3D Gaussian Descriptor for Video-based Person Re-Identification

Chirine Riachy[1], Noor Al-Maadeed[2], Daniel Organisciak[1], Fouad Khelifi[1], Ahmed Bouridane[1]
[1]Northumbria University, Newcastle Upon Tyne, UK
[2]Qatar University, Doha, Qatar
{chirine.riachy, daniel.organisciak, fouad.khelifi,
ahmed.bouridane}@northumbria.ac.uk
n.alali@qu.edu.qa

ABSTRACT

Despite being often considered less challenging than image-based person re-identification (re-id), video-based person re-id is still appealing as it mimics a more realistic scenario owing to the availability of pedestrian sequences from surveillance cameras. In order to exploit the temporal information provided, a number of feature extraction methods have been proposed. Although the features could be equally learned at a significantly higher computational cost, the scarce nature of labelled re-id datasets encourages the development of robust hand-crafted feature representations as an efficient alternative, especially when novel distance metrics or multi-shot ranking algorithms are to be validated. This paper presents a novel hand-crafted feature representation for video-based person re-id based on a 3-dimensional hierarchical Gaussian descriptor. Compared to similar approaches, the proposed descriptor (i) does not require any walking cycle extraction, hence avoiding the complexity of this task, (ii) can be easily fed into off-shelf learned distance metrics, (iii) and consistently achieves superior performance regardless of the matching method adopted. The performance of the proposed method was validated on PRID2011 and iLIDS-VID datasets outperforming similar methods on both benchmarks.

Keywords

Person Re-identification, Spatio-temporal Descriptor, Feature Extraction, Gaussian Distribution, Surveillance.

1 INTRODUCTION

With the rise in the need for smart surveillance applications, person re-identification (re-id) has attracted the attention of many researchers within the computer vision community. The problem entails finding a match for a given person image or sequence of images (the probe) among a set of gallery person instances captured using a different non-overlapping camera view. The aim is to track the person under a multi-camera setting. The challenges associated reside mainly in the large intra-class variations caused by significant changes in viewpoint angle, pose and illumination, added to the presence of background clutter and occlusions [1, 2, 3]. Furthermore, small inter-class variations caused by clothes similarities between different people render the task even more challenging.

To mitigate the effect of these challenging attributes on re-id performance, one could exploit the rich visual cues and temporal information provided by person video sequences available from surveillance cameras. For this purpose, the past few years have witnessed the development of several video person re-id algorithms. As few of the descriptors were specially designed for the video-based problem, most of them rely on image-based low-level representations to develop multi-shot ranking algorithms based on set-to-set distance measures [4, 5, 6], or to perform frame selection and weighting [7, 8]. When it comes to descriptors utilising spatio-temporal information, they are often accompanied by walking cycle extraction [9, 10], which is not trivial especially under severe noise and occlusions.

A robust hand-crafted spatio-temporal descriptor that can be efficiently computed and used for matching with common distance metrics has not been yet proposed, which motivates the current work. Such a descriptor presents the advantage of considerably boosting metric learning or multi-shot ranking accuracy, while being equally suited for use in unsupervised settings as no learning is required.

Benefitting from the advances achieved in image-based person re-id feature design, the state-of-the-art Gaussian of Gaussian (GOG) [11] feature is here extended to 3 dimensions integrating temporal information. The suggested extension coupled with existing metric learning approaches achieved significant accuracy improve-

https://doi.org/10.24132/CSRN.2019.2901.1.20

ISBN 978-80-86943-37-4

ment on two widely tested video-based person re-id benchmarks, PRID2011 and iLIDS-VID.

Briefly, this paper proposes a robust person descriptor for video-based re-id leveraging both spatial and temporal cues. The proposed method is based on local Gaussian distributions of 3-dimensional pixel features that are subsequently projected into the Euclidean space. This allows the extracted feature to be learning-free and flexible to use with any matching method.

The remainder of the paper is organised as follows. Section 2 highlights recent related work. Section 3 explains the proposed approach. The experiments conducted are described in detail in Section 4. Finally, the paper is concluded in Section 5.

2 RELATED WORK

Tracking a person in a multi-camera setting involves three main steps: person detection, tracking, and retrieval. The latter consists of searching a person captured in one view, in a different non-overlapping camera view. It is commonly known as person re-id. For that purpose, two tasks should be fulfilled: pedestrian description and matching. The former involves representing persons by a set of features describing their physical appearance. The most popular have been colour and texture features [11, 12, 13, 14, 15, 16, 17]. On the other hand, distance metric learning has emerged as the most prevalent matching method [15, 6, 14, 18, 19] due to its efficiency and promising accuracy.

Early research in person re-id focused mostly on the single-shot scenario where each person is represented by a pair of images that need to be matched [17, 20, 21]. Other scenarios were also investigated including multi-shot re-id and video-based re-id [4, 5, 6, 9, 10, 22]. In that case, each subject is represented by multiple images or video sequences in both probe and gallery views.

Hand-crafted features have been predominant in person re-id until very recently when deep-learning popularity started growing [22, 23, 24, 25]. This was mainly triggered by the release of large-scale datasets such as CUHK03 [24], Market-1501 [26], and MARS [25]. A common practice for hand-crafted methods is to divide the person image into small patches and several horizontal stripes on which features are subsequently extracted. The most prominent methods in this category are ELF [20], gBiCov [12], HistLBP [14], Densecolor-SIFT [27], and LOMO [15]. They all leverage colour and texture information to describe the person's appearance.

More recently, a high-performing image person descriptor called GOG [11] has been proposed. It divides the image into small overlapping patches and a number of horizontal stripes. It then leverages both mean and covariance information by encoding each patch using a Gaussian distribution. Towards the goal, each pixel i is initially described by a feature vector p_i given by $p_i = [y, M_{0^\circ}, M_{90^\circ}, M_{180^\circ}, M_{270^\circ}, R, G, B]^T$ where y is the y-coordinate of pixel i, M_{0° to M_{270° are the orientations along which the gradient is quantised and multiplied by the gradient magnitude, and R, G, B are the RGB colour channels. Patches belonging to the same horizontal stripe are in turn summarised using another Gaussian distribution that is flattened into the Euclidean space. The final image representation is the concatenation of all stripes' features.

As for the matching process, distance metric learning attempts to find a subspace that brings positive samples (feature vectors of the same person) closer to each other while pushing negative samples apart. Various methods were proposed in this category of which the most prominent are KISSME [18], Cross-view Quadratic Discriminant Analysis (XQDA) [15], and kernelised versions of Local Fisher Discriminant Analysis (kLFDA) and Marginal Fisher Analysis (kMFA) [14]. More recently, Zhang *et al.* [19] exploited the kernel Null Foley-Sammon Transform (kNFST) to learn a subspace where the within-scatter is zero and the between-scatter is positive. By enforcing such a strict condition, the learned subspace is more discriminative and exhibits better separability of the data. Distance metric learning methods have been a great success in person re-id given their efficiency and accuracy especially on small datasets where deep learning usually fails. These reasons motivate their use in this work to additionally boost the performance of the proposed person descriptor.

Compared to image-based re-id, video-based re-id represents a more intuitive scenario owing to the availability of pedestrian videos from surveillance cameras. The early trend was to treat the task as a multi-shot matching problem. That is, features were extracted from still images on which the person appearance was built, and temporal cues were completely ignored [5, 21]. This was compensated by exploiting the multiple person instances in the matching process. More recently, Wang *et al.* [9] leveraged temporal information by proposing a spatio-temporal descriptor based on HOG3D features [28], they later combined it with mean colour values [29]. These features were computed on fragments extracted from the person sequence using the Flow Energy Profile (FEP) signal as an approximation of walking cycles. A ranking function, DVR, was learned for matching. Liu *et al.* [10] subsequently proposed a spatio-temporal descriptor based on Fisher vectors extracted from video-fragments representing body-action units after performing walking cycle extraction similarly to [9].

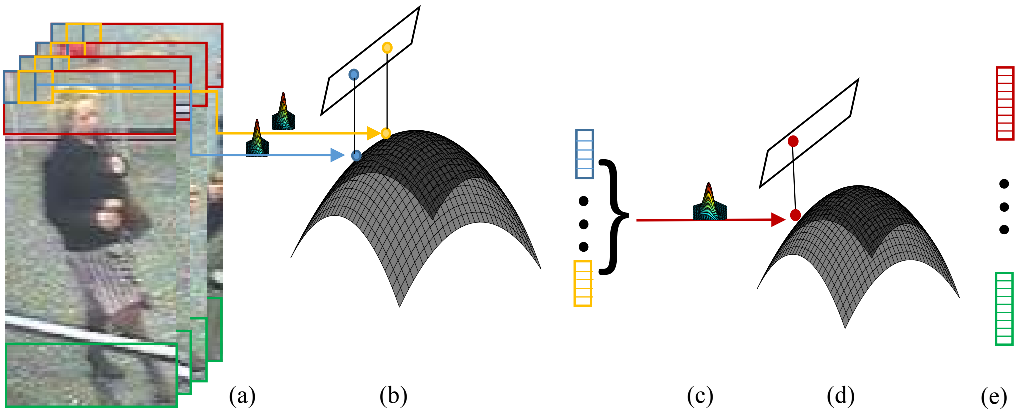

Figure 1: Representative diagram of GOG3D feature. (a) Patch Gaussians are computed. (b) They are then flattened into the Euclidean space. (c) Region Gaussians are formed using patches in the same horizontal region. (d) They are also projected into the Euclidean space. (e) Region feature vectors are concatenated to form the final image feature. Finally, average-pooling is performed over image features of the same person.

Following these works, temporal information was mainly exploited in deep learning methods such as RNN [22] or ASTPN [23] where spatial and temporal pooling layers were added for that purpose. The remaining video-based re-id systems either design a set-to-set distance metric such as DRAH [4] that models the distance between two sequences as the minimum distance between their respective affine hulls, or assign a signature to each person using the corresponding frames by fitting a GMM [30] for instance.

The proposed method is a spatio-temporal descriptor that benefits from temporal correlation to give a richer representation for each person. However, each person sequence is eventually described by a single feature vector which is in essence similar to the single-shot scenario. This is very convenient as it speeds up the matching process significantly and broadens the applicability of the descriptor with any off-shelf distance metric. It is worth noting here that extracting the proposed feature from sampled walking cycles is also possible. However, in this work we highlight the advantages presented by using it in its simplest setting.

3 PROPOSED METHOD

This section details the proposed spatio-temporal descriptor for video-based person re-id which we call GOG3D. Similar to GOG [11], a part-based model is adopted where each person image is divided into R overlapping horizontal stripes, roughly representing different human body-parts. Each stripe is also divided into small overlapping patches, and each patch is modelled by a Gaussian distribution using its pixel feature information. Patch Gaussians are subsequently embedded into the space of Symmetric Positive Definite (SPD) matrices and flattened into the Euclidean space forming the patch feature vector. Patches belonging to the same horizontal region are in turn summarised by a Gaussian that is projected into the Euclidean space forming the region feature vector. Region feature vectors are concatenated to form the final image feature. Finally, image features of the same person are averaged to form the final representation of a person sequence. A diagram summarising GOG3D is shown in Fig. 1.

3.1 Pixel Features

By first considering each of the patches, local information is described using a 10-dimensional pixel feature vector summarising the spatial position of the pixel, its gradient magnitudes along four directions, and the intensity values of some colour channels. Namely, for a pixel i, the pixel feature vector p_i is given by:

$$p_i = [x, y, M_{0^\circ}, M_{90^\circ}, M_{180^\circ}, M_{270^\circ}, |I_t|, L, A, B]^T, \quad (1)$$

where x and y are the x- and y-coordinates of pixel i taken from the top-left of the image, $M_0{}^\circ$ through M_{270° are the orientations along which the gradient is quantised and multiplied by its magnitude, $|I_t|$ is the gradient magnitude in the temporal direction, and finally L, A, B correspond to the Lab colour channels. Each dimension of p_i is scaled to the range [0, 1] before further processing can take place. The computation of these pixel features is detailed in the following.

For simplicity, let us consider one person sequence of images given by $S = \{Q_k | k = 1, ..., N\}$, where N is the number of frames in this video sequence. By taking a pixel $i(x,y,t)$ in frame Q_t, the gradients I_x, I_y and I_t in the horizontal, vertical and temporal directions can be computed as:

$$I_x(x,y,t) = i(x+1,y,t) - i(x-1,y,t), \quad (2)$$

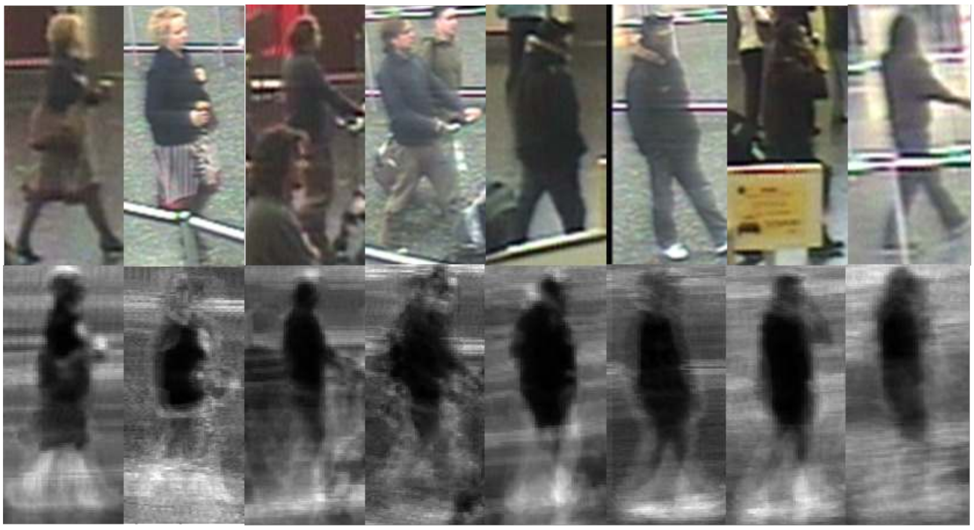

Figure 2: The top row represents example images sampled from person sequences of iLIDS-VID dataset, each 2 adjacent images represent a correct match. The bottom row includes the temporal gradient computed for these sequences and averaged over the frames involved.

$$I_y(x,y,t) = i(x,y+1,t) - i(x,y-1,t), \quad (3)$$

$$I_t(x,y,t) = i(x,y,t+1) - i(x,y,t-1), \quad (4)$$

where x,y and t are the x-, y- and t-coordinates of pixel $i(x,y,t)$, respectively.

Although, it is possible to compute the gradient orientation in 3D and quantise it using a regular polyhedron in a manner similar to HOG3D in spirit [28], this is suboptimal in our case due to the following reasons. Firstly, binning in 3D while preserving the distinctive power of the descriptor requires the use of a dodecahedron (12 bins) or icosahedron (20 bins) [28], thus raising significantly the dimensionality of the pixel feature vector p_i. A high-dimensional p_i will definitely cause numerical problems upon the computation of covariance matrices in small patches. Moreover, it is favourable to use spatial gradients separately as a texture descriptor, while motion information is encoded via the temporal gradient. For this purpose, hereby the orientation of the spatial gradient given by I_x and I_y is exploited separately to the temporal gradient I_t.

The gradient orientation O is given by $O = \arctan(I_y/I_x)$ and its magnitude is defined as $M = \sqrt{(I_x^2 + I_y^2)}$. It has been proved in [31] that quantisation into vector angles rather than using magnitude and orientation raw values is essential to enhance the discriminative power of the descriptor. Therefore, soft voting is used to quantise the values of O into two neighbouring bins to account for the loss of information caused by quantisation while maintaining some rotation invariance. Since four bins are considered in this case, the reference points are 0°, 90°, 180°, and 270°. Given $\alpha \leq O < \beta$ where $\alpha, \beta \in \{0°, 90°, 180°, 270°\}$ are the boundaries of the bin containing O, the distances (positive differences) $d_\alpha = |O - \alpha|$ and $d_\beta = |O - \beta|$ from O to the bin boundaries are computed, and the voting weights are assigned as $w_\alpha = d_\beta/(d_\alpha + d_\beta)$, $w_\beta = d_\alpha/(d_\alpha + d_\beta)$ and $w_{(\theta \neq \alpha,\beta)} = 0$. These weights are finally multiplied by the gradient magnitude M to obtain $M_\theta, \theta \in \{0°, 90°, 180°, 270°\}$. As for the temporal gradient I_t, the magnitude of I_t is found by taking its absolute value $|I_t|$. Some examples of the information provided by computing $|I_t|$ for all sequence images and taking their average can be seen in Fig. 2. This highlights the type of information added by computing the gradient in the temporal direction.

The choice of the colour channels is crucial for any appearance descriptor, especially for person re-id where individuals are mainly distinguished by their clothes' colours and texture. Performing feature fusion by extracting the features four times, each with different colour channels such as RGB, HSV, Lab and normalised RGB (nRGB), and concatenating them similarly to GOG$_{\text{fusion}}$ [11] and moM [16], is highly inefficient. The reason is that the features have to be extracted four times for each person, and the resulting vector has a high dimension which in turn slows down the matching process. For this purpose, it is convenient to select the most discriminative colour channels that can better deal with illumination changes. In this work, these are found to be the Lab colour channels.

It has been previously argued that person images are aligned vertically but not horizontally [13], therefore only the y-coordinate has been used in GOG algorithm. However, as it will be detailed thereafter, we find that the orderless representation of patches in the same horizontal stripe as a mean to deal with viewpoint angle variations may cause the loss of some important spatial information that could be useful for re-identification.

This could be avoided by including the horizontal location of the pixel represented by its x-coordinate.

3.2 Patch and Region Gaussians

As both mean and covariance features have proved successful in person re-id [9, 12], a promising way to leverage both types of information is to summarise them using a Gaussian distribution. As discussed in [11], it is definitely possible to use a Gaussian Mixture Model (GMM) instead for a more accurate representation. However, given the small patch size, a simple Gaussian model should be sufficient to describe the patches. More importantly, unimodal Gaussians can be efficiently projected into the Euclidean space which renders the matching process with the resulting feature much easier, as any off-shelf distance metric can thus be exploited. Therefore, for each patch H, the mean μ_H and covariance Σ_H are estimated as $\mu_H = \frac{1}{n_H}\sum_{i\in H} p_i$ and $\Sigma_H = \frac{1}{n_H-1}\sum_{i\in H}(p_i-\mu_H)(p_i-\mu_H)^T$ where n_H is the number of pixels in patch H and p_i is the feature vector of pixel i defined in Section 3.1. Subsequently, the patch Gaussian $\mathcal{N}(p;\mu_H,\Sigma_H)$ is given by:

$$\mathcal{N}(p;\mu_H,\Sigma_H) = \frac{\exp\left(-\frac{1}{2}(p-\mu_H)^T\Sigma_H^{-1}(p-\mu_H)\right)}{(2\pi)^{d/2}|\Sigma_H|}, \tag{5}$$

where $|\cdot|$ is the matrix determinant operator and d is the dimension of pixel feature vector p.

Once all patch Gaussians are computed, an algorithm (see Section 3.3) is used to project these Gaussians into the Euclidean space transforming them into patch feature vectors f_H. To account for background clutter, patches are weighted similarly to SDALF algorithm [17] according to their distance from the central vertical axis of the image such that $w_H = \exp\left(-\frac{(x_H-x_C)^2}{2\sigma^2}\right)$ where $x_C = W/2$, $\sigma = W/4$ and x_H is the x-coordinate of the central pixel in patch H. By W we denote the image width. According to this definition, it is easy to see that more weight is assigned to patches closer to the central vertical axis of the image where the person is expected to be centred.

In a similar manner to patch Gaussian computation, the patches belonging to the same horizontal region are in turn summarised into a region Gaussian using their mean and covariance information. Based on the above defined weights, the region mean μ_R and covariance Σ_R are defined for region R as:

$$\mu_R = \frac{1}{\sum_{H\in R} w_H}\sum_{H\in R} w_h f_h, \tag{6}$$

$$\Sigma_R = \frac{1}{\sum_{H\in R} w_H}\sum_{H\in R} w_H(f_H-\mu_R)(f_H-\mu_R)^T, \tag{7}$$

where f_H is the patch feature vector for patch H. The region Gaussians are consequently computed according to 5 and projected into the Euclidean space before concatenation, to form the final representation of an image. Finally, frame-wise features are averaged over a person's sequence to form the final representation of that sequence. It is also worth noting that covariance matrices are regularised by adding a small value ε to diagonal entries to prevent them from becoming singular.

3.3 Euclidean Space Projection

Projecting patch and region Gaussians into the Euclidean space is essential for GOG3D, primarily to obtain a final descriptor that can be used with off-shelf distance metrics. Towards the goal, the following two steps need to be applied.

It is well known that multivariate Gaussian distributions lie on a Riemannian manifold that can be embedded into the space of SPD matrices [32]. Such embedding is favoured as the SPD space endowed with the log-Euclidean metric can be locally flattened into the tangent Euclidean space through matrix logarithm. More specifically, consider a d-dimensional multivariate Gaussian $\mathcal{N}(\mu_H,\Sigma_H)$, this can be embedded into a $(d+1)$-dimensional SPD matrix P_H as follows:

$$\mathcal{N}(p;\mu_H,\Sigma_H) \sim P_H = |\Sigma_H|^{-\frac{1}{d+1}}\begin{bmatrix}\Sigma_H+\mu_H\mu_H^T & \mu_H\\ \mu_H^T & 1\end{bmatrix}. \tag{8}$$

P_H can subsequently be mapped into the Euclidean tangent space by computing $\Gamma_H = \log(P_H)$ where $\log(\cdot)$ is the matrix logarithm operator. Noting that Γ_H is symmetric, only the upper triangular part needs to be stored resulting in the final vector f_H being $m = (d^2+3d)/2+1$ dimensional. Thus, $f_H = \mathrm{vec}(\Gamma_H) = [\Gamma(1,1),\sqrt{2}\Gamma(1,2),...,\sqrt{2}\Gamma(1,d+1),\Gamma(2,2),\sqrt{2}\Gamma(2,3),...,\Gamma(d+1,d+1)]$. Note that off-diagonal entries are multiplied by $\sqrt{2}$ upon half-vectorisation to ensure that the Frobenius norm of Γ_H remains equal to the ℓ_2-norm of f_H, that is $||\Gamma_H||_F = ||f_H||_2$.

4 EXPERIMENTS

4.1 Datasets

The proposed GOG3D feature is evaluated on the two most widely tested benchmarks for video-based person re-id, PRID2011 and iLIDS-VID.

iLIDS-VID [9] includes 600 person sequences created from two non-overlapping camera views captured by a CCTV network in an airport arrival hall. 300 different subjects are sampled in this dataset with two sequences per person. Sequences have variable lengths ranging from 23 to 192 frames with an average number of 73. iLIDS-VID is a very challenging dataset in terms of illumination changes, viewpoint angle variations and occlusions.

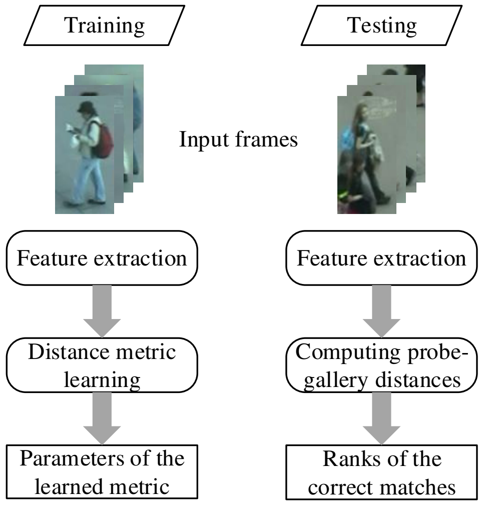
Training
Testing
Input frames
Feature extraction
Feature extraction
Distance metric learning
Computing probe-gallery distances
Parameters of the learned metric
Ranks of the correct matches

Figure 3: General pipeline of the re-id system employed in this work.

PRID2011 [33] consists of 400 image sequences of 200 different subjects captured by two adjacent cameras. Sequences lengths vary from 5 to 675 frames with an average number of 100 frames per sequence. PRID2011 is less challenging than iLIDS-VID as no occlusions or background clutter are involved. However, significant changes in illumination are exhibited.

4.2 Implementation Details

The original GOG algorithm divides the image into 7 overlapping horizontal regions and employs a patch size of 5×5 pixels. Patches are extracted at 2 pixels interval, and the regularisation parameter is $\varepsilon = 0.001$. For the first experiment reported in Section 4.3, and for fair comparison with GOG, a similar setting is used for GOG3D. This results in the patch feature vector being of size $(10^2+3\times10)/2+1 = 66$, and the region feature vector being $(66^2+3\times66)/2+1 = 2,278$ dimensional, finally the person descriptor obtained is $7 \times 2,278 = 15,945$ dimensional. When comparing to state of the art (Section 4.4), the parameters of GOG3D are further tuned through extensive experimentation which results in a better performance for a patch size of 9×9 pixels and $\varepsilon = 0.0001$. The patch extraction interval is kept at 2 pixels, and the results obtained on both datasets are presented using 10 and 15 overlapping horizontal regions denoted GOG3D^{10} and GOG3D^{15}, respectively. In that case, the resulting feature vector is 22,780 and 34,170 dimensional for $R = 10$ and $R = 15$, respectively.

Similar to GOG, mean removal and ℓ_2-normalisation are applied. Dimension reduction using PCA is performed with KISSME metric [18], and the dimension of the reduced feature is set to 100. A linear kernel is used with kNFST [19], kLFDA [14], and kMFA [14] in all experiments. The code provided by the authors is employed for GOG, and frame-wise features are also averaged for a person sequence in a similar manner to proposed GOG3D. The performance is evaluated using the Cumulative Matching Characteristic (CMC) curve as follows. Given a probe (query) instance, the gallery elements are ranked according to their distance from the probe, and at each given rank of the CMC curve, the probability of the correct match appearing at a similar or higher rank is computed. The general pipeline of the re-id system employed in this work can be seen in Fig. 3.

4.3 Components Analysis and Comparison to GOG

To highlight the consistent performance gain by employing GOG3D over GOG algorithm for video-based re-id, we compare the results in top-matching rates of the CMC curve using five state-of-the-art metric learning methods: XQDA [15], KISSME [18], NFST [19], kLFDA and kMFA [14]. For this purpose, each dataset was divided into two subsets, half for training and half for testing. The experiments were repeated over 10 trials and the average results are reported.

Table 1 shows the performance obtained by adding different pixel features to baseline GOG. Specifically, GOG_I$_t$ is obtained by adding the $|I_t|$ component, GOG_I$_t$_x also involves the x-coordinate, and GOG3D is the final feature obtained by replacing RGB colour channels by the Lab channels. A detailed evaluation of these components is performed using XQDA distance metric since the latter was initially used with GOG in [11]. Moreover, results comparing GOG directly to GOG3D with four other metrics can be seen in Table 2 for further validation.

It is clear from these results that GOG3D presents a remarkable advantage over GOG for video-based person re-id. Using XQDA, the performance gain was gradual and consistent by adding different pixel feature components to reach a maximum of almost 7% for iLIDS-VID and 3% for PRID2011 from baseline GOG to GOG3D. This margin undergoes some fluctuations when employing different metrics. It reaches a maximum of almost 13% with kLFDA on iLIDS-VID and 2% on PRID2011 with most other metrics.

4.4 Comparison to State of the Art

Since available spatio-temporal descriptors are mostly designed to be applied to extracted walking cycles or fragments sampled from person sequences [9, 10], employing them directly with a common distance metric in the same evaluation protocol to ours is unfair. When possible, it will in fact cause their performance to downgrade. Therefore, video-based re-id systems were

Dataset	**iLIDS-VID**			**PRID2011**		
Rank	1	5	20	1	5	20
GOG + XQDA	66.6	87.2	96.9	86.6	97.6	99.6
GOG_I$_t$ + XQDA	70.6	90.6	98	87.2	97.2	99.8
GOG_I$_t$_x + XQDA	72.9	91.3	**98.7**	87.7	97.8	99.9
GOG3D + XQDA	**73.7**	**92**	98.3	**89.9**	**97.9**	**100**

Table 1: Components analysis of GOG3D. Best results in top-matching rates are in bold.

Dataset	**iLIDS-VID**			**PRID2011**		
Rank	1	5	20	1	5	20
GOG + KISSME	49.5	73.4	88.1	81	94.2	99.2
GOG3D + KISSME	**55.1**	**78.9**	**92.7**	**83.3**	**94.9**	**99.2**
GOG + kNFST	63.9	85.9	95	87.8	97.4	99.9
GOG3D + kNFST	**74.3**	**92.2**	**98.9**	**89.6**	**97.8**	**100**
GOG + kLFDA	54.6	85.5	97.5	82.8	**96.8**	99.6
GOG3D + kLFDA	**67.4**	**90.7**	**98.8**	**85.1**	96.5	**99.9**
GOG + kMFA	56.3	85.3	97.6	83.3	**96.7**	99.5
GOG3D + kMFA	**66.5**	**90.9**	**98.9**	**85.4**	96.6	**99.9**

Table 2: Comparison to GOG. Best results in top-matching rates for each distance metric are in bold.

compared to our method in their original setting. For this purpose, the same evaluation protocol employed by most algorithms was adopted. More specifically, for PRID2011 dataset, only sequences from 178 persons consisting of more than 27 frames were retained. Half of each dataset was used for training and the remaining half for testing. Experiments were repeated over 10 trials and average results in CMC top-matching rates are reported in Table 3. The distance metric used in this experiment was kNFST due to its superior performance. It is worth noting that the results here are different from those of the previous subsection because the evaluation protocol of PRID2011 and the parameters (patch size, ε, and number of stripes) have been changed as previously detailed in Section 4.2.

ColHOG3D [9] and STFV3D [10] are state-of-the-art spatio-temporal descriptors that fall in the same category with GOG3D. The difference in performance between GOG3D and these descriptors is very obvious, even when using the same distance metric KISSME as shown in the second row of Table 2. The gap in rank1 matching-rate with the better performing STFV3D is over 10% on iLIDS-VID and around 20% on PRID2011.

When compared to deep learning techniques, GOG3D + kNFST outperforms RNN [22], CNN + XQDA [25] and ASTPN [23] by at least 18% on iLIDS-VID and 17% on PRID2011 in terms of rank1 accuracy. It also outperforms multi-shot ranking methods DRAH [4] and SPW [8] by around 10% on iLIDS-VID for SPW and around 5% on PRID2011 for DRAH. The only method that exhibits comparable or slightly worse performance than GOG3D on both datasets is PAM+KISSME [30]. However, while not being in the same category with GOG3D, PAM requires (i) extracting low-level features, (ii) fitting GMMs to person sequences with many parameters to learn, and (iii) the final person representation does not fall in a Euclidean space. Hence, special care needs to be taken for matching. This also means that the final person signature does not exhibit the flexibility of use with other distance metrics.

It is finally noteworthy that GOG3D not only achieves outstanding results on two challenging benchmarks, it is also simple, flexible and computationally efficient. The low computational complexity is derived from omitting additional tasks like walking cycle extraction and fragment selection used by similar space-time descriptors, or feature clustering and frame weighting required for multi-shot ranking methods that may involve further learning. Moreover, the high computational cost needed to train deep neural networks is also avoided. Finally, the flexibility of GOG3D feature is granted by the possibility of its use with any matching method in both supervised and unsupervised settings since it does not involve any learning.

4.5 Computational Cost

GOG3D is implemented in MATLAB and experiments are run on a desktop PC equipped with Intel Xeon X5550 @2.67GHz CPU. The average time to extract GOG3D features per frame is 0.44 seconds. It is computed on 10 video sequences from PRID2011 dataset

Dataset	**iLIDS-VID**			**PRID2011**		
Rank	**1**	**5**	**20**	**1**	**5**	**20**
ColHOG3D+DVR [9]	39.5	61.1	81	40	71.7	92.2
STFV3D+KISSME [10]	44.3	71.7	91.7	64.1	87.3	92
RNN [22]	58	84	96	70	90	97
CNN+XQDA [25]	53	81.4	95.1	77.3	93.5	99.3
ASTPN [23]	62	86	98	77	95	99
DRAH [4]	64	86	96.3	88.7	97.9	99.7
PAM+KISSME [30]	**79.5**	95.1	99.1	92.5	**99.3**	**100**
SPW [8]	69.3	89.6	98.2	83.5	96.3	**100**
GOG3D^{10}+kNFST (proposed)	**80**	**95.3**	**99.5**	**93.6**	**99.4**	**100**
GOG3D^{15}+kNFST (proposed)	**79.5**	**95.4**	**99.5**	**94**	99.1	**100**

Table 3: Comparison to state-of-the-art methods. Best and second best results in top-matching rates are in bold.

and averaged over the number of frames constituting these sequences. Under the same setting, the time taken to compute GOG features is 0.35 seconds per frame. It is intuitive for GOG3D to be slightly slower than GOG since it uses additional pixel features and more horizontal stripes. However, compared to other video person re-id descriptors [9, 10], GOG3D is evidently more efficient since it omits the walking cycle extraction step and any further post-processing. Moreover, frame-wise feature pooling employed with GOG3D renders the matching process very efficient. For instance, the average time taken to train the kNFST metric on PRID2011 dataset over 10 trials is 0.034 seconds, and the testing time on the same dataset is 0.008 seconds which is exceptionally fast for video-based re-id methods.

5 CONCLUSION

A novel spatio-temporal descriptor for video-based person re-id based on hierarchical Gaussian distributions was presented in this paper. The proposed algorithm leverages the gradient in the temporal direction to describe the temporal correlation between consecutive frames yielding significant improvement in accuracy. Unlike available spatio-temporal re-id descriptors, the proposed method does not require any complex walking cycle extraction of frame selection and weighting. It also does not involve any learning. By simply averaging the frame-wise computed features over a person sequence, robust representation can be achieved and consequently fed to most off-shelf distance metrics.

A thorough analysis of the proposed descriptor was conducted on 2 widely used benchmarks using 5 distance metrics, highlighting the advantages brought by exploiting temporal cues. Extensive experiments showed that the performance achieved surpasses similar methods by a large margin. It also outperforms a number of existing deep learning and multi-shot ranking techniques.

ACKNOWLEDGEMENT

This publication was made possible NPRP grant #NPRP 8-140-2-065 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

6 REFERENCES

[1] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, R. Radke, A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets., IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (3) (2019) 523–536.

[2] L. Zheng, Y. Yang, A. G. Hauptmann, Person re-identification: Past, present and future, arXiv preprint arXiv:1610.02984.

[3] C. Riachy, A. Bouridane, Person re-identification: Attribute-based feature evaluation, in: IEEE World Symposium on Applied Machine Intelligence and Informatics (SAMI), 2018.

[4] S. Karanam, Z. Wu, R. J. Radke, Learning affine hull representations for multi-shot person re-identification, IEEE Transactions on Circuits and Systems for Video Technology.

[5] S. Karanam, Y. Li, R. J. Radke, Sparse re-id: Block sparsity for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2015.

[6] X. Zhu, X.-Y. Jing, X. You, X. Zhang, T. Zhang, Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics, IEEE Transactions on Image Processing 27 (11) (2018) 5683–5695.

[7] Y.-J. Cho, K.-J. Yoon, Improving person re-identification via pose-aware multi-shot matching, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[8] W. Huang, C. Liang, Y. Yu, Z. Wang, W. Ruan, R. Hu, Video-based person re-identification via self paced weighting, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[9] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in: European Conference on Computer Vision (ECCV), 2014.

[10] K. Liu, B. Ma, W. Zhang, R. Huang, A spatio-temporal appearance representation for video-based pedestrian re-identification, in: International Conference on Computer Vision, 2015.

[11] T. Matsukawa, T. Okabe, E. Suzuki, Y. Sato, Hierarchical gaussian descriptor for person re-identification, in: Computer Vision and Pattern Recognition (CVPR), 2016.

[12] B. Ma, Y. Su, F. Jurie, Covariance descriptor based on bio-inspired features for person re-identification and face verification, Image and Vision Computing 32 (6-7) (2014) 379–390.

[13] B. Ma, Q. Li, H. Chang, Gaussian descriptor based on local features for person re-identification, in: Asian Conference on Computer Vision (ACCV), 2014.

[14] F. Xiong, M. Gou, O. Camps, M. Sznaier, Person re-identification using kernel-based metric learning methods, in: European Conference on Computer Vision (ECCV), 2014.

[15] S. Liao, Y. Hu, X. Zhu, S. Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[16] M. Gou, O. Camps, M. Sznaier, mom: Mean of moments feature for person re-identification, in: International Conference on Computer Vision (ICCV), 2017.

[17] L. Bazzani, M. Cristani, V. Murino, Symmetry-driven accumulation of local features for human characterization and re-identification, Computer Vision and Image Understanding 117 (2) (2013) 130–144.

[18] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: Computer Vision and Pattern Recognition (CVPR), 2012.

[19] L. Zhang, T. Xiang, S. Gong, Learning a discriminative null space for person re-identification, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[20] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: European conference on computer vision (ECCV), 2008.

[21] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, V. Murino, Custom pictorial structures for re-identification., in: British Machine Vision Conference (BMVC), 2011.

[22] N. McLaughlin, J. Martinez del Rincon, P. Miller, Recurrent convolutional network for video-based person re-identification, in: Computer Vision and Pattern Recognition (CVPR), 2016.

[23] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, P. Zhou, Jointly attentive spatial-temporal pooling networks for video-based person re-identification, in: International Conference on Computer Vision (ICCV), 2017.

[24] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in: Computer Vision and Pattern Recognition (CVPR), 2014.

[25] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, Q. Tian, Mars: A video benchmark for large-scale person re-identification, in: European Conference on Computer Vision (ECCV), 2016.

[26] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: International Conference on Computer Vision (ICCV), 2015.

[27] R. Zhao, W. Ouyang, X. Wang, Unsupervised salience learning for person re-identification, in: Computer Vision and Pattern Recognition (CVPR), 2013.

[28] A. Klaser, M. Marszałek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, in: British Machine Vision Conference (BMVC), 2008.

[29] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by discriminative selection in video ranking, IEEE transactions on pattern analysis and machine intelligence 38 (12) (2016) 2501–2514.

[30] F. M. Khan, F. Brèmond, Multi-shot person re-identification using part appearance mixture, in: Winter Conference on Applications of Computer Vision (WACV), 2017.

[31] T. Kobayashi, N. Otsu, Image feature extraction using gradient local auto-correlations, in: European Conference on Computer Vision (ECCV), 2008.

[32] M. Lovrić, M. Min-Oo, E. A. Ruh, Multivariate normal distributions parametrized as a riemannian symmetric space, Journal of Multivariate Analysis 74 (1) (2000) 36–48.

[33] M. Hirzer, C. Beleznai, P. M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: Scandinavian Conference on Image Analysis (SCIA), 2011.