Supervised Person Re-ID based on Deep Hand-crafted and CNN Features

Salma Ksibi[1], Mahmoud Mejdoub[1,2] and Chokri Ben Amar[1]
[1]*REGIM: Research Groups on Intelligent Machines, University of Sfax, ENIS, 3038, Sfax, Tunisia*
[2]*Department of Computer Science, College of AlGhat, Majmaah University, 11952, Riyadh, Saudi Arabia*

Keywords: Person Re-identification, Fisher Vector, Gaussian Weight, Deep Hand-crafted Feature, Deep CNN, XQDA.

Abstract: Gaussian Fisher Vector (GFV) encoding is an extension of the conventional Fisher Vector (FV) that effectively discards the noisy background information by localizing the pedestrian position in the image. Nevertheless, GFV can only provide a shallow description of the pedestrian features. In order to capture more complex structural information, we propose in this paper a layered extension of GFV that we called LGFV. The representation is based on two nested layers that hierarchically refine the FV encoding from one layer to the next by integrating more spatial neighborhood information. Besides, we present in this paper a new rich multi-level semantic pedestrian representation built simultaneously upon complementary deep hand-crafted and deep Convolutional Neural Network (CNN) features. The deep hand-crafted feature is depicted by the combination of GFV mid-level features and high-level LGFV ones while a deep CNN feature is obtained by learning in a classification mode an effective embedding of the raw pedestrian pixels. The proposed deep hand-crafted features produce competitive accuracy with respect to the deep CNN ones without needing neither pre-training nor data augmentation, and the proposed multi-level representation further boosts the re-ID performance.

1 INTRODUCTION

Person re-identification (re-ID) (Guo et al., 2006; Zheng et al., 2016) is a challenging task in the camera surveillance field (Wali et al., 2010), since it addresses the problem of matching people across potentially multiple non-overlapping cameras. Many re-ID works (Mahmoud Mejdoub and Koubaa, 2017; Ksibi et al., 2016b; Ksibi et al., 2016a; Ksibi et al., 2016c) used either shallow or deep representations coupled with supervised metric learning. Shallow methods are built upon the low-level and the mid-level appearance features. We can cite as the most successful low-level appearance features the Local Maximal Occurrence (LOMO) (Liao et al., 2015), the Symmetry-Driven Accumulation of Local Features (SDALF) and the eSDC (Zhao et al., 2013). Concerning the mid-level features (Mejdoub et al., 2009; Mejdoub et al., 2009; Ben Aoun et al., 2014; Mejdoub et al., 2008) that demonstrated their robustness in the image classification field (M. El Arbi et al., 2011; Mejdoub et al., 2015b; Mejdoub et al., 2008), the Bag of visual Words (BOW) model (Ksibi et al., 2012; Zheng et al., 2015) that quantifies the low-level features into a dictionary of visual words, has been presented for the person re-ID task in (Zheng et al., 2015). Locality-constrained linear coding (LLC) was proposed in (Li et al., 2015) by using a soft quantization. It considers the locality information in the feature encoding process by taking into account only the k-nearest basis vectors from each local feature. Fisher Vector (FV) (Ma et al., 2012; Messelodi and Modena, 2015; Liu et al., 2015; Wu et al., 2017; Sekma et al., 2015a; Sekma et al., 2015b; Mejdoub and Ben Amar, 2013) is another encoding method that learns a Gaussian Mixture model (GMM) on the local descriptors, in order to compute the visual words. B. Ma et al. (Ma et al., 2012) were the first to introduce the FV encoding scheme in person re-ID task. They employed a spatial representation that divides the pedestrian image into 4×3 fixed regions and used a simple 7-d local descriptor. These local descriptors are turned into FVs and these are employed to measure the similarity between two persons using the Euclidean distance between their representations. In (Messelodi and Modena, 2015), the authors introduced a boosting method that learns a scoring function taking into account the likelihood between the local FVs of the same identity. Regarding the deep representations, most of the current state-of-the-art methods used a Convolutional Neural Network (CNN) (M. El Arbi et al., 2006; Bouchrika et al., 2014) verification model. This infers

Ksibi, S., Mejdoub, M. and Amar, C.
Supervised Person Re-ID based on Deep Hand-crafted and CNN Features.
In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) - Volume 4: VISAPP*, pages 63-74
ISBN: 978-989-758-290-5

positive image pairs and negative ones as input to the CNN, owing to the lack of training data (Boughrara et al., 2017; Othmani et al., 2010) per pedestrian identity. However, the recognition accuracy is generally badly influenced by the absence of the intra-class similarity and inter-class dissimilarity information. To tackle this problem, the ID-discriminative Embedding (IDE) deep CNN feature was presented in (Zhong et al., 2017; Zheng et al., 2016) to learn an embedded feature space in a classification mode. It was stated in (Wu et al., 2016a) that IDE performs better than the previously used verification model. Pedestrian matching is then operated in the learned feature space. Wu et al. presented another classification CNN model (Wu et al., 2017). They built a hybrid network by moving the input FVs on the fully connected layers and enforcing the linear discriminative analysis (LDA) as an objective function to produce embeddings that have low intra-class variance and high inter-class variance. In (Wu et al., 2016b; Xiong et al., 2014) low-level hand-crafted features are combined with high-level CNN features. Afterwards, metric learning is applied to the obtained combination. The good re-ID results obtained by the concatenation between low-level hand-crafted features and CNN ones provide support on their complementary nature. To enhance the discriminative ability of the appearance features, supervised metric learning, such as the Keep It Simple and Straightforward MEtric learning (KISSME) (Köstinger et al., 2012), the locally adaptive decision functions (LADF) (Li et al., 2013), the Null Space (NS) metric learning (Zhang et al., 2016), and the Cross-view Quadratic Discriminant Analysis (XQDA) (Liao et al., 2015) are often applied upon the generated features in order to learn an optimal distance allowing to increase the intra-similarity and decrease the inter-similarity. Among them, XQDA achieves good re-ID results (Zheng et al., 2016). This is mainly due to the fact that XQDA has the ability to simultaneously learn a discriminative subspace as well as a distance in the low dimensional subspace. Indeed, in this paper, we propose to encode the appearance features throughout a rich histogram representation well adapted to the person re-ID field. In this sense, an extension of the traditional FV encoding method is introduced, namely the Gaussian weighted FV (GFV). This consists in weighting the histogram encoding process (Dammak et al., 2014b; Mejdoub et al., 2015a; Dammak et al., 2014c; Mejdoub et al., 2015b; Dammak et al., 2015; Dammak et al., 2014a) via the pedestrian Gaussian template. This latter fosters the locations that lie nearby the pedestrian in the image. GFV provides a shallow representation of the pedestrian. To further describe the complex spatial structural information that can be present in a pedestrian image, we propose in this paper a layered version of GFV called layered GFV (LGFV). This latter is based on two nested layers that hierarchically refine the FV encoding from one layer to the next by integrating more spatial neighborhood information. We separately apply LGFV on three low-level local appearance features (Color Name (CN), Color Histogram (CHS) and 15-d descriptors). In the first layer, we densely sample a set of spatial sub-windows on the pedestrian image. We then perform a local FV encoding within every sub-window on the low-level features describing the sub-window patches. Thus we obtain for each sub-window a local FV that depicts the rich spatial structural information. The second layer applies GFV on the set of the local FVs, within the whole image. After that, by combining the GFV and the LGFV global image signatures, we obtain a deep hand-crafted feature per pedestrian image. The latter is further combined with the deep CNN feature (IDE) to provide a rich multi-level representation. Consequently, we obtain seven pedestrian global histograms (see Figure 1). Finally, the pedestrian are matched by combining the XQDA distance learned upon these seven histograms. It is worth mentioning that all images are pre-treated with Retinex transform (Liao et al., 2015), to reduce the illumination variation before the application of the encoding methods. Besides, both GFV and LGFV are applied upon a stripe representational scheme in order to consider the spatial alignment information between pedestrian parts. Our main contributions in this work are:

- We propose a new deep hand-crafted feature based on the combination of the GFV and LGVF representations. LGFV explores how the performance of the shallow hand-crafted features can be improved with increased structural spatial information depth. The experimental results have shown that the proposed deep hand-crafted feature can produce competitive results with the deep CNN feature: IDE (Zhong et al., 2017; Zheng et al., 2016), without needing neither pre-training nor data augmentation.
- We propose to combine the deep hand-crafted features with the well performing IDE deep CNN features. The resulting multi-level representation is semantically rich since it exploits the complementary power between the spatial structural information generated by the two deep representations.

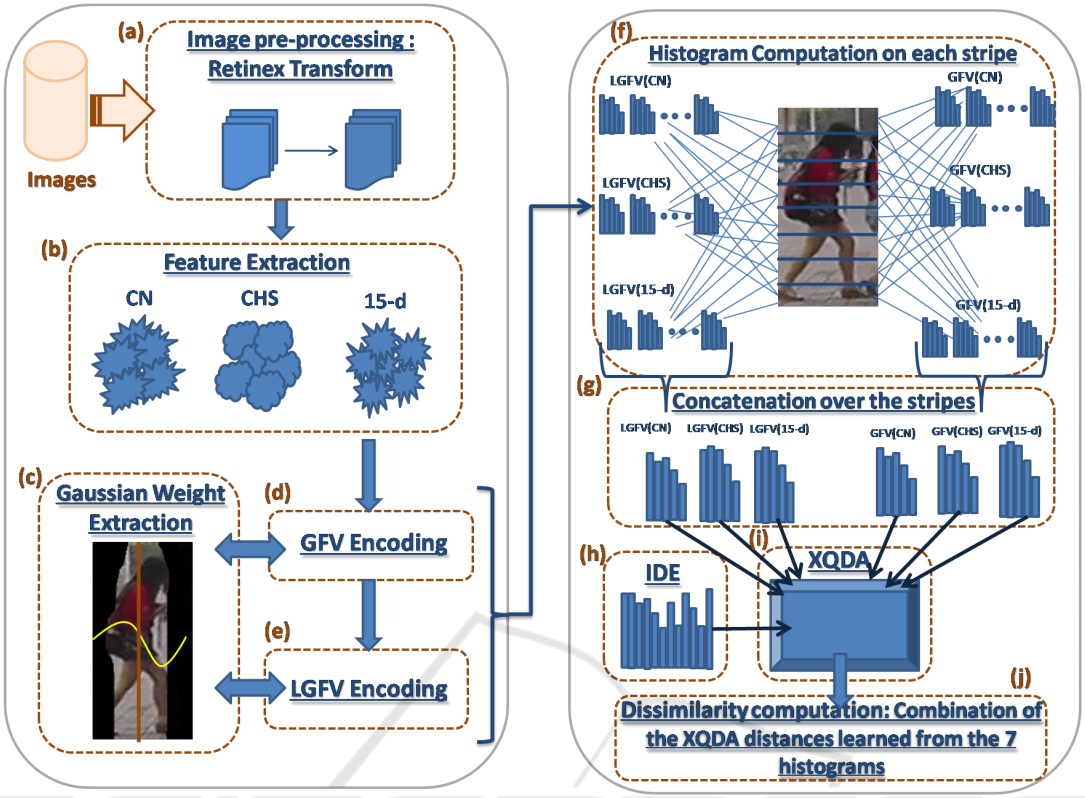

Figure 1: Overview of the proposed method pipeline. (a) Image pre-processing by Retinex Transform to deal with illumination variation. (b) Low-level feature extraction: CN, CHS and 15-d descriptors. (c) Extraction of the Gaussian template weights. (d) Gaussian weighted FV (GFV) Encoding. (e) Layered Gaussian weighted FV (LGFV) Encoding. (f) GFV and LGFV are calculated separately on each given low-level color feature (CN, CHS and 15-d), at each stripe. (g) The generated histograms are concatenated over the stripes producing the final GFV and LGFV representations (one histogram per low-level feature). (h) Extraction of the deep CNN feature: IDE. (i) Computation of the XQDA distance learned separately on the IDE features and each kind of the previously generated histograms. (j) Dissimilarity computation: combination of the XQDA distances learned from the seven histograms.

2 THE PROPOSED METHOD

2.1 Dealing with Illumination Variations

We apply in this paper the Multi-scale Retinex transform with Color Restoration (MSRCR) (Jobson et al., 1997) to handle illumination variations. Single Scale Retinex algorithm (SSR) is the basic Retinex algorithm which uses a single scale. The original image is processed in the logarithmic space in order to highlight the relative details. Besides, a $2D$ convolution operation with Gaussian surround function is applied to smooth the image. Afterwards, the smooth part is subtracted from the image to obtain the final enhanced image. SSR can either provide dynamic range compression (small scale), or tonal rendition (large scale), but not both simultaneously. The MSRCR algorithm bridges the gap between color images and the human observation by combining effectively the dynamic range compression of the small-scale Retinex and the tonal rendition of the large scale with a color restoration function. In the experiments, we used two scales of the Gaussian surround function ($\sigma = 5$ and $\sigma = 20$).

2.2 Low-level Feature Extraction

In this work, the pedestrian image is sampled with dense patches, using a size of 4×4, and a stride of 4 pixels, respectively. For each patch three kinds of color low-level descriptors are extracted (CN, CHS and 15-d). The latter are chosen underlying their good compromise between efficiency and re-ID accuracy (Zheng et al., 2015; Ma et al., 2012). Indeed, their small dimensionality as compared to other state of the art descriptors such as the global LOMO descriptor (Liao et al., 2015) and the local dColorSift one (Zhao

et al., 2013) makes them well adapted to the efficiency factor required by the person re-ID task.

2.2.1 Color Names (CN)

Authors in (Kuo et al., 2013) have demonstrates that the color description based on color names ensures a good robustness against the photometric variance. In this paper, as was done in (Kuo et al., 2013), we use the 11 basic color terms of the English language, i.e. black, blue, brown, gray, green, orange, pink, purple, red, white, and yellow. First, the CN feature vector of each pixel is calculated by making a mapping from HSV pixel values to an 11 dimensional CN vector. Afterwards, we apply a sum pooling on the CN pixel features related to each patch. Finally, the resulting histogram undergoes a square rooting operation followed by $l1$ normalization. The size of the generated CN descriptor is then equal to 11.

2.2.2 Color Histogram (CHS)

For each patch, a 16-bin color histogram is computed in each HSV color space channel. For each color channel, the patch color histogram is square-rooted and subsequently $l1$ normalized. The three obtained histograms are then concatenated, generating a color descriptor of size 16×3.

2.2.3 15-d Descriptor

Inspired by (Ma et al., 2012), we design a simple 15-d descriptor. First the pedestrian image is split into 3 color channels (HSV). For each channel C, each pixel is converted into a 5-d local feature, which contains the pixel intensity, the first-order and second-order derivative of this pixel. The description is on the following equation:

$$f(x,y,C) = (C(x,y), C_x(x,y), C_y(x,y), C_{xx}(x,y), C_{yy}(x,y)) \quad (1)$$

where $C(x,y)$ is the raw pixel intensity at position (x,y), C_x and C_y are the first-order derivatives with respect to pixel coordinates x and y, and C_{xx}, C_{yy} are the second-order derivatives. Then, we apply, for each color channel, a sum-pooling operation over the 15-d descriptors of the pixels located within each patch. Each of the three obtained patch descriptors undergoes a square root operation followed by $l1$ normalization. Afterwards, we horizontally concatenate the three normalized descriptors into one single signature.

2.3 Extraction of the Gaussian Template Weights

In (Farenzena et al., 2010), the authors proposed to separate the foreground from the background of the pedestrian image, and that by using segmentation. However, it was difficult to obtain an aligned bounding box, and an accurate segmentation, especially in the presence of cluttered backgrounds. This makes the extraction of reliable features describing the person of interest hard. We propose in this paper a simple solution by employing a 2-D Gaussian template on the pedestrian image, in order to remove the noisy background. Consider $p_{w,h}$ the patch whose spatial center is located at the w-th row and h-th column in the image, $I = \{p_{h,w}, h = 1 \dots H, w = 1 \dots W\}$ of width W and height H. Inspired by (Farenzena et al., 2010; Zheng et al., 2015), the Gaussian function is defined by $N(\mu_x \sigma)$, where μ_x is the mean value of the horizontal coordinates, and σ is the standard deviation. We set μ_x to the image center ($\mu_x = W/2$), and $\sigma = W/4$. This method uses a prior knowledge on the person position, which assumes that the pedestrian lies in the image center. Therefore, the Gaussian template works by weighting the locations near the vertical image center with higher probabilities. This permits to discard the noise surrounding the person's silhouette, and thus to keep meaningful parts of the images and eliminate needless ones. Explicitly, we endow each patch $p_{h,w}$ with a Gaussian weight $G(p_{h,w})$, given by:

$$G(p_{h,w}) = exp(-(w-\mu_x)^2/2\sigma^2) \quad (2)$$

2.4 Proposed Gaussian Weighted Fisher Vector (GFV) Encoding Method

We propose in this paper a rich extension of the traditional FV encoding method. It consists in the incorporation of the Gaussian weight in the encoding process of the latter. The proposed encoding operation is made within three steps. Indeed, as operated in the traditional FV, the first step consists in learning a Gaussian Mixture model (GMM) represented by K components, on the local descriptors extracted from all training pedestrian images. Then, the second step implies the computation the mixture weights, means, as well as the diagonal covariance of the GMM, which are respectively denoted as π_k, μ_k, σ_k. For concise clarity, we omit hereinafter the patch index (h,w) used in the previously notation (subsection 2.3), and we replace it by i. Thus, the Gaussian weight of the image patch p_i is noted $G(p_i)$. Consider M local descriptors d_i corresponding to the M patches p_i

of an image I. The proposed encoding incorporates the Gaussian weight in the traditional FV encoding, as given by the following Equation:

$$u_k = \frac{1}{M\sqrt{2\pi_k}} \sum_{i=1}^{M} G(p_i) \times \alpha_k(d_i) \left(\frac{d_i - \mu_k}{\sigma_k} \right) \quad (3)$$

$$v_k = \frac{1}{M\sqrt{2\pi_k}} \sum_{i=1}^{M} G(p_i) \times \alpha_k(d_i) \left(\frac{(d_i - \mu_k)^2}{\sigma_k^2} - 1 \right) \quad (4)$$

where, $\alpha_k(di)$ is the soft assignment weight of the i-th descriptor d_i to the k-th Gaussian $G(p_i)$ that represents the Gaussian weight. For each GMM component, the sum-pooling operation aggregates the M descriptors in the image, into a single encoded feature vector, given by the concatenation of u_k and v_k for all K components:

$$FV = [u_1 \dots u_K, v_1 \dots v_K] \quad (5)$$

Finally, we apply power normalization to each FV component before normalizing them jointly. Such normalization demonstrates a good performance in previous works (Sapienza et al., 2014). We note that the proposed GFV encoding is applied separately to the three proposed low-level descriptors: CN, CHS and 15-d descriptor.

2.5 Proposed Deep Hand-crafted Feature

The deep hand-crafted feature is obtained by combining the shallow GFV with its layered extension. The process is described hereinafter:

2.5.1 Proposed Layered GFV (LGFV) Representation

The proposed shallow GFV robustly encodes the local features of a pedestrian, aggregating them by performing a sum-pooling operation over the entire pedestrian image. The obtained representation achieves the encoding directly from the flat local feature space, without considering the complex spatial structure that can be present in a pedestrian image. In order to incorporate the spatial information in GFV encoding, we propose to design a LGFV representation that describes the pedestrian with higher level structures extracted from the spatial pedestrian neighborhood. This idea is an adaptation of the layered FV encoding presented in the context of image and action recognition (Sekma et al., 2015a; Peng et al., 2014; Simonyan et al., 2013) to the person re-ID case. The layered encoding is performed over two layers (see Figure 2).

First Layer: In the first layer, we perform local FV encoding within each sub-window. Toward this end, we first perform the FV encoding ($M = 1$ in Eqs. 3, 4 without accounting the weighting) on each image local feature using the same GMM codebook prelearned in GFV. This produces for each local feature a tiny high dimensional FV. After the generation of these tiny FVs, we aggregate them by applying sum-pooling within every sub-window. This is comparable to the traditional FV. The difference is that the encoding is achieved in the sub-window rather than the whole image. The sub-windows are obtained by scanning in a dense way the image using a stride of 4 pixels (the size of a patch). Each sub-window corresponds to the spatial neighborhood constituted by 3×3 patches. The obtained local FVs related to the sub-windows are subsequently power-$l2$ normalized. As a result, instead of a unique FV, representing the whole image, the latter is described by a set of dense local FVs, each of which reflects the local spatial information among spatially adjacent local features. Therefore, the generated representation can reflect a rich image structural information.

Second Layer: Since the local FVs are too high-dimensional to be straightly used as the inputs of the second encoding layer, we adopt XQDA to reduce the local FVs in a discriminative supervised manner. As the only provided annotation is the identity label, we exploit this information by (1) averaging the local FVs over the whole image, (2) applying XQDA to the resulting intermediate vectors, and (3) projecting each local FV on the learned XQDA projection matrix. This gives a good compromise between accuracy and efficiency, and de-correlates the local FVs in order to make them suitable to their further FV encoding (Sekma et al., 2015a). After learning a GMM on the set of the reduced local FVs, we perform GFV encoding upon them. The weight used in GFV for each reduced local FV is given by the weight of its corresponding sub-window. The latter is computed by averaging the weights of each individual patch in the sub-window. The output vector is subsequently power-$l2$ normalized to form the final LGFV of the image.

2.5.2 Final Deep Representation

The tiny FVs obtained in the first layer can be further exploited to form a global pedestrian FV by performing a weighted sum-pooling over the entire pedestrian image. This is then equivalent to the GFV representation. Therefore, both global GFV and LGFV histograms are combined to constitute the outputs of the

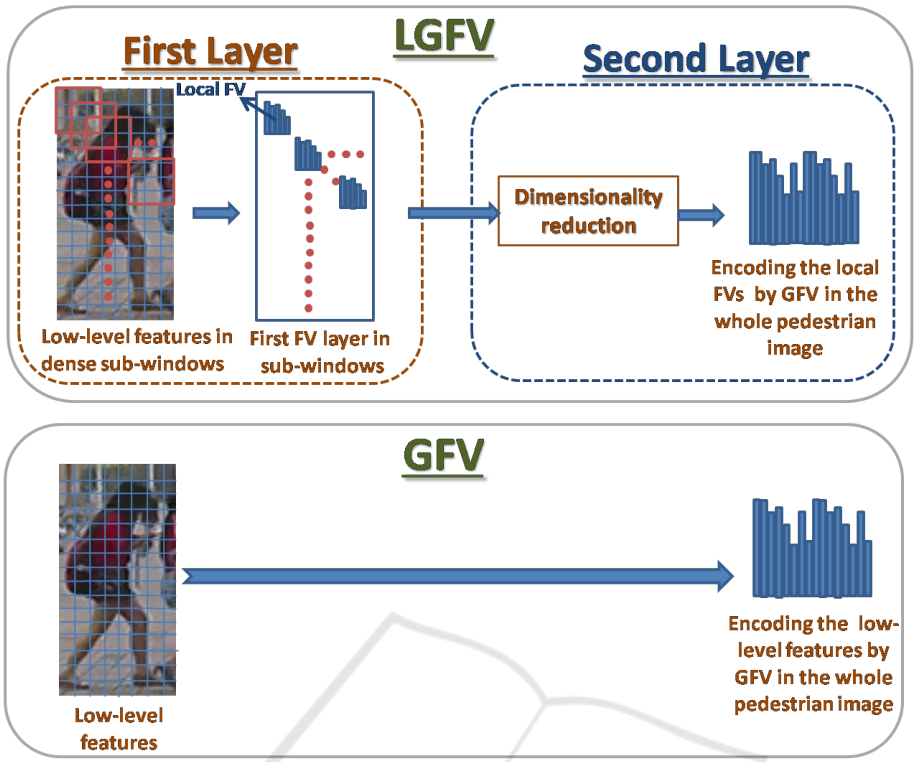

Figure 2: Comparison between the proposed LGFV and GFV methods. Top: The pipeline of the proposed LGFV with two layers. Bottom: pipeline of GFV. The image representation in LGFV is constructed based on sub-windows of size of 3×3 patches.

proposed hand-crafted deep representation.

2.6 Histogram Computation based on Pedestrian Image Partition

In order to take benefit of the spatial alignment information among the different body parts in the persons images, appearance modeling typically exploits the spatial pyramidal model (Zheng et al., 2015; Sekma et al., 2014) to treat the appearance of different body parts independently. Inspired by these works, we propose to sub-divide the pedestrian into a set of stripes. Since, the spatial information of the horizontal y-axis exhibits greater intra-class variance than the vertical x-axis due to viewpoint and pose variations, we choose to divide the silhouette according to the y-axis. Indeed, the image is split into $N_S = 8$ stripes as it has shown a good compromise between accuracy and efficiency in (Zheng et al., 2015). The proposed GFV method is applied separately in every single stripe rather than the whole pedestrian image. Afterwards, histograms corresponding to each stripe are $l2$ normalized separately prior to stacking. Since we use, in this paper, three low-level descriptors (CN, CHS and 15-d) for GFV and LGFV, we obtain six global histograms. Finally, every global histogram is further $l2$ normalized to ensure the linear separability of the data.

2.7 Deep CNN Feature: IDE

In this paper, the IDE feature introduced in (Zhong et al., 2017) is employed since it has been shown to outperform many other deep CNN models. Specifically, we use CaffeNet (Krizhevsky et al., 2012) to train the CNN in a classification mode. In the training phase, images are resized to 227×227 pixels, and they are passed to the CNN model, along with their respective identities. The CaffeNet network contains five convolutional layers with the same original architecture, and two globally connected layers each with $1,024$ neurons. The number of neurons in the final fully connected layer is defined by the number of training identities in each dataset. In the testing phase, 1024 dimensional CNN features are extracted, for each pedestrian image, throughout the 7-th layer of CaffeNet. The CNN features are then subsequently $l2$ normalized.

2.8 Dissimilarity Computation

After the generation of the six global histograms ba-

ed on GFV and LGFV, as well as the IDE feature, we separately learn on each of them an XQDA (Liao et al., 2015) distance in a supervised way. XQDA learns a reduced subspace from the original training data, and at the same time learns a distance function in the resulting subspace for the dissimilarity measure. Once the distances are learned, they are summed-up to derive the final dissimilarity function. Given a probe, dissimilarity scores are assigned to all gallery items. The gallery set is then ranked according to the dissimilarity to the probe. It is worth mentioning that, as performed in (Liao et al., 2015), we select as subspace components the eigenvectors corresponding to the eigenvalues of $S_w^{-1} S_b$ that are larger than 1, where S_w and S_b refer to the within and the between scatter matrices, respectively.

2.9 Multiple Queries

When each identity has multiple queries (MultiQ) in a single camera, we could merge them into a single query to reformulate the MultiQ problem to one query (OneQ) one. In this way, the intra-class variation is taken into account, and the method will be more robust to the pedestrian variations over the gallery images. We apply an average pooling on the GFV and LGFV related histograms over the multiple queries. As for the IDE feature, we use max pooling since the latter has shown better re-ID results than average pooling in (Zheng et al., 2016). The resulting pooled vectors are then used to perform the matching process with the probe set.

3 EXPERIMENTS

3.1 Datasets

CUHK03 (Li et al., 2014). contains 13,164 Deformable Part Model (DPM) (Arandjelovic and Zisserman, 2012) bounding boxes, of 1,467 different identities of the training set. Each single identity is observed through two different cameras and there are on average 4.8 images, for each view and each identity. We follow the experimentation protocol proceeded in (Zheng et al., 2015). Indeed, we select 100 persons randomly and for each person, all the DPM bounding boxes are taken as queries in turns. After that, a cross camera search is performed. This test process is repeated 20 times and statistics are reported next. Note that the dataset comes with manual (Labelled) and algorithmically (Detected) pedestrian bounding boxes.

Market-1501 (Zheng et al., 2015). contains 32,643 fully annotated boxes of 1501 pedestrians, making it one of the largest person re-ID image datasets. This dataset is captured with 6 different cameras placed in front of a supermarket, and contains 32,643 bounding boxes of 1501 different identities. Actually, each single identity is captured by at most 6 cameras and at least 2. Each identity may have multiple images under each camera, and even if images of same identity are captured by the same camera, they are totally distinct and different. Market-1501 is randomly divided into training and testing sets, containing respectively 750 and 751 identities. In the testing phase, for each single identity, there is one query image selected in each camera. The search is processed in a cross-camera mode, i.e. we discard from the re-ID process images that belong to the same camera as the query. Note that there are 3,368 queries in the gallery, 19,732 images used for testing and 12,936 images for training. We use the provided fixed training and test set, under both OneQ and MultiQ evaluation settings.

3.2 Experimental Settings

In this paper, we use a codebook of 256 GMM components for GFV and LGFV since it yields a good compromise between accuracy and efficiency. Unless otherwise stated, all results generated by our proposed method are given for the supervision case obtained by XQDA and the one query setting. We also note that the Cosine distance is used for the unsupervised case.

3.3 Evaluation Metrics

In this paper, we use the Cumulative Matching Characteristics (CMC) curve in the but of evaluating performances of the proposed methods and comparing with the re-ID state-of-art ones, on all datasets. Every probe image is matched with every gallery one, and ranks of the correct matches are obtained. Indeed, the rank-k recognition rate is the expectation of a correct match at the rank-k, and the cumulative values of recognition rates at all ranks, are recorded as one-trial CMC result. We also use the mean average precision (mAP) to evaluate the performances of the proposed methods. In fact, for each query, we calculate the area under the Precision-Recall curve, called average precision (AP). After that, the mean value of the APs (mAP) of all queries is calculated by taking into account both precision and recall, and thus provides a more comprehensive evaluation.

Table 1: Impact of the GFV, LGFV and IDE combination. Results (rank-1 matching rate and on mAP) are reported on CUHK03 and Market-1501 datasets for the different encoding methods.

Methods	CUHK03		Market-1501	
	r=1	mAP	r=1	mAP
FV	36.43	37.86	52.12	23.38
GFV	43.61	45.08	58.85	31.88
LGFV	51.15	55.71	65.46	40.16
GFV + LGFV	58.97	64.75	71.92	46.45
IDE(C)	58.91	64.92	57.72	35.95
GFV + LGFV + IDE(C)	63.60	69.51	75.18	53.88

Table 2: Comparison of the proposed unsupervised (GFV_U+LGFV_U) method with the state-of-the-art methods in the case of unsupervised (first table part), and the proposed supervised GFV, (GFV+LGFV) and (GFV+LGFV+IDE(C)) methods with the supervised methods (second table part), on the Market-1501 dataset. Note that (+MultiQ) and (+re) refer to the Multi Query setting and the re-ranking method, respectively. '-' means that corresponding results are not available.

Methods	OneQ		MultiQ	
	r=1	mAP	r=1	mAP
SDALF (Farenzena et al., 2010)	20.53	8.20	-	-
eSDC (Zhao et al., 2013)	33.54	13.54	-	-
BOW (Zheng et al., 2015)	34.40	14.10	42.14	19.20
Ours(GFV_U+LGFV_U+IDE(C)_U)	65.11	40.08	71.86	46.19
LOMO (Liao et al., 2015)	26.07	7.75	-	-
PersonNet (Wu et al., 2016a)	37.21	18.57	-	-
KISSME(BOW) (Zheng et al., 2015)	44.42	20.76	-	-
KISSME(LOMO) (Liao et al., 2015)	40.50	19.02	-	-
XQDA(LOMO) (Liao et al., 2015)	43.79	22.22	54.13	28.41
kLDFA(LOMO) (Liao et al., 2015)	51.37	24.43	52.67	27.36
FVdeepLDA (Wu et al., 2017)	48.15	29.94	-	-
SCSP (Chen et al., 2016)	51.90	26.35	-	-
NS(LOMO) (Zhang et al., 2016)	55.43	29.87	67.96	41.89
NS(fusion) (Zhang et al., 2016)	61.02	35.68	71.56	46.03
NS-CNN (Li et al., 2016)	59.56	34.44	69.95	44.82
XQDA(IDE (C)) (Zhong et al., 2017)	57.72	35.95	-	-
XQDA(IDE(C))+re (Zhong et al., 2017)	61.25	46.79	-	-
SCNN (Varior et al., 2016)	65.88	39.55	76.04	48.45
Ours(GFV)	58.85	29.78	66.49	41.63
Ours(LGFV)	65.46	40.16	72.78	49.12
Ours(GFV+LGFV)	71.92	46.45	78.90	56.26
Ours(GFV+LGFV+IDE(C))	75.18	53.88	82.72	60.46
Ours(GFV+LGFV+IDE(C)) + re	77.42	63.54	84.92	70.52

3.4 Empirical Analysis of the Proposed Method

3.4.1 Impact of the Gaussian Weight

As can be shown by Table 1, weighting the traditional FV via the Gaussian weight (GFV) considerably increase the matching rates in all datasets. Indeed, when applying the Gaussian template, we eliminate the background noise located around the pedestrian and its harmful effects on the re-ID accuracy.

3.4.2 Impact of the GFV, LGFV and IDE Combination

We propose in this paper a rich multi-level representation, issued of the combination of the deep handcrafted (GFV + LGFV) and the deep CNN (IDE) features. We start by studying the impact of the GFV and LGFV combination (GFV+LGFV). Results in Table 1 show good improvements in accuracy in both datasets when combining the shallow GFV with the layered LGFV. This achievement proves the complementary of the shallow description (GFV), and the layered one (LGFV) that integrates richer structural information.

Table 3: Comparison of the proposed unsupervised methods (GFV_U + LGFV_U) with the state-of-the-art methods in the case of unsupervised (first table part), and the proposed supervised methods GFV, (GFV + LGFV) and (GFV + LGFV + IDE(C)) with the supervised methods (second table part), on CUHK03 dataset, on the labelled and detected cases.

Methods	CUHK03 (detected)				CUHK03 (manual)			
	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
SDALF(Farenzena et al., 2010)	4.87	-	-	-	5.60	23.45	36.09	51.96
eSDC (Zhao et al., 2013)	7.68	-	-	-	8.76	24.07	38.28	53.44
BOW (Zheng et al., 2015)	22.95	-	-	-	24.33	58.42	71.28	84.91
Ours (GFV_U+LGFV_U)	48.58	79.51	88.92	94.73	54.43	84.65	93.76	96.82
ITML (Davis et al., 2007)	5.14	-	-	-	5.53	18.89	39.96	44.20
LMNN (Sun and Chen, 2011)	-	-	-	-	7.29	21.00	32.06	48.94
KISSME (Köstinger et al., 2012)	11.70	-	-	-	14.17	41.12	54.89	70.09
DeepReid (Li et al., 2014)	19.89	50.00	64.00	78.50	20.65	51.50	66.50	80.00
Improved Deep (Ahmed et al., 2015)	44.96	76.01	83.47	93.15	54.74	86.50	93.88	98.10
XQDA(LOMO) (Liao et al., 2015)	46.25	78.90	88.55	94.25	52.20	82.23	92.14	96.25
NS(LOMO) (Zhang et al., 2016)	53.70	83.05	93.00	94.80	58.90	85.60	92.45	96.30
NS(fusion) (Zhang et al., 2016)	54.70	84.75	94.80	95.20	62.55	90.05	94.80	98.10
Metric Ens.(Paisitkriangkrai et al., 2015)	-	-	-	-	62.10	87.81	92.30	97.20
FVdeepLDA (Wu et al., 2017)	-	-	-	-	62.23	89.95	92.73	97.55
PersonNet (Wu et al., 2016a)	-	-	-	-	64.80	89.40	94.92	98.20
IDE(C)+XQDA (Zhong et al., 2017)	58.90	-	-	-	61.70	-	-	-
IDE(C)+XQDA+re (Zhong et al., 2017)	58.50	-	-	-	61.60	-	-	-
Ours (GFV+LGFV)	58.97	88.08	94.26	98.84	62.38	91.82	96.95	99.88
Ours (GFV+LGFV+IDE(C))	63.60	91.59	96.71	99.73	67.77	95.05	99.78	100
Ours (GFV+LGFV+IDE(C))+re	65.88	94.78	99.61	100	69.97	98.05	100	100

Besides, while comparing the deep hand-crafted representation (GFV + LGFV) with the high-level deep CNN descriptors IDE(C) we notably notice that the propsed deep hand-crafted features achieve better results than the deep CNN ones. Finally, when combining the deep hand-crafted and the CNN features, the accuracy increasingly rises proving thus the complementarity of these latter features. Indeed, the combination provide two different views of the re-ID process since they exploit the structural arrangements among the hand-crafted features and the raw pixels, respectively.

3.4.3 Comparison with the State-of-the-art Methods on Market-1501 Dataset

First, we find out in Table 2 that the proposed unsupervised GFV_U+ LGFV_U method outperforms all the recent state-of-the-art unsupervised methods (Farenzena et al., 2010; Zheng et al., 2015; Zhao et al., 2013) in the literature, thus proving thus the forcefulness of the proposed multi-level representation. Moreover, the proposed supervised GFV method gives a better result than (Wu et al., 2016a; Liao et al., 2015; Zheng et al., 2015; Chen et al., 2016; Zhong et al., 2017; Zhang et al., 2016) in the Market-1501 dataset. For example, GFV obviously outperforms the deep PersonNet method (Wu et al., 2016a) (r1=58.85% and mAP=9.78% versus r1=37.21% and mAP=18.57%). This can be due to the verification model employed in (Wu et al., 2016a) that does not take into consideration the similarity context i.e. ignores the intra-similarity and inter-similarity among the different identities. GFV also outperforms the FVdeepLDA (Wu et al., 2017) which passes on the FVs as inputs to a deep LDA metric learning and SCSP (Chen et al., 2016) which imposes spatial constraints on the pedestrian image. This proves the importance of the localization of the pedestrian position in the image. Otherwise, the proposed deep hand-crafted representation (GFV+LGFV) achieves a rank-1= 71.92% and mAP=46.45%. It considerably surpass the shallow representation methods NS(LOMO) and NS(fusion) (Zhang et al., 2016). Besides, (GFV+LGFV) gives better rank-1 matching rates than the method proposed in (Zhong et al., 2017) that uses the IDE deep CNN feature, the XQDA and the re-ranking technique, but the latter slightly achieves a higher mAP (46.79%). Indeed, the re-ranking scheme has also brought a considerable improvement to the mAP obtained by the proposed method. Moreover, (GFV + LGFV) notably surpasses NS-CNN (Li et al., 2016) which consists in the fusion of deep CNN features and low-level features (LBP, HOG, CN and LOMO) and achieves comparable results with the SCNN (Varior et al., 2016) which uses deep CNN features.

This proves that deep hand-crafted features can be a good alternative to the deep learning without needing neither pre-training neither data augmentation, at a lower training computational cost. When combining to the hand-crafted (GFV+ LGFV) with the deep CNN features (IDE(C)), we considerably outperform all state-of-the-art methods on Market-1501 dataset (rank-1=75.18% and mAP= 53.88%). This proves the complementarity of these two representations.

3.4.4 Comparison with the State-of-the-art Methods on CUHK03 Dataset

As shown in Table 3, the proposed unsupervised GFV_U + LGFV_U is compared with the unsupervised state-of-the-art methods (Farenzena et al., 2010; Zhao et al., 2013; Zheng et al., 2015) on CUHK03. Also, the proposed supervised (GFV + LGFV) method achieves better results than the supervised state-of-the-art methods (Sun and Chen, 2011; Davis et al., 2007; Köstinger et al., 2012; Li et al., 2014; Ahmed et al., 2015; Liao et al., 2015; Zhang et al., 2016). This good performance is due to the effectiveness of the proposed deep hand-crafted features. We remark that the proposed (GFV + LGFV) achieves competitive results with NS(fusion) (Zhang et al., 2016), Metric Ensembles (Paisitkriangkrai et al., 2015), FVdeepLDA (Wu et al., 2017) and PersonNet (Wu et al., 2016a). This can be explained by the small number of training data provided in this dataset, which can have a bad effect on the GMM learning process. In revenge, the proposed multi-level representation (GFV + LGFV + IDE(C)) achieves higher re-ID accuracy (in all ranks for the Labelled and Detected dataset, respectively) with respect to all the state-of-art methods, and specifically outperforms the re-ranking (IDE(C) + XQDA + re) method (Zhong et al., 2017). This proves the complementarity power of both the deep hand-crafted and deep CNN features.

4 CONCLUSION

In this paper, we proposed a new deep hand-crafted feature that exploit the richness of the spatial structural information The proposed deep hand-crafted feature is competitive with the deep CNN features without needing neither pre-training nor data augmentation. Moreover, we designed a multi-level feature representation built upon the complementary deep hand-crafted and deep CNN features. The experimental results in the two benchmarking image datasets (Market-1501 and CUHK03), have shown the good performances of the proposed methods. In feature research, we will investigate more sophisticated deep CNN architectures.

REFERENCES

Ahmed, E., Jones, M. J., and Marks, T. K. (2015). An improved deep learning architecture for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3908–3916.

Arandjelovic, R. and Zisserman, A. (2012). Multiple queries for large scale specific object retrieval. In *British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012*, pages 1–11.

Bouchrika, T., Zaied, M., Jemai, O., and C. Ben Amar (2014). Neural solutions to interact with computers by hand gesture recognition. *Multimedia Tools Appl.*, 72(3):2949–2975.

Boughrara, H., Chtourou, M., and Ben Amar, C. (2017). MLP neural network using constructive training algorithm: application to face recognition and facial expression recognition. *IJISTA*, 16(1):53–79.

Chen, D., Yuan, Z., Chen, B., and Zheng, N. (2016). Similarity learning with spatial constraints for person re-identification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1268–1277.

Dammak, M., Mejdoub, M., and Amar, C. B. (2014a). Extended laplacian sparse coding for image categorization. In *Neural Information Processing - 21st International Conference, ICONIP 2014, Kuching, Malaysia, November 3-6, 2014. Proceedings, Part III*, pages 292–299.

Dammak, M., Mejdoub, M., and Amar, C. B. (2015). Histogram of dense subgraphs for image representation. *IET Image Processing*, 9(3):184–191.

Dammak, M., Mejdoub, M., and Ben Amar, C. (2014b). Laplacian tensor sparse coding for image categorization. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, pages 3572–3576.

Dammak, M., Mejdoub, M., and Ben Amar, C. (2014c). A survey of extended methods to the bag of visual words for image categorization and retrieval. In *VISAPP 2014 - Proceedings of the 9th International Conference on Computer Vision Theory and Applications, Volume 2, Lisbon, Portugal, 5-8 January, 2014*, pages 676–683.

Davis, J. V., Kulis, B., Jain, P., Sra, S., and Dhillon, I. S. (2007). Information-theoretic metric learning. In *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, pages 209–216.

Farenzena, M., Bazzani, L., Perina, A., Murino, V., and Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In

The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010, pages 2360–2367.

Guo, Y., Wu, L., Lu, H., Feng, Z., and Xue, X. (2006). Null foley-sammon transform. *Pattern Recognition*, 39(11):2248–2251.

Jobson, D. J., Rahman, Z., and Woodell, G. A. (1997). A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Trans. Image Processing*, 6(7):965–976.

Köstinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., and Bischof, H. (2012). Large scale metric learning from equivalence constraints. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 2288–2295.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'12, pages 1097–1105, USA. Curran Associates Inc.

Ksibi, A., Dammak, M., Ammar, A. B., Mejdoub, M., and Amar, C. B. (2012). Flickr-based semantic context to refine automatic photo annotation. In *3rd International Conference on Image Processing Theory Tools and Applications, IPTA 2012, 15-18 October 2012, Istanbul, Turkey*, pages 377–382.

Ksibi, S., Mejdoub, M., and Ben Amar, C. (2016a). Extended fisher vector encoding for person re-identification. In *2016 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2016, Budapest, Hungary, October 9-12, 2016*, pages 4344–4349.

Ksibi, S., Mejdoub, M., and Ben Amar, C. (2016b). Person re-identification based on combined gaussian weighted fisher vectors. In *13th IEEE/ACS International Conference of Computer Systems and Applications, AICCSA 2016, Agadir, Morocco, November 29 - December 2, 2016*, pages 1–8.

Ksibi, S., Mejdoub, M., and Ben Amar, C. (2016c). Topological weighted fisher vectors for person re-identification. In *23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, December 4-8, 2016*, pages 3097–3102.

Kuo, C.-H., Khamis, S., and Shet, V. D. (2013). Person re-identification using semantic color names and rankboost. In *IEEE Workshop on Applications of Computer Vision*, pages 281–287.

Li, J., Yang, Z., and Xiong, H. (2015). Encoding the regional features for person re-identification using locality-constrained linear coding. In *2015 International Conference on Computers, Communications, and Systems (ICCCS)*, pages 178–181.

Li, S., Liu, X., Liu, W., Ma, H., and Zhang, H. (2016). A discriminative null space based deep learning approach for person re-identification. In *4th International Conference on Cloud Computing and Intelligence Systems, CCIS 2016, Beijing, China, August 17-19, 2016*, pages 480–484.

Li, W., Zhao, R., Xiao, T., and Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 152–159.

Li, Z., Chang, S., Liang, F., Huang, T. S., Cao, L., and Smith, J. R. (2013). Learning locally-adaptive decision functions for person verification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 3610–3617.

Liao, S., Hu, Y., Zhu, X., and Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 2197–2206.

Liu, K., Ma, B., Zhang, W., and Huang, R. (2015). A spatio-temporal appearance representation for video-based pedestrian re-identification. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3810–3818.

Ma, B., Su, Y., and Jurie, F. (2012). Local descriptors encoded by fisher vectors for person re-identification. In *ECCV Workshops*, volume 7583, pages 413–422.

Mahmoud Mejdoub, Salma Ksibi, C. r. B. A. and Koubaa, M. (2017). Person re-id while crossing different cameras: Combination of salient-gaussian weighted bossanova and fisher vector encodings. *International Journal of Advanced Computer Science and Applications(IJACSA)*, 8(9):399–410.

Mejdoub, M., Dammak, M., and Ben Amar, C. (2015a). Extending laplacian sparse coding by the incorporation of the image spatial context. *Neurocomputing*, 166:44–52.

Mejdoub, M., Fonteles, L. H., Ben Amar, C., and Antonini, M. (2008). Fast indexing method for image retrieval using tree-structured lattices. In *International Workshop on Content-Based Multimedia Indexing, CBMI 2008, London, UK, June 18-20, 2008*, pages 365–372.

Mejdoub, M., Fonteles, L. H., Ben Amar, C., and Antonini, M. (2009). Embedded lattices tree: An efficient indexing scheme for content based retrieval on image databases. *J. Visual Communication and Image Representation*, 20(2):145–156.

Mejdoub, M. and Ben Amar, C. (2013). Classification improvement of local feature vectors over the KNN algorithm. *Multimedia Tools Appl.*, 64(1):197–218.

Mejdoub, M., Ben Aoun, N., and Ben Amar, C. (2015b). Bag of frequent subgraphs approach for image classification. *Intell. Data Anal.*, 19(1):75–88.

Messelodi, S. and Modena, C. M. (2015). Boosting fisher vector based scoring functions for person re-identification. *Image Vision Comput.*, 44:44–58.

Othmani, M., Bellil, W., Ben Amar, C., and Alimi, A. M. (2010). A new structure and training procedure for multi-mother wavelet networks. *IJWMIP*, 8(1):149–175.

Paisitkriangkrai, S., Shen, C., and van den Hengel, A. (2015). Learning to rank in person re-identification with metric ensembles. *CoRR*, abs/1503.01543.

Peng, X., Zou, C., Qiao, Y., and Peng, Q. (2014). Action recognition with stacked fisher vectors. In *European Conference on Computer Vision (ECCV)*, pages 581–595.

Ben Aoun, N., Mejdoub, M., and Ben Amar, C. (2014). Graph-based approach for human action recognition using spatio-temporal features. *J. Visual Communication and Image Representation*, 25(2):329–338.

M. El Arbi, Koubàa, M., Charfeddine, M., and Ben Amar, C. (2011). A dynamic video watermarking algorithm in fast motion areas in the wavelet domain. *Multimedia Tools Appl.*, 55(3):579–600.

M. El Arbi, C. Ben Amar, and Nicolas, H. (2006). Video watermarking algorithm based on neural network. In *IEEE International Conference on Multimedia and Expo (ICME 2006), Toronto Ontario, Canada, July 9-12, 2006*, pages 1577–1580.

Sapienza, M., Cuzzolin, F., and Torr, P. H. S. (2014). Learning discriminative space-time action parts from weakly labelled videos. *International Journal of Computer Vision*, 110(1):30–47.

Sekma, M., Mejdoub, M., and Amar, C. B. (2014). Spatio-temporal pyramidal accordion representation for human action recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, pages 1270–1274.

Sekma, M., Mejdoub, M., and Ben Amar, C. (2015a). Human action recognition based on multi-layer fisher vector encoding method. *Pattern Recognition Letters*, 65:37–43.

Sekma, M., Mejdoub, M., and Ben Amar, C. (2015b). Structured fisher vector encoding method for human action recognition. In *15th International Conference on Intelligent Systems Design and Applications, ISDA 2015, Marrakech, Morocco, December 14-16, 2015*, pages 642–647.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep fisher networks for large-scale image classification. In *Advances in Neural Information Processing Systems (NIPS)*, pages 163–171.

Sun, S. and Chen, Q. (2011). Hierarchical distance metric learning for large margin nearest neighbor classification. *IJPRAI*, 25(7):1073–1087.

Varior, R. R., Haloi, M., and Wang, G. (2016). Gated siamese convolutional neural network architecture for human re-identification. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 791–808.

Wali, A., N. Ben Aoun, Karray, H., Ben Amar, C., and Alimi, A. M. (2010). A new system for event detection from video surveillance sequences. In *Advanced Concepts for Intelligent Vision Systems - 12th International Conference, ACIVS 2010, Sydney, Australia, December 13-16, 2010, Proceedings, Part II*, pages 110–120.

Wu, L., Shen, C., and van den Hengel, A. (2016a). Personnet: Person re-identification with deep convolutional neural networks. *CoRR*, abs/1601.07255.

Wu, L., Shen, C., and van den Hengel, A. (2017). Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification. *Pattern Recognition*, 65:238–250.

Wu, S., Chen, Y., Li, X., Wu, A., You, J., and Zheng, W. (2016b). An enhanced deep feature representation for person re-identification. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7-10, 2016*, pages 1–8.

Xiong, F., Gou, M., Camps, O. I., and Sznaier, M. (2014). Person re-identification using kernel-based metric learning methods. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII*, pages 1–16.

Zhang, L., Xiang, T., and Gong, S. (2016). Learning a discriminative null space for person re-identification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1239–1248.

Zhao, R., Ouyang, W., and Wang, X. (2013). Unsupervised salience learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593.

Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., and Tian, Q. (2016). Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*. Springer.

Zheng, L., Shen, L., Tian, L., Wang, S., Bu, J., and Tian, Q. (2015). Person re-identification meets image search. In *CoRR*, volume abs/1502.02171, pages 2360–2367.

Zhong, Z., Zheng, L., Cao, D., and Li, S. (2017). Re-ranking person re-identification with k-reciprocal encoding. *CoRR*, abs/1701.08398.