian distribution to encode foreground human appearance topics and a Uniform distribution to encode more spread-out background topics.

Model Description – Our model is a generalisation of the Latent Dirichlet Allocation (LDA) model [2] 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. Among these K topics, K^{ha} topics are used to model foreground human appearance, and $K^{cb} = K - K^{ha}$ topics represent background within the bounding boxes from the entire training dataset. In this work we set $K^{cb} = K^{ha} = 20$. Suppose Dir, $Multi$, $\mathcal{NW}$, $\mathcal{N}$ denote respectively Dirichlet, Multinomial, Normal-Wishart and Normal distributions, the generative process of this model is:

1. For each topic $t_k \in \{t_1, t_2, \ldots, t_K\}$, draw its appearance distribution $\beta_k \sim Dir(\beta_k^0)$.
2. For each image $I_m \in \{I_1, I_2, \ldots, I_M\}$, draw the human appearance and camera background topic distribution $\theta_m \sim Dir(\alpha)$. Each human appearance topic $t_k \in T^{ha}$ is as-

signed with a Gaussian distribution parameters to reflect the spatial location and size of the human appearance on I_m: $\{\mu_{km}, \sigma_{kj}\} \sim \mathcal{NW}(\mu_0^k, \lambda_0^k, W_0^k, v_0^k)$.

3. For each patch $P_{nm} \in \{P_{1m}, P_{2m}, \ldots, P_{N_m m}\}$, draw its topic $z_{nm} \sim Multi(\theta_m)$, draw its vocabulary $w_{nm} \sim Multi(\beta_{z_{nm}})$ and draw its location $\mathbf{l}_{nm}$. If z_{nm} is a human appearance topic then its location is Gaussian distributed, $\mathbf{l}_{nm} \sim \mathcal{N}(\mu_{z_{nm}m}, \sigma^{-1}_{z_{nm}m})$; if z_{nm} is a camera background topic then its location is Uniformly distributed, $\mathbf{l}_{nm} \sim Uniform$.

Model Learning – The learning task for this model is to infer the following quantities: (1) The vocabulary distribution of each human appearance and background topics β_k, (2) all topics' word proportion θ_m and their locations $\{\mu_{mk}, \sigma_{mk}\}$ in each image, and (3) each patch's topic assignment z_{nm}. The joint distribution of observed data set O, latent variables set L and hyper-parameters set H is given by:

$$Pr(O,L|H) = \prod_m^M \prod_k^K \Bigg[Pr(\mu_{mk}, \sigma_{mk}|\mu_0^k, \lambda_0^k, W_0^k, v_0^k) Pr(\theta_m|\alpha_m) \left(\prod_n^{N_m} Pr(w_{nm}|z_{nm}, \theta_m) Pr(z_{nm}|\theta_m) \right) \Bigg] Pr(\beta_k|\beta_k^0) \tag{1}$$

This model is intractable by exact solutions. An approximate solution can be learned by the EM algorithm with a variational inference strategy, through introducing a Dirichlet parameter γ and a multinomial parameter φ as variational parameters. Under this variational inference framework, γ is learned for each image, with γ_{mk} modelling the proportion of patches which belong to topic t_k in image I_m. φ is learned for each patch, with φ_{nmk} modelling the probability of patch P_{nm} on image I_m being generated by topic t_k. The hyper-parameter α is set to 1 for all human appearance and camera background topics because our method is completely unsupervised and thus all topics may appear in any images.

3.3 Determining Patch Prevalence

A key objective of our model is to discover 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. To compute such a saliency value, let us first consider to compute a 'prevalence' value of each patch and define saliency as the inverse of prevalence, as the former is more naturally computable by the topic model. Specifically, for a patch P_A on image I_A, its saliency value is measured by 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. P_A's saliency score is the inverse of its prevalence value in $\mathcal{I}^R$. For computing patch prevalence value, suppose the learned latent variables set is L and their hyper-parameter set is H. The topic appearance vector β_{vk} reflects the probability that vocabulary (the collection of words in the codebook) v is generated under topic t_k. The multinomial parameter φ_{nmk} refers to the probability that patch P_{nm}'s topic is t_k given the learned model parameters:

$$\beta_{kv} = Pr(w = v|t_k, L, H),\ v = 1,2,\ldots,N_v; \quad \varphi_{nmk} = Pr(z_{nm} = t_k|L,H),\ k = 1,2,\ldots,K \tag{2}$$

Based on the Bayes' Theorem, combining the two equations in Eqn. (2) gives the joint likelihood of observed word w_{nm} and its topic z_{nm} as:

$$Pr(w_{nm} = v, z_{nm} = t_k|L,H) = Pr(w = v|t_k, L, H) Pr(z_{nm} = t_k|L,H) \tag{3}$$

By margining out the topic variable z_{nm} over t_1 to t_K, we obtain the likelihood of patch P_{nm}'s vocabulary w_{nm}. This likelihood value reflects our model's confidence for the visual word w_{nm} to be vocabulary v: $(v = 1, 2, \ldots, N_v)$:

$$\mathcal{L}(w_{nm}) = Pr(w_{nm} = v|L,H) = \sum_{z_{nm}=t_1}^{t_K} Pr(w_{nm} = v, z_{nm} = t_k|L,H) \tag{4}$$

To measure the probability of patch P_A appearing in image I_m, we impose a simple but reasonable human prior knowledge on people's images, that is, a person's position within a bounding box is relatively stable, and a patch's horizontal shift caused by viewpoint change is far larger than its vertical shift. This assumption is typically valid for a pedestrian captured in a bounding box. Based on this assumption, in each image I_m in $\mathcal{I}^R$ we build a patch set $\hat{P}_m^A$ by taking all the patches in the same horizontal row as P_A. The elements in $\hat{P}_m^A$ are referred as $P_{m,r}^A$, with r as the row index. Given P_A's vocabulary $w_{P_A} = v_0$, the probability that patch P_A repeatedly appears in image I_m of $\mathcal{I}^R$ is measured by the maximum probability for $\hat{P}_m^A$ patches' vocabulary equalling to v_0:

$$\mathcal{P}(P_A \text{ in } I_m) = \max\left(Pr\left(w_{P_{m,r}^A} = v_0|L,H\right)\right), \quad P_{m,r}^A \in \hat{P}_m^A \tag{5}$$

Patch P_A's prevalence level is computed by accumulating $\mathcal{P}(P_A \text{ in } I_m)$ for all the images I_m in $\mathcal{I}^R$:

$$Prevalence(P_A) = \sum_{I_m} \mathcal{P}(P_A \text{ in } I_m), \quad I_m \in \mathcal{I}^R \tag{6}$$

3.4 Computing Saliency

Given the prevalence value of each patch (Eqn. (6)), its saliency score is initialised by applying an inverse function $h(x)$ on its prevalence value. These saliency scores are then further refined by two basic principles as follows. First, a patch with high probability of belonging to background topics should have low saliency scores. Second, 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.

The learned Dirichlet parameter γ_{mk} reveals the proportion of patches on I_m belonging to topic t_k, which can be treated as a pseudo count for the amount of patches falling into each topic on I_m. We then model the popularity of topic t_k by accumulating γ_{mk} over all images in the probe set $\mathcal{I}^p$ and gallery set $\mathcal{I}^g$:

$$Popularity(t_k) = \sum_{I_m} \gamma_{mk}, \quad I_m \in \{\mathcal{I}^p, \mathcal{I}^g\}, t_k \in T^{ha} \tag{7}$$

The M foreground topics with highest *Popularity* values is treated as popular human appearance topics, and deployed to form a topic set T^{pop}. In practice, we take $M = K^{ha}/2$, i.e. 50% of all human appearance topics with higher popularity scores are considered to be statistically common/typical. The final saliency score of patch P_A is computed by combining its prevalence level, the probability of its topic *not* belonging to a background topic, and being less popular (atypical) among foreground appearance topics, i.e.

$$\begin{aligned} Saliency(P_A) = h(Prevalence(P_A)) &- \eta_1 \cdot \sum_{t_k \in T^{cb}} Pr(z_A = t_k|L,H) \\ &- \eta_2 \cdot \sum_{t_k \in T^{pop}} Pr(z_A = t_k|L,H), \qquad 0 < \eta_1, \eta_2 < 1 \end{aligned} \tag{8}$$

where $h(x)$ is a inverse function defined as taking the additive inverse and normalising the result into the $[0,1]$ interval. $Prevalence(P_A)$ is given by Eqn. (6). The last two terms can be calculated through Eqn. (2), whilst η_1, η_2 are their weights to affect the saliency score, determined by cross-validation during our experiment. If one considers that $Prevalence(P_A)$ simply measures how likely the exact same patch appears repeatedly across images, its topic's *popularity* 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. It is worth pointing out that the model of [27] selects two independent reference training datasets (one for the gallery camera view and another for the probe camera view) and trains many patch-specific and view-specific discriminative models: a different model for every patch of every probe image and every gallery image in order to match the probe image against a set of gallery images for re-id. In contrast, our method only requires to train a *single* model for each camera view given an independent training dataset from that view. Then only two models are required for all patches of all the probe images and all the gallery images respectively. Some examples of the saliency maps obtained using our method are shown in Figure 2.

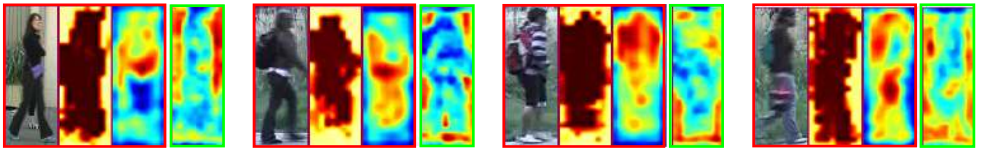

Figure 2: Saliency maps comparison (left to right): A person image in detected bounding box, GTS-computed background map, GTS-computed saliency map, saliency map computed by the model of [27] (green bounding box).

3.5 Re-id Matching

Given the saliency score, we adopt the same patch-based image matching scheme of [27] to compute a matching score between a set of gallery images and a probe image from an independent test set. First we build a corresponding pairwise relationship for all the patches in a probe image I_A and a gallery image I_B. In each patch pair (image location indexed), one patch P_1 is from I_A and the other P_2 from I_B. More precisely, a pair of (P_1, P_2) patch is the nearest neighbour match searched in the proximity of P_1 in I_B or vice versa P_2 in I_A. The matching similarity distance metric is given by $s = exp(-d^2/2a^2)$, where d is the Euclidean distance between two patch feature vectors and a is the bandwidth of a Gaussian function. The overall similarity between the two images is computed by a weighted sum accumulating all the patch pairs' similarities weighted by the saliency scores of patches in each pair, i.e. an accumulation over the quantity $Saliency(P_1) \cdot s(P_1,P_2) \cdot Saliency(P_2)$, where P_1 and P_2 are two patches in one pair. It is worth pointing out that the published code of [27] utilizes foreground masks to remove background patches in VIPeR images. The similarity score between a pair of images is only computed in the foreground region. A similar process of background removal is adopted by many existing works [6, 15, 27]. Body parts information are not explored in our experiments.

4 Experiments

Datasets and Settings – We evaluate our method on two widely used benchmark datasets, VIPeR [9] and iLIDS [28]. The VIPeR dataset contains 632 pedestrian image pairs. Each pair of images contain the same individual, but were taken from different camera views. Following the experimental setting of [6, 8], we randomly choose half of the dataset, i.e. 316 image pairs, as our training sets. On this training set, we train two topic models, one for each camera view. Among the 316 pairs of training images, we choose 100 pairs as our reference sets for computing saliency and use one reference set per camera view, same as [27]. The iLIDS dataset contains 476 images of 119 people. We followed the same single shot experiment protocol as [29], i.e. randomly choose all images of $p = 50$ people as test set, and use the other images as training set. In the test set, one image per person is chosen to form a gallery set, while all the remaining images compose a probe set. We run our experiments for 10 trials with different splits, and report the average of these 10 trials as our final result. The performance is evaluated using the Cumulated Matching Characteristics (CMC) curves.

Comparison to non-learning and unsupervised learning models – We first compare our GTS model against non-learning based methods, i.e. template matching with a distance measure. L1-norm and L2-norm distances are used as the baseline models for comparison. Figures 3 and 4 show respectively the results on VIPeR and iLIDS. It is evident that our method significantly outperforms the baseline non-learning methods, e.g. Rank-1 about 150% (VIPeR) and 14% (iLIDS) relative improvement over L1-norm. This suggests that the unlabelled data indeed helps improve re-id matching accuracy.

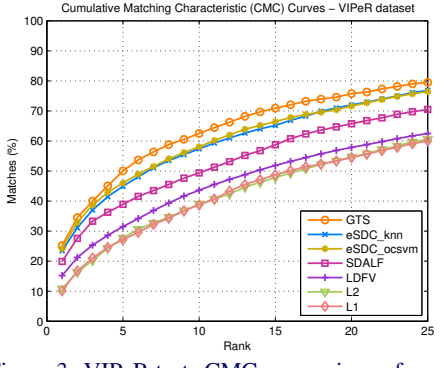

Figure 3: VIPeR test: CMC comparison of unsupervised learning based re-id models.

Method	r=1	r=5	r=10	r=20
ELF [8]	12.00	31.50	44.00	61.00
PRDC [31]	15.66	38.42	53.86	70.09
PCCA [19]	19.27	48.89	64.91	80.28
LMNN-R [5]	20.00	49.00	66.00	79.00
KISSME [12]	19.46	48.10	62.50	78.32
RPLM [11]	27.00	-	69.00	83.00
LF [21]	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.

Next we compare GTS to a number of contemporary unsupervised learning methods including eSDC_knn [27], eSDC_ocsvm [27] [1], LDFV [18] and SDALF [6]. Figures 3 and 4 show that our model is clearly superior to LDFV and SDALF, e.g. Rank-1 27% (VIPeR) relative improvement over SDALF. These results show that modelling human saliency gives the GTS model an advantage over the feature-design based unsupervised learning approaches. Comparing with eSDC_knn and eSDC_ocsvm, which are also patch based unsupervised saliency learning methods, the GTS model still shows a notable improvement, e.g. Rank-1 5% (VIPeR) and 15% (iLIDS) relative improvement over eSDC_ocsvm. Figure 2 sheds

[1] The results of KNN and OCSVM in our experiments are obtained by running the author published code under our experiment settings. The results are thus slightly different from those reported in [27].

some light into why the GTS model outperforms these two models in [27]. It is evident that a better saliency map is obtained using the GTS model. This is mainly because our topic model explicitly models human appearance as well as background so that the background cannot be mistaken as distractions to true foreground local salient region discovery. In contrast, the model of [27] can give false high saliency scores due to confusion with background regions, while the saliency scores for those real salient regions on those image are pulled down due to the interference of backgrounds, thus cannot be utilised in the re-id process. Computationally, the GTS model is also twice as fast to compute when compared to [27].

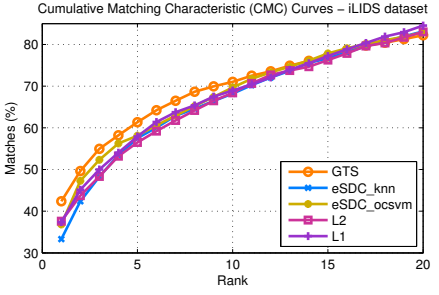

Figure 4: iLIDS test: CMC comparison of unsupervised learning based re-id models.

Method	r=1	r=5	r=10	r=20
SDC_knn [27]	33.31	57.55	68.22	83.13
SDC_ocsvm [27]	36.81	58.10	69.69	82.94
PRDC [31]	37.83	63.70	75.09	88.35
LMNN [26]	27.97	53.75	66.14	82.33
PLS [23]	22.10	46.04	59.95	78.68
ITM [4]	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.

Comparison to supervised learning models – We also compared our GTS model against some recently proposed supervised learning based re-id models. In general, supervised learning of discriminative models are expected to provide better re-id performance due to the use of labelled information for learning strong discriminative functions, with a high price for labelling the data. Tables 1 and 2 show results on VIPeR and iLIDS respectively. It is clear that without using any labelled data for model training, the GTS model is competitive against these supervised learning methods *without* the benefit from learning strong discriminative functions using labelled data. Moreover, the GTS model is able to outperform a number of the supervised learning models by some notable margins, e.g. Rank-1 20% (VIPeR) and 13% (iLIDS) relative improvement over PRDC, LMNN and KISSME (Tables 1 and 2). This suggests that the GTS model is scalable to large scale applications when manual annotations of identity labels across camera views are not available or feasible.

5 Conclusion

We proposed a novel unsupervised generative saliency learning framework for person re-identification. The core of this framework is a probabilistic topic model specifically designed for modelling jointly typical human appearance and the surrounding background appearance. The model can be deployed to simultaneously learn a saliency map and foreground segmentation for a more accurate and scalable person re-identification model. 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.

References

[1] Tamar Avraham, Ilya Gurvich, Michael Lindenbaum, and Shaul Markovitch. Learning implicit transfer for person re-identification. In *ECCV workshop on Person Re-id*, 2012.

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

[3] Liangliang Cao and Fei-Fei Li. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *ICCV*, 2007.

[4] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. Information-theoretic metric learning. In *ICML*, 2007.

[5] Mert Dikmen, Emre Akbas, Thomas S. Huang, and Narendra Ahuja. Pedestrian recognition with a learned metric. In *ACCV*, 2010.

[6] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[7] Shaogang Gong, Marco Cristani, Change Loy Chen, and Timothy M. Hospedales. The re-identification challenge. In *Person Re-Identification*. Springer, 2014.

[8] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[9] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition and tracking. In *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, 2007.

[10] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011.

[11] Martin Hirzer, Peter M. Roth, Martin Koestinger, and Horst Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, 2012.

[12] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[13] Ryan Layne, Timothy M. Hospedales, and Shaogang Gong. Domain transfer for person re-identification. In *ACM Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Stream*, 2013.

[14] Li-Jia Li, Richard Socher, and Fei-Fei Li. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, 2009.

[15] Chunxiao Liu, Shaogang Gong, and Chenge Loy Chen. On-the-fly feature importance mining for person re-identification. *Pattern Recognition*, (4):1602.

[16] Andy J. Ma, Pong C. Yuen, and Jiawei Li. Domain transfer support vector ranking for person re-identification without target camera label information. In *ICCV*, 2013.

[17] Bingpeng Ma, Yu Su, and Frédéric Jurie. Bicov: a novel image representation for person re-identification and face verification. In *BMVC*, 2012.

[18] Bingpeng Ma, Yu Su, and Frédéric Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *ECCV workshop*, 2012.

[19] Alexis Mignon and Frédéric Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012.

[20] Zhenxing Niu, Gang Hua, Xinbo Gao, and Qi Tian. Context aware topic model for scene recognition. In *CVPR*, 2012.

[21] Sateesh Pedagadi, James Orwell, Sergio A. Velastin, and Boghos A. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013.

[22] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by support vector ranking. In *BMVC*, 2010.

[23] William Robson Schwartz and Larry S. Davis. Learning discriminative appearance-based models using partial least squares. In *SIBGRAPI*, 2009.

[24] Zhiyuan Shi, Timothy M. Hospedales, and Tao Xiang. Bayesian joint topic modelling for weakly supervised object localisation. In *ICCV*, 2013.

[25] Daniel M. Steinberg, Oscar Pizarro, and Stefan B. Williams. Synergistic clustering of image and segment descriptors for unsupervised scene understanding. In *ICCV*, 2013.

[26] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

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

[28] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *BMVC*, 2009.

[29] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011.

[30] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Transfer re-identification: From person to set-based verification. In *CVPR*, 2012.

[31] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Re-identification by relative distance comparison. *TPAMI*, 2013.