Multi-shot Person Re-identification using Part Appearance Mixture

Furqan M. Khan and François Brèmond
INRIA Sophia Antipolis-Mediterranee
{furqan.khan, francois.bremond}@inria.fr

Abstract

Appearance based person re-identification in real-world video surveillance systems is a challenging problem for many reasons, including ineptness of existing low level features under significant viewpoint, illumination, or camera characteristic changes, to robustly describe a person's appearance. One approach to handle appearance variability is to learn similarity metrics or ranking functions to implicitly model appearance transformation between cameras for each camera pair, or group, in the system. The alternative, that this paper follows, is the more fundamental approach of improving appearance descriptors, called signatures*, to cater for high appearance variance and occlusions. The novel signature representation for* multi-shot *person re-identification presented in this paper uses multiple appearance models, each describing appearance as a probability distribution over some low-level feature for a certain portion of individual's body. Combined with metric learning, it achieves rank-1 recognition rates of* 92% *and* 79% *on PRID2011 [12] and iLIDS-VID [33] datasets, respectively.*

1. Introduction

The goal of person re-identification (Re-ID) is to identify a person at distinct times, locations, or in different camera views. The problem often arises in the context of search of individuals or long term tracking in a multi-camera visual surveillance system with non-overlapping views. In a real-world system, Re-ID of a person is very challenging because of significant alteration in an individual's appearance due to changes in camera properties, illumination, viewpoint and pose. On the contrary, inter-person appearance similarity is generally high in absence of biometric (facial or iris) cues, due to low resolution imaging or viewpoint (Fig 1). Occlusions may impede visibility, and because a Re-ID system is driven by automatically acquired person tracks in practice, the individual may be only partially visible or improperly localized. These are significant challenges for appearance based Re-ID algorithms.

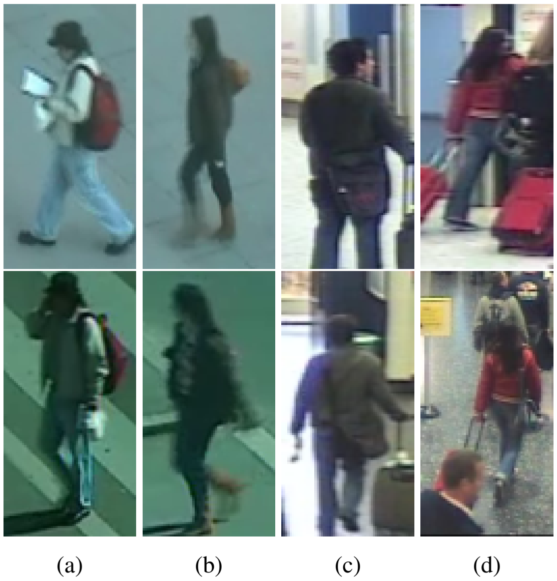

Figure 1: Associated challenges with Re-ID: occlusion, absence of biometric cues, variance in viewpoint, illumination, pose, and camera properties.

Re-ID is often treated as a retrieval problem; *i.e.*, given one or more images of an unknown person (*probe*) and a dataset (*gallery*) that consists of (images or image sets of) a number of unknown persons, the goal is to rank the persons in the gallery based on their similarity from the probe. Hence, the Re-ID process is often divided into two stages: i) representing each person using his appearance **signature** acquired from image(s), and ii) sorting candidate matches using a matching function (similarity metric or a ranking function) of appearance signatures. For good performance, signatures should be invariant under scene (viewpoint, illumination, camera, *etc.*) variance, and discriminative given high inter-person similarity. Significant amount of literature is available related to description of appearance for *single-shot* scenario, *i.e.* when only one image is available to learn appearance signature. However, the work focusing on signature representation for *multi-shot* scenario, *i.e.* when multiple images are available to learn each signature, is limited despite that multi-shot scenario is more relevant to video surveillance systems due to availability of multiple images of a person that are grouped using an object tracking algorithm. This paper concentrates on multi-shot Re-ID.

Availability of multiple images per person in multi-shot case can be helpful in learning robust appearance signatures; however, the set of images belonging to a person may have variable illumination, occlusion, person's orientation, pose and alignment. Thus aggregating information from multiple images require careful consideration. For instance, the trivial solution to use the mean of appearance descriptors obtained from different images of a person as multi-shot signature is adversely affected by large variance in a person's appearance. Therefore, the alternative is to represent a multi-shot signature as a set of image-wise (spatial) descriptors. Thus the space complexity becomes linear in number of images instead of number of identities. More importantly, this significantly increases the time complexity of matching, as computation complexity of employed set metrics grows quadratically in terms of average cardinality of the sets. To reduce associated high space and compute cost, most Re-ID methods use a small random subset instead of all images of a person. Hence a trade-off is required between signature robustness and computational/storage cost. This paper proposes a novel signature representation for multi-shot scenario to automatically trade-off signature robustness with computational/storage cost. The proposed approach explicitly deals with variability in illumination, person's orientation and pose, and implicitly deals with occlusion and alignment problems.

Specifically, appearance of a person is modeled as a multi-channel appearance mixture, where each channel corresponds to a particular region (part) of the body - *full*, *upper*, or *lower*. We call the representation Part Appearance Mixture (*PAM*). Appearance of each part is defined as a *multi-modal parametric probability distribution* of low-level features. Coarse parts localization and dense feature grids are used to enable computation sharing between different channels for computational efficiency. Independently for each person and part, model selection is used to find a compact appearance model by trading off signature variance. Since part models are probability distributions, f-Divergence based distance is used to define similarity between two signatures. Furthermore, we define a learn-able metric to compute similarity between two signatures. For the proposed metric, KISSME [15] algorithm is adopted to learn feature transformations between different scenes by directly learning transformation between probability distributions. The increase in computational cost of signature computation is compensated by the decrease in time complexity of signature matching and metric learning. More importantly, despite decreased storage complexity, signature robustness is increased that leads to significantly better performance than current state-of-the-art.

To summarize, the main contributions of this paper is a novel signature representation for multi-shot Re-ID to cater for high variance in a person's appearance by automatically trading compactness with variability. A signature is acquired over coarse body regions of a person in a computationally efficient manner instead of reliance on fine body part localization. The representation has probabilistic interpretation of appearance signatures that allows for application of information theoretic similarity measures. We also define a Mahalanobis based distance measure to compute similarity between two signatures. The metric is also amenable to existing metric learning methods and appearance transformation between different scenes can be learned directly using proposed signature representation.

2. Related work

Earlier literature can be broadly divided along two main aspects of person Re-ID: signature modeling and matching function learning. Matching function learning approaches [8, 9, 13, 15, 17, 19, 21, 24, 27, 29, 30, 33, 37, 40] focus on improving Re-ID performance regardless of the underlying signature representation used to model appearance of individuals. Given training data, their objective is to learn a model that minimizes intra-class variance and maximizes inter-class variance of signatures. Most of the approaches use supervised training of models, hence require significant annotation effort which may not be attractive for real-world applications, however, significant improvements are possible for underlying representation. On the other hand, signature modeling approaches focus on creating invariant and discriminative representations for individual's appearance, which are robust to viewpoint, illumination, and camera changes, as well as, occlusions and localization errors.

Single-shot approaches such as [2, 11, 16, 20, 24, 25, 28, 31, 35, 39, 40] construct signature using only one image. These approaches use a mixture of color, texture, edge or "deep" features to identify persons given only one image. Hence they do not explicitly use additional information available in multi-shot case.

Other approaches such as [3, 4, 5, 6, 10, 14, 19, 22, 23, 36, 38, 41] are used for **multi-shot** Re-ID; however, they extract features from each image independently and then aggregate information. For instance, after extracting descriptors of input images independently, [3] uses Karcher mean to accumulate information, [41] uses max-pooling and [36] uses LSTM to aggregate information from multiple images of a person into a single feature vector. Hence they fail to capture multi-modality of an individual's appearance. Other methods represent a multi-shot signature as a set of image descriptors acquired from different images of a person and use a set based similarity metric such as minimum point-wise distance (MPD), or average point-wise distance (APD). For computational and storage efficiency, instead of

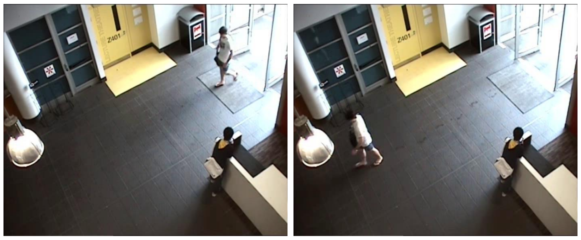

Figure 2: Appearance variation of an individual in one scene. As the person walks across the room with illumination variance, number of images are not uniformly distributed between darker and brighter regions.

using all images, a smaller subset is selected. [4, 19, 38] use a fixed number of randomly selected images; however, uninformed random sampling is prone to losing valuable information unless a large number of samples are drawn. For instance, in Fig. 2, for the person walking across the space with illumination variance, number of images are not uniformly distributed between darker and brighter regions. Thus brighter images may be left out from a small random subset. Thus, matching the person with images collected from an outdoor camera becomes more difficult.

On the other hand, [6, 10] group images of a person into a fixed number of clusters using their HSV histograms and randomly select one image from each cluster to extract multiple feature descriptors. This falsely assumes that appearance modalities in different feature domains are aligned with HSV histogram feature, which limits efficiency of multiple feature fusion. Nonetheless, [5] uses viewpoint cues and [33] use motion cues to discover track segments for different appearance modalities and extract spatiotemporal descriptors such as HOG3D. In addition to temporal segmentation, [21] also uses body-part segmentation to extract more localized spatiotemporal features. Consequently, results in [5, 21, 33] show that a sophisticated multi-shot representation can outperform trivial aggregation of image based (spatial) descriptors; for instance multi-shot SDALF [10] that uses color and texture features is outperformed using HOG3D features [21]. However, all of [5, 21, 33] focus on segmentation using either orientation of a person or the walk cycle and completely ignore the effect of lighting and other factors on the appearance.

Alternatively, we propose that appearance modalities in multi-shot case should be discovered: i) independently for each descriptor space (feature or body-part), and ii) using variance of respective features as cues instead of "external" ones like orientation or pose, because most low-level features are *not* robust to arbitrary transformations, such as pose or orientation changes; thus variance based cues subsume pose and orientation cues. We achieve these goals by independently learning appearance model of each body region and feature descriptor given the set of images of a person. Furthermore, by modeling appearance as parametric distribution we retain more information about appearance of a person with high compression that allows us to better discriminate between persons with similar appearance.

3. Part Appearance Mixture

3.1. Signature model

A scene may have variable illumination and a person may have arbitrary pose, or orientation w.r.t. camera. This requires that low-level appearance descriptors be invariant to illumination, viewpoint, and/or pose changes. Varma and Ray [32] show that low-level appearance descriptors trade-off invariance with discriminative power. Therefore, instead of relying on invariant appearance descriptors, we explicitly deal with variance in a person's appearance by modeling it as a multi-modal probability distribution of descriptors. In particular, we use Gaussian Mixture Model (GMM) to represent appearance. Furthermore, to add robustness to occlusion, signature of a person consists of three part independently learned appearance models, one each for full, upper and lower body region of the person.

Formally, given a set of N_q images $\{I_n^q | n = 1 : N_q\}$ of person q, the corresponding PAM signature, Q, is defined as a set of appearance models $\mathcal{M}_p^q$, one for each part p: $Q = \{\mathcal{M}_p^q | p \in \{full, upper, lower\}\}$. Each appearance model defines distribution of a low level feature for part p of person q using a multivariate GMM, $\mathcal{M}_p^q = \{\pi_{p,k}^q, \mathcal{G}_{p,k}^q | k = 1 : K_p^q\}$ with K_p^q components, where $\pi_{p,k}^q$ is the prior probability of the k^{th} Gaussian $\mathcal{G}_{p,k}^q \sim \mathcal{N}(\boldsymbol{\mu}_{p,k}^q, \boldsymbol{\Sigma}_{p,k}^q)$ having mean $\boldsymbol{\mu}_{p,k}^q$ and covariance $\boldsymbol{\Sigma}_{p,k}^q$. Fitting GMM with full covariance matrix is difficult when number of points is limited and dimensionality of feature is high. To address this concern, we restrict covariance matrices to be diagonal, hence reducing the number of free parameters. This also significantly improves computational efficiency of signature learning and matching.

3.2. Parameter learning

Parameters of each appearance mixture $\mathcal{M}_p^q$ are estimated independently for each person q and part p. The number of modes of a person's appearance is not known *a priori*. Therefore, both the problems of "mode discovery" - finding the number of modes, and "mode description" - appearance description using low-level features, need to be solved. Our strategy is to use variance in low-level feature descriptors of images as a cue to solve both problems of mode discovery and mode description together.

For a fixed number of components K_p^q, given the set of feature descriptors $S_p^q = \{\boldsymbol{s}_{p,n}^q | n = 1 : N_q\}$ corresponding to images $\{Im_n^q : n = 1 : N_q\}$ and part p, the parameters of each appearance model $\mathcal{M}_p^q$ can be easily estimated using Expectation-Maximization algorithm. However, each person may have different number of images and may require different number of GMM components to correctly represent appearance. Therefore, the number of components K_p^q cannot be determined *a priori*. To automatically find optimal number of components for each appearance model, we use Alkaline Information Criterion based model selection to learn an appearance mixture.

3.3. Efficient computation of part models

We use part models to handle occlusions. Instead of feature pruning, which is non-trivial and scenario specific, we build multiple models with some redundancy and accumulate their results. For computational efficiency, coarse localization of upper and lower body is used. Specifically, upper (or lower) body region is defined as $\sim 60\%$ of the total height from the top (or bottom) of the bounding box localizing the person. Although it is possible to use different low-level features for each part (or have redundant parts with different features), for computational efficiency we suggest use of same feature for all three parts. Furthermore, by using descriptors that describe an image region using concatenation of features computed locally over dense spatial grid, such as HOG, MCSH [38], LOMO [19], computation can be shared between full-body and other parts. Specifically, upper-body descriptor can be constructed by concatenating only the local features corresponding to the top $\sim 60\%$ of the bounding box. Lower-body descriptor can be constructed similarly. Thus, only additional computational cost is that of model selection, which can be performed independently in parallel.

4. Similarity metric for PAM

Similarity metric is an important element of a Re-ID method. In our case, a signature is a set of part appearance mixtures. Given the definition of a distance measure between two appearance mixtures, $d(\mathcal{M}_1, \mathcal{M}_2)$, similarity between two signatures Q and G is defined as the sum of similarities of different part models. To convert distance between two appearance mixtures into similarity, we use Gaussian similarity kernel.

$$Sim(Q,G) = \sum_{p\in\mathbb{P}} exp\left(-\frac{\overline{d(\mathcal{M}_p^q, \mathcal{M}_p^g)} - \gamma_{p,g}}{\frac{1}{3}(\beta_{p,g} - \gamma_{p,g})}\right) \quad (1)$$

where $\mathbb{P} = \{full, upper, lower\}$, $\overline{d(\mathcal{M}_p^q, \mathcal{M}_p^g)} = d(\mathcal{M}_p^q, \mathcal{M}_p^g)/\max_{\hat{g}\in Gal} d(\mathcal{M}_p^q, \mathcal{M}_p^{\hat{g}})$ is the distance between a query person q and a gallery person g max normalized over gallery set Gal; and $\beta_{p,g}, \gamma_{p,g}$ are the maximum and minimum normalized distances, respectively, between person g in Gal and any other person $\hat{q}$ in $Query$ set. The factor of $\frac{1}{3}$ above makes Gaussian similarity kernel goes to zero for $\hat{q}$ that has maximum normalized distance from g.

Earlier set based representations often define similarity between two appearance models as average or minimum distance between their elements. Similarly, we define distance between two GMMs as the sum of distance between their components, weighted by the product of their priors.

$$d(\mathcal{M}_1, \mathcal{M}_2) = \sum_{i=1:K_1, j=1:K_2} \pi_{1i}\pi_{2j} d(\mathcal{G}_{1i}, \mathcal{G}_{2j}) \quad (2)$$

where, π_{nk} is the prior for component k of GMM n.

Popular point distance choices are Euclidean and Mahalanobis distance; however, some methods such as [38, 7] use coding theory to compute distance. We evaluated one static and one adaptive (learn-able) definition of $d(\mathcal{G}_i, \mathcal{G}_j)$. The details of the two distance measures are as follows:

4.1. Static distance between signatures

The elements of our signature representation are GMMs, so we use f-Divergence based distances to define similarity between two signatures. In particular, Jeffrey's Divergence (*JDiv*), *i.e.* symmetric KL Divergence, has closed form solution for Gaussian densities and since we restrict covariance matrices to be diagonal, it can be computed efficiently.

$$\begin{aligned}\text{JDiv}(\mathcal{G}_i, \mathcal{G}_j) = &\frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Psi}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \\ &\frac{1}{2} tr\{\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_j + \boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Sigma}_i - 2\boldsymbol{I}\}\end{aligned} \quad (3)$$

where $\boldsymbol{\Psi} = \boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_j^{-1}$.

An alternative to using Jeffrey's Divergence is to use Bhattacharyya distance (*BD*) between two Gaussians.

$$\begin{aligned}\text{BD}(\mathcal{G}_i, \mathcal{G}_j) = &\frac{1}{8}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Gamma}^{\mathbf{-1}}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \\ &\frac{1}{2} ln\{|\boldsymbol{\Sigma}_i|^{-1/2}|\boldsymbol{\Sigma}_j|^{-1/2}|\boldsymbol{\Gamma}|\}\end{aligned} \quad (4)$$

where $\boldsymbol{\Gamma} = \frac{1}{2}(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j)$.

However, we empirically found *JDiv* to work better than *BD*; therefore, when training data is not available we use *JDiv* to define static distance between two Gaussian components: $d(\mathcal{G}_i, \mathcal{G}_j) = \text{JDiv}(\mathcal{G}_i, \mathcal{G}_j)$

4.2. Adaptive distance between signatures

To take advantage of metric learning techniques, we use Mahalanobis-Riemannian Distance (*MRD*) between two Gaussian distributions that can be learned using existing metric learning algorithms, such as KISSME [15] or XQDA [19]. MRD is defined based on the observation that f-Divergence based measures, such as Jeffrey's Divergence (Eq.3) and Bhattacharyya distance (Eq.4) have closed form solutions for Gaussian distributions that can be factored into two terms corresponding to distance between their first and second order moments, *i.e.* mean and covariance. In both cases, the term corresponding to distance between means of Gaussian distributions has the form of Mahalanobis distance. Then by replacing the term corresponding to distance between covariances of Gaussian distributions with Riemannian metric for symmetric positive definite matrices, MRD is proposed by [1] as follows:

$$\mathrm{MRD}(\mathcal{G}_i, \mathcal{G}_j) = \alpha(\boldsymbol{u}^T \boldsymbol{M} \boldsymbol{u})^{\frac{1}{2}} + (1-\alpha) d_{\mathcal{R}}(\boldsymbol{\Sigma_i}, \boldsymbol{\Sigma_j}) \quad (5)$$

where, $\boldsymbol{u} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ is the difference of mean vectors; α controls the weight between the two distance components, and $d_{\mathcal{R}}(,)$ is the Riemannian metric between the two covariance matrices defined as follows:

$$d_{\mathcal{R}}(\boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j) = \left(\sum_{e=1}^{E} log^2 \lambda_e \right)^{\frac{1}{2}} \quad (6)$$

where, $dig(\lambda_1, \lambda_2, ..., \lambda_E) = \boldsymbol{\Lambda}$ is the generalized eigenvalue matrix for the generalized eigenvalue problem: $\boldsymbol{\Sigma}_i \boldsymbol{V} = \boldsymbol{\Lambda} \boldsymbol{\Sigma}_j \boldsymbol{V}$, and $\boldsymbol{V}$ is the column matrix of its generalized eigenvectors. Solving above equation is computationally efficient for diagonal covariance matrices.

We estimate parameters of matrix $\boldsymbol{M}$ using KISSME[15], *i.e.* $\boldsymbol{M} = \boldsymbol{\Sigma}_+^{-1} - \boldsymbol{\Sigma}_-^{-1}$, where $\boldsymbol{\Sigma}_+$ and $\boldsymbol{\Sigma}_-$ are feature-difference covariance matrices of positive and negative classes, respectively. As Mahalanobis distance term in Eq. 5 is defined between the means of two Gaussian distributions, $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$, given ground truth similarity labels, $y_{ij} \in \{+, -\}$, between pairs of Gaussian distributions, $(\mathcal{G}_i, \mathcal{G}_j)$, the positive and negative class covariance matrices are defined as:

$$\boldsymbol{\Sigma}_+ = \sum_{y_{ij}=+} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \quad (7)$$

$$\boldsymbol{\Sigma}_- = \sum_{y_{ij}=-} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \quad (8)$$

Similarity labels between Gaussian pairs are trivially derived from similarity labels of person pairs. For each part, corresponding Mahalanobis metric is learned independently from the appearance models for that part. Alternatively, matrix $\boldsymbol{M}$ can be estimated using XQDA [19] in similar spirit.

Our initial attempts to use the learned metric did not lead to good results. The reason is that by learning a Mahalanobis metric, distance between Gaussian means is computed in a transformed feature space related to learned matrix $\boldsymbol{M}$; however, distance between Gaussian covariances is computed in the original feature space. Therefore, we modify the above definition of MRD to share information of learned matrix $\boldsymbol{M}$ with Riemannian metric between Gaussian covariances. Specifically, we project the covariance matrices of Gaussian distributions using decomposition $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{M}$ of learned matrix $\boldsymbol{M}$. This gives us the following modified definition of MRD, called *Covariance Projected MRD* (CPMRD), where we set $\alpha = 0.5$:

$$\mathrm{CPMRD}(\mathcal{G}_i, \mathcal{G}_j) = \boldsymbol{u}^T \boldsymbol{M} \boldsymbol{u} + d_{\mathcal{R}}(\boldsymbol{L}^T \boldsymbol{\Sigma}_i \boldsymbol{L}, \boldsymbol{L}^T \boldsymbol{\Sigma}_j \boldsymbol{L}) \quad (9)$$

Thus, when training data is available, we use $d(\mathcal{G}_i, \mathcal{G}_j) = \mathrm{CPMRD}(\mathcal{G}_i, \mathcal{G}_j)$ to compute the distance between two Gaussian mixtures using equation 2. To address overfitting concerns, we limit the number of metric parameters by projecting Gaussian distributions (both mean and covariance) to a lower dimensional subspace using PCA before metric learning. This additionally makes metric learning and signature matching computationally efficient.

5. Implementation details

To emphasize on the contribution of proposed representation, we used PAM with two different image descriptors: *HOG* - normalized after concatenation of independently computed 8-bin, 3 channel (RGB), histogram of unsigned oriented gradients descriptor for local regions using an overlapping 3 × 11 grid of 32 × 32 pixels with 16 pixels overlap for the input image re-scaled to 64 × 192 pixels, and *LOMO* [19]. As explained earlier, for efficient computation, we only compute the image descriptor for the full-body. Given the full-body descriptor, we extract the upper and lower descriptors of reduced dimensions from the full-body descriptor. For HOG, upper- and lower-body descriptors correspond to 3 × 6 grids aligned with top and bottom of the bounding box of the person, respectively. A full-body HOG descriptor has 792 dimensions, whereas upper- and lower-body descriptors have 432 dimensions. On the other hand, full-body LOMO descriptor with 26960 dimensions is computed over 3 scales, by dividing an image in 24, 11 and 5 horizontal bands. To extract upper- and lower-body LOMO descriptors, we aggregate information over all 3 scales from 12, 6 and 3 horizontal bands aligned respectively with top or bottom of bounding box of the person. When doing PCA before metric learning, we keep enough components to retain 95% of the variance in original data, but a minimum of 200 and a maximum of 1000.

6. Experiments and results

To demonstrate effectiveness of our Part Appearance Mixture representation, *PAM*, we performed experiments on two publicly available datasets, iLIDS-VID [33] and PRID2011 [12]. These datasets were chosen because they provide multiple images per individual collected in realistic visual surveillance settings using two cameras. Both datasets offer viewpoint variations. In addition, PRID2011 has significant color inconsistency between two cameras, whereas iLIDS-VID has significant occlusions. For each experiment on either dataset, we follow evaluation protocol of [33] for fair comparison with other approaches. Specifically, performance is average over 10 trials of random train-test splits of non-overlapping person IDs, using only the identities with at least 21 images, even though our approach does not impose any restriction on minimum number of im-

	iLIDS-VID		PRID2011	
Model Complexity	HOG	LOMO	HOG	LOMO
Median	2	3	2	3
Maximum	4	7	5	7

Table 1: Complexity of appearance model of a person is measured in number of GMM components.

ages to learn a signature. Additionally, for experiments using static distance measure *JDiv*, which does not require supervised learning, the training split is not used at all.

Table 1 presents statistics about the optimal number of components learned for different full-body appearance models. Maximum model complexity (number of GMM components) per appearance model is 5 and 7, whereas the median is 2 and 3, for HOG and LOMO descriptors, respectively. The average size of input image set is 73 frames in iLIDS-VID; hence, on average, the compression ratio for LOMO descriptor is $(3\times2)/73 \sim 1/12$ (including diagonal covariance matrix). Model selection to find optimal appearance model takes less than 20ms for HOG and 230ms for LOMO descriptor on average. Note that appearance learning cost is paid one time per person, whereas matching cost is paid for each query person per gallery item. Thus appearance learning cost is amortized over time.

6.1. Signature quality

It is important to develop an insight about the semantics or quality of PAM signature model. Therefore, we visualize each GMM component by constructing a corresponding composite image. Specifically, given appearance model of a part of a person, we find likelihood of an image belonging to a model component using its appearance descriptor. Then composite image for a GMM component is generated by summing images of the corresponding person weighted by their likelihood. Thus in the composite image, images with descriptors having high likelihood of belonging to a model component are weighted more. This allows for visualization of appearance models in image space instead of feature space. A selected sample of such signature visualizations is presented in Figure 3. It is easy to note that in describing appearance, proposed signature representation is able to: i) reduce effect of background as information is aggregated over multiple images, hence it does not rely on explicit person segmentation; ii) implicitly deal with transient occlusions due to smoothing effect; and iii) explicitly deal with variance in person's pose, orientation, or illumination by finding optimal number of distinct appearance modes using variance of low-level feature as cue.

6.2. Effectiveness of representation

Since, most authors report results for end-to-end Re-ID methods using a mix of features and learning methods, a direct comparison with our complete Re-ID system (provided in Section 6.3) is not sufficient to highlight improvement brought by better signature modeling. Therefore, we first present a comparison of our complete PAM signature model and only Full Body Mixture (*FBM*) with a baseline that represents a signature as a set of HOG or LOMO descriptors of N randomly selected images of a person, called *RN*. Further, to facilitate comparison of storage and computational complexity, we use three different subset cardinalities, $N \in \{10, Median, Max\}$, where *Median*, and *Max*, are median and maximum complexity (number of GMM components) of a person's full-body appearance model. Moreover, since covariance of descriptors is not computed in baseline method, we use Euclidean distance instead of Jeffrey's divergence to measure distance between two elements of the set. Recognition rates obtained at rank-1 using these models are presented in Table 2

	Rank-1 recognition rate (%)				Avg. Distance	
	PRID2011		iLISD-VID		Time (ms)	
Model	HOG	LOMO	HOG	LOMO	HOG	LOMO
RMedian	23.3	56.3	8.4	21.1	0.005	0.12
RMax	32.4	63.7	12.3	27.4	0.01	0.5
R10	34.8	63.1	14.9	27.1	0.05	1.15
FBM	44.6	67.4	18.0	29.2	0.02	0.67
PAM	50.6	70.6	22.9	33.3	0.03	1.50

Table 2: Performance of different multi-shot representations using HOG and LOMO image descriptors.

Understandably, performance of baseline representation improves as the cardinality of the set increases; however, the matching time also significantly increases with it. In comparison, *FBM* performs significantly better than all three variants of baseline under all variations of datasets and features, while being twice as efficient to *R10* in terms of computational time. The performance improvement for HOG is considerably noticeable, which performs relatively poorly than LOMO on both datasets. Additional improvement was achieved for all cases when we used *PAM*, for signature representation, while still taking less time for matching on average than *R10* with HOG descriptor. Similar trends in performance were also observed when metric learning was used with these representations. Results for learned metrics with *PAM* are discussed in section 6.3.

6.3. Comparison with state-of-the-art

For comparison with existing Re-ID methods, we classify them as *unsupervised* or *supervised* based on whether they require supervised learning of matching functions or features. Table 3 shows recognition rates of methods that are unsupervised. Methods are grouped based on whether they use spatial (treat images independently) or spatiotemporal (ST) features. All competing methods,

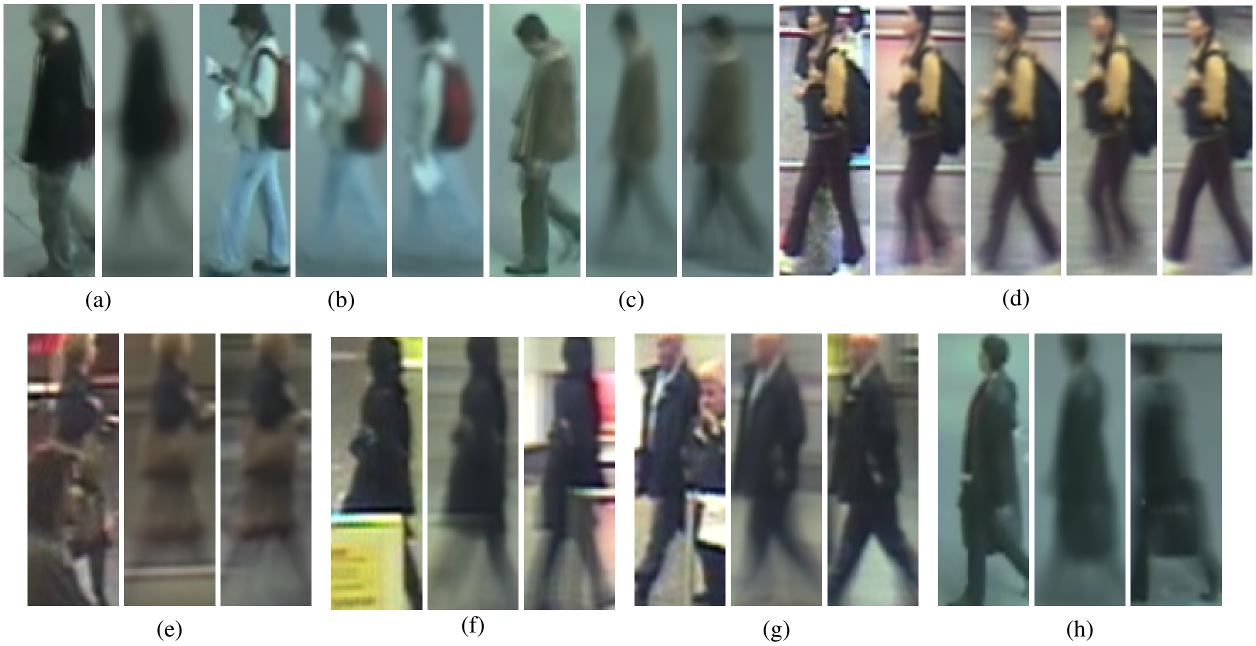

Figure 3: Visualization of full-body appearance mixtures of HOG descriptor. For each person, first image is one of the input images used to learn appearance model. The input image is followed by the composite images, one for each component of the GMM mixture. Optimal number of GMM components for each appearance model varies between persons. (a)-(d) GMM components focus on different pose and orientation of person. (e)-(g) Transient occlusions are implicitly dealt with in appearance models as GMM components focus on pose and/or orientation. (h) GMM components focus on different person alignment within the bounding box.

		PRID2011				iLIDS-VID			
Features	Method	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
Spatiotemporal	HOG3D[21]	20.7	44.5	57.1	76.8	8.3	28.7	38.3	60.7
	FV3D[21]	38.7	71.0	80.6	90.3	25.3	54.0	68.3	87.7
	STFV3D[21]	42.1	71.9	84.4	91.6	**37.0**	**64.3**	**77.0**	**86.9**
Spatial	SDALF[10]	5.2	20.7	32.0	47.9	6.3	18.8	27.1	37.3
	eSDC[39]	25.8	43.6	52.6	62.0	10.2	24.8	35.5	52.9
	FV2D[23]	33.6	64.0	76.3	86.0	18.2	35.6	49.2	63.8
	PAM-HOG	50.6	72.2	83.6	93.0	22.9	44.3	55.7	69.3
	PAM-LOMO	**70.6**	**90.2**	**94.6**	**97.1**	33.3	57.8	68.5	80.5

Table 3: Recognition rates (%) at different ranks for *unsupervised* methods.

except STFV3D[21], were evaluated using set based signature representation. Among spatial approaches, only SDALF[10] uses informed selection of images via HSV histogram clustering. In other cases, either all images are used or a small set is selected randomly. Among spatiotemporal approaches, HOG3D and FV3D features are computed over spatiotemporal volumes using optical flow energy based segmentation of [33], while STFV3D[21] is concatenation of Fisher vector feature descriptors computed from *action units* after video segmentation.

Even with simple 2d HOG features, PAM outperforms all other existing methods on PRID2011. On iLIDS-VID, PAM-HOG is better than all methods using spatial features and HOG3D based spatiotemporal model. Note that performance improvement of PAM-HOG is significant over other spatial features based models such as SDALF, eSDC[39] and FV2D[23] even though HOG computation is relatively simple in the sense that it does not use symmetry or saliency information, nor higher order Fisher information. This shows that PAM is a more efficient representation of appearance models than existing ones. Moreover, PAM with LOMO achieves even better performance on both datasets. On PRID2011, PAM-LOMO outperforms state-of-the-art by 28% and is inferior only to spatiotemporal Fisher vector based representation on iLIDS-VID. This shows generality of PAM representation for different feature descriptors.

		PRID2011				iLIDS-VID			
Learning	Method	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
Dictionary or Feature	DVDL[14]	40.6	69.7	77.8	85.6	25.9	48.2	57.3	68.9
	Color+LFDA[26]	43.0	73.1	82.9	90.3	28.0	55.3	70.6	88.0
	AFDA[18]	43.0	72.7	84.6	91.9	37.5	62.7	73.0	81.8
	MTL-LORAE[30]	-	-	-	-	43.0	60.1	70.3	85.3
	ColorLBP+RFA-Net+RankSVM[36]	58.2	85.8	93.4	97.9	49.3	76.8	85.3	90.0
Metric or Rank	HOG3D+RankSVM[33]	19.4	44.9	59.3	77.2	12.1	29.3	41.5	56.3
	Color+RankSVM[33]	29.7	49.4	59.3	71.1	16.4	37.3	48.5	62.6
	ColorLBP[13]+RankSVM	34.3	56.0	65.5	77.3	23.2	44.2	54.1	68.8
	DVR[33]	28.9	55.3	65.5	82.8	23.3	42.4	55.3	68.6
	DSVR[34]	40.0	71.1	84.5	92.2	39.5	61.1	71.7	81.0
	STFV3D+KISSME[21]	64.1	87.3	89.9	92.0	43.8	69.3	80.0	90.0
	LOMO+XQDA[19]	-	-	-	-	53.0	78.5	86.9	93.4
	LOMO+SBSR+XQDA[7]	-	-	-	-	68.5	87.9	93.0	96.3
	CNN+KISSME[41]	69.9	90.6	-	98.2	48.8	75.6	-	92.6
	CNN+XQDA[41]	77.3	93.5	-	99.3	53.0	81.4	-	95.1
	PAM-HOG+KISSME	55.3	80.7	90.2	95.6	33.9	60.0	70.2	79.1
	PAM-LOMO+KISSME	**92.5**	**99.3**	**100.0**	**100.0**	**79.5**	**95.1**	**97.6**	**99.1**

Table 4: Recognition rates (%) at different ranks for methods using adaptive metrics or rank functions.

Furthermore, to demonstrate suitability of our representation to metric learning and end-to-end Re-ID pipeline, a comparison of PAM+KISSME algorithm with other supervised approaches is provided in Table 4. The methods are classified based on whether they learn matching functions - KISSME[15], XQDA[19], RankSVM[27], DVR[33], and DVSR[34], or representations (features or dictionaries) - DVDL[14], LFDA[26], AFDA[18], MTL-LORAE[30], and RFA-Net[36], or both CNN+KISSME and CNN+XQDA[41].

PAM with LOMO descriptor achieves 79.5% and 92.5% rank-1 recognition rates on iLIDS-VID and PRID2011, respectively. This is significant improvement over current state-of-the-art on both datasets including LOMO and CNN based methods [19, 7, 41] on iLIDS-VID, which shows that our adaptation of KISSME is effective. Notice that unlike in unsupervised case, PAM outperforms STFV3D on both datasets quite significantly. Moreover, PAM with HOG outperforms other HOG based Re-ID methods, HOG3D+RankSVM and DVR[33], which further strengthens our claim that PAM is a more robust signature representation for multi-shot Re-ID. Thus, improvement in performance can be attributed to both improvement in representation and its amenability to metric learning. In addition, it can achieve high recognition rates without using projection on large dictionaries [14, 18, 7], or storing large number of image descriptors. For instance, results for LOMO+XQDA on iLIDS-VID were obtained by using all images of a person and the performance decreases as the number of images is decreased.

7. Conclusion

Person Re-ID is significantly challenging due to high intra-class variance and inter-class similarity. This paper presents a novel representation to model appearance of a person using coarsely localized body parts (regions). Appearance of each part is modeled as a Gaussian Mixture Model to explicitly deal with variance in scene illumination, and pose and orientation of a person. By automatically discovering number of GMM components, the model trades-off signature variance with robustness. Our model visualizations indicate that different GMM components automatically focus on different poses or orientations of a person. Moreover, aggregation of information from multiple images adds robustness to transient occlusions and background clutter. As a result of this improved appearance modeling, new state-of-the-art is achieved on two publicly available datasets.

Acknowledgement: The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement No. 324359 and from project ANR-Movement.

References

[1] K. T. Abou-Moustafa and F. P. Ferrie. A note on metric properties of some divergence measures: The gaussian case. In *ACML*, 2012. 5

[2] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 2

[3] S. Bak, G. Charpiat, E. Corvee, F. Bremond, and M. Thonnat. Learning to match appearances by correlations in a covariance metric space. In *ECCV*, 2012. 2

[4] S. Bak, R. Kumar, and F. Bremond. Brownian descriptor: a rich meta-feature for appearance matching. In *WACV*, 2014. 2, 3

[5] S. Bak, S. Zaidenberg, B. Boulay, and F. Bremond. Improving person re-identification by viewpoint cues. In *AVSS*, 2014. 2, 3

[6] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino. Multiple-shot person re-identification by hpe signature. In *ICPR*, 2010. 2, 3

[7] S. Chan-Lang, Q. Pham, and C. Achard. Bidirectional sparse representations for multi-shot person re-identification. In *AVSS*, 2016. 4, 8

[8] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang. Similarity learning on an explicit polynomial kernel feature map for person. In *CVPR*, 2015. 2

[9] M. Dikmen, E. Akbas, T. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *ACCV*, 2010. 2

[10] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 2, 3, 7

[11] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008. 2

[12] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis*, pages 91–102. Springer, 2011. 1, 5

[13] M. Hirzer, P. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, 2012. 2, 8

[14] S. Karanam, Y. Li, and R. J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV*, 2015. 2, 8

[15] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 2, 4, 5, 8

[16] R. Layne, T. Hospedales, and S. Gong. Person re-identification by attributes. In *BMVC*, 2012. 2

[17] W. Li, Y. Wu, M. Mukunoki, Y. Kuang, and M. Minoh. Locality based discriminative measure for multiple-shot human re-identification. *Neurocomputing*, 2015. 2

[18] Y. Li, Z. Wu, S. Karanam, and R. Radke. Multi-shot human re-identification using adaptive fisher discriminant analysis. In *BMVC*, 2015. 8

[19] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 2, 3, 4, 5, 8

[20] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: What features are important? In *ECCV Workshops and Demonstrations*, 2012. 2

[21] K. Liu, W. Zhang, and R. Huang. A spatio-temporal appearance representation for video-based pedestrian re-identification. In *ICCV*, 2015. 2, 3, 7, 8

[22] X. Liu, M. Song, Q. Zhao, D. Tao, C. Chen, and J. Bu. Attribute-restricted latent topic model for person re-identification. *Pattern Recognition*, pages 4204–4213, 2012. 2

[23] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher descriptors for person re-identification. In *ECCV Workshops*, 2012. 2, 7

[24] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury. Temporal model adaptation for person re-identification. In *ECCV*, 2016. 2

[25] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 2016. 2

[26] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013. 8

[27] B. Prosser, W.-S. Z. S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *BMVC*, 2010. 2, 8

[28] W. R. Schwartz and L. S. Davis. Learning discriminative apperance-based models using partial least squares. In *Brazilian Symposium on Computer Graphics and Image Processing*, 2009. 2

[29] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang. Person re-identification with correspondence structure learning. In *ICCV*, 2015. 2

[30] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao. Multi-task learning with low aank attribute embedding for person re-identification. In *CVPR*, 2015. 2, 8

[31] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016. 2

[32] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *ICCV*, 2007. 3

[33] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*, 2014. 1, 2, 3, 5, 7, 8

[34] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by discriminative selection in video ranking. *T-PAMI*, 2016. 8

[35] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV*, 2007. 2

[36] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang. Person re-identification via recurrent feature aggregation. In *ECCV*, 2016. 2, 8

[37] J. You, A. Wu, X. Li, and W. Zheng. Top-push video-based person re-identification. In *CVPR*, 2016. 2

[38] M. Zeng, Z. Wu, C. Tian, L. Zhang, and L. Hu. Efficient person re-identification by hybrid spatiogram and covariance descriptor. In *CVPR Workshop*, 2015. 2, 3, 4

[39] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, 2013. 2, 7

[40] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identfiation. In *CVPR*, 2014. 2

[41] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016. 2, 8