Fusion schemes for image-to-video person re-identification
Thuy-Binh Nguyen\textsuperscript{a,b,c}, Thi-Lan Le\textsuperscript{a} and Nam Pham Ngoc\textsuperscript{b}
\textsuperscript{a}International Research Institute MICA, Hanoi University of Science and Technology, Hanoi, Vietnam; \textsuperscript{b}School of Electronics and Telecommunications, Hanoi University of Science and Technology, Hanoi, Vietnam; \textsuperscript{c}Faculty of Electrical and Electronics, University of Transport and Communications, Hanoi, Vietnam

ABSTRACT
This paper focuses on improving the performance of image-to-video person re-identification through feature fusion. In this study, image-to-video person re-identification is formulated as a classification-based information retrieval in which a pedestrian appearance model is learned in the training phase, and the identity of an interested person is determined based on the probability his/her probe image belongs to the model. Four state-of-the-art features belonging to two categories: hand-designed features and learned features are investigated for person image representation. They are Kernel Descriptor, Gaussian of Gaussian, features extracted from two famous convolutional neural networks (GoogleNet and ResNet). Furthermore, three fusion schemes that are early fusion, product-rule and query-adaptive late fusion are proposed. To evaluate the performance of the chosen features for person appearance representation as well as their combination in three proposed fusion schemes, 114 experiments on two public benchmark datasets (CAVIAR4REID and RAiD) have been conducted. The experiments confirm the robustness and effectiveness of the proposed fusion schemes. The proposed schemes obtain improvement of +7.16\%, +5.42\%, and +6.30\% at rank-1 over those of single feature in case A, case B of CAVIAR4REID, and RAiD, respectively.

1. Introduction
Person re-identification, the task of recognizing people in a non-overlapping camera network, has attracted the especial attention of the computer vision and pattern recognition community because of its widespread applications in video surveillance. The existing person re-identification methods are classified into two main approaches: image-based and video-based (Zheng, Yang, & Hauptmann, 2016). In the first approach, an individual has only a single image on both the probe and the gallery sets while an individual in the second approach contains a set of images. As a consequence, the studies belonging to the first approach focus mainly on image content analysis and matching while those of the second approach can exploit different information such as temporal and motion. One special case of the video-based approach is the image-to-video person re-identification in which the probe is an image while the gallery has a set of images (Pham, Le, Vu, © 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Dao, & Nguyen, 2017; Wang, Lai, & Xie, 2017; Zhang et al., 2017). This reflects situations in real life such as criminal or suspect search where one sole query image is available. The image-to-video person re-identification shares the same common challenges with the image-based person re-identification (i.e. occlusion, low resolution, large variation in poses and viewpoints, similar appearance) and has its own challenge of matching two different modalities that are an image and a video.

Taking into account that each feature can reflect certain characteristics of the person image, the fusion of these features is an important step to secure a high performance. In Nguyen, Le, Nguyen, and Pham (2018), we have proposed three fusion schemes including early fusion, product rule and query-adaptive late fusions for image-to-video person re-identification. In this work, image-to-video person re-identification is formulated as a classification-based information retrieval problem where a person appearance model is learned from the gallery images in the training phase and the identity of interested person is determined by the probability of his/her probe image belonging to the model. However, the choice of features as well as their combination is not considered in the previous work. This paper is an extended version of Nguyen et al. (2018). In this study, we investigate in details the choice of features and their combination for image-to-video person re-identification. Two extra features are considered in this work: Gaussian of Gaussian (GOG) and learned features from ResNet for representing a person image. GOG is evaluated as one of the most effective hand-designed features for image-to-image person re-identification (Matsukawa, Okabe, Suzuki, & Sato, 2016). While ResNet is known as a residual leaning network and can provide extracted features at deeper layers (He, Zhang, Ren, & Sun, 2016). A huge number of experiments (114 vs. 30 in the previous work) on two common datasets (CAVIAR4REID and RAiD) have been performed in this study. Results obtained from these experiments again confirm the effectiveness of the three fusion schemes. The proposed schemes obtain improvement of +7.16%, +5.42%, and +6.30% at rank-1 over those of single feature in case A, case B of CAVIAR4REID, and RAiD, respectively. Furthermore, they can be used as a recommendation for feature selection/combination in image-to-video person re-identification.

The remainder of this paper is organized as follows. Section 2 describes the related work on the feature extraction and image-to-video person re-identification. The proposed framework is presented in Section 3. And finally, experimental results and conclusion are shown in Section 4 and Section 5, respectively.

2. Related work

In this section, we briefly review two kinds of studies that are related to our work: feature extraction and image-to-video person re-identification. Feature extraction plays an important role in person re-identification. Building a robust and discriminative descriptor for person appearance representation is a non-trivial task. To the best of our knowledge, there are only a few studies on image-to-video person re-identification due to the challenges as mentioned in Section 1.

In the literature, there are many features that have been proposed for person appearance representation (Liu, Gong, Loy, & Lin, 2014). Taking into account that each feature has its own advantages and drawbacks, several studies aim at fusing different features in order to get a good person re-identification performance. The feature fusion approach is divided
into two categories: early fusion (also named feature-level fusion) and late fusion (named score-level fusion). In the first category, features are concatenated to generate a larger dimension vector for representing an image while the methods belonging to the second category compute the weight (score) for each feature in the similarity function. Gao, Ai, and Bai (2016) propose an early fusion method for the image-to-image person re-identification. A combined feature is created by concatenating a high-dimensional low-level Weighted Histograms of Overlapping Stripes features that are formed by combining colour histograms and Histogram of Oriented Gradients and a low-dimensional mid-level colour name descriptor. The authors have proved that the recognition rate is improved when using the combined features with experiments conducted on several datasets. Eisenbach, Kolarow, Vorndran, Niebling, and Gross (2015) claim that late fusion methods can achieve better results than early fusion ones for the image-to-image person re-identification. Lejbølle, Nasrollahi, and Moeslund (2017) improve the performance of person re-identification by combining three different features including local maximal occurrence (Liao, Hu, Zhu, & Li, 2015), a cross-view projective dictionary learning (Li, Shao, & Fu, 2015), and feature fusion network (Wu et al., 2016) extracted at different abstraction levels, namely low-, mid-, and high-level, respectively. In this work, late fusion strategy can be performed in two different ways by using either scores or rank aggregation of the identities. The matching rates at rank-1 corresponding to two late fusion ways are 45.63% and 45.24% with an increase of 7.52% and 17.4% compared to those in the case of exploiting a single feature. Taking into account that features are not equally important for all queries, in Zheng et al. (2015), the authors propose to learn adaptively weight for each query. The obtained results of this technique for the image-to-image person re-identification are very promising.

Concerning the image-to-video person re-identification, Pham et al. (2017) have proposed a fully automated person re-identification that also formulates the image-to-video person re-identification as a classification-based problem. The authors propose to employ Kernel Descriptor (KDES) as a descriptor and SVM as a classifier. This method has outperformed many state-of-the-art works on different benchmark datasets. However, this method uses one sole kind of feature for person representation. In this study, we compare the proposed framework with the method provided by Pham et al. (2017) on two benchmark datasets. Zhang et al. (2017) introduce a novel temporally memorized similarity learning neural network to solve the challenge in the image-to-video person re-identification. In this work, CNN features are extracted at image-level and these feature vectors are fed into Long Short-Term Memory (LSTM) network to generate a unified signature to represent a sequence of images by concatenating all feature vector at each node. Finally, feature vectors of a probe image and sequence images are forwarded to the similarity sub-network for distance metric learning. Therefore, the image-to-video problem is turned into matching two feature vectors by learning a metric in a new sub-network. By using LSTM, the authors focus on the case that images of the same person in training set have a temporal constraint. In the case that the images of the training set are not temporal-related, LSTM cannot show its advantage.

As analysed, several studies have been dedicated to the feature fusion and image-to-video person re-identification. However, most of feature fusion approaches focus on the image-to-image person re-identification. While working with the image-to-video person re-identification, fusion feature is not simply a similarity function with appropriate scores/weights.
3. Proposed framework

3.1. Overall framework

The proposed framework for the image-to-video person re-identification is shown in Figure 1. In comparison with the previous framework (Nguyen et al., 2018), two extra features that are GOG and CNN-ResNet features are added in the feature extraction stage. GOG has been evaluated as a robust feature for single shot person re-identification problem (Karanam et al., 2016; Matsukawa, Okabe, Suzuki, & Sato, 2017). CNN-ResNet is a kind of learned features which are extracted from ResNet, a residual learning network that has ability to learn characteristics in deeper layers. Two added blocks are marked in red line in Figure 1. By adding two extra features, the purpose of our study is to investigate the most effectiveness of these features as well as their combination for person re-identification. For feature fusion, three fusion schemes that are early fusion, late fusion with product rule and query-adaptive late fusion are proposed.

3.2. Feature extraction

3.2.1. Hand-designed features

For hand-designed features, GOG and KDES are proposed for person image representation. The robustness and effectiveness of the two kinds of features are proved in Figure 1.

Figure 1. The proposed framework for the image-to-video person re-identification. Two new blocks that are Extracting GOG feature and Extracting CNN-ResNet feature are added.
some recent works (Karanam et al., 2018; Matsukawa et al., 2016, 2017; Pham, Le, Dao, Le, & Nguyen, 2015; Pham et al., 2017). The common point of the two methods is that features are extracted at different levels to ensure information lossless. The dimension of a GOG feature vector is much smaller in comparison with that of a KDES feature vector for describing a person image. This is an advantage of GOG feature compared to the other one in the training phase. The way to extract each kind of features is shown in more detail in the next paragraphs.

GOG feature is introduced by Matsukawa et al. (2016) to build a robust and discriminative descriptor for feature extraction in person re-identification problem. The novel point in this work is that Gaussian distribution is applied twice, once at patch-level and the other at region-level, this is the reason why this descriptor is called GOG. Figure 2 illustrates some step to generate a GOG descriptor for representing a given image. Firstly, each image is divided into several overlapped horizontal stripes to form different separated regions and local patches, represented by a Gaussian distribution of pixel features, are extracted densely on each stripe. A feature vector at pixel level contains information about the location of the pixel in the vertical direction, magnitudes of the pixel intensity gradient along four orientations and three colour channel values. Any colour space such as RGB, HSV and Lab can be employed for providing colour information for GOG descriptors. We have conducted experiments in three colour spaces including RGB, HSV, and Lab. However, the best matching rates are obtained with Lab colour space. Therefore, we decided to use Lab colour space in our experiments. Each pixel is represented by an 8-dimensional feature vector as indicated in the expression (1):

$$f_i = [y, M_{0^\circ}, M_{90^\circ}, M_{180^\circ}, M_{270^\circ}, L, a, b]^T,$$

where $f_i$ is the feature vector at pixel $i$, $y$ is the pixel location in the vertical direction, $M_{\theta \in \{0^\circ, ..., 270^\circ\}}$ are the magnitudes of pixel intensity gradient along four orientations, and $L, a, b$ are colour channel values.

In the next step, these patches Gaussians are flattened and vectorized by considering the underlining geometry of Gaussians. Then, patch Gaussians which are in the same region are combined together to generate a region Gaussian. After that, this region Gaussian is also flattened to form a unique signature for a given region. Finally, these signatures are concatenated for representing a person image. Obviously, by applying Gaussian distribution at both patch and region levels, information at the centre region is highlighted which results in the effectiveness of GOG features in the feature extraction stage for

![Figure 2. GOG descriptor (Matsukawa et al., 2016).](image)
person re-identification problem. As mentioned above, each pixel is represented by an eight-dimensional feature vector. Therefore, the dimension of a patch Gaussian is \((8^2 + 3 \times 8)/2 + 1 = 45\) and a region Gaussian is represented by a \((45^2 + 3 \times 45)/2 + 1 = 1081\)-dimensional vector. As a result, each image which is divided into seven overlapping horizontal regions is represented by a \(7 \times 1081 = 7567\)-dimensional feature vector.

KDES is firstly introduced by Bo, Ren, and Fox (2010) with the main idea is to build compact patch-level features from pixel attributes such as gradient, colour, and texture. This feature has been proved to be robust for object recognition in general and person re-identification in particular. Pham et al. (2017) have improved this feature to make it be invariant to scale change and have shown that this feature outperforms many hand-designed features on a number of benchmark datasets. In this study, we employ the improved version of KDES computed with three kernels provided by Pham et al. (2017). KDES descriptor is computed at three levels: pixel-, patch-, and image-level. Three features that are gradient, colour, and texture are calculated at the pixel level. Then, the patch-level feature is determined by matching kernels via kernel approximation. Then, image level is computed by concatenating features from three layers-pyramid. By this way, each image is represented by a 63,000-dim vector.

3.2.2. Deep learning features

Recent years have witnessed various impressive results of deep learning in computer vision and pattern recognition fields. It is firstly applied in person re-identification problem by Zheng et al. (2016), in which each person is considered as a separate class and a convolutional neural network is trained for a classification objective. Among numerous architectures, two kinds of networks, called GoogleNet (Szegedy et al., 2015) and ResNet (He et al., 2016), are utilized to extract deep learning features of a person image. Both models are pre-trained on ImageNet dataset, and after that, these models are fine-tuned by using evaluated datasets in our work. During the fine-tuning process, the weights of all layers of the network are modified. For both networks, image is resized to 224 \(\times\) 224 pixels according to Simonyan and Zisserman (2014). For GoogleNet, feature vectors are extracted from ‘pool5/7x7 s1’ layer to provide a 1024-dimensional vector for a given person image. In our work, we use and fine-tune the pre-trained GoogleNet on ImageNet dataset – a large visual database designed for pattern recognition with the following parameters. First, the parameter ‘base learning rate (lr)’ for fine-tune process is set to 0.001 to achieve the convergence point as well as the best accuracy. Second, parameter max iter depends on a number of epochs we would like to train CNN network, here, we chose epochs=10, and max iter=10,000 to ensure each image is trained at least 10 times to get a high performance. Finally, parameter ‘test interval’ that depends on the number of images for testing phase is set to 1000 which allows to test all validations in each testing procedure. For ResNet, we chose the value of parameter epochs based on the evaluated datasets, and weights corresponding to the best epoch are saved and used in the test phase. In our work, we used ResNet_101 structure and feature vectors that are extracted at the average pooling layer giving 2048-dimensional vectors. In general, deep learning is not effective on a small dataset, therefore, in order to enrich the
training set for fine-tuning process, several data augmentation techniques are exploited (e.g., rotation, translation, adjustment in colour, illumination, and contrast). For both GoogleNet and ResNet structures, we apply the same data augmentation techniques. By this way, from one image, three images are generated by rotating the original image five degree in left and right sides, and translating this image by 1% of image width and 4% of image height.

3.3. Feature fusion schemes
The image-to-video person re-identification is described as follows. Given a probe (query) image $q$, its identity is determined by:

$$i^* = \arg\max_{i \in 1, 2, \ldots, N_g} \text{sim}(q, G_i^{N_j})$$

where $i^*$ is the identity of probe $q$, and $\text{sim}(\cdot, \cdot)$ is some kinds of similarity functions. $N_g$ is the number of persons in the database. $G_i = \{g_{ij} \}_{j=1}^{N_j}$ is the $j$th person in the gallery set who is represented by a set of $n_j$ instance images.

3.3.1. Feature-level fusion
In the early fusion scheme, a larger dimension feature is created by concatenating the features (GOG, KDES, CNN-Googlenet, CNN-ResNet). This complementary combination takes advantages of both the above features for representing a person image. After that, extracted feature vectors are forwarded to the SVM classifier to match a query image to a set of images in the gallery. In this case, the identity of the query is determined by the identity of the person in gallery having the highest score.

3.3.2. Score-level fusion
In the late fusion scheme, for each feature, a list of retrieved people and its corresponding score is determined. Then, the final list is determined by the following strategy. Denote $s^{(k)}_{q,G_j}$ the similarity score between query $q$ and $j$th person in the gallery set by using $k$th feature. $K$ is the number of features ($K= 2$ or $4$ in our experiments).

- Product rule-based late fusion: In this scheme, the final similarity is determined by product rule (Kittler, Hatef, Duin, & Matas, 1998) as follows:

$$\text{sim}(q, G_j)_{\text{product--rule}} = \prod_{k=1}^{K} s^{(k)}_{q,G_j}.$$  

- Query-adaptive late fusion: Zheng et al. (2015) have observed that a feature may be effective for a given query but ineffective for others. Based on this observation, they try to estimate the effectiveness of a feature based on the score curve. A feature is considered as an effective feature for a given image if its score curve has an L-shape. It means the score corresponding to the first rank is much higher than those of the next ranks. The authors propose to compute feature weight according to its shape of score. However, the authors have only applied the technique to the image-to-image approach. Inspired by this work, we
propose to apply it to the image-to-video person re-identification.

$$\text{sim}(q, G)_{\text{query-adaptive}} = \prod_{k=1}^{K} \left( s_{q, G}^{(k)} \right)^{\omega_{q}^{(k)}}, \quad \text{with} \quad \sum_{k=1}^{K} \omega_{q}^{(k)} = 1,$$

where $\omega_{q}^{(k)}$ is the feature weighting for the $k$th feature. This weight is adaptively determined according to the score's shape of feature for each query. The detail of weight computation is presented in Zheng et al. (2015).

4. Experiments and results

4.1. Datasets and evaluation measure

In order to evaluate the proposed framework, two benchmark datasets that are CAVIAR4-REID (Cheng, Cristani, Stoppa, Bazzani, & Murino, 2011) and RAiD (Das, Chakraborty, & Roy-Chowdhury, 2014) are used.

CAVIAR4REID dataset contains multiple images for 72 pedestrians in two non-overlapping camera views. However, only 50 of them have images on both cameras. This dataset is one of the most challenging ones in benchmark datasets because of occlusion, low resolution and strong variation in illumination. The resolution of these images varies from $17 \times 32$ to $72 \times 144$. With CAVIAR4REID dataset, there exist different ways to create training and testing sets. Hence, in this paper, two scenarios are setup named case A and case B. In case A, a person has five images for testing and training sets while in case B, each individual has five images for testing sets and the rest of images are used for training. These setups allow to evaluate the effect of the number of samples in training sets on the person re-identification accuracy.

RAiD dataset (Das et al., 2014) includes 6920 images of 43 people appeared in two indoor and two outdoor cameras. In this paper, the training and testing sets are set in the same way as in Pham et al. (2017), in which images from one indoor camera are set for training (210 images) and the other for testing (6710 images). All images in this dataset have the same size of $64 \times 128$. The strong variation in illumination of images in this dataset is one of the difficulties for person re-identification problem.

We employ Cumulative Matching Characteristic curve (CMC) as an evaluation measure for person re-identification. The horizontal axis shows ranks while the vertical axis describes the rate of the true matching corresponding to each rank. The value of CMC curve at rank $k$ presents the rate of true matching in the first $k$ images are ranked. The higher the CMC curve is, the better the person re-identification method is.

4.2. Experimental results

To improve the performance of person re-identification, in the proposed framework, two kinds of features including hand-designed and deep learning features are applied for the feature extraction stage. For the hand-designed features, we propose to use GOG and KDES, two robust and effective features for representing a person image in person re-identification problem. And, for deep learning features, GoogleNet and ResNet-101 networks are employed to extract feature vector of each image. As mentioned above, these convolutional neural networks are trained on ImageNet dataset, and after that,
they are fine-tuned on the evaluated datasets. Feature vectors are extracted at the layer which is before the classifier one.

To evaluate the performance of the four chosen features as well as their combinations with the three fusion schemes, we have performed 114 experiments: 19 experiments for each case × 6 cases; 4 cases for CAVIAR4REID (case A with/without data augmentation, case B with/without data augmentation) and 2 cases for RAiD (with/without augmentation). It is worth to note that the number of experiments has been greatly increased in comparison with that of the previous work (Nguyen et al., 2018) in which the number of experiments is 30 (Figure 3).

4.2.1. Evaluating the performance of the features used for person image representation

The four features are evaluated through four cases of CAVIAR4ReID and two cases for RAiD dataset. The results of CAVIAR4REID-case A (with/without data augmentation) are shown in Figure 4 while those of case B are shown in Figure 5. The results obtained for RAiD dataset are illustrated in Figure 6. The values of their accuracy at the five important ranks (1, 5, 10, 15, and 20) are reported in Tables 1–6. From the obtained results, some interesting points are extracted as follows:

First, the chosen features can represent well the person image in image-to-video person re-identification. It is worth to note that in Pham et al. (2015), the authors have shown that KDES outperforms all of state-of-the-art studies for image-to-video person re-identification. KDES still shows its high performance in case B with data augmentation (see Figure 3).
Table 4). However, GOG keeps the first place in case A (with and without data augmentation) (see Tables 2 and 1) while CNN-Googlenet outperforms the others features in case B without data augmentation (Table 3) and CNN-ResNet obtains the best performance for RAiD dataset without data augmentation.

Second, when we compare the performance of the features in the same group (handcrafted and learned features), we can observe that the performance of CNN-ResNet is lower than that of CNN-Googlenet. When considering the matching rates at rank-1 in four cases of CAVIAR4REID, these values corresponding to CNN-Googlenet are 21.11%, 5.83%, 5.56%, 7.70% higher than those of the case using CNN-ResNet in case A (without/with), case B (without/with) data augmentation techniques, respectively. This can be explained by the fact that ResNet network requires a large dataset for the training phase to achieve a good performance. Concerning the features in hand-crafted category, GOG increases the rank-1 accuracies by 6.12% and 4.17% compared to KDES in case A-

![Figure 4](image1.png)

**Figure 4.** Evaluation of the performance of an independent feature (a) CAVIAR4REID-case A without data augmentation (b) CAVIAR4REID-case A with data augmentation.

![Figure 5](image2.png)

**Figure 5.** Evaluation of the performance of an independent feature (a) CAVIAR4REID-case B without data augmentation (b) CAVIAR4REID-case B with data augmentation.
CAVIAR4REID without and with data augmentation techniques, respectively. Inversely, in case B-CAVIAR4REID, KDES provides higher matching rates at rank-1 by 1.94% (without data augmentation) and 4.72% (with data augmentation) compared to GOG.

Third, we can observe the effect of data split from the results obtained for CAVIAR4REID with cases A and B. It is obvious that on the same dataset different data split will provide different performance. In comparison between case A and case B, the accuracies in case B are higher than those in the other case because the number of training images in case B is larger. Finally, the proposed data augmentation allows to increase the performance of person re-identification. After applying the data augmentation techniques, the accuracies at rank-1 in case A are 81.39%, 77.22%, 75.00%, and 69.17% which are 3.33%, 5.28%, 0.28%, and 15.56% higher than those in the case of without data augmentation with GOG, KDES, CNN-GoogleNet, and CNN-ResNet, respectively.

Figure 6. Evaluation of the performance of an independent feature (a) without data augmentation (b) with data augmentation on RAiD dataset.

Table 1. Summary of all experimental results for CAVIAR4REID in case A without data augmentation.

| Methods                      | Rank=1 (%) | Rank=5 (%) | Rank=10 (%) | Rank=15 (%) | Rank=20 (%) |
|------------------------------|------------|------------|-------------|-------------|-------------|
| Single feature               |            |            |             |             |             |
| GOG                          | 78.06      | 90.83      | 95.56       | 97.22       | 97.78       |
| KDES                         | 71.94      | 86.94      | 91.39       | 94.72       | 95.83       |
| CNN-GoogleNet                | 74.72      | 88.06      | 93.89       | 95.83       | 97.50       |
| CNN-ResNet                   | 53.61      | 76.39      | 85.00       | 90.00       | 92.78       |
| Early fusion                 |            |            |             |             |             |
| GOG+CNN-GoogleNet            | 83.61      | 94.44      | 97.22       | 97.78       | 98.89       |
| GOG+CNN-ResNet               | 76.11      | 91.67      | 95.28       | 96.39       | 98.33       |
| KDES+CNN-GoogleNet           | 77.22      | 91.67      | 94.44       | 96.94       | 97.78       |
| KDES+CNN-ResNet              | 69.17      | 84.17      | 90.56       | 93.33       | 95.56       |
| All features                 | 77.22      | 90.28      | 94.17       | 96.11       | 96.94       |
| Late fusion-Product          |            |            |             |             |             |
| GOG+CNN-GoogleNet            | 84.17      | 94.72      | 97.22       | 98.33       | 99.44       |
| GOG+CNN-ResNet               | 73.89      | 89.72      | 95.56       | 96.94       | 98.06       |
| KDES+CNN-GoogleNet           | 80.83      | 92.50      | 96.11       | 97.22       | 98.33       |
| KDES+CNN-ResNet              | 76.39      | 90.28      | 93.06       | 95.28       | 96.67       |
| All features                 | 84.44      | 93.33      | 95.56       | 98.33       | 99.17       |
| Late fusion-Adaptive         |            |            |             |             |             |
| GOG+CNN-GoogleNet            | 81.94      | 92.50      | 95.28       | 96.94       | 98.06       |
| GOG+CNN-ResNet               | 76.67      | 90.00      | 93.06       | 95.28       | 96.94       |
| KDES+CNN-GoogleNet           | 83.89      | 95.00      | 96.94       | 98.33       | 99.44       |
| KDES+CNN-ResNet              | 74.44      | 89.72      | 95.56       | 96.94       | 97.50       |
| All features                 | 83.33      | 93.33      | 95.83       | 98.89       | 99.44       |

Note: The best result at each group for each considered rank is in bold.
CNN-GoogleNet and CNN-ResNet, respectively. In case B, by using data augmentation techniques the matching rate at rank-1 are 0.28% (GOG), 3.06% (KDES), 5.00% (CNN-GoogleNet), and 3.06% (CNN-ResNet) higher than those in the case of without data augmentation. However, the improvement of data augmentation is not really significant for RAiD dataset. This can be explained as follows. There are only 210 images for training phase, and by applying the above data augmentation techniques, the number of training images is multiplied by four times, i.e. 840 images. However, this number is still too small for the fine-tuning step in deep learning. This obtained results can be seen more clearly in Tables 5 and 6.

From the aforementioned observations, we recommend to use GOG feature for person representation in image-to-video person re-identification for the applications where the

| Table 2. Summary of all experimental results for CAVIAR4REID in case A with data augmentation. |
|---------------------------------------------------------------|
| Methods | Rank=1 | Rank=5 | Rank=10 | Rank=15 | Rank=20 |
| Single feature | GOG | 81.39 | 91.94 | 96.11 | 97.50 | 98.06 |
|  | KDES | 77.22 | 88.89 | 91.94 | 94.44 | 95.00 |
|  | CNN-GoogleNet | 75.00 | 91.11 | 94.72 | 96.67 | 97.50 |
|  | CNN-ResNet | 69.17 | 87.50 | 92.22 | 94.44 | 95.00 |
| Early fusion | GOG+CNN-GoogleNet | 85.28 | 93.89 | 96.39 | 97.50 | 98.33 |
|  | GOG+CNN-ResNet | 83.61 | 93.89 | 96.39 | 96.94 | 97.50 |
|  | KDES+CNN-GoogleNet | 80.28 | 92.50 | 96.11 | 97.22 | 97.78 |
|  | KDES+CNN-ResNet | 78.33 | 91.11 | 94.44 | 96.39 | 96.67 |
|  | All features | 83.61 | 93.89 | 96.11 | 96.67 | 97.78 |
| Late fusion-Product | GOG+CNN-GoogleNet | 84.72 | 93.89 | 97.22 | 98.06 | 99.17 |
|  | GOG+CNN-ResNet | 83.06 | 93.89 | 96.39 | 97.22 | 97.78 |
|  | KDES+CNN-GoogleNet | 83.61 | 93.33 | 95.83 | 96.67 | 97.22 |
|  | KDES+CNN-ResNet | 80.83 | 93.33 | 95.00 | 96.39 | 96.94 |
|  | All features | 86.39 | 95.00 | 96.11 | 97.78 | 98.61 |
| Late fusion-Adaptive | GOG+CNN-GoogleNet | 85.00 | 93.06 | 95.28 | 96.11 | 97.50 |
|  | GOG+CNN-ResNet | 81.11 | 93.33 | 95.28 | 96.11 | 97.50 |
|  | KDES+CNN-GoogleNet | 85.00 | 94.17 | 97.22 | 98.06 | 99.17 |
|  | KDES+CNN-ResNet | 83.06 | 93.61 | 96.39 | 97.22 | 97.50 |
|  | All features | 86.39 | 95.28 | 96.39 | 97.78 | 98.33 |

Note: The best result at each group for each considered rank is in bold.

| Table 3. Summary of all experimental results for CAVIAR4REID in case B without data augmentation. |
|---------------------------------------------------------------|
| Methods | Rank=1 | Rank=5 | Rank=10 | Rank=15 | Rank=20 |
| Single feature | GOG | 82.50 | 94.17 | 96.39 | 98.61 | 99.44 |
|  | KDES | 84.44 | 94.72 | 98.06 | 98.89 | 99.44 |
|  | CNN-GoogleNet | 85.56 | 98.06 | 98.61 | 98.89 | 99.17 |
|  | CNN-ResNet | 80.00 | 94.17 | 96.67 | 97.50 | 98.61 |
| Early fusion | GOG+CNN-GoogleNet | 89.72 | 97.78 | 99.17 | 99.72 | 99.72 |
|  | GOG+CNN-ResNet | 88.06 | 96.67 | 98.33 | 98.89 | 99.44 |
|  | KDES+CNN-GoogleNet | 89.72 | 98.61 | 99.17 | 99.17 | 99.17 |
|  | KDES+CNN-ResNet | 82.78 | 95.28 | 97.22 | 98.61 | 98.89 |
|  | All features | 86.94 | 96.67 | 98.61 | 98.61 | 99.17 |
| Late fusion-Product | GOG+CNN-GoogleNet | 91.39 | 98.61 | 99.44 | 99.44 | 99.44 |
|  | GOG+CNN-ResNet | 88.06 | 97.22 | 98.61 | 99.17 | 99.72 |
|  | KDES+CNN-GoogleNet | 91.39 | 98.33 | 99.17 | 99.72 | 99.72 |
|  | KDES+CNN-ResNet | 88.33 | 97.22 | 98.61 | 99.44 | 100.00 |
|  | All features | 92.22 | 98.89 | 99.72 | 99.72 | 99.72 |
| Late fusion-Adaptive | GOG+CNN-GoogleNet | 91.39 | 98.06 | 99.44 | 99.72 | 99.72 |
|  | GOG+CNN-ResNet | 87.78 | 97.50 | 99.17 | 99.44 | 99.72 |
|  | KDES+CNN-GoogleNet | 91.67 | 98.89 | 99.44 | 99.44 | 99.44 |
|  | KDES+CNN-ResNet | 86.94 | 96.94 | 98.61 | 99.17 | 99.72 |
|  | All features | 92.22 | 98.89 | 99.72 | 99.72 | 99.72 |

Note: The best result at each group for each considered rank is in bold.
In our work, extensive experiments are conducted to evaluate the different combination of features when applying the three fusion schemes. The discussion for each fusion scheme is provided as follows:

When applying early fusion scheme, for case A in CAVIAR4REID, as seen in Figure 7 and Tables 1 and 2, the best pairwise combination is (GOG, CNN-GoogleNet) with the recognition rates at rank-1 are 83.61% and 85.28% that are 6.39% and 1.67% higher than the other cases, KDES and CNN-GoogleNet are suitable choices.

4.2.2. Evaluate the proposed fusion schemes

| Methods | Rank=1 | Rank=5 | Rank=10 | Rank=15 | Rank=20 |
|---------|--------|--------|---------|---------|---------|
| Single feature | GOG | 82.78 | 92.22 | 97.50 | 98.33 | 98.89 |
| | KDES | 87.50 | 96.94 | 98.61 | 98.61 | 99.17 |
| | CNN-GoogleNet | 90.56 | 96.67 | 98.33 | 98.33 | 99.17 |
| | CNN-ResNet | 83.06 | 94.17 | 95.83 | 97.22 | 98.06 |
| Early fusion | GOG+CNN-GoogleNet | 90.56 | 98.06 | 98.89 | 99.72 | 99.72 |
| | GOG+CNN-ResNet | 85.56 | 96.67 | 99.17 | 99.44 | 99.44 |
| | KDES+CNN-GoogleNet | 92.22 | 97.78 | 99.44 | 99.72 | 99.72 |
| | KDES+CNN-ResNet | 86.39 | 95.56 | 98.06 | 98.33 | 98.89 |
| | All features | 86.94 | 96.94 | 99.17 | 99.44 | 99.44 |
| Late fusion-Product | GOG+CNN-GoogleNet | 91.67 | 98.61 | 99.72 | 100.00 | 100.00 |
| | GOG+CNN-ResNet | 88.61 | 96.94 | 98.89 | 99.44 | 100.00 |
| | KDES+CNN-GoogleNet | 93.06 | 99.44 | 99.72 | 99.72 | 99.72 |
| | KDES+CNN-ResNet | 89.44 | 96.94 | 98.89 | 98.89 | 99.44 |
| | All features | 92.5 | 98.89 | 100.00 | 100.00 | 100.00 |
| Late fusion-Adaptive | GOG+CNN-GoogleNet | 93.33 | 99.44 | 99.72 | 99.72 | 100.00 |
| | GOG+CNN-ResNet | 88.89 | 96.67 | 99.17 | 99.17 | 99.44 |
| | KDES+CNN-GoogleNet | 92.22 | 98.89 | 99.72 | 99.72 | 100.00 |
| | KDES+CNN-ResNet | 86.94 | 96.39 | 98.06 | 99.44 | 99.72 |
| | All features | 91.94 | 98.89 | 100.00 | 100.00 | 100.00 |

Note: The best result at each group for each considered rank is in bold.

collection of huge training dataset is unfeasible or the high computing platform is unaffordable. In other cases, KDES and CNN-GoogleNet are suitable choices.

Table 5. Summary of all experimental results for RAiD without data augmentation.

| Methods | Rank=1 | Rank=5 | Rank=10 | Rank=15 | Rank=20 |
|---------|--------|--------|---------|---------|---------|
| Single feature | GOG | 86.93 | 96.57 | 98.69 | 99.36 | 99.70 |
| | KDES | 85.92 | 96.21 | 98.36 | 99.18 | 99.52 |
| | CNN-GoogleNet | 83.68 | 92.97 | 95.32 | 96.71 | 97.88 |
| | CNN-ResNet | 84.92 | 96.39 | 98.82 | 99.61 | 99.93 |
| Early fusion | GOG+CNN-GoogleNet | 89.60 | 97.18 | 99.24 | 99.70 | 99.85 |
| | GOG+CNN-ResNet | 90.15 | 98.29 | 99.67 | 99.84 | 99.90 |
| | KDES+CNN-GoogleNet | 86.65 | 94.63 | 97.12 | 98.36 | 99.05 |
| | KDES+CNN-ResNet | 89.03 | 97.78 | 99.49 | 99.84 | 99.96 |
| | All features | 90.49 | 98.63 | 99.70 | 99.87 | 99.93 |
| Late fusion-Product | GOG+CNN-GoogleNet | 90.28 | 96.51 | 98.30 | 98.85 | 99.37 |
| | GOG+CNN-ResNet | 90.83 | 98.27 | 99.66 | 99.88 | 99.99 |
| | KDES+CNN-GoogleNet | 89.39 | 96.26 | 98.08 | 98.84 | 99.23 |
| | KDES+CNN-ResNet | 91.06 | 98.17 | 99.57 | 99.76 | 99.87 |
| | All features | 92.58 | 98.03 | 99.25 | 99.57 | 99.69 |
| Late fusion-Adaptive | GOG+CNN-GoogleNet | 89.14 | 96.66 | 98.29 | 99.21 | 99.49 |
| | GOG+CNN-ResNet | 90.95 | 98.09 | 99.48 | 99.73 | 99.82 |
| | KDES+CNN-GoogleNet | 90.43 | 96.90 | 98.63 | 99.15 | 99.57 |
| | KDES+CNN-ResNet | 90.89 | 98.18 | 99.64 | 99.85 | 99.97 |
| | All features | 92.41 | 98.29 | 99.36 | 99.72 | 99.82 |

Note: The best result at each group for each considered rank is in bold.
next best results in the case of without and with data augmentation, respectively. Meanwhile, two best pairwise combinations in case B are (GOG, CNN-GoogleNet) and (KDES, CNN-GoogleNet). The performance in these two cases is comparable. It is interesting to see that combining all evaluated features does not provide the best performance in this case (Figures 8 and 9).

Concerning the late fusion strategy, the experiments are performed for two cases: product-rule and query-adaptive-based schemes. Figures 10–12 indicate the CMC curves in the case of applying the product-rule late fusion scheme on case A, case B of CAVIAR4REID, and RAiD dataset, respectively. In case A, the best results are provided by combination (GOG, CNN-GoogleNet) and all features. The matching rates at considered ranks when using two above strategies are similar, the best results at rank-1 are 84.44% and 86.39% which are 0.27%, 1.67% higher than the next best results. While in case B,

Table 6. Summary of all experimental results for RAiD with data augmentation.

| Methods                  | Rank=1 | Rank=5 | Rank=10 | Rank=15 | Rank=20 |
|--------------------------|--------|--------|---------|---------|---------|
| Single feature           |        |        |         |         |         |
| GOG                      | 87.41  | 96.32  | 98.41   | 99.14   | 99.57   |
| KDES                     | 87.02  | 96.54  | 98.49   | 99.37   | 99.52   |
| CNN-GoogleNet            | 87.51  | 95.02  | 96.87   | 97.94   | 98.64   |
| CNN-ResNet               | 84.25  | 95.77  | 97.94   | 98.90   | 99.42   |
| Early fusion             |        |        |         |         |         |
| GOG+CNN-GoogleNet        | 91.45  | 97.91  | 99.17   | 99.55   | 99.72   |
| GOG+CNN-ResNet           | 90.40  | 98.45  | 99.58   | 99.90   | 99.96   |
| KDES+CNN-GoogleNet       | 90.25  | 96.30  | 97.88   | 98.54   | 98.99   |
| KDES+CNN-ResNet          | 88.45  | 97.63  | 99.11   | 99.67   | 99.79   |
| All features             | 90.77  | 98.46  | 99.69   | 99.96   | 99.99   |
| Late fusion-Product      |        |        |         |         |         |
| GOG+CNN-GoogleNet        | 91.77  | 97.18  | 98.69   | 99.37   | 99.54   |
| GOG+CNN-ResNet           | 90.15  | 97.97  | 99.23   | 99.60   | 99.76   |
| KDES+CNN-GoogleNet       | 91.13  | 97.12  | 98.73   | 99.28   | 99.52   |
| KDES+CNN-ResNet          | 91.01  | 98.08  | 99.24   | 99.63   | 99.79   |
| All features             | 92.86  | 98.36  | 99.37   | 99.67   | 99.85   |
| Late fusion-Adaptive     |        |        |         |         |         |
| GOG+CNN-GoogleNet        | 90.98  | 97.23  | 98.94   | 99.37   | 99.58   |
| GOG+CNN-ResNet           | 90.94  | 98.03  | 99.18   | 99.67   | 99.81   |
| KDES+CNN-GoogleNet       | 91.76  | 97.27  | 98.85   | 99.37   | 99.55   |
| KDES+CNN-ResNet          | 89.84  | 97.85  | 99.20   | 99.58   | 99.79   |
| All features             | 92.80  | 98.39  | 99.37   | 99.70   | 99.84   |

Note: The best result at each group for each considered rank is in bold.
one of three combinations (GOG, CNN-GoogleNet), (KDES, CNN-ResNet) and all features achieve comparable accuracy. For RAiD dataset, the performance of all combinations in product-rule late fusion scheme is slightly different.

For query-adaptive late fusion schemes, Figures 13–15 show the performance of the query-adaptive late fusion scheme in case A, case B of CAVIAR4REID, and RAiD, respectively. Different from early fusion case, the performance of person re-identification can achieve the highest accuracy when fusing all chosen features. Furthermore, the combination between a hand-crafted feature and CNN-GoogleNet one still provides a higher recognition rate compared to the case of using CNN-ResNet feature. Actually, when observing obtained results in more detail in Tables 1–4 we realize that by combining GOG/KDES with CNN-GoogleNet we can achieve results which are very similar to those when using all four features. This is clearly shown that in the case of without utilizing data augmentation, the best accuracies at rank-1 in the case of using the fusion schemes are 84.44% and 92.22%
which are increased by 6.38% and 6.66% compared to those in the case of exploiting only single feature in case A and case B on CAVIAR4REID dataset, respectively. In the case of using data augmentation techniques, the fusion schemes still confirm their effectiveness, however, the matching rates are not increased as strongly as those in the case of without data augmentation. By utilizing the fusion schemes, the matching rates at rank-1 are increased by 5% and 2.77% compared to the best results when using the individual feature in case A and case B. For the RAiD dataset, the accuracies at rank-1 while applying fusion schemes are improved by 6.65% and 5.35% in the case of without and with data augmentation techniques. Overall, all fusion schemes allow to improve the accuracy of person re-identification. However, with the early fusion, all features have to be concatenated to feed to the classifier, it may require time for classification training especially when the dimension of feature vector is high. Concerning two late fusions, some parallel mechanisms can be applied.

Figure 10. Evaluation of the performance of the product-rule late fusion schemes (a) CAVIAR4REID-case A without data augmentation (b) CAVIAR4REID-case A with data augmentation.

Figure 11. Evaluation of the performance of the product-rule late fusion schemes (a) CAVIAR4REID-case B without data augmentation (b) CAVIAR4REID-case B with data augmentation.
Moreover, we evaluate the consuming time for the proposed framework including two main stages that are feature extraction and classification ones. It is worth to note that fine-tune process in the deep learning feature extraction step as well as training phase in the classification step is performed off-line. Therefore, consuming time in both steps is calculated for only tested images. Table 7 shows the consuming time in both steps for a given image. It takes 219, 226, 14, and 525 ms to extract GOG, KDES, CNN-GoogleNet, and CNN-ResNet for an image with Processor Intel(R) Xeon(R) CPU E5-2420 v2 @ 2.20 GHz, Ram 16 G, GPU 1080 Ti, respectively. Because GOG and KDES are implemented in Matlab and do not take benefit of the GPU. Among these features, ResNet spends the longest time in feature extraction due to its deep architecture. From this Table, we realize that time for classification step is much smaller than that of feature extraction. It spends only 0.047, 0.519, 0.015, 0.014 ms in case of using separated feature GOG, KDES, CNN-GoogleNet, and CNN-
ResNet, respectively. Early fusion spends 0.655 ms for the prediction process which is a little longer than the prediction time in case of using KDES. For the two late fusion schemes, fusion strategies are performed when the scores corresponding to all extracted features are provided. As provided in the Table 7, the computational time for two late fusion is relatively small. As conclusions, feature extraction is the most time-consuming step in the proposed method.

![Figure 14](image1.png)

**Figure 14.** Evaluation of the performance of the query-adaptive late fusion schemes (a) CAVIAR4REID-case B without data augmentation (b) CAVIAR4REID-case B with data augmentation.

![Figure 15](image2.png)

**Figure 15.** Evaluation of the performance of the query-adaptive late fusion schemes (a) without data augmentation (b) with data augmentation on the RAiD dataset.

![Table 7](image3.png)

**Table 7.** Summary of computation time for feature extraction and classification steps.

| Feature extraction time (ms) | Classification time (ms) |
|-----------------------------|--------------------------|
| GOG | KDES | CNN- GoogleNet | CNN- ResNet | GOG | KDES | CNN- GoogleNet | CNN- ResNet | Early | Product- rule Adaptive |
|---|---|---|---|---|---|---|---|---|---|
| 219 | 226 | 14 | 525 | 0.047 | 0.519 | 0.015 | 0.014 | 0.655 | ~0 | 0.003 |
5. Conclusions and future work

This paper is an extended version of Nguyen et al. (2018). In this study, we investigated in detail the choice of features and their combination for person re-identification. Compared to our previous work (Nguyen et al., 2018), two extra features that are GOG and CNN-ResNet are proposed for person image representation. A huge number of extensive experiments are conducted on two benchmark datasets (CAVIAR4REID and RAiD) to show the effectiveness of the proposed framework. From the obtained results, we can conclude that for a small dataset, GOG feature is really effective to represent person images. Nevertheless, for a larger dataset, the other features (KDES, CNN-GoogleNet, CNN-ResNet) can be chosen for representing person images in person re-identification. Furthermore, although four evaluated features are all robust and effective in person re-identification, the matching rates can be further improved by exploiting these fusion schemes. When data augmentation is not considered, the matching rates at rank-1 are improved by 6.38%, 6.66%, and 6.65% in case A, case B of CAVIAR4REID, and RAiD, respectively. While applying data augmentation, these values are 6.66%, 2.77%, and 5.35% in case A, case B of CAVIAR4REID and RAiD, respectively. By observing the obtained results we realize that the best pairwise are (GOG/KDES, CNN-GoogleNet) for our experiments. Furthermore, in this study, CNN-ResNet does not show its benefit compared to the others. In future work, we will perform extensive experiments on a larger dataset to investigate these features on different cases. In addition, different late fusion strategies can be applied.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01-2017.12.

Notes on contributors

Thuy-Binh Nguyen graduated at the School of Telecommunications and Electronics from Hanoi University of Science and Technology, Vietnam. She obtained her MS degree in Electronic Engineering from the University of Transport and Communications. She is a PhD student of HUST. Her researches include person re-identification, person search in images and video databases.

Thi-Lan Le graduated in Information Technology from the Hanoi University of Science and Technology (HUST), Vietnam. She obtained her MS in Signal Processing and Communication from the HUST, Vietnam. In 2009, she received her PhD in Video Retrieval at the INRIA Sophia Antipolis, France. She is currently a lecturer/researcher at the Computer Vision Department, HUST, Vietnam. Her research interests include computer vision, content-based indexing and retrieval, video understanding and human-robot interaction.

Nam Pham Ngoc received her BEng in Electronics and Telecommunications from the Hanoi University of Science and Technology (Vietnam) and MSc in Artificial Intelligence from the K.U. Leuven (Belgium) in 1997 and 1999, respectively. He obtained his PhD in Electrical Engineering from the K.U. Leuven in 2004. From 2004 until now he has been working at the Hanoi University of Science and Technology, Vietnam. His research interests include artificial intelligence, QoS/QoE management for multimedia applications, reconfigurable embedded systems and low-power embedded system design.
References

Bo, L., Ren, X., & Fox, D. (2010). Kernel descriptors for visual recognition. *Advances in neural information processing systems* (pp. 244–252).

Cheng, D. S., Cristani, M., Stoppa, M., Bazzani, L., & Murino, V. (2011). Custom pictorial structures for re-identification. *British machine vision conference (BMVC)*.

Das, A., Chakraborty, A., & Roy-Chowdhury, A. K. (2014). Consistent re-identification in a camera network. *European conference on computer vision* (pp. 330–345). Springer.

Eisenbach, M., Kolarow, A., Vonrordan, A., Niebling, J., & Gross, H.-M. (2015). Evaluation of multi feature fusion at score-level for appearance-based person re-identification. *2015 international joint conference on neural networks (IJCNN)* (pp. 1–8). IEEE.

Gao, M., Ai, H., & Bai, B. (2016). A feature fusion strategy for person re-identification. *2016 IEEE international conference on image processing (ICIP)* (pp. 4274–4278). IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Karanam, S., Gou, M., Wu, Z., Rates-Borras, A., Camps, O., & Radke, R. J. (2016). A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *arXiv preprint arXiv:1605.09653*.

Karanam, S., Gou, M., Wu, Z., Rates-Borras, A., Camps, O., & Radke, R. J. (2018). A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1, 1–1.

Kittler, J., Hatef, M., Duin, R. P., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3), 226–239.

Lejbole, A. R., Nasrollahi, K., & Moeslund, T. B. (2017). Enhancing person re-identification by late fusion of low-, mid- and high-level features. *IJET Biometrics*, 7, 125–135.

Li, S., Shao, M., & Fu, Y. (2015). Cross-view projective dictionary learning for person re-identification. *IJCAI* (pp. 2155–2161).

Liao, S., Hu, Y., Zhu, X., & Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2197–2206).

Liu, C., Gong, S., Loy, C. C., & Lin, X. (2014). Evaluating feature importance for re-identification. *Person re-identification* (pp. 203–228). Springer.

Matsukawa, T., Okabe, T., Suzuki, E., & Sato, Y. (2016). Hierarchical gaussian descriptor for person re-identification. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1363–1372).

Matsukawa, T., Okabe, T., Suzuki, E., & Sato, Y. (2017). Hierarchical gaussian descriptors with application to person re-identification. *arXiv preprint arXiv:1706.04318*.

Nguyen, T.-B., Le, T.-L., Nguyen, D.-D., & Pham, D.-T. (2018). A reliable image-to-video person re-identification based on feature fusion. *Asian conference on intelligent information and database systems* (pp. 433–442). Springer.

Pham, T. T. T., Le, T. L., Dao, T. K., Le, D. H., & Nguyen, V. (2015). A robust model for person re-identification in multi-modal person localization. *UBICOMM 2015* (p. 51).

Pham, T. T. T., Le, T.-L., Vu, H., Dao, T. K., & Nguyen, V. T. (2017). Fully-automated person re-identification in multi-camera surveillance system with a Robust kernel descriptor and effective shadow removal method. *Image and Vision Computing*, 59, 44–62.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).

Wang, G., Lai, J., & Xie, X. (2017). P2snet: Can an image match a video for person re-identification in an end-to-end way? *IEEE Transactions on Circuits and Systems for Video Technology*. 
Wu, S., Chen, Y.-C., Li, X., Wu, A.-C., You, J.-J., & Zheng, W.-S. (2016). An enhanced deep feature representation for person re-identification. *2016 IEEE winter conference on applications of computer vision (WACV)* (pp. 1–8). IEEE.

Zhang, D., Wu, W., Cheng, H., Zhang, R., Dong, Z., & Cai, Z. (2017). Image-to-video person re-identification with temporally memorized similarity learning. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., & Tian, Q. (2016). Mars: A video benchmark for large-scale person re-identification. *European conference on computer vision* (pp. 868–884). Springer.

Zheng, L., Wang, S., Tian, L., He, F., Liu, Z., & Tian, Q. (2015). Query-adaptive late fusion for image search and person re-identification. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1741–1750).

Zheng, L., Yang, Y., & Hauptmann, A. G. (2016). Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*. 