Person Re-identification using group context

Yiqiang Chen, Stefan Duffner, Andrei Stoian, Jean-Yves Dufour, Atilla Baskurt

Person Re-identification using group context

Yiqiang Chen[1], Stefan Duffner[1], Andrei Stoian[2], Jean-Yves Dufour[2], and Atilla Baskurt[1]

[1] Univ Lyon, INSA-Lyon, CNRS, LIRIS, F-69621, Villeurbanne, France
[2] Thales Services, ThereSIS, Palaiseau, France

Abstract. The person re-identification task consists in matching person images detected from surveillance cameras with non-overlapping fields of view. Most existing approaches are based on the person's visual appearance. However, one of the main challenges, especially for a large gallery set, is that many people wear very similar clothing. Our proposed approach addresses this issue by exploiting information on the *group* of persons around the given individual. In this way, possible ambiguities are reduced and the discriminative power for person re-identification is enhanced, since people often walk in groups and even tend to walk alongside strangers. In this paper, we propose to use a deep convolutional neural networks (CNN) to extract group feature representations that are invariant to the relative displacements of individuals within a group. Then we use this group feature representation to perform group association under non-overlapping cameras. Furthermore, we propose a neural network framework to combine the group cue with the single person feature representation to improve the person re-identification performance. We experimentally show that our deep group feature representation achieves a better group association performance than the state-of-the-art methods and that taking into account group context improves the accuracy of the individual re-identification.

Keywords: Person re-identification, Convolutional Neural Network, Group association

1 Introduction

Person re-identification consists in matching identities across images captured from disjoint camera views. Increasing attention has been dedicated to person re-identification algorithms in the past few years as they have important applications in video surveillance and are necessary for cross-camera tracking, multi-camera event detection, and pedestrian retrieval.

Despite the recent progress in the performance of person re-identification methods, several difficulties remain since a person's appearance often undergoes large variations across different cameras due to varying view points and illumination conditions. Different human poses and partial occlusions further make the task challenging. Moreover, the problem becomes increasingly difficult when

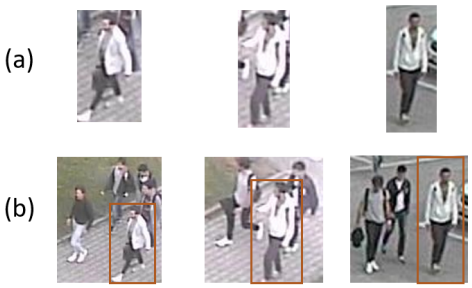

Fig. 1. (a) Single person images. (b) Corresponding group images of (a). Even for a human, it may be difficult to tell if the three top images belong to the same person or not. Using the context of the surrounding group, it is easier to see that the middle and right images belong to the same person and the left image belongs to another person.

there is a large number of candidate persons since, in practice, many persons have similar appearance as they share the same visual attributes or wear very similar clothes and colours.

Most existing approaches only use the visual appearance of a *single* person for its re-identification in different images. However, this can lead to strong ambiguities, for example when people wear similar clothes, as shown in Fig. 1. To address this problem, context information about the surrounding group of persons can be used. In realistic settings, people often walk in groups rather than alone. Thus, the appearance of these groups can serve as visual context and help to determine whether two images of persons with similar clothing belong to the same individual.

However, matching the surrounding people in a group in different views is also challenging. On the one hand, it undergoes the variations of single person appearances. On the other hand, the number of persons and their relative position within the group can vary over time and across cameras. Further, partial occlusions among individuals are very likely in groups.

In this paper, we propose to extract group feature representations using a deep convolutional neural network. First, we train the model with single-person re-identification data. Then, transfer it to the group association problem. In order to cope with the relative displacements of persons in a group, we applied a Global Max-Pooling (GAP) operation of CNN activations to achieve translation invariance in the resulting representation. Furthermore, we measure a group context distance with this representation and then combine it with the distance measure based on single-person appearance to enhance the re-identification accuracy.

The main contributions of this paper are the following:

– we learn a deep feature representation with displacement invariance and apply it to the group association problem. Our experiments show that this approach outperforms the state-of-the-art on group association.

– we propose a novel way to combine group context and single-person appearance and experimentally show that the group information can improve the person re-identification performance.

2 Related Work

Person re-identification Existing person re-identification approaches mainly concentrate on two aspects: building a robust feature representation for a person's visual appearance and learning a distance metric. Some specifically designed features are proposed in order to address the changes in view-point and pose, for example ELF [4] , SADALF [3], LOMO features [10]. The main metric learning methods like KISSME [7], LFDA [12] and XQDA [10]. Recently, many methods based on deep convolutional neural network(CNN) have been proposed for person re-identification, like [9, 1, 16]. The advantage of these deep learning approaches is that they incorporate feature representations and the distance metric into an integrated model that is jointly learnt from the data.

Group association In the literature, there are several group association (or re-identification) approaches. Zheng et al. [19] extracted visual words which are the clusters of SIFT+RGB features in a group image. Then they built two descriptors that describe the ratio information of visual words between local regions to represent group information. Cai et al. [2] used covariance descriptor to encode group context information. And Lisanti et al. [11] proposed to learn a dictionary of sparse atoms using patches extracted from single person images. Then the learned dictionary is exploited to obtain a sparsity-driven residual group representation. These approaches can be severely affected by background clutter, and thus a preprocessing stage is necessary. For example, in [19, 2] a background subtraction is performed before feature extraction. And in [11], three pedestrian detectors based on respectively deformable part models, aggregated channel features and RCNN were used to weight the contribution of each pixel in the histogram computation.

Some other approaches use trajectory features to describe group information. Wei et al. [17], for example, presented a group extraction approach by clustering the persons' trajectories observed in a camera view. They introduced person-group features composed of two parts: SADALF features [3], extracted after background subtraction and representing the visual appearance of the accompanying persons of a given individual, and a signature encoding the position of the subject within the group. Similarly, Ukita et al. [15] determined for each pair of pedestrians whether they form a group or not, using spatio-temporal features of their trajectories like relative position, speed and direction. Then, the group features composed of the trajectory features (position, speed, direction) of individuals in each group, the number of persons as well as the mean colour histograms of the individual person images. However, when people walk in group, the position and speed are not always uniform. Thus, the trajectory-based features may not be precise and change significantly over time.

Unlike these methods, the advantage of our approach is that there is no need for a pre-processing stage of person detection or background subtraction. Our model is pre-trained on single-person re-identification data to learn the discriminative features that distinguish identities in images. The applied global max-pooling operation captures maximum activations over feature maps,which correspond to salient discriminative patches in the input image. Thus the proposed model is, by design, invariant to displacements of individuals within a group. Moreover, the deep neural network that we employed can provide a richer feature representation to describe groups than the colour and texture features used by existing methods.

3 Proposed method

In this section, we first describe our group association method and further introduce how we use the group information to improve person re-identification performance.

3.1 Group association

In the first step, we train a neural network predicting the identities of the images, given an input image resized to 64x124. The model can be a CNN pre-trained on ImageNet, like Alexnet [8] or ResNet-50 [5]. The final fully-connected (FC) layer is replaced by another FC layer with an output dimension of N, with N being the number of identities in the training set.

Then, the CNN is fine-tuned in a supervised way, using images and identity labels from a separate person re-identification dataset. To this end, we minimise the following softmax cross-entropy loss on the given classification task:

$$E_{identification} = -\sum_{k=1}^{N} y_k log(P(y_k = 1|x)) \ , \tag{1}$$

$$\text{with} \quad P(y_j = 1|x) = \frac{e^{W_j^T x + b_j}}{\sum_{k=1}^{N} e^{W_k^T x + b_k}} \ , \tag{2}$$

where y is the one-hot coded identity label, x is the input to the last fully-connected layer, W and b are weights and bias of the last fully-connected layer and $P(y_j = 1|x)$ is the predicted probability that the input x corresponds to identity j. The intuition of this supervised training is that the resulting feature representation can be used for learning similarities between arbitrary pedestrian images and thus be transferred to the task of group re-identification.

After training the model, we discard the FC layer and represent the activation map of the last convolutional layer as a set of K 2D feature channel responses $\mathcal{X} = \mathcal{X}_i, i = 1...K$, where $\mathcal{X}_i$ is the 2D map representing the responses of the i^{th} feature channel. A ReLU activation function is applied as a last step to guarantee that all elements are non-negative. A final location-invariant representation,

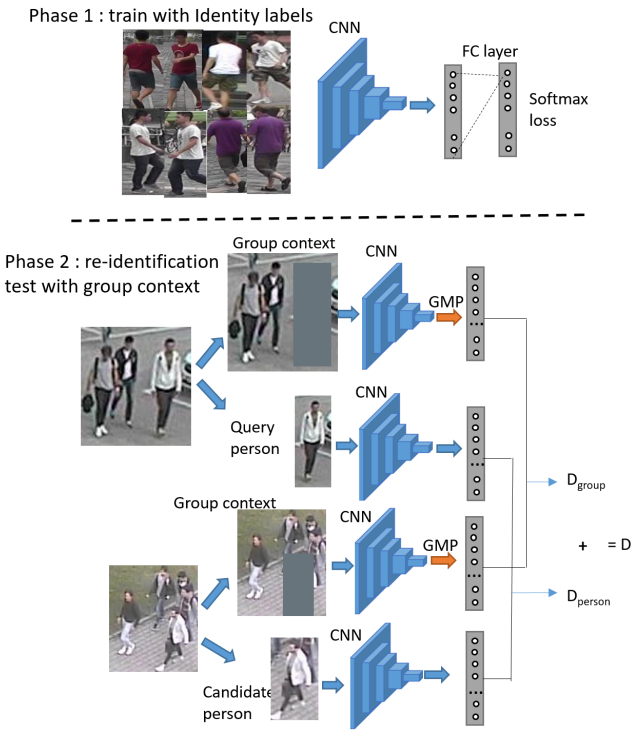

Fig. 2. Overview of our group association assisted re-identification method. The CNN is first trained with person images. Then, group context distance and single-person distance are computed and summed to obtain the final distance.

called Maximum Activations of Convolutions (MAC) [14], is constructed by a spatial max-pooling over all locations concatenated in a K-dimensional vector:

$$f = [f_1 ... f_i ... f_K]^\top, with f_i = \max_{x \in \mathcal{X}_i}(x) \ . \tag{3}$$

When applying the model for group association, a group image resized to 224x244 pixels is given as input (corresponding to a single-person size of roughly 64x128 pixels in the image). The distance between two images is measured with the cosine distance between the feature vectors produced as described above. This feature representation does not encode the location of the activations unlike the activations of fully connected layers, due to the max-pooling operated over the whole last convolution layer feature map. It encodes the maximum local response of each of the convolutional filters. Thus, it offers translation invariance to the resulting representation.

3.2 Group-assisted person re-identification

In our setting, the input data is composed of group images with annotated individual identities and corresponding bounding boxes. As shown in Fig. 2, to

OGRE dataset

Ilidis-group dataset

Fig. 3. Some example images from people group datasets.

explicitly capture both *person* and *group context* features, we divide the input image **I** into two input images to process them separately.

First, a query person image **P** is obtained from the raw group image by using the given annotated bounding box. Second, its group context image **G** is obtained from the raw group image by covering the query person image region with the pixels of the mean image colour. Then, these two images are given to the CNN model explained in Section 3.1. Two parallel branches of this network are employed to extract the feature embeddings for respectively the group context input image and the person image. The two branches are almost identical, except for the last layer, where the MAC feature representation is used for the (larger) group context image and the input vector of the discarded FC layer is used for the single-person image. As illustrated in Fig. 2, after processing the query image, the same procedure is applied to the gallery images in order to compute a distance measure on the two representation between query and candidate image (resulting in 4 feature vectors in total).The advantage of this method is that it can be easily combined with any CNN-based single person re-identification approach. For a given query and candidate image, the cosine distance is used to separately compute a group context distance D_{gr} between the two group images and a person distance D_{id} between the two single-person images. The final distance measure is simply the sum:

$$D(I_i, I_j) = D_{id}(P_i, P_j) + D_{gr}(G_i, G_j) \; . \tag{4}$$

This equation can also be formulated as a weighted sum. Since in our test, there is not a validation set with group images to determine the weight which is a hyper-parameter, we use just a simple sum to combine these two distances.

4 Experiments

4.1 Datasets

The **Market-1501 Dataset** [18] is one of the largest publicly available datasets for human re-identification. It contains 32668 images of 1501 subjects captured

on a campus. The dataset is split into 751 identities for training and 750 identities for test. In our experiments, we uses only the training set of Market-1501 to train the CNN model.

The **Ilidis-group Dataset** is extracted by Zheng et al. [19] from the i-LIDS MCTS dataset. It contains 274 images of 64 groups taken from airport surveillance cameras. Most of the groups have 4 images, either from different camera views or at different times. Some example images are shown in Fig. 3.

The **OGRE Dataset** [11] contains 1279 images of 39 groups acquired by three disjoint cameras pointing at a parking lot. This is a challenging dataset with many different viewpoints and self-occlusions. We manually annotated a subset of this dataset with 450 bounding boxes and 75 identities.

The Cumulative Match Curve (CMC) is employed as evaluation measure for both group association and person re-identification. The CMC curve shows the probability that a query identity appears in the top-k of the ordered candidate list with varying k.

For the group association test, we follow the test protocol in [19, 11]. That is, for each group, one randomly selected image is included in the gallery, all the remaining images form the probe set. The test is repeated 10 times, then the average scores are computed. For the person re-identification test, the images with person bounding boxes are used. We take each person bounding box as query image in turn, and the rest of the images as gallery set. The final result is the average CMC score over all queries.

4.2 Experimental setting

We used ResNet-50 as the CNN model and the weights pre-trained on the ImageNet dataset are used as initialization. For training, data augmentation is performed by randomly flipping the images and cropping central regions with random perturbation. Dropout is applied to the fully connected layers to reduce the risk of over-fitting. The optimization is performed by Stochastic Gradient Descent with a learning rate of 0.001, a momentum of 0.9 and a batch size of 50.

4.3 Group association results

The comparison with the state-of-the-art method on the Ilids-group and OGRE dataset is shown in the Table 1. We compared not only with the group association methods in [19, 11], but also with two encoding techniques, namely IFV [13] and VLAD [6], applied by [11] in group association as well a CNN model with global average pooling (GAP). Our method outperforms the best state-of-the-art method PREF [11] in terms of the Rank 1 score by a margin of 5.6% and 6.1% points on Ilidis-group and OGRE datasets, respectively. This clearly shows the effectiveness of the deep feature representation. Compared to the GAP-based model, using GMP increased the Rank 1 score on the two datasets by 3.9% and 2.9% points, respectively. This demonstrates the benefit of the invariance property of the GMP for group association.

Method	Ilids-Group			OGRE		
	Rank 1	Rank 10	Rank 25	Rank 1	Rank 10	Rank 25
CRRRO+BRO [19]	22.5*	57.0*	**76.0***	-	-	-
IFV [13]	26.1	60.2	75.8	14.6	43.3	76.8
VLAD [6]	26.0	57.0	75.0	13.0	41.1	74.3
PREF [11]	31.1	60.3	75.5	15.1	41.6	75.8
Ours with GAP	32.8	56.0	70.7	18.3	46.8	79.7
Ours with GMP	**36.7**	**60.5**	73.7	**21.2**	**50.4**	**82.2**

Table 1. Comparison with group association state-of-the-art methods on the Ilids-group and OGRE dataset. *: figures extracted from a curve.

Variant	Rank 1	Rank 5	Rank 10
Single person only	47.2	69.3	78.8
Group context only	26.2	57.2	66.3
Sum features	41.1	69.9	77.7
Concatenate features	51.9	**75.1**	81.1
Distance sum without mean image cover	54.1	73.7	80.8
Distance sum with mean image cover	**56.8**	73.7	**81.7**

Table 2. Person re-identification accuracy (in %) on the OGRE dataset.

4.4 Group-assisted person re-identification results

The result of person re-identification is shown in Table 2. We compare the person re-identification results with some variants of our method. *Sum feature* and *Concatenate feature* represent variants that first sum or concatenate the single-person feature representation and the group feature representation and then compute the distance measure on these vectors. We compared also to a variant that retains the query or candidate person image in the group image without covering the corresponding region with the mean colour. The results show that the method proposed in this paper (i.e. covering the person in the group image and summing the person and group distance) achieved the best re-identification results. Covering the person image improves the Rank 1 score by 2.7% points. Since some persons from the same group share very similar context, covering the query or candidate person can better discriminate persons in the same or similar group context. Finally, the combined distance achieves better results than only using the single person distance and the group distance. Overall, our proposed method increases the result by 9.6% points with respect to only using single-person images. This clearly shows that group context has the ability to considerably reduce the appearance ambiguity.

5 Conclusion

In this paper, we presented an effective deep learning-based group association and group-assisted person re-identification approach. The deep group feature

representation is extracted by a CNN and global max-pooling is applied to achieve location-invariance of individuals in group images. We also proposed a method improving single-person re-identification by incorporating the group context, defining a combined distance metric. This method can be combined with any CNN-based single person re-identification approach. We experimentally showed that our method outperforms the state-of-the-art in group association and that the deep group feature representation considerably enhances the person re-identification performance.

Acknowledgement

This work was supported by the Group Image Mining (GIM) which joins researchers of LIRIS Lab. and THALES Group in Computer Vision and Data Mining. We thank NVIDIA Corporation for their generous GPU donation to carry out this research.

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: Proceedings of CVPR. pp. 3908–3916 (2015)
2. Cai, Y., Takala, V., Pietikainen, M.: Matching groups of people by covariance descriptor. In: Pattern Recognition (ICPR), 2010 20th International Conference on. pp. 2744–2747. IEEE (2010)
3. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: Proceedings of CVPR. pp. 2360–2367 (2010)
4. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proceedings of ECCV. pp. 262–275 (2008)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: CVPR. pp. 3304–3311 (2010)
7. Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: Proceedings of CVPR. pp. 2288–2295 (2012)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
9. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid:deep filter pairing neural network for person re-identification. In: Proceedings of CVPR. pp. 152–159 (2014)
10. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of CVPR (2015)
11. Lisanti, G., Martinel, N., Del Bimbo, A., Foresti, G.L.: Group re-identification via unsupervised transfer of sparse features encoding. In: Proceedings of ICCV (2017)
12. Pedagadi, S., Orwell, J., Velastin, S., Boghossian, B.: Local fisher discriminant analysis for pedestrian re-identification. In: Proceedings of CVPR. pp. 3318–3325 (2013)

13. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the fisher vector: Theory and practice. International journal of computer vision 105(3), 222–245 (2013)
14. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of cnn activations. In: International Conference on Learning Representations (2016)
15. Ukita, N., Moriguchi, Y., Hagita, N.: People re-identification across non-overlapping cameras using group features. Computer Vision and Image Understanding 144, 228–236 (2016)
16. Varior, R.R., Haloi, M., Wang, G.: Gated siamese convolutional neural network architecture for human re-identification. In: Proceedings of the ECCV. pp. 791–808 (2016)
17. Wei, L., Shah, S.K.: Subject centric group feature for person re-identification. In: CVPR Workshops. pp. 28–35 (2015)
18. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Computer Vision, IEEE International Conference on (2015)
19. Zheng, W.S., Gong, S., Xiang, T.: Associating groups of people. In: BMVC (2009)