TELKOMNIKA, Vol.15, No.2, June 2017, pp. 726~732
ISSN: 1693-6930, accredited **A** by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v15i2.6121

Human Re-identification with Global and Local Siamese Convolution Neural Network

K. B. Low*[1], U. U. Sheikh[2]
Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310, Johor Bahru, Malaysia
Corresponding author, e-mail: kblow2@live.utm.my*[1], usman@fke.utm.my[2]

Abstract

Human re-identification is an important task in surveillance system to determine whether the same human re-appears in multiple cameras with disjoint views. Mostly, appearance based approaches are used to perform human re-identification task because they are less constrained than biometric based approaches. Most of the research works apply hand-crafted feature extractors and then simple matching methods are used. However, designing a robust and stable feature requires expert knowledge and takes time to tune the features. In this paper, we propose a global and local structure of Siamese Convolution Neural Network which automatically extracts features from input images to perform human re-identification task. Besides, most of the current human re-identification tasks in single-shot approaches do not consider occlusion issue due to lack of tracking information. Therefore, we apply a decision fusion technique to combine global and local features for occlusion cases in single-shot approaches.

Keywords: *siamese network, convolution neural network, human re-identification, surveillance system*

1. Introduction

In recent years, human re-identification task has acquired a lot of attention from researchers because it is a core task in surveillance system. It can be divided into two approaches which are biometric based approaches and appearance based approaches. Biometric based approaches such as iris, fingerprint, face, ears and gait recognition require specific sensors to perform the recognition task. Besides, biometric based approaches hardly perform well in wide area surveillance system due to the low resolution of images. Hence, appearance based approaches are popular to be used in human re-identification task because it is less constrained than biometric approaches and are suitable for surveillance system. Appearance based approaches can be divided into two categories which are single-shot approaches [1-11] and multiple-shot approaches [8, 9], [12-17]. These approaches are determined by the number of images that are used to create their representation. Since single-shot approaches only use single image per person in each camera, therefore it does not contain tracking information like multiple-shot approaches. However, more complex algorithms are required to exploit information from multiple images for multiple-shot approaches.

Human re-identification is a challenging task due to the viewpoint, human pose and orientation, illumination and background variations. Most of the researchers are focusing on the design of hand-crafted features to perform a stable and robust human representation and then apply a simple matching such as Bhattacharyya distance and L1-norm, L2-norm based distance to determine the similarity. For example, shape and appearance context model is proposed by Wang et al. [3]. In their proposed method, the human image is segmented into regions and their spatial color information is registered into a co-occurrence matrix. Their approach works well when human is seen from similar viewpoints. Cai et al [6] proposed a patchbased approach by using Canny edge detection algorithm. Extracted patches are represented by most dominant color and frequency within the patch. Sequence-to-sequence matching is used in their approach. However, their algorithm is computationally expensive. In [7], Bak et al performed person re-identification task by using spatial covariance regions of their body parts. Four human body parts which are torso, left arm, right arm and legs are detected by Histogram of Oriented Gradients (HOG). Cheng et al [9] applied pictorial structures model for part based human detection. They computed HSV histograms and MSCRs based on different body parts. However, the approach proposed by [7] and [9] performs well only when the pose estimators

Received February 5, 2017; Revised April 18, 2017; Accepted May 3, 2017

work accurately. Farenzena et al [8] extracted three types of features which are weighted color histograms, maximally stable color regions (MSCR) and recurrent high-structured patches (RHSP) to model human appearance. Besides, they exploit symmetry property in human images to minimize the effect of view variation. Although their proposed method achieves certain robustness, it is quite time consuming to extract these three features. To deal with pose variations, Gheissari et al [12] fit a decomposable triangulated graph to capture the spatial distribution of the local descriptors over time. However, the drawback of their approach is similar to [3] that is, it is only applicable for human seen from similar viewpoints.

Besides hand-crafted feature design approach, metric learning based method is another research direction for researchers to perform human re-identification task. For metric learning based method, simplefeatures are extracted and similarity distance is measured by using a distance metric. The goal of metric learning based method is to minimize the similarity distance for similar pairs and maximize the distance for dissimilar pairs. Dikmen et al [10] proposed the Large Margin Nearest Neighbor (LMNN-R) algorithm to learn the most effective metric. Their metric learning method typically requires enormous labeled target pairs. Zheng et al [11] presented a Probabilistic Relative Distance Comparison (PRDC) model which focused on maximizing the probability that a similar pair have smaller distance. Nevertheless, in their method, noise information of features was not taken into consideration.

In a nutshell, designing and learning a robust and stable feature still remains an open problem to perform human re-identification task. One direction is to use deep learning methods as it integrates both feature extraction and metric learning in a single framework. Until now, there are only several papers on reidentification task by using deep learning methods. Li et al [18] proposed deep filter pairing neural network (FPNN) to handle photometric and geometric transformations. Their proposed method was the first work that applied deep learning in human re-identification task. Another deep learning work which used Siamese Convolution Neural Network (SCNN) is presented in [19]. SCNN architecture consists of two sub-networks which are connected by similarity layers. Three overlapping body parts are used to extract features and compute similarity metric separately. Final similarity score is obtained by summing up these three similarity metrics.

2. Research Approach

Convolution Neural Network (CNN) consists of multiple layers with combination of convolution layers, pooling layers and fully connected layer. Convolution layer is used to detect same features at different locations in input image. Activation function is applied on feature maps which are computed from convolution layer. Pooling layer reduces the spatial resolution of each input feature map to achieve certain degree of shift, distortion and small tranformations invariance. Fully connected layer is applied after convolution and pooling layers. CNN is one kind of deep neural network. It works based on three basic ideas which are local receptive fields, shared weights and pooling. It has less parameters than fully connected networks due to its shared weights and local receptive field properties. Therefore, training a CNN is faster than fully connected networks. The advantage of CNN is that it can automatically extract features from 2D input images.

Subjects in training set are generally different than in testing set, therefore SCNN is used to make the person re-identification task as binary classification which is “sample pair with label” mode that is shown in Figure 1. In this paper, we propose a SCNN for global and local structures to perform person reidentification task.

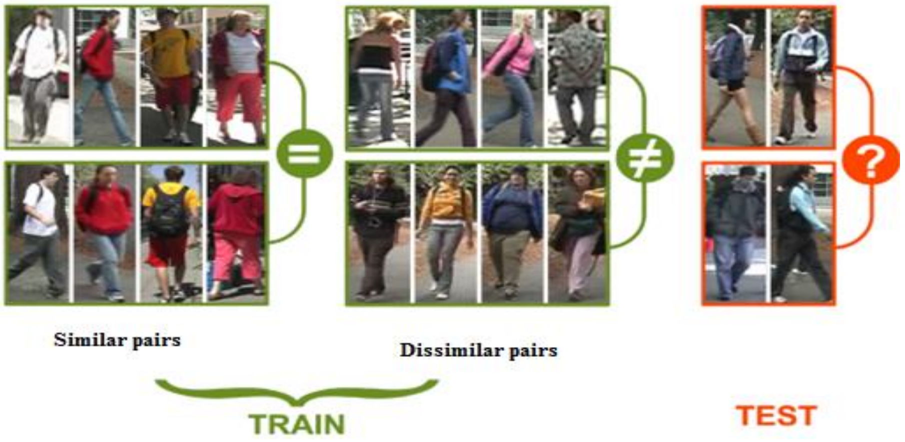

Figure 1. Example of similar and dissimilar pairs which are used for training set, test set are formed like training set for binary classification

2.1. Architecture

In our proposed architecture, the whole image of 128x48 pixels is used for global representation while input image is divided into 4 horizontal stripes with 32x48 pixels and 2 vertical stripes with 128x24 pixels for local representation which is shown in Figure 2. Each part is used as input to the CNN. Image pair from different cameras is passed through the SCNN which consists of 2 identical CNNs with common parameters. Contrastive cost function is applied to decrease the distance for similar pair and increase the distance for dissimilar pair. At the end, seven features are extracted from global and local parts. These features are used to compute similarity metrics between probe and gallery set.

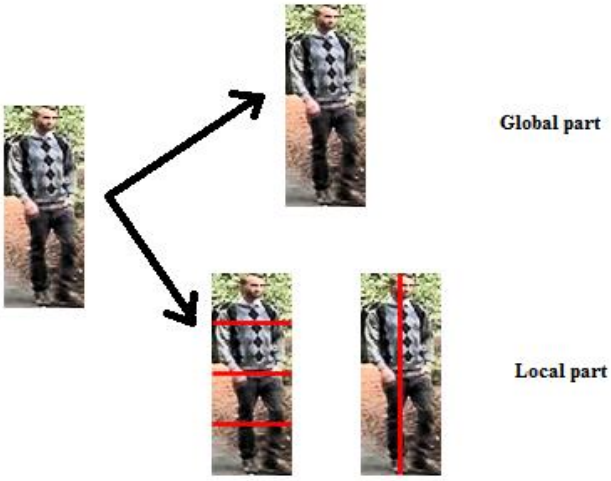

Figure 2. Global and Local parts from input image

2.2. Convolution Neural Network

Figure 3 illustrates the proposed CNN architecture for the global part and it is composed of 7 layers which are 3 convolution layers (C1, C3 and C5), 3 pooling layers (S2, S4 and S6) and 1 fully connected layer (F7). The feature maps in C1, C3 and C5 are 30, 60 and 100 respectively. The number of feature maps in pooling layers is exactly the same with their previous convolution layers, but the size is reduced to half. Number of output neurons for fully connected layer is 120 dimensions which represents the final feature vector. The filter size for C1, C3 and C5 are 5x5, 7x7 and 3x3.

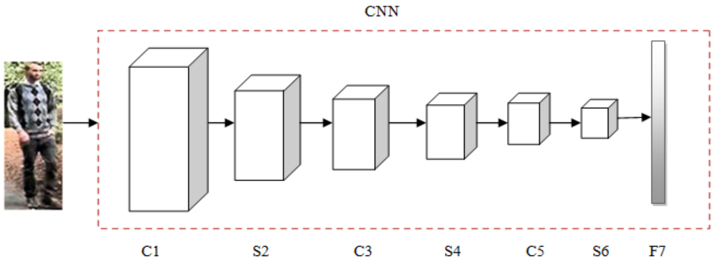

Figure 3. CNN architecture for global part

Figure 4 illlustrates the proposed CNN architecture for the local part, whereby the horizontal stripe is composed of 2 convolution layers (C1 and C3) and 2 pooling layers (S2 and S4) and 1 fully connected layer (F5). The number of feature maps in C1 and C3 are 20 and 40 while the final fully connected layer is a one dimensional layer, with 60 neurons. The filter size for C1 and C3 for local part is 5x5 and 3x3. The CNN for local part with vertical stripe is similar to the horizontal stripe; however the only difference is that the filter for C3 is 5x5.

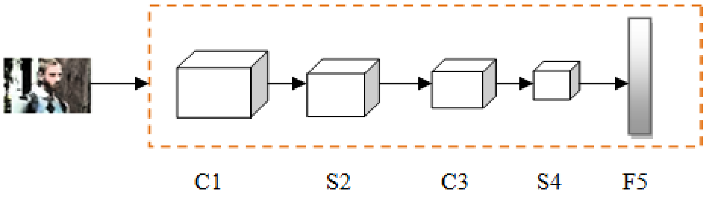

Figure 4. CNN architecture for local part

2.3. Siamese Convolution Neural Network

SCNN is applied on image pairs as shown in Figure 5. Both of the images will pass through the same CNN architecture with common parameters to obtain their feature vectors. L1 norm or Manhattan distance is used to compute similarity metrics as in Equation 1.

$$E_w(X_1, X_2) = || \; G_w(X_1) - \; G_w(X_2) \; || \qquad (1)$$

Where $G_w(X_1)$ and $G_w(X_2)$ are feature vectors after passing through the CNN, || || represents the similarity measure which is Manhattan distance, E_w corresponds to the energy between images. When the image pair is a similar image pair, then the energy will be lower, otherwise the energy level will be high (as in dissimilar image pair).

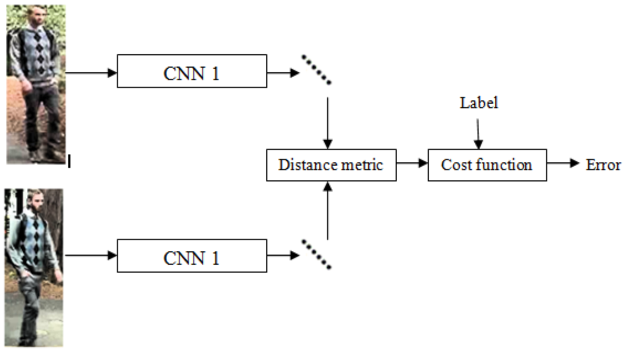

Figure 5. Illustration of Siamese Convolution Neural Network proposed in this work

2.4. Cost Function

Contrastive cost function shown in Equation 2 is applied to discriminate between similar and dissimilar pairs. By using contrastive cost function, distance between similar pair will be decreased while distance between dissimilar pair will be increased.

$$L_0(W,Y,X_1,X_2) = (Y)\frac{2}{Q}\,(E_w)^2 + (1-Y)2Qe^{\frac{-2.7726E_W}{Q}} \qquad (2)$$

Where Q represents the total number of output neurons and Y is the label of the image pair, if the image pair is similar pair, then Y = 1. Y = 0 when image pair is dissimilar pair.

2.5. The Proposed Decision Fusion

A set of feature vectors are computed from positive pairs and a set of feature vectors are computed from negative pairs. These feature vectors are stored in a database. Decision fusion is applied after feature vectors are formed for global and local part. Decision fusion involves with distance computation, weight computation, and combined decision to get the final distances.

For distance computation step, distances of feature vectors between training and testing set are computed by using Euclidean distance, then minimum distances for similar class and dissimilar class are determined by Equation 3.

$$D = \min d\big(F_v(test), F_v(train)\big) \qquad (3)$$

Where $F_v(test)$ is the *v*-th feature vector of local parts for testing images , $F_v(train)$ is *v*-th feature vector of local parts for training images and $d(.)$ is distance measure between two feature vectors. Distance matrix is formed based on Equation 4.

$$D_1 = \begin{bmatrix} D_{1,similar} & D_{1,dissimilar} \\ D_{2,similar} & D_{2,dissimilar} \\ & . \\ & . \\ & . \\ D_{7,similar} & D_{7,dissimilar} \end{bmatrix} \qquad (4)$$

For weight computation step, weight of the *v*-th local feature vector is computed by using Equation 5. In each row of D_1, the lower distance value is represented by $D'_{v,1}$ and another distance value is represented by $D'_{v,2}$. The degree of importance of the *v*-th local feature vector is represented by weight λ_v, therefore the higher the λ_v, the importance of *v*-th local feature is lower. This strategy ensures that less discriminant feature vectors are assigned with higher weights. Weight λ_v is normalized as λ'_v. Finally, final distance for similar class and dissimilar class is computed by using Equation 6.

$$\lambda_v = \frac{D'_{v,1}}{D'_{v,2}} \qquad (5)$$

$$D_L = \lambda'_v D_1 \qquad (6)$$

$$D_2 = \begin{bmatrix} D_{L,similar} & D_{L,dissimilar} \end{bmatrix} \qquad (7)$$

3. Results and Analysis

Viewpoint Invariant Pedestrian Recognition (VIPeR) dataset is used in our experiments. VIPeR is the most widely used benchmark in the field of human re-identification because it is quite challenging, since it suffers from viewpoint and illumination changes between the two cameras, giving a disjointed view. We utilize 100 pairs of human images as training images and testing on 20 pairs of human images in our experiment, due to the constraint on available computation power. Inside the 100 pairs of human images, there are 50 pairs of positive training images, which are the same human images captured with the same cameras. The 50 pairs of negative training images are randomly paired between two cameras with different humans.

3.1. Implementation Details

Input images that are going to be passed to the SCNN architecture will undergo simple pre-processing steps. Color equalization method is applied on the raw image to reduce illumination variations between different cameras. After that, input image after color equalization is normalized so that it is in the range of 0 and 1. There are many activation functions which can be used in CNN such as sigmoid, tanh, hypertanh and Rectified Linear Unit (ReLu) activation function. In our proposed work, we chose ReLu as the activation function because it does not face gradient vanishing problem as with sigmoid and tanh function. ReLu is the most popular activation function in deep networks nowadays.

A pre-trained model is learned by using CUHK-02 dataset. Parameters of deep network are initialized by the pre-trained model and then whole network is fine tuned by backpropagation with VIPeR dataset.

3.2. Experimental Results

Cumulative Matching Characteristic (CMC) curve and normalized Area Under Curve (nAUC) score for CMC curve are used to represent the performance of our proposed method. Expectation of finding correct match in the top n matches is represented by CMC and how well our method performs overall is represented by the nAUC.

In Table 1, nAUC for global and local SCNN is 95.75% which is better than using local part SCNN (95.50%) and global part SCNN (91%). For occlusion case, the nAUC for global and local SCNN (77.5%) is still better than local SCNN (75.50%) and global SCNN (76.5%) (Table 2). Global SCNN represents feature vectors that are extracted from the whole image while local SCNN represents feature vectors that are extracted from part of the image. Feature vectors that are extracted from global and local SCNN are complementary to each other.

Table 3 shows that nAUC for our proposed method (95.75%) is better than DML method (82%) when there is no occlusion. In occlusion cases, nAUC of our proposed method (77.5%) still performs better than DML method (47.50%) significantly. Our proposed method performs better when occlusion occurs due to our decision fusion part for global and local SCNN. Decision fusion is used to fuse global and local features and weighting scheme is applied to balance the importance of global and local information. When there is occlusion in the image, feature vectors extracted from occlusion part contains less discriminant feature vectors, therefore it should be assigned with higher weights. More reliable discriminant feature vectors will be assigned with a lower weight.

Table 1. The performance of our proposed method when there is no occlusion.

No occlusion	CMC										nAUC
Global and Local SCNN	65	75	85	90	100	100	100		100	100	**95.75**
Local SCNN	60	80	90	95	95	95	95		100	100	95.50
Global SCNN	45	70	75	75	90	90	95	95		95	91.00

Table 2. The performance of our proposed method perform when occlusion occur.

Occlusion	CMC										nAUC
Global and Local SCNN	30	45	55		60	70	70	75	75	75	**77.50**
Local SCNN	25	40	40		55	55	60	70	75	80	75.50
Global SCNN	25	35	45	55	60	70		75	75	80	76.50

Table 3. Comparison between the proposed method and DML method.

	CMC										nAUC
No Occlusion											
Global and Local SCNN	65	75	85		90	100	100	100	100	100	**95.75**
DML [19]	35	45	50	50	60	65	80	85	90		82.00
Occlusion											
Global and Local SCNN	30	45	55		60	70	70	75	75	75	**77.50**
DML	0	5	5	10	20	30	35	35	40		47.50

4. Conclusion

In a nutshell, our proposed method which integrates both global and local structures of SCNN and the proposed decision fusion has shown good result in handling human re-identification task. Our experiments have shown that fusing both global and local SCNN is better than only applying local part or global part alone. The proposed method was also evaluated in the case of occlusion and performed better than existing work.

Acknowledgements

The authors would like to express their gratitude to Universiti Teknologi Malaysia (UTM) for supporting this research work under GUP grant number Q.J130000.2523.11H73.

References

[1] J Kang, I Cohen, G Medioni. *Object reacquisition using invariant appearance model.* Proc. - Int. Conf. Pattern Recognit. 2004; 4: 759-762.

[2] U Park, AK Jain, I Kitahara, K Kogure, N Hagita. *ViSE: Visual Search Engine Using Multiple Networked Cameras.* 18th Int. Conf. Pattern Recognit. 2006; 3: 1204-1207.

[3] X Wang, G Doretto, T Sebastian, J Rittscher, P Tu. Shape and Appearance Context Modeling. *Int. Conf. Comput. Vis.* 2007: 1-8.

[4] Y Yu, D Harwood, K Yoon, LS Davis. Human appearance modeling for matching across video sequences. *Mach. Vis. Appl*. 2007; 18(3-4): 139-149.

[5] AC Gallagher, T Chen. *Clothing cosegmentation for recognizing people.* 26th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR. 2008.

[6] Y Cai, Huang, T Tan. *Human appearance matching across multiple non-overlapping cameras*. Icpr. 2008: 1-4.

[7] S Bak, E Corvee, F Bremond, M Thonnat. *Person Re-identification Using Spatial Covariance Regions of Human Body Parts.* Int. Conf. Adv. Video Signal Based Surveill. 2010: 435-440.

[8] M Farenzena, L Bazzani, A Perina, V Murino, M Cristani. *Person re-identification by symmetry-driven accumulation of local features.* Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2010: 2360-2367.

[9] VM Dong Seon Cheng, Marco Cristani, Michele Stoppa. *Custom Pictorial Structures for Re-identification.* Bmvc. 2011: 1-11.

[10] M Dikmen, E Akbas, TS Huang, N Ahuja. *Pedestrian recognition with a learned metric.* Proc. Asian Conf. Comput. Vision. 2011: 501-512.

[11] WS Zheng, S Gong, T Xiang. Person Re-identification by Probabilistic Relative Distance Comparison. *IEEE Trans. Pattern Anal. Mach. Intell.* 2014; 35(3): 653-668.

[12] N Gheissari, TB Sebastian, PH Tu, J Rittscher, R Hartley. *Person reidentification using spatiotemporal appearance.* Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2006; 2: 1528-1535.

[13] O Hamdoun, F Moutarde, B Stanciulescu, B Steux. Person Re-identification In Multi-camera System By Signature Based On Interest Point Descriptors Collected On Short Video Sequences. *Robotics*. 2008: 1-5.

[14] L Bazzani, M Cristani, A Perina, M Farenzena, V Murino. *Multiple-shot person re-identification by HPE signature. Proc. - Int. Conf. Pattern Recognit.* 2010: 1413-1416.

[15] A Bedagkar-Gala, SK Shah. Part-based spatio-temporal model for multi-person re-identification. *Pattern Recognit. Lett.* 2012; 33(14): 1908-1915.

[16] B Ma, Y Su, F Jurie. *BiCov: a novel image representation for person re-identification and face verification.* Procedings Br. Mach. Vis. Conf. 2012: 1-11.

[17] S Bąk, E Corvee, F Brémond, M Thonnat. *Multiple-shot human Re-identification by Mean Riemannian Covariance Grid.* 2011 8th IEEE Int. Conf. Adv. Video Signal Based Surveillance, AVSS 2011. 2011: 179-184.

[18] W Li, R Zhao, T Xiao, X Wang. *DeepReID: Deep filter pairing neural network for person re-identification.* Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2014: 152-159.

[19] H Shi, X Zhu, S Liao, Z Lei, Y Yang, SZ Li. *Deep Metric Learning for Person Re-identification.* Proc. Int. Conf. Pattern Recognit. 2014: 2666-2672.