Person re-ID while Crossing Different Cameras: Combination of Salient-Gaussian Weighted BossaNova and Fisher Vector Encodings

Mahmoud Mejdoub
Department of Computer Science,
College of AlGhat,
Majmaah University, 11952
Riyadh, KSA

Salma Ksibi and Chokri Ben Amar
REGIM: Research Groups
on Intelligent Machines,
University of Sfax,
ENIS, Tunisia

Mohamed Koubaa
Computer Science
Shaqra University,
Riyadh, KSA

***Abstract*—Person re-identification (re-ID) is a challenging task in the camera surveillance field, since it addresses the problem of re-identifying people across multiple non-overlapping cameras. Most of existing approaches have been concentrated on: 1) achieving a robust and effective feature representation; and 2) enforcing discriminative metric learning to predict if two images represent the same identity. In this context, we present a new approach for person re-ID built upon multi-level descriptors. This is achieved by combining three complementary representations: salient-Gaussian Fisher Vector (SGFV) encoding method, salient-Gaussian BossaNova (SGBN) histogram encoding method and deep Convolutional Neural Network (CNN) features. The two first methods adapt the histogram encoding framework to the person re-ID task. This is achieved by integrating the pedestrian saliency map and the spatial location information, in the histogram encoding process. On one hand, human saliency is reliable and distinctive in the person re-ID task, since it can model the uniqueness of the identity. On the other hand, localizing a person in the image can effectively discard noisy background information. Finally, one of the most advanced metric learning in person re-ID: the Cross-view Quadratic Discriminant Analysis (XQDA) is applied on the top of the resulting description. The proposed method yields promising person re-ID results on two challenging image-based person re-ID benchmarks: CUHK03 and Market-1501.**

***Keywords*—Person re-identification; histogram encoding; fisher vector; BossaNova; Convolutional Neural Network (CNN); salient weight; Gaussian weight**

I. Introduction

The person re-identification (re-ID) [1, 2, 3, 4, 5, 6, 7, 8, 9] goal is to retrieve gallery images containing the same person as the probe (query) in a video surveillance [10, 11] cross-camera mode. In general, images containing human subjects who are pre-captured by a detector or a tracker, form the input of a person re-ID system. Though biometrics are effective to identify a person and specially the face, these are not always available in the person re-ID filed. This is mainly due to the low resolution and pose variation problems faced in this domain. Therefore, in such situations, the solution was to focus on the body appearance features, especially the colour ones, supposing that each identity keeps the same clothes while switching from a camera to another. Indeed, actual person re-ID works are divided according to two main categories: shallow and deep methods. Shallow methods are specifically based on the appearance hand-crafted features [1, 2, 3, 4, 5, 6, 7, 12, 13, 14, 15]. In this context, two types of features are distinguished: low-level as well as mid-level ones. Low-level features may be global or local. Global features such as the Local Maximal Occurrence (LOMO) [16] measure the holistic appearance characteristics from the whole person's image. Local feature based approaches [2, 5, 6, 17, 18] are those in which characteristics are extracted from small regions of interest. These are proven to be effective in the person re-ID task. However many of them [2, 17, 18] rely on brute-force feature matching technique, which can badly influence the retrieval efficiency. In eSDC [17] and SalMatch [18], the salience information has been investigated for person re-ID, leading to an improved discriminative representation. Indeed, the saliency-based methods take into account the contextual information present in the feature space to derive the saliency information. More specifically, salience means distinct features that are discriminative and unique in the feature space. The pedestrian retrieval integrates then the saliency weights into the brute-force sequential matching between the patches of the query and the gallery images. Despite the good recognition rate, the expensive computational cost of the brute-force matching limits the potential application of these saliency-based methods in large-scale person re-ID datasets. Hence, the proposed work is motivated by studying the ability of incorporating the saliency information into the histogram encoding technique to replace efficiently the brute-force matching scheme. Regarding the mid-level features, they are extracted in the person re-ID field, by applying a histogram encoding method such as Bag of visual Words (BOW) [3] and Fisher Vector (FV) [4, 19, 20, 21, 22] on the local features. This leads to the quantification of the local features into a set of visual words that form a codebook. Then, in the pooling step, the visual words are aggregated to generate the final histogram representation. Indeed, our work is specially motivated by studying the effectiveness of the histogram encoding methods [23] in the person re-ID task, since they can simplify the matching process between persons, with respect to the brute force matching methods. Bag Of Statistical Sampling Analysis BossaNova (BN) [24, 25] is a pooling method that demonstrated its effectiveness in the image classification field [11, 26, 27, 28, 29, 30, 31, 32, 33, 34]. It consists in discretizing the patches to clusters' assignments into several Bins.

For each Bin in a given cluster, the discretized assignments are sum-pooled over the patches. This considerably improves the pooling operation, since this latter is performed by taking into account the local distributions of the features around each cluster.

Regarding the second category of methods i.e the deep CNN learning methods, it was stated in [35], that the ID-discriminative Embedding (IDE) feature performs better than the previously used verification models [36, 37, 38]. Therefore, the IDE feature is adopted in this paper. IDE feature is obtained by learning a discriminative embedding in a classification mode. The learned model is obtained by categorizing the training features into pre-defined identity classes. The IDE feature generated by the last convolutional layer is used for pedestrian matching.

To enhance the discrimination of the features, supervised metric learning, such as Keep It Simple and Straightforward Metric Learning (KISSME) [39], locally adaptive decision functions (LADF) [40], the Null space (NS) metric learning [41], and the Cross-view Quadratic Discriminant Analysis (XQDA) [16] 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 [35]. 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.

In this paper, we propose to encode colour descriptors throughout a rich histogram representation well adapted to the person re-ID field. In this sense, two extensions of the traditional FV and BN encoding methods are introduced (see Fig. 1), namely, the Salient-Gaussian FV (SGFV) and the Salient-Gaussian BN (SGBN). This consists in weighting the histogram encoding process via the Gaussian and the saliency weights. The injected weights take into consideration two important aspects in person re-ID: the elimination of the background noise around the pedestrian via the Gaussian weight [42], and the emphasize on the salient regions in the image. The Gaussian weighting is related to the pedestrian spatial location in the image. Indeed, it fosters the locations that lie nearby the pedestrian in the image. The saliency weighting is inspired from the patch-to-patch brute-force matching method of [17] and adapted to our work in the case of the histogram encoding scheme. It enhances the encoding process by highlighting meaningful parts of the images and eliminating needless ones. Specifically, SGFV and SGBN are applied on three low-level local colour features (Colour Name (CN), Colour Histogram (CHS) and 15-d descriptor). The resulted histograms are further combined with the deep CNN feature to provide a rich multi-level representation. Thus, we obtain seven histograms (see Fig. 1). Finally, the pedestrian are matched by combining the XQDA distances learned upon these seven histograms. It is worth mentioning that all images are pre-treated with Retinex transform [16], to reduce the illumination variation before the application of the encoding methods. Besides, SGFV and SGBN are applied upon a spatial stripe representational scheme in order to consider the spatial alignment information between pedestrian parts. Indeed, this work makes several contributions:

- We propose new Salient-Gaussian weighted histogram encoding methods (SGFV and SGBN), well adapted to the person re-ID task, since they take into account the location of the pedestrian in the image as well as its uniqueness. Gaussian and saliency weights respectively remove background clutters surrounding the person silhouette and, emphasize the most distinctive regions in the person images in order to highlight the uniqueness of each pedestrian. Also, to the best of our knowledge, we are the first that apply the BN encoding framework in the context of person re-ID.
- We propose to combine the two mid-level representations SGFV and SGBN, and the high-level IDE representation captured by the deep CNN, taking profit from their complementarity. Indeed, the SGFV and SGBN representations are complementary. This complementarity is actually due to two facts: 1) On one hand, FV may lack locality during pooling, whereas BN does not since it handles the local distribution of the descriptors around each cluster. 2) On the other hand, FV is more accurate during the coding step than BN, since it provides higher high order statics about the difference between the low-level descriptors and the GMM components. Besides, the mid-level representation is complementary with the semantic one depicted by the IDE CNN feature. This rich multi-level representation is fed into one of the advanced metric learnings: XQDA [16], which considerably enhances the re-ID accuracy.

The paper is structured as follows:

After introducing the context of our work in Section I, and presenting the related state-of-art methods in Section II, the proposed approach is described in Section III. In this context, the Retinex method is first introduced (Subsection III-A), to deal with the illumination variation. Then, the low-level features (Subsection III-B) used in this work are further presented. Afterwards, we describe the proposed weighted histogram encoding scheme (Subsection III-C). We further explain how we computed the proposed histograms based on the pedestrian image partition (Subsection III-D1). The deep CNN features adapted in this paper are next presented (Subsection III-D2). After that, the combination of the two proposed complementary weighted histogram encoding methods between two pedestrian images via the XQDA distances is explained in details (Subsection III-E), and present the Multi Query process devoted in this work (Subsection III-F). Finally, the originality of the proposed method is justified in the Section IV via the promising experimental results obtained on two challenging benchmarking datasets.

II. Related Works

A. Histogram Encoding Methods

The Bag of visual Words (BOW) model [9, 43] has been used for person re-ID in several state-of-art works [3, 44]. In [44], authors built groups of descriptors by integrating the visual words into concentric spatial structures and by enriching the BOW description of a person by the contextual information coming from people that surround it. Moreover, in [3], L. Zheng et al. have designed an unsupervised BOW representation. In order to include geometric constraints, they

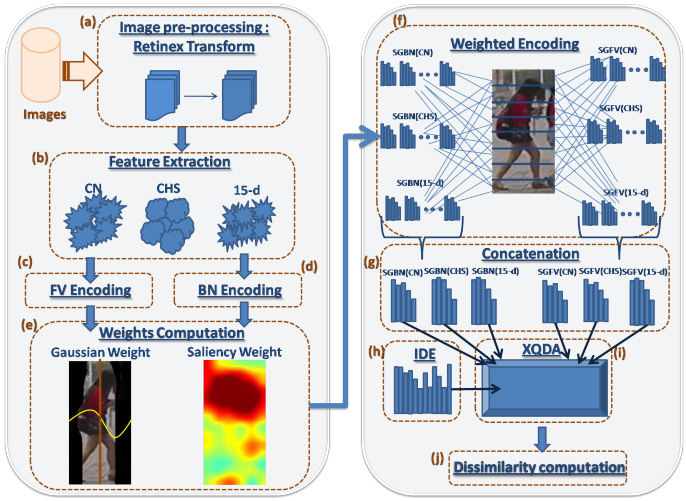

Fig. 1. Overview of the proposed method pipeline. (a) Image pre-processing through Retinex Transform to deal with the illumination variation problem. (b) Low-level feature extraction: CN, CHS and 15-d descriptors. (c) Fisher Vector (FV) Encoding. (d) BossaNova (BN) Encoding. (e) Computation of the Gaussian and the saliency weights. (f) Salient-Gaussien weighted FV (SGFV) and Salient-Gaussian weighted BN (SGBN) are calculated separately on each given descriptor (CN, CHS and 15-d), at each stripe. (g) The generated histograms are concatenated over the stripes producing the final SGFV and SGBN representations (one histogram per low-level feature). (h) IDE features. (i) Computation of the XQDA distances separately on each generated histogram and on the IDE deep CNN feature. (j) Dissimilarity computation: combination of the XQDA distances learned from the seven histograms.

incorporated the Spatial Pyramid Representation (SPR) [16, 45] into the BOW model. They used the colour Names (CN) and HS Histogram (HS) descriptors and employed Multiple Assignment (MA) for each descriptor. They also applied a Gaussian mask in order to remove the noisy background of the person image. Locality-constrained linear coding (LLC) [46] is a sparse encoding method [47, 48, 49] that provides a soft quantified histogram. It considers the locality information in the feature encoding process by taking into account only the k-nearest basis vectors from each local feature. Indeed, Z. Yang et al. have proposed a colour histogram based on LLC feature representation for person re-ID. They integrated LLC and colour histogram and employed SPR to reflect global geometric constraints. They also tested the performance of their method while comparing the performances of five different features with five different metric learning methods. Fisher Vector (FV) [4, 20, 50, 51] is another encoding method representation that learns a Gaussian Mixture model (GMM) model on the local descriptors in order to compute the visual words. It was applied in several person re-ID methods. B. Ma et al. [4] were the first to introduce the FV scheme in person re-ID task. In fact, they employed a spatial representation that divides the pedestrian image into 4×3 fixed regions and used a new very simple 7-d local descriptor. These local descriptors are turned into FVs and these latter are employed to measure the similarity between two persons using the Euclidean distance between their representations. In [51], the authors introduced a boosting method that learns a scoring function taking into account the likelihood between the local Fisher vectors of the same identity.

In the context of image classification, BN [52, 53] extends the BOW method by applying a richer pooling operation. It enhances considerably the traditional sum pooling operated in BOW. While BOW compacts all information related to a visual word into a single scalar, BN operates a more significant statistical analysis that estimates the distribution of the features around each visual word. Compared to these methods, this work presents a histogram encoding method better suited to the person re-ID filed. In this regard, the proposed saliency and Gaussian weighting allows to improve the encoding process by focusing on the discriminative parts of the pedestrian. In our earlier works [5, 6], we proposed an unsupervised version of the weighted FV encoding for the image-based person re-ID task. [5] proposed a Gaussian weighted encoding FV version with Retinex transform and the combination of CN, CHS and 15-d low-level features, while [6] introduced a salient weighted FV version with CN and CHS features. With respect to our earlier works [5, 6], a supervised combined Salient-Gaussian weighted histogram encoding based both on FV and BN is presented in this paper.

B. Deep CNN Learning

CNN-based deep learning models have been popular and shown great success in many fields [54, 55, 56]. Nevertheless, the study of the CNN model has started only recently in the person re-ID task [7, 35, 50, 55, 57, 58, 59] and that is due to the small scale of the existent re-ID datasets. Verification models treat person re-ID as a two-class recognition task, by taking a pair of images as input and determining whether they belong to the same person or not. Indeed, image pairs [36, 37] or triplets [60] are passed as input to CNN, rather than single training images. In this way, the training set is enlarged and the shortage of the training images is avoided. Yet, the most recent large scale datasets (e.g., Market-1501 [3]) provide richer training samples per class. In this sense, it was shown in [56] that the classification model performs better than a verification one in large scale datasets, since it can exploit more adequately the correlation between the pedestrians. Therefore, we chose to train a classification CNN model in this paper. In [35, 61], the IDE feature is also extracted in a classification mode.

In [62, 63], low-level hand-crafted features are combined with high-level CNN features. Afterwards, metric learning is applied on the obtained combination. In [62], the CNN model is first learned by adding a fusion layer that combines CNN features with the hand-crafted low-level ELF [64] features. Afterwards, the high-level resulting feature is concatenated with LOMO, and subsequently presented to the KMFA [65] metric learning, which considerably boosts the re-ID accuracy. The good re-ID results obtained by the concatenation between low-level hand-crafted features and CNN ones provide support on their complementary nature. However, mid-level features generated by histogram encoding methods provide richer information than low-level hand-crafted features. In this sense, in this work, the discriminative power of the mid-level features is exploited to further boost the complementary aspect of both hand-crafted and CNN features. While for the Discriminative Null Space based Deep Learning Approach Deep Learning approach [7], authors adapted NFST metric learning approach to their method and combined low-level, mid-level and high-level features, all learned by SCNN in a new discriminative null space.

III. Proposed Method

A. Dealing with Illumination Variations

In this paper, the Multi-scale Retinex transform with colour Restoration (MSRCR) [66] is used in order to handle the 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 colour 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 colour restoration function. In the experiments, two scales of the Gaussian surround function are used ($\sigma = 5$ and $\sigma = 20$).

B. 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 colour low-level descriptors are extracted (CN, CHS and 15-d). These latter ones are chosen underlying their good compromise between efficiency and re-ID accuracy [3, 4]. Indeed, their small dimensionality as compared to other state of the art descriptors such as the global LOMO descriptor [16] and the local dColourSift one [17] makes them suitable to the efficiency factor required by person re-ID task.

1) Colour names (CN): The authors in [13] demonstrate that colour description based on colour names has a good robustness against photometric variance. In this paper, as is done in [13], we use the 11 basic colour terms of the English language: black, blue, brown, grey, green, orange, pink, purple, red, white, and yellow. First, the CN feature vector of each pixel is calculated by performing a mapping from the HSV pixel value to a 11 dimensional colour names vector. Afterwards, a sum pooling is applied on the CN pixel features related to each patch. Finally, the resulting histogram undergoes a square rooting operation followed b $l1$ normalization. The size of the generated CN descriptor is then equal to 11.

2) 15-d Descriptor: Inspired by [4], a simple 15-d descriptor is designed. First, the pedestrian image is split into 3 colour 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, a sum-pooling operation is applied for each colour channel, 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, the three normalized descriptors are horizontally concatenated into one single signature.

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

C. Proposed Salient-Gaussian Histogram Encoding Methods

1) Saliency extraction: In [17] the authors proposed a saliency model that we adopt in this work due to its capability to boost the discriminative power of the person re-ID. The assumption behind the saliency computation is that a patch derived from an input image is ascribed to a high saliency if a great number of people in the training set do not share similar patches with it. Salient patches are therefore defined as those that possess property of uniqueness among a reference set taken from the learning set. Consider $p_{w,h}$ the patch whose spatial centre 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, the input image to the saliency extractor, and R the reference set that corresponds to the Nr training images. For the saliency extractor input image, a nearest neighbour set of size Nr is built for every patch $p_{h,w}$. This is carried out by searching for the most similar patch to $p_{h,w}$ in every v-th reference image in the training set. When seeking a patch $p_{h,w}$ in the v-th training image $I^{R,v}$, the search space is restricted to the adjacency set $S(p_{h,w}, I^{R,v})$ (see (2)). This latter one corresponds to the horizontal region centred on the h-th row. This is established in order to avoid misalignment (which manifestly occurs on the horizontal direction) and relax the search space.

$$S(p_{h,w}, I^{R,v}) = \left\{ p_{i,j}^{R,v}, i \in \Delta(h), j = 1 \dots W \right\} \quad (2)$$

where $\Delta(h) = \{max(0, h - l), \dots, h, \dots, min(h + l, H)\}$. The parameter l defines the width of the adjacency set. Thus, we compute for each input patch $p_{h,w}$, the matching set $XNN(p_{h,w})$ defined by (3):

$$\begin{aligned} XNN(p_{h,w}) = \{ & p_{i,j}^{R,v} | \underset{p_{i,j}^{R,v}}{\text{argmin}}\, dist(p_{h,w}, p_{i,j}^{R,v}), \\ & p_{i,j}^{R,v} \in S(p_{h,w}, I^{R,v}), v = 1 \dots N_r \} \end{aligned} \quad (3)$$

where $dist(\cdot)$ is the Euclidean distance between two patch features. Afterwards, the computed matching set is used to define the patch salience score as the distance to the k-th nearest neighbour denoted by $dist_k$ ($k = \alpha Nr$):

$$score(p_{h,w}) = dist_k(XNN(p_{h,w})) \quad (4)$$

and the saliency weight of $p_{h,w}$:

$$S(p_{h,w}) = 1 - \exp(-score(p_{h,w})^2 / \sigma_0^2) \quad (5)$$

where σ_0 is a salient scores' bandwidth parameter. Actually, the higher the patch saliency weight, the more discriminative it is. As achieved in [17], we set $k = \alpha Nr$ with $\alpha = 1/2$ in the salience learning scheme with an empirical assumption

that a patch is considered to have a special appearance when more than half of the people in the reference set do not share similar patches with it. To build the saliency map, [17] relies on high dimensional local dColorSIFT features (672 dimensions) to describe the patches. In this work, to reduce the saliency extraction computation time, the stacking of the three descriptors CN, CHS and 15-d is use as patch low-level feature, that corresponds to a total size of 74 dimensions. Besides, the patch-to-patch matching is computed via an approximate nearest neighbour (ANN) algorithm (we use as ANN the randomized best-bin-first KD-tree forest introduced in [67]). In the saliency computation, for CUHK03 and Market-1501 datasets, one pedestrian image is picked from each camera and the reference set is built from the training images belonging to the other cameras.

2) Background noise elimination: In [2], 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. In this paper a simple solution is proposed by employing a 2-D Gaussian template on the pedestrian image, in order to remove the noisy background. Inspired by [2, 3], 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 allows to discard the noise surrounding person's silhouette, and thus to keep meaningful parts of the images and eliminate needless ones. Explicitly, each patch $p_{h,w}$ is endowed with a Gaussian weight $G(p_{h,w})$, given by:

$$G(p_{h,w}) = exp(-(w - \mu_x)^2/2\sigma^2) \tag{6}$$

3) Proposed Salient-Gaussian weighted BN (SGBN) encoding method: BoSSA [52, 53] (Bag Of Statistical Sampling Analysis) extends the BOW method, by applying a richer pooling operation that robustly integrate the feature space locality information. Specifically, a local histogram of B bins is computed by quantifying the distances between each visual word and the local features assigned to it. Toward this end, the average number of cluster assigned features, whose quantified distances fall into a given range, is counted. The local histograms are then horizontally concatenated over all visual words. The resulting vector is further $l1$ normalized and combined with the traditional BOW histogram. BoSSA is further extended to BN [53] replacing the $l1$ normalization by the power-$l2$ one and the hard quantization by a soft one. Hereinafter, the proposed adaptation of the BN representation for person re-ID is introduced. This adaptation consists in weighting the classical representation by both Gaussian and saliency weights. As explained in subsection III-C2, the Gaussian weight is taken into account the spatial location of the person in the image. Indeed, high Gaussian weights are accorded for patches located nearby the center of the image and low Gaussian weights are accorded for those that are far. Regarding the saliency weighting, as described in subsection III-C1, this latter one aims to emphasize the more discriminant and significant parts in the image. For concise clarity, we omit hereinafter the patch index (h, w) used in the previously notation (subsections III-C1 and III-C2), that will be replaced by i. Thus, the saliency and Gaussian weights of the image patch p_i are noted S_i and G_i, respectively. Consider M local descriptors d_i corresponding to the M patches p_i of an image I. The assignment $a_{i,k}$ of d_i to the k-th cluster, by the soft BOW, is given by:

$$a_{i,k} = \frac{exp - \beta_k \times dis(d_i, c_k)}{\sum_{k'=1}^{K} exp - \beta_{k'} \times dis(d_i, c_{k'})} \tag{7}$$

where c_k is its k-th closet codeword, $dis(d_i, c_k)$ the Euclidean distance between c_k and d_i, β_k is a parameter that regulates the softness of the assignment, i.e. the bigger it is, the hardest is the assignment. In fact, β_k varies for each codeword c_k. It is given by the standard deviation of each cluster c_k: $\beta_k = \sigma_k^{-2}$. In the traditional BoSSA, the distribution of the feature-to-cluster assignments around each visual word c_k is computed by 1) discretizing, for each visual word c_k, each assignment $a_{i,k}$ $(1 \leq i \leq M)$ over B bins; and 2) computing the sum of assignments falling into each bin. To enhance the BoSSA pooling and to adapt it to the case of person re-ID, the sum of the assignments is weighted by the Gaussian and saliency weights. Thus, for each visual word c_k, a local histogram f_k^I is obtained where $f_{k,b}^I$ corresponds to the weighted sum of the cluster assignments $a_{i,k}$ that fall into the b^{th} bin. Formally, f_k^I can be expressed as follows:

$$f_{k,b}^I = \sum_i (G_i \times S_i \times a_{i,k},\ d_i \in I \ and \ a_{i,k} \in [r_1, r_2]) \tag{8}$$

where

$$r_1 = (v_k^{min} + s \times b)$$

and

$$r_2 = (v_k^{min} + s \times (b+1))$$

$b \in [0, \ldots, B-1]$, $v_k^{\min}$ and $v_k^{\max}$ limit the range of the activated clusters' weights $a_{i,k}$ over all descriptors d_i extracted from the images of the learning set. The step $s = \frac{v_k^{\max} - v_k^{\min}}{b}$ corresponds to the length of the bin. The final representation is given by:

$$f^I = \left[\left[f_{k,b}^I\right], o_k\right] \ ; (k,b) \in \{1, \ldots, K\} \times \{1, \ldots, B\} \tag{9}$$

where o_k corresponds to the Salient-Gaussian weighted BOW histogram, and it is computed as depicted by the following equation:

$$o_k = \sum_{i=1}^{M} G_i \times S_i \times a_{i,k} \tag{10}$$

$f_{k,b}^I$ and o_k separately undergoes then a power-$l2$ normalization. Indeed, the effect of power normalization is to smooth the sum pooled histogram to avoid the bad influence

of the frequent yet uninformative descriptors. The proposed SGBN encoding is applied separately to the three low-level descriptors: CN, CHS and 15-d descriptor with respective dimensions.

D. Proposed Salient Weighted Gaussian FV (SGFV) Encoding Method

In this paper, a rich extension of the traditional FV encoding method is proposed. It consists of the incorporation of the Gaussian and saliency weights in the encoding process of this latter. The construction of the proposed encoding starts, as operated in the traditional FV, by (1) learning a Gaussian Mixture model (GMM) model represented by K components, on the local descriptors extracted from all training pedestrian images, than by (2) computing the mixture weights, means, and diagonal covariance of the GMM respectively denoted as π_k, μ_k, σ_k. In a further step, the traditional FV encoding is weighted via both the saliency and Gaussian weights, as given by the following equation:

$$u_k = \frac{1}{M}\sum_{i=1}^{M} G_i \times S_i \times \alpha_k(d_i)\left(\frac{d_i - \mu_k}{\sigma_k}\right) \quad (11)$$

$$v_k = \frac{1}{M}\sum_{i=1}^{M} G_i \times S_i \times \alpha_k(d_i)\left(\frac{(d_i - \mu_k)^2}{\sigma_k^2} - 1\right) \quad (12)$$

where, $\alpha_k(di)$ is the soft assignment weight of the *i*-th descriptor d_i to the *k*-th Gaussian, G_i and S_i are respectively the Gaussian and the saliency weights. 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:

$$FV = [u_1 \dots u_K, v_1 \dots v_K] \quad (13)$$

Finally, a power normalization is applied to each FV component before normalizing them jointly. Note that, as performed in SGBN, the proposed SGFV encoding is applied separately to the three proposed low-level descriptors: CN, CHS and 15-d descriptor.

1) Histogram computation based on pedestrian image partition: In order to alleviate the misalignment caused by the pose variations problem in the person's images, appearance modelling typically exploits part-based body models to take into account the non-rigid shape of the human body and treat the appearance of different body parts independently [68, 69]. 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 which is as shown in [3] a good compromise between accuracy and efficiency. Each proposed histogram encoding method (SGFV or SGBN) is applied separately in every single stripe. Afterwards, histograms corresponding to each stripe are $l2$ normalized separately prior to stacking. As there are three low-level descriptors (CN, CHS and 15-d) for each histogram encoding method, six global histograms are obtained. Finally, every global histogram is further $l2$ normalized to ensure the linear separability of the data. The size of the final SGFV and SGBN representations for each low-level feature is given by $[2 \times K_{SGFV} \times dim_i \times N_s]\, and\, [\times(B+1) \times K_{SGBN} \times N_s]$, respectively, where K_{SGFV} and K_{SGBN} stand for the codebook size of SGFV and SGBN, respectively, N_s corresponds to the total number of stripes, and $dim_1 = 11$, $dim_2 = 48$ and $dim_2 = 15$ are the respective dimensions of CN, CHS and 15-d descriptors.

2) Deep CNN features: The Convolutional Neural Network (CNN) has achieved state-of-the-art accuracy a number of vision tasks. In person re-ID, the majority of current CNN methods uses a verification model [57]. This latter infers positive image pairs and negative ones as input to the CNN, owing to the lack of training data per pedestrian identity. However, the recognition accuracy is generally badly influenced by the absence of the intra-class similarity and inter-class dissimilarity information. In this paper, the ID-discriminative Embedding (IDE) feature introduced in [61] is employed to tackle the aforementioned problem of the verification model. Specifically, CaffeNet [54] and ResNet-50 [70] are used to train the CNN in 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, two globally connected layers each with $1,024$ neurons, and a fully connected classifier layer. The number of neurons in the final fully connected layer is defined by the number of training identities in each dataset. The deep residual ResNet-50 network is constituted by 5 convolutional blocks (conv1, conv2-x, conv3-x, conv4-x, conv5-x) and a classifier block. The conv5-x block ends with 2048 convolutional filters of size 1×1 each one. We note that the CNN model is pre-trained on ImageNet [54] dataset before fine-tuning on the target dataset (all the CNN layer weights are fine-tuned, while the classifier layer weights are trained from scratch). In testing phase, 1024 and 2048 dimensional CNN features are extracted, for each pedestrian image, throughout the 7-th layer of CaffeNet and the conv5-x block of ResNet-50, respectively. The CNN features are then subsequently $l2$ normalized.

E. Dissimilarity Computation

After the generation of the six global histograms based on SGFV and SGBN, as well as the IDE feature, an XQDA [16] distance is learned separately on each histogram, in a supervised way. Next, the obtained distances are summed up forming what we called the dissimilarity score, in order to combine the corresponding histograms. Indeed, 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. XQDA metric learning is adopted in this work, since it has shown good compromise between efficiency and accuracy in many works [16, 35]. It is worth mentioning that, as performed in [16], the eigenvectors corresponding to the eigenvalues of $S_w^{-1} - S_b$ that are larger than 1 are selected as subspace components, where S_w and S_b design the within and the between scatter matrices, respectively.

F. Multiple Queries

The usage of multiple queries is shown to yield superior results in image search [71] and re-ID [2]. When each identity has multiple queries in a single camera, they could be merged into a single query. In this paper, the multiple queries problem is reformulated to a one query problem, by applying average pooling on each of the SGFV and SGBN related histograms over the multiple queries. As for the IDE feature, max pooling is used over the multiple queries. The resulting pooled vectors are then used to perform the matching process with the probe set. In this way, the intra-class variation is taken into account, and the method will be more robust to pedestrian variations over the gallery images.

IV. Experiments

A. Datasets

In this section, the proposed method is evaluated on two challenging image benchmarks CUHK03 [37] and Market-1501 [3]. These datasets are very challenging for the person re-ID task because they contain many important variations on viewpoints, poses, and illuminations; also their images have low resolutions, with occlusions and background clutters. Also, Market-1501 dataset is the largest person re-ID datasets currently available for image-based, currently.

1) CUHK03 [37]: contains $13,164$ Deformable Part Model (DPM) [71] bounding boxes, of $1,467$ different identities of the training set. Each single identity is observed by two different cameras and for each view, there are average 4.8 images, for each identity. We follow the experimentation protocol in [3]. In fact, 100 persons are selected randomly and for each person, all the DPM bounding boxes are taken as queries in turns. Then, a cross camera search is performed. The test process is repeated 20 times and then statistics are reported.

2) Market-1501 [3]: contains $32,643$ fully annotated boxes of 1501 pedestrians, making it among the largest image person re-ID dataset to date. It is captured with 6 cameras placed in front of a supermarket. This dataset contains $32,643$ bounding boxes of 1501 identities. Each identity is captured by at most 6 cameras and at least 2. Even though images of the same identity are captured by the same camera, they are distinct and different. The dataset is randomly divided into training and testing sets, containing 750 and 751 identities, respectively. During testing, for each identity, one query image is selected in each camera. The search is processed in a cross-camera mode, i.e. images that belong to the same camera as the query are discarded from the re-ID process. Note that there are $3,368$ queries in the gallery. Each identity may have multiple images under each camera. We use the provided fixed training and test set, under both the OneQ and MultiQ evaluation settings. There are $19,732$ images used for testing and $12,936$ images used for training.

B. Experimental Settings

- In this paper, a codebook of 256 GMM components and $1,000$ visual words are used for SGFV and SSGBN, respectively. This yields a good compromise between accuracy and efficiency.
- For both SGFV and SGBN, the dimensionality of the descriptors are reduced to 100 via PCA, since this can effectively de-correlate the feature before their introduction to the encoding step.
- The adjacency set for the saliency computation is defined by $l = 2$, i.e three patches in the vertical direction and all patches in the horizontal one, since it allows a good localization of the matched patches.
- Unless otherwise stated, all results generated by our proposed method are given for the supervision case obtained by XQDA and the one query (OneQ) setting.

TABLE I. Impact of the Variation of the Number of Bins B on the Market-1501 Dataset. Note that the Reported Results are Those of the Proposed SGFV+SGBN+IDE(C) Method

B	2	3	4
r=1 (%)	80.45	81.86	81.31
mAP (%)	54.91	56.82	55.08

C. Evaluation Metrics

In this paper, the Cumulative Matching Characteristics (CMC) curve is used in order to evaluate the performances of the person re-ID methods for all datasets in this paper. Every probe image is matched with every image in gallery, and the rank of correct match is obtained. Rank-k recognition rate is the expectation of correct match at rank-k, and the cumulative values of recognition rate at all ranks, are recorded as a one-trial CMC result. For Market-1501 dataset, there are on average 14.8 cross-camera ground-truths for each single query. Therefore, the mean average precision (mAP) is also used in this paper in order to evaluate the performances. In fact, for each query, the area under the Precision-Recall curve called average precision (AP), is computed. Then, the mean value of the APs of all queries (denoted mAP), is calculated while taking into consideration both precision and recall, and thus providing a more comprehensive evaluation.

D. Empirical Analysis of the Proposed Method

1) Impact of Bin quantization : Here, we investigate how the re-ID performance is affected by the variation of the numbers of bins. In fact, the number of bins determines the compromise between accuracy and histogram size. The smaller it is, the less the representation is accurate, but the faster it is. Using a codebook of size $1,000$, the number of bins vary from 2 to 4. As shown in Table I, $B = 3$ produces a good trade-off between the histogram size and the classification accuracy. For all further experiments, the number of bins is set to $B = 3$.

2) Impact of the Gaussian and Saliency weights : As is shown in Table II, both the Gaussian and saliency weights have shown important improvements when applied to our work. Indeed, when weighting the traditional FV and BN via the Gaussian weight (GFV-U and GBN-U), the matching rates considerably increase for both BN and FV, on all datasets. This is due the elimination of the background noise effects when applying the Gaussian mask. Furthermore, the saliency weighting has a significant impact on the re-ID accuracy. This proves the effectiveness of the saliency map in stressing the identity uniqueness. Specially, we notice this improvement on

TABLE II. IMPACT OF WEIGHTING AND SUPERVISION LEVEL ON THE PROPOSED HISTOGRAM ENCODING METHODS. RESULTS (RANK-1 MATCHING RATE AND ON MAP) ARE REPORTED ON CUHK03 AND MARKET-1501 DATASETS FOR DIFFERENT ENCODING METHODS, I.E., THE UNSUPERVISED PROPOSED FV (FV-U), GAUSSIAN WEIGHTED FV (GFV-U), SALIENT-GAUSSIAN WEIGHTED FV (SGFV-U), SUPERVISED SALIENT-GAUSSIAN WEIGHTED FV (SGFV), THE UNSUPERVISED PROPOSED BN (BN-U), GAUSSIAN WEIGHTED BN (GBN-U), SALIENT-GAUSSIAN WEIGHTED BN (SGBN-U) AND SUPERVISED SALIENT-GAUSSIAN WEIGHTED BN (SGBN). THE COSINE DISTANCE IS USED FOR THE UNSUPERVISED CASE

Methods	CUHK03		Market-1501	
	r=1	mAP	r=1	mAP
FV-U (OneQ)	31.83	32.78	42.02	17.98
FV-U (MultiQ)	35.85	37.25	50.82	24.21
GFV-U (OneQ)	35.61	37.08	49.92	23.22
GFV-U (MultiQ)	39.68	40.42	59.01	33.09
SGFV-U (OneQ)	39.84	41.08	57.81	31.28
SGFV-U (MultiQ)	44.19	45.17	66.02	41.06
SGFV (OneQ)	47.87	49.28	62.10	36.88
SGFV (MultiQ)	52.05	53.32	70.62	45.76
BN-U (OneQ)	30.58	32.12	40.63	17.70
BN-U (MuliQ)	34.85	36.71	49.14	23.52
GBN-U (OneQ)	34.37	36.77	48.26	22.83
GBN-U (MultiQ)	38.91	40.82	56.68	30.62
SGBN-U (OneQ)	38.58	40.38	56.84	30.88
SGBN-U (MultiQ)	42.67	44.79	65.15	39.17
SGBN (OneQ)	46.66	48.78	60.95	34.67
SGBN (MultiQ)	50.88	53.12	69.24	44.84

the results for all datasets. Indeed, results on mAP increase from 32.78% and 17.98% to 37.08% and 23.22% for FV and from 32.12% and 17.70% to 36.77% and 22.83% for BN, when weighting our histograms via Gaussian weight; and by $+4.23\%$ and $+4.52\%$ for FV and $+3.61\%$ and $+8.05\%$ for BN, when adding the saliency weight, for CUHK03 and Market-1501 datasets, respectively. Similarly the Gaussian weight, the saliency weight remarkably improves the results in both datasets (see Table II).

3) Impact of SGFV, SGBN and IDE combination : In this paper, a rich multi-level representation is proposed by the combining mid-level (SGFV and SGBN) and high-level (IDE) features. Indeed, on one hand, the combination of the two complementary histogram encoding methods SGFV and SGBN shows good improvements in accuracy by reaching 54.83% and 68.74% at rank-1 matching rates, for CUHK03 and Market-1501 dataset, respectively (see Table III). This achievement is due to the complementarity of these two encoding methods. In fact, SGBN takes into account the locality in the feature space during pooling whereas SGFV produces a more robust representation that reflect higher order statics (the average first order (mean) and second order (standard deviation) of the differences between the image local features and every visual word). On another hand, while combining these mid-level representations with the high-level descriptors IDE(C) (learned on the CaffeNet dataset), the accuracy increasingly rises to 61.06% and 74.88% at rank-1 for or CUHK03 and Market-1501, respectively. When combining with IDE(R) (learned on the ResNet-50 dataset), we achieve a rank-1 matching rates of 72.12% and 81.61% on the respective datasets.

4) Comparison with the state-of-the-art methods: The comparison of the proposed methods with other state-of art methods is detailed in Tables IV and V. The state-of-the-art methods could be divided into two categories: high dimensional descriptors based methods and local descriptors based

TABLE III. IMPACT OF THE COMBINATION OF THE CNN FEATURES AND THE MID-LEVEL ONES. RESULTS (RANK-1 MATCHING RATE AND ON MAP) ARE REPORTED ON CUHK03 AND MARKET-1501 DATASETS FOR THE PROPOSED METHODS, I.E. SGFV+SGBN AND SGFV+SGBN+IDE(C)

Methods	CUHK03		Market-1501	
	r=1	mAP	r=1	mAP
SGFV	47.87	49.28	62.10	36.88
SGBN	46.66	48.78	60.95	34.67
SGFV+SGBN	54.83	61.13	68.74	43.01
IDE (C)	58.91	64.92	57.72	35.95
IDE (R)	66.20	71.10	71.41	48.89
SGFV+SGBN+IDE (C)	61.06	69.01	74.88	52.72
SGFV+SGBN+IDE (R)	72.12	78.22	81.61	60.08

ones. We first compare our results with some methods based on high dimensional global signatures such as LOMO [16]. Actually, we achieve for example on Market-1501 a rank-1 matching rate of 62.10% for the proposed unsupervised method SGFV-U and 60.95% for SGBN-U, versus and 26.07% for LOMO. Although LOMO apply a higher dimensional global descriptor ($26,960$ dimensional descriptor for LOMO), we obviously obtain better results. The same for CUHK03 and Market-1501 datasets). This is due to the fact that the proposed representation is much richer then global descriptors based ones, which are generally not enough sensitive to the affine deformations and to the cross-view pose, illumination, background changes, and space misalignment. Otherwise, the category of methods based on local features can be divided into two sub-categories: brute-force based methods [2, 16, 17] and encoding histogram based ones [3, 4].

Indeed, the method outperforms SDALF [2], eSDC [17] and LOMO [16] that rely on brute-force feature-feature matching, are much more computational complex and in the same time ensure lower results then the histogram encoding ones (the proposed methods). As well, we experiment and compare the proposed methods in the supervised and unsupervised case. Therefore, we first compare the unsupervised proposed method with some popular unsupervised methods in the person re-ID field. In point of fact, it's notably clear that the proposed SGFV-U+SGBN-U achieve better results then this latter ones, in both datasets. This is obviously due to the robust proposed weighted encoding. In our earlier works [5, 6], we proposed an unsupervised version of the weighted FV encoding for the image-based person re identification task. [5] proposed a Gaussian weighted encoding FV version with Retinex transform and the combination of CN, CHS and 15-d low-level features, while [6] introduced a salient weighted FV version with CN and CHS features.

As shown in table (the unsupervised case), we considerably outperform [5, 6]. This is due to the successful combination of saliency and Gaussian weighting as well as the weighted BN and FV encodings. Regarding the group of the supervised methods, i.e. for example SalMatch [18], BOW [3], Improved Deep [36], Kissme (LOMO) [16], XQDA (LOMO) [16], Metric Ensembles [41], etc.; the proposed supervised method SGFV+SGBN outperforms these latter methods in both CUHK03 and Market-1501 datasets (see Tables IV and V).

Actually, compared to the cited supervised methods, we proposed a robust weighted encoding scheme and projected the proposed histograms on a learned discriminative NS. For

TABLE IV. COMPARISON OF THE PROPOSED UNSUPERVISED METHODS (SGFV-U+SGBN-U) AND (SGFV-U+SGBN-U+IDE-U) WITH THE STATE OF THE ART METHODS IN THE CASE OF UNSUPERVISED (FIRST TABLE PART), AND THE PROPOSED SUPERVISED METHODS (SGFV+SGBN) AND (SGFV+SGBN+IDE) WITH THE SUPERVISED METHODS (SECOND TABLE PART), ON CUHK03 DATASET. NOTE THAT '-' MEANS THAT CORRESPONDING RESULTS ARE NOT AVAILABLE

Methods	CUHK03 (detected)				CUHK03 (manual)			
	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
SDALF[2]	4.87	-	-	-	5.60	23.45	36.09	51.96
eSDC [17]	7.68	-	-	-	8.76	24.07	38.28	53.44
BOW [3]	22.95	-	-	-	24.33	58.42	71.28	84.91
TWFV [5]	35.26	48.23	62.86	81.22	-	-	-	-
STWF [6]	32.54	46.12	54.23	70.81	-	-	-	-
Ours (SGFV-U+SGBN-U)	37.58	50.02	63.91	82.13	40.83	53.95	65.86	85.12
Ours(SGFV-U+SGBN-U+IDE(C)-U)	46.11	58.31	70.77	88.11	49.77	61.85	72.41	91.62
Ours (SGFV+SBN+IDE(R))	50.12	82.75	74.26	91.21	53.21	65.32	82.28	94.54
ITML [72]	5.14	-	-	-	5.53	18.89	39.96	44.20
LMNN [73]	6.25	-	-	-	7.29	21.00	32.06	48.94
KISSME [39]	11.70	-	-	-	14.17	41.12	54.89	70.09
XQDA(LOMO) [16]	52.20	82.23	92.14	96.25	46.25	78.90	88.55	94.25
NS(LOMO) [41]	53.70	83.05	93.00	94.80	58.90	85.60	92.45	96.30
NS(fusion)[41]	54.70	84.75	94.80	95.20	62.55	90.05	94.80	98.10
Metric Ensembles [74]	-	-	-	-	62.10	87.81	92.30	97.20
DeepReid [37]	19.89	50.00	64.00	78.50	20.65	51.50	66.50	80.00
Improved Deep [36]	44.96	76.01	83.47	93.15	54.74	86.50	93.88	98.10
FVdeepLDA [50]	-	-	-	-	62.23	89.95	92.73	97.55
PersonNet [57]	-	-	-	-	64.80	89.40	94.92	98.20
IDE(C)+XQDA[35]	58.90	-	-	-	61.70	-	-	-
IDE(C)+XQDA+re [35]	58.50	-	-	-	61.60	-	-	-
PIE (R) [55]	61.50	89.30	94.50	97.60	-	-	-	-
Ours (SGFV+SGBN)	54.83	85.18	95.36	96.71	58.59	87.05	96.55	97.72
Ours (SGFV+SGBN+IDE(C))	61.06	89.29	94.11	98.03	66.21	92.05	97.72	98.75
Ours (SGFV+SGBN+IDE(R))	72.12	91.75	95.86	99.01	75.42	96.12	98.88	99.34

TABLE V. COMPARISON OF THE PROPOSED UNSUPERVISED METHODS (SGFV-U+SGBN-U) AND (SGFV-U+SGBN-U+IDE-U) WITH THE STATE OF THE ART METHODS IN THE CASE OF UNSUPERVISED (FIRST TABLE PART), AND THE PROPOSED SUPERVISED METHODS AND (SGFV+SGBN) AND (SGFV+SGBN+IDE) WITH THE SUPERVISED METHODS (SECOND TABLE PART), ON THE MARKET-1501 DATASET

Methods	OneQ		MultiQ	
	r=1	mAP	r=1	mAP
SDALF [2]	20.53	8.20	-	-
eSDC [17]	33.54	13.54	-	-
BOW [3]	34.40	14.10	42.14	19.20
LOMO [16]	26.07	7.75	-	-
TWFV [5]	49.64	23.01	57.51	29.32
STWFV [6]	54.45	24.73	59.15	27.93
Ours (SGFV-U+SGBN-U)	60.63	35.33	68.48	42.17
Ours(SGFV-U+SGBN-U+IDE(C)-U)	66.26	40.82	74.75	48.02
Ours(SGFV-U+SGBN-U+IDE(R)-U)	72.37	46.91	80.42	54.96
ITML(BOW) [3]	38.21	17.05	-	-
KISSME(BOW) [3]	44.42	20.76	-	-
KISSME(LOMO) [16]	40.50	19.02	-	-
XQDA(LOMO) [16]	43.79	22.22	54.13	28.41
kLDFA(LOMO) [16]	51.37	24.43	52.67	27.36
MFA(LOMO) [16]	45.67	18.24	-	-
NS(LOMO) [41]	55.43	29.87	67.96	41.89
NS(fusion) [41]	61.02	35.68	71.56	46.03
PersonNet [57]	37.21	18.57	-	-
FVdeepLDA [50]	48.15	29.94	-	-
NS-CNN [7]	59.56	34.44	69.95	44.82
IDE(C)+XQDA[35]	57.72	35.95	-	-
IDE(C)+XQDA+re [35]	61.25	46.79	-	-
SCNN [58]	65.88	39.55	76.04	48.45
SOMAnet [59]	73.87	47.89	81.29	56.98
PIE(R) [55]	79.33	55.95	-	-
Ours(SGFV+SGBN)	68.74	42.41	76.61	49.81
Ours(SGFV+SGBN+IDE(C))	74.88	48.62	81.86	56.82
Ours(SGFV+SGBN+IDE(R))	81.61	58.88	87.78	64.88

example, although XQDA (LOMO) method [16] applies a higher dimensional global descriptor and also uses a sophisticated metric learning (XQDA), we obtain better results (for example: r1=54.83% for SGFV+SGBN versus 52.20% for XQDA(LOMO) on CUHK03 dataset). Also, when comparing to NS(fusion) [8], we achieve much better matching rates results (for example: r1=68.74% for SGFB+SGBN versus 55.43% and 61.02% for NS(LOMO) and NS(fusion) on Market-1501) while both methods learn a discriminative NS.

In spite of that, the proposed complementary representation (SGFB+SGBN) achieves better results on both datasets. Also, (SGFV+SGBN+IDE(C)) highly outperforms all the supervised methods. Naturally, the proposed multi-level representation is much richer and meaningful than the descriptor one, so the explanation. As well, when compared to the deep learning methods [7, 35, 36, 37, 50, 55, 57, 58, 59] on Market-1501 and CUHK03 datasets, the proposed method achieves better re-ID rates without needing neither data augmentation nor drop out or GPU computing.

Indeed, the proposed method SGFV+SGBN+IDE(C) considerably outperforms most deep learning based methods on both datasets, by achieving for example, a rank-1 matching rate of 61.06% versus 19.89%, 44.96% and 58.90% for [36, 37] and [35], respectively, on CUHK03. Moreover, we similarly achieve on Market-1501 a mAP of 48.62% versus 18.57%, 29.94%, 34.44%, 35.95% and 39.55%, respectively for [7, 35, 50, 57, 58]; and 58.88% for the proposed SGFV+SGBN+IDE(R) versus 55.95% for PIE(R) [55], for example. We also remarkably achieve a higher result (mAp=48.62%) than IDE(C)+XQDA+re [35] that is based on the powerful and effective re-ranking approach (mAP=56.79%).

With respect to PersonNet [57], we notice that the improvement of the proposed method is most significant in the challenging Market-1501 dataset. This proves the the robustness of our method due to the rich proposed multi-level representation. From this point, we also deduct that deep learning methods are prone over-fitting. To tackle this problem, drop out and data augmentation are used. Besides, training is carried out by GPU implementation to cope with the massive deep learning computations.

V. Conclusion

In this paper, a new approach based on multi-level feature representation is proposed. Mid-level features are generated by weighting two complementary histogram encoding methods: FV and BN, with saliency and Gaussian weights. This introduces a robust extension of the traditional histogram encoding methods to the person re-ID field. More specifically, Gaussian and saliency weights respectively remove background clutters surrounding the person silhouette and, emphasize the most distinctive regions in the person images in order to highlight the uniqueness of each pedestrian. As high-level features, the IDE deep CNN feature is computed over a classification mode CNN. Finally, the well-performing XQDA metric learning is learned on the top of the resulting representations. The experimental results demonstrate the good performances of the proposed method. In future research, the investigation of more sophisticated deep CNN architectures is conceivable. Also, it seems to be interesting to further explore the motion cues in the video-based person re-ID task. Moreover, re-ranking methods could be considered in future work in order to take profit more adequately of the rich similarity context.

VI. Aknowledgment

The authors acknowledge the support from Deanship of Scientific Research, Majmaah University under project N. 37/39.

References

[1] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features." in *ECCV (1)*, 2008, pp. 262–275.

[2] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "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*, 2010, pp. 2360–2367.

[3] L. Zheng, L. Shen, L. Tian, S. Wang, J. Bu, and Q. Tian, "Person re-identification meets image search," in *CoRR*, vol. abs/1502.02171, 2015, pp. 2360–2367.

[4] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *ECCV Workshops*, vol. 7583, 2012, pp. 413–422.

[5] S. Ksibi, M. Mejdoub, and C. Ben Amar, "Topological weighted fisher vectors for person re-identification," in *23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, December 4-8, 2016*, 2016, pp. 3097–3102.

[6] ——, "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*, 2016, pp. 4344–4349.

[7] S. Li, X. Liu, W. Liu, H. Ma, and H. Zhang, "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*, 2016, pp. 480–484.

[8] Y. Guo, L. Wu, H. Lu, Z. Feng, and X. Xue, "Null foley-sammon transform," *Pattern Recognition*, vol. 39, no. 11, pp. 2248–2251, 2006.

[9] L. Ma, H. Liu, L. Hu, C. Wang, and Q. Sun, "Orientation driven bag of appearances for person re-identification," *CoRR*, vol. abs/1605.02464, 2016.

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

[11] A. Wali, N. Ben Aoun, H. Karray, C. Ben Amar, and A. M. Alimi, "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*, 2010, pp. 110–120. [Online]. Available: https://doi.org/10.1007/978-3-642-17691-3_11

[12] Z. Yang, L. Jin, and D. Tao, "A comparative study of several feature extraction methods for person re-identification," in *Biometric Recognition - 7th Chinese Conference, CCBR 2012, Guangzhou, China, December 4-5, 2012. Proceedings*, 2012, pp. 268–277.

[13] C.-H. Kuo, S. Khamis, and V. D. Shet, "Person re-identification using semantic color names and rankboost," in *IEEE Workshop on Applications of Computer Vision*, 2013, pp. 281–287.

[14] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, 2014, pp. 536–551.

[15] N. Ben Aoun, M. Mejdoub, and C. Ben Amar, "Graph-based approach for human action recognition using spatio-temporal features," *J. Visual Communication and Image Representation*, vol. 25, no. 2, pp. 329–338, 2014. [Online]. Available: https://doi.org/10.1016/j.jvcir.2013.11.003

[16] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "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*, 2015, pp. 2197–2206.

[17] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3586–3593.

[18] ——, "Person re-identification by saliency learning," vol. 39, no. 2, 2017, pp. 356–370.

[19] F. Perronnin and C. R. Dance, "Fisher kernels on visual vocabularies for image categorization." in *CVPR*. IEEE Computer Society, 2007.

[20] S. Pedagadi, J. Orwell, S. A. Velastin, and B. A. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, 2013, pp. 3318–3325.

[21] M. Sekma, M. Mejdoub, and C. Ben Amar, "Human action recognition based on multi-layer fisher vector encoding method," *Pattern Recognition Letters*, vol. 65, pp. 37–43, 2015. [Online]. Available: https://doi.org/10.1016/j.patrec.2015.06.029

[22] ——, "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*, 2015, pp. 642–647. [Online]. Available: https://doi.org/10.1109/ISDA.2015.7489193

[23] M. Mejdoub, M. Dammak, and C. Ben Amar, "Extending laplacian sparse coding by the incorporation of the image spatial context," *Neurocomputing*, vol. 166, pp. 44–52, 2015. [Online]. Available: https://doi.org/10.1016/j.neucom.2015.03.086

[24] M. Dammak, M. Mejdoub, and C. Ben Amar, "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*, 2014, pp. 3572–3576.

[25] S. E. F. de Avila, N. Thome, M. Cord, E. Valle, and A. de Albuquerque Araújo, "Pooling in image representation: The visual codeword point of view," *Computer Vision and Image Understanding*, vol. 117, no. 5, pp. 453–465, 2013.

[26] M. Mejdoub and C. Ben Amar, "Classification improvement of local feature vectors over the KNN algorithm," *Multimedia Tools Appl.*, vol. 64, no. 1, pp. 197–218, 2013. [Online]. Available: https://doi.org/10.1007/s11042-011-0900-4

[27] M. Mejdoub, L. H. Fonteles, C. Ben Amar, and M. Antonini, "Embedded lattices tree: An efficient indexing scheme for content based retrieval on image databases," *J. Visual Communication and Image Representation*, vol. 20, no. 2, pp. 145–156, 2009. [Online]. Available: https://doi.org/10.1016/j.jvcir.2008.12.003

[28] M. Sekma, M. Mejdoub, and C. Ben Amar, "Bag of graphs with geometric relationships among trajectories for better human action recognition," in *Image Analysis and Processing - ICIAP 2015 - 18th International Conference, Genoa, Italy, September 7-11, 2015, Proceedings, Part I*, 2015, pp. 85–96. [Online]. Available: https://doi.org/10.1007/978-3-319-23231-7_8

[29] M. Mejdoub, N. Ben Aoun, and C. Ben Amar, "Bag of frequent subgraphs approach for image classification," *Intell. Data Anal.*, vol. 19, no. 1, pp. 75–88, 2015. [Online]. Available: https://doi.org/10.3233/IDA-140697

[30] M. Mejdoub, L. H. Fonteles, C. Ben Amar, and M. Antonini, "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*, 2008, pp. 365–372. [Online]. Available: https://doi.org/10.1109/CBMI.2008.4564970

[31] M. El Arbi, M. Koubàa, M. Charfeddine, and C. Ben Amar, "A dynamic video watermarking algorithm in fast motion areas in the wavelet domain," *Multimedia Tools Appl.*, vol. 55, no. 3, pp. 579–600, 2011. [Online]. Available: https://doi.org/10.1007/s11042-010-0580-5

[32] T. Bouchrika, M. Zaied, O. Jemai, and C. Ben Amar, "Neural solutions to interact with computers by hand gesture recognition," *Multimedia Tools Appl.*, vol. 72, no. 3, pp. 2949–2975, 2014. [Online]. Available: https://doi.org/10.1007/s11042-013-1557-y

[33] H. Boughrara, M. Chtourou, and C. Ben Amar, "MLP neural network using constructive training algorithm: application to face recognition and facial expression recognition," *IJISTA*, vol. 16, no. 1, pp. 53–79, 2017. [Online]. Available: https://doi.org/10.1504/IJISTA.2017.10002246

[34] M. Othmani, W. Bellil, C. Ben Amar, and A. M. Alimi, "A new structure and training procedure for multi-mother wavelet networks," *IJWMIP*, vol. 8, no. 1, pp. 149–175, 2010. [Online]. Available: https://doi.org/10.1142/S0219691310003353

[35] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," *CoRR*, vol. abs/1701.08398, 2017.

[36] E. Ahmed, M. J. Jones, and T. K. Marks, "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*, 2015, pp. 3908–3916.

[37] W. Li, R. Zhao, T. Xiao, and X. Wang, "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*, 2014, pp. 152–159.

[38] I. Filkovic, Z. Kalafatic, and T. Hrkac, "Deep metric learning for person re-identification and de-identification," in *39th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2016, Opatija, Croatia, May 30 - June 3, 2016*, 2016, pp. 1360–1364.

[39] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, 2012, pp. 2288–2295.

[40] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "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*, 2013, pp. 3610–3617.

[41] L. Zhang, T. Xiang, and S. Gong, "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*, 2016, pp. 1239–1248.

[42] S. Ksibi, M. Mejdoub, and C. Ben Amar, "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*, 2016, pp. 1–8.

[43] M. Shahiduzzaman, D. Zhang, and G. Lu, in *ACCV (4)*, vol. 6495, 2010, pp. 449–459.

[44] W. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *British Machine Vision Conference, BMVC 2009, London, UK, September 7-10, 2009. Proceedings*, 2009, pp. 1–11.

[45] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 1565–1573.

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

[47] A. Sharma and K. K. Paliwal, "Linear discriminant analysis for the small sample size problem: an overview," *Int. J. Machine Learning & Cybernetics*, vol. 6, no. 3, pp. 443–454, 2015.

[48] W. Li and X. Wang, "Locally aligned feature transforms across views," in *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, 2013, pp. 3594–3601.

[49] W. Li, R. Zhao, T. Xiao, and X. Wang, "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*, 2014, pp. 152–159.

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

[51] S. Messelodi and C. M. Modena, "Boosting fisher vector based scoring functions for person re-identification," *Image Vision Comput.*, vol. 44, pp. 44–58, 2015.

[52] S. E. F. de Avila, "Extended bag-of-words formalism for image classification," Ph.D. dissertation, Pierre and Marie Curie University, Paris, France, 2013.

[53] S. E. F. de Avila, N. Thome, M. Cord, E. Valle, and A. de Albuquerque Araújo, "BOSSA: extended bow formalism for image classification," in *18th IEEE International Conference on Image Processing, ICIP 2011, Brussels, Belgium, September 11-14, 2011*, 2011, pp. 2909–2912.

[54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 1097–1105.

[55] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose invariant embedding for deep person re-identification," *CoRR*, vol. abs/1701.07732, 2017.

[56] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang, "Improving person re-identification by attribute and identity learning," *CoRR*, vol. abs/1703.07220, 2017.

[57] L. Wu, C. Shen, and A. van den Hengel, "Personnet: Person re-identification with deep convolutional neural networks," *CoRR*, vol. abs/1601.07255, 2016.

[58] R. R. Varior, M. Haloi, and G. Wang, "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*, 2016, pp. 791–808.

[59] I. B. Barbosa, M. Cristani, B. Caputo, A. Rognhaugen, and T. Theoharis, "Looking beyond appearances: Synthetic training data for deep cnns in re-identification," *CoRR*, vol. abs/1701.03153, 2017.

[60] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification." *Pattern Recognition*, vol. 48, pp. 2993–3003, 2015.

[61] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian, "Person re-identification in the wild," *CoRR*, vol. abs/1604.02531, 2016.

[62] S. Wu, Y. Chen, X. Li, A. Wu, J. You, and W. Zheng, "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*, 2016, pp. 1–8.

[63] F. Xiong, M. Gou, O. I. Camps, and M. Sznaier, "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*, 2014, pp. 1–16.

[64] W. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, 2013.

[65] Y. Chen, W. Zheng, and J. Lai, "Mirror representation for modeling view-specific transform in person re-identification," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, 2015, pp. 3402–3408.

[66] D. J. Jobson, Z. Rahman, and G. A. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Trans. Image Processing*, vol. 6, no. 7, pp. 965–976, 1997.

[67] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *In VISAPP International Conference on Computer Vision Theory and Applications*, 2009, pp. 331–340.

[68] B. Prosser, W. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *British Machine Vision Conference, BMVC 2010, Aberystwyth, UK, August 31 - September 3, 2010. Proceedings*, 2010, pp. 1–11.

[69] Y. Xu, B. Ma, R. Huang, and L. Lin, "Person search in a scene by jointly modeling people commonness and person uniqueness," in *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*, 2014, pp. 937–940.

[70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 770–778.

[71] R. Arandjelovic and A. Zisserman, "Multiple queries for large scale specific object retrieval," in *British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012*, 2012, pp. 1–11.

[72] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Machine Learning, Proceedings of the*

Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007, 2007, pp. 209–216.

[73] S. Sun and Q. Chen, "Hierarchical distance metric learning for large margin nearest neighbor classification," *IJPRAI*, vol. 25, no. 7, pp. 1073–1087, 2011.

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