DOI: 10.21917/ijivp.2020.0305

VIDEO BASED PERSON RE-IDENTIFICATION USING SUPPORT VECTOR MACHINE AND LONG SHORT TERM MEMORY

Jyoti Nigam

School of Computing and Electrical Engineering, Indian Institute of Technology, Mandi, India

Abstract

We address the problem of re-identifying a person in videos captured from cameras with disjoint field of view in different background. The varying pose, scale, and motion pattern of a person makes this task challenging. In this work we propose two frameworks wherein the first one exploits only the appearance information whereas the second incorporates motion information as well. The first framework utilizes Support Vector Machine (SVM) and inner product for measuring the similarity between query and gallery videos. In the next framework we present a neural network based system for re-identifying persons in videos. We employ an LSTM as a classifier and train it over output vectors from the memory cells corresponding to different persons obtained from another LSTM. Both the proposed methods outperform existing state-of-the-art methods.

Keywords:

Video based Re-identification, Support Vector Machine (SVM), Classification Score, Convolution Neural Network (CNN), Long Short Term Memory (LSTM), Sequence Classifier

1. INTRODUCTION

The aim of video based person re-identification is to recognize the same person across the videos captured from different cameras. It is a fundamental visual recognition task in video surveillance which has various applications [1]. The task of person re-identification, on the large-scale video data, such as surveillance videos, has gained significant attention in recent years [2] [6]. The prime challenges arise due to the non-overlapping camera views, variation in the temporal transition time between cameras, and the different lighting conditions or the person who poses across cameras [3] - [5].

To solve this problem, several recent approaches attempt to localize body parts explicitly and combine the representations over them [7] [8] [11]. Some existing methods employ a static strategy to determine the quantity of selected pseudo-labeled data for further training.

For example, in [12] and [13] the prediction confidences of pseudo-labeled samples are compared with a pre-defined threshold. The samples with higher confidence over the fixed threshold are then selected for the subsequent training. During iterations, these algorithms select a fixed and large number of pseudo-labeled data from starting to end [24] [29].

In this paper we present two different frameworks to re-identify a person in a video-based re-identification dataset. The first method relies only on appearance information of the persons, whereas, in the second method, motion information is also incorporated along with the appearance information. More precisely our appearance based method utilizes Support Vector Machine (SVM) and generates positive and negative support vectors from the person and the background, respectively. The positive support vectors are used to identify a particular person. The second method is a neural network based model which consists of Convolution Neural Network (CNN) and Long Short Term Memory (LSTM). CNN is responsible for extracting appearance information which is fed to LSTM for extracting the motion information. One additional LSTM is employed as sequence classifier to identify a particular person.

In our experimentation we do not consider the samples for training and testing separately. Instead, our framework is based on online training approach. For instance, we have different videos from different cameras (views) corresponding to a particular identity as shown in Fig.1. One video is considered as training video and rest of the videos from the same as well as different identities are used as testing videos.

In the first framework we generate and maintain the positive support vectors corresponding to each identity. Then the features are extracted from query (testing) video and matching is performed. Similarly, in the second framework we exploit LSTM features as the discriminative feature and classification is done using another LSTM classifier. In other words, a single LSTM classifier is used to memorize different features of various identities. We propose a gallery free method for re-identification as there is no need to keep the images/feature for all identities. Furthermore, we do not require any distance computation among features to identify a particular person. Through extensive experimentation, we verify that our approach outperforms the existing methods over a standard video based dataset DukeMTMC.

The contributions of this work are

- We propose an appearance based framework for identifying a particular person.
- We propose an appearance as well as motion based model for re-identification of a person.
- Gallery and distance metric free re-identification system is proposed.

The outline of the paper is presented: section 2 provides related work. Section 3 discusses the proposed algorithm. Section 4 provides experiments and results followed by conclusions in section 5.

2. RELATED WORK

The task of re-identifying persons attracts great attention due to its important application values. The initial solutions of person re-identification mainly relied on hand-crafted features [14], probabilistic patch matching algorithms [15] [16] and metric learning techniques [17] [18] to deal with resolution or light or view or pose changes. Recently, re-ranking [19] attributes [20], and human-in-the-loop learning [21], have also been studied. The details can be found in the survey [22]. In the following, we review recent spatial partition-based and part-aligned representations, matching techniques, and some works using bilinear pooling.

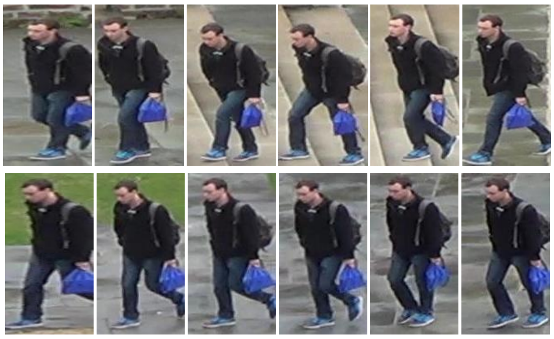


Fig.1. DukeMTMCReID video based re-identification dataset - first row frames show gallery images (Camera-2) and second row show query images (Camera-5)

Prior to deep learning methods, there are many works which explore designing hand-crafted features [23] [25] that are robust to changes in person pose and image condition. In addition to this there are also many works that make efforts to utilize robust distance metrics like Mahalanobis distance function, KISSME metric learning [26], etc. Recently, DCNN is widely used in the field of ReID [14] [27].

A huge number of researcher's design different DCNN structures to learn effective features. Zhao et al. [9] has developed a DCNN named as Spindle Net to fuse entire body feature and body region feature, and Li et al. [10] has proposed a Multi-Scale Context-Aware Network to extract small visual cues that may be very useful to distinguish the pedestrian pairs. Some researchers combine DCNN with metric learning. In the second framework we do not employ any distance computation method. Instead, a single LSTM discriminates the persons.

3. PROPOSED SYSTEM

In the proposed system for video based re-identification, corresponding to each identity the video from one camera is considered as training video and the video from the rest of the cameras are considered as testing. In our method the training and testing identities are common because we follow the online training method, where the system needs to observe the sample from an identity to recognize it later.

3.1 APPEARANCE BASED FRAMEWORK

In video based person re-identification task corresponding to each identity, we have a video which contains the images of a particular person. An instance is shown in Fig.1. For creating a feature representation corresponding to an identity, we generate support vectors from the images. By repeating this process for each identity we gather a set of positive support vectors for each person which we denote as gallery.

3.1.1 Re-identification:

The proposed system can re-identify a person in a query video, by computing the similarity score between support vectors generated from query and gallery. The identity with maximum classification score (ρ) is assigned for that query. Similarity is calculated as follows: let P be the set of positive support vectors of query, S be the set of support vectors of gallery, the classification score (ρ) is defined as follows:

$$\rho = (s,p), \text{ where } s \in (S) \text{ and } p \in (P),$$

where (a,b) represents inner product of vectors a and b.

Generally in tracking literature, an object is considered to be matched with ground truth if $IOU \geq 0.5$ [28] where IOU is Intersection over Union between prediction and ground truth. Thus, keeping this aspect in mind we have empirically chosen a stricter threshold value θ of 0.4 to predict no match corresponds to a particular person.

The score going below a particular threshold indicates no match for a query video. The score going higher than threshold for a query shows that the set of positive support vectors corresponds to that particular person.

3.2 APPEARANCE AND MOTION BASED FRAMEWORK

In second framework the input is the query video which is fed as pair of frames to the CNN. The CNN extracts the appearance features which is further fed to adjacent LSTM. The overall architecture is shown in Fig.2. The recurrent parameters of the LSTM learns the motion information of that person. The output vector of the LSTM is exploited as a discriminative feature and further fed to the LSTM classifier. The classifier LSTM re-identifies a person without maintaining a gallery for each person.

The CNN is employed with skip connections to obtain visual features at various levels. Under the assumption of a limited motion between two consecutive frames, it extracts features from both the frames. Note that this assumption increases the chances of failure when the resultant (target and wearer) motion is high.

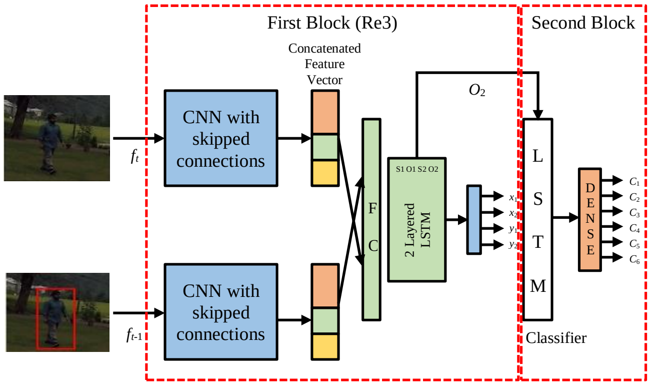


Fig.2. Overview - Consecutive Frames are fed to the first block of the network. The output of convolutional layers is propagated to LSTM. Input for LSTM in second block is the cells output vector $O_2 \in R^{1024}$ of LSTM in the first block. The last layer of the network classifies the input sequence to assign identity to the person

A 2-layered LSTM is used to capture the target motion information and to estimate the target location in the next frame. The extracted features from the first component is fed as an input to this LSTM and it looks back up to 32 timestamps for learning the motion information.

3.2.1 Classifier Single LSTM as Sequence Classifier:

Sequence classification involves predicting a class label for a given input sequence. In this section, we describe an LSTM-based network that learns the classification of persons from training data. An LSTM uses its memory cells to memorize long-range information and in tracking procedure it keeps track of moving pattern and varying features of a target. The activations of these cells are stored in the recurrent parameters of the network which can hold long-range temporal information.

3.2.2 Cell Output Vector as Target Specific Sequence:

Karpathy et al. [29] proposed that in an LSTM different cells turn on to remember different types of information. Similarly, in the proposed framework a single LSTM is used to observe multiple persons one by one in one iteration and it learns appearance as well motion information of each person. Thus, in this experiment the values of multiple cells are exploited as a sequence to represent a specific target. It may be noted that a sequence is generated at a single timestamp, but its ON/OFF pattern is observed across sequences to identify a target (different cells act as indicators for different targets).

3.2.3 Unrolling during Training:

The system generates a 1024 dimensional sequence (O_2) from LSTM output values corresponding to each identity. This sequence is fed as input to the LSTM classifier with the ground truth, i.e. person identity. The LSTM classifier looks back through the entire sequence of 1024 size to learn the specific acting pattern of memory cells corresponding to each person.

3.2.4 Training Methodology - A Single LSTM as Multi-Class Classifier:

The LSTM output vectors generated for the video from one camera corresponding to each identity is used to train the LSTM classifier. We consider all videos from gallery folder of DukeMTMC4ReID dataset where all the 702 identities are present. In training we involve all the frames which are captured from one angle but it possesses sufficient variation in pose as well as scale in persons. This results in proper training data i.e. sequences with correct class labels.

A particular identity is concatenated with a particular LSTM cell's output vector as its class label. These concatenated sequences are shown to LSTM for training and enable it to discriminate these sequences. 25% of training data is split as validation set. Categorical cross entropy loss is used with softmax nonlinearity at the final layer. The number of neurons in the last layer are decided by considering the number of different identities.

3.2.5 Re-identification:

Generally, a person re-identification system maintains a gallery of subject or target images for matching with probe/query image. Whereas, in our proposed method, different cells of tracker's single LSTM encapsulate the appearance and motion information of different persons which is useful for re-identifying a person. As soon as a query video is given in the form of pair of frames to CNN, it extracts the features and sends it to LSTM which generates output sequence. The classifier LSTM predicts the probability values against each class, and the label with highest probability is adjudged the correct class label.

4. EXPERIMENTAL ANALYSIS AND RESULTS

In order to evaluate the proposed system for video based re-identification in the cases of varying pose and partial occlusion, we use the DukeMTMC4ReID publicly available dataset. Note that without demanding the maintenance of a gallery our second framework predicts the correct identity of a query video.

4.1 DATASET

This dataset uses eight disjoint surveillance cameras capturing parts of the Duke University campus. It contains 1,404 identities appearing in more than two cameras, and 408 identities that appear in only one camera are used as distractors. 702 identities are reserved for training and 702 for testing. There are three different folders in the dataset: train, query and gallery.

4.2 IMPLEMENTATION DETAILS

In our implementation for the first framework we use the publicly available C++ code of STRUCK [33] for generating the support vectors and computing the similarity score. The experimentations are done on a computer system with configuration of 8GB RAM, quad core Intel i5 processor and 2.20 GHz speed.

In the LSTM based framework, the CNN along with LSTM for feature extraction is taken from Gordon et al. [34]. We employ Keras and Tensorflow (at back end) for training the LSTM classifier. The LSTM model constitutes of a single layer with 256 cell units with 0.2 dropout. The ADAM gradient optimizer is used with the default momentum and learning rate of 10^{-5}. We set the batch size to 32 as in Re3, at the time of testing they update the LSTM states at every 32 number of time stamps. Hence, this setting allows us to capture a continuous flow of LSTM cells output vectors. We use categorical cross entropy loss for training which goes down till 0.0002 value in 500 number of epochs. All training and testing were carried out using Nvidia 1080 Txi GPU @ 2.20GHz.

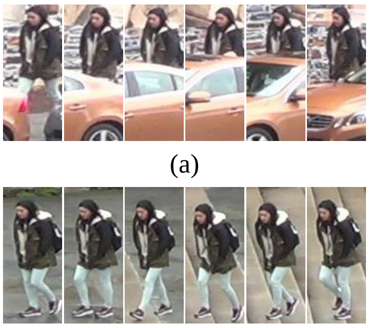
4.3 RESULTS FOR VIDEO BASED RE-IDENTIFICATION DATASET

In this section we assess the re-identification performance of our proposed method on a publicly available large-scale real-world re-id dataset, DukeMTMC4ReID [30].

Table.1. Results for re-identification on DukeMTMC-ReID

Method	**Rank 1 Accuracy**
AWTL (2 Stream)	79.80
Appearance based framework	82.30
Appearance with motion based framework	88.70

Since we create an online target model in our method, the videos from train folder are not being used for creating target models unlike the recent deep learning methods [31], [32]. We use videos from gallery to create target models. Sample images are shown in Fig.3(a) (first row images), and for testing the query videos are used as shown in Fig.3(b) (second row images).



(a)

(b)

Fig.3. DukeMTMCReID video based re-identification dataset: first row frames show gallery images (Camera-1) and second row show query images (Camera-2)

In these frames the target is same but with different background, pose. Although these frames are having different poses and contain partial occlusion, our method re-identifies the target across different camera views. We provide the comparison of our Rank-1 accuracy with the recent deep learning based method [31] in Table.1. Note that the proposed method does not use a train/test procedure. But we provide it here since AWTL is the state-of-the-art and our results are on test videos.

5. CONCLUSION

This work has been carried out with the aim of developing a system to re-identify a person in videos. We propose two different frameworks where the first is inspired by the certainty that in online training method we generate support vectors corresponding to each person which can be used for re-identifying that person in later videos. In the second framework, we focus on the use of LSTMs as sequence classifier to re-identify the persons in different camera videos. We conclude that the appearance and motion features extracted by one LSTM internal states can be used effectively to train another LSTM to detect the persons. In general, re-identification methods maintain a gallery set for matching, but this type of learning offers a gallery free re-identification as a single LSTM keeps information about all the persons. The proposed method performs way better than the state of-the-art in conventional re-identification problems.

ACKNOWLEDGEMENT

This publication is an outcome of the R&D work project that was done under the Visvesvaraya PhD Scheme of Ministry of Electronics and Information Technology, Government of India, being implemented by Digital India Corporation

REFERENCES

[1] S. Tang, M. Andriluka, B. Andres and B. Schiele, “Multiple People Tracking by Lifted Multicut and Person Re-Identification”, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3539-3548, 2017.

[2] A. Hermans, L. Beyer and B. Leibe, “In Defense of the Triplet Loss for Person Re-Identification”, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2333-2339, 2017.

[3] Y. Liu, J. Yan and W. Ouyang, “Quality Aware Network for Set to Set Recognition”, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5790-5799, 2017.

[4] T. Wang, S. Gong, X. Zhu and S. Wang, “Person Re-Identification by Discriminative Selection in Video Ranking”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38, No. 12, pp. 2501-2514, 2016.

[5] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang and P. Zhou, “Jointly Attentive Spatial-Temporal Pooling Networks for Video-Based Person Re-Identification”, *Proceedings of IEEE Conference on Computer Vision*, pp. 4733-4742, 2017.

[6] W. Zhang, S. Hu and K. Liu, “Learning Compact Appearance Representation for Video-based Person Re-Identification”, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4660-4669, 2017.

[7] C. Su, J. Li, S. Zhang, J. Xing, W. Gao and Q. Tian, “Pose Driven Deep Convolutional Model for Person Re-Identification”, *Proceedings of IEEE Conference on Computer Vision*, pp. 3960-3969, 2017.

[8] L. Zheng, Y. Huang, H. Lu and Y. Yang, “Pose Invariant Embedding for Deep Person Re-Identification”, *IEEE Transactions on Image Processing*, Vol. 28, No. 9, pp. 4500-4509, 2019.

[9] X. Tang, “Spindle Net: Person Re-Identification with Human Body Region Guided Feature Decomposition and Fusion”, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1077-1085, 2017.

[10] D. Li, X. Chen, Z. Zhang and K. Huang, “Learning Deep Context-Aware Features Over Body and Latent Parts for Person RE- Identification”, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 384-393, 2017.

[11] L. Zhao, X. Li, Y. Zhuang and J. Wang, “Deeply-Learned Part- Aligned Representations for Person Re-Identification”, *Proceedings of IEEE Conference on Computer Vision*, pp. 3219–3228, 2017.

[12] H. Fan, L. Zheng, C. Yan and Y. Yang, “Unsupervised Person Re- Identification: Clustering and Fine-Tuning”, *ACM Transactions on Multimedia Computing, Communications, and Applications*, Vol. 14, No. 4, pp. 83-93, 2018.

[13] M. Ye, A.J. Ma, L. Zheng, J. Li and P. C. Yuen, “Dynamic Label Graph Matching for Unsupervised Video Re-Identification”, *Proceedings of IEEE Conference on Computer Vision*, pp. 5142-5150, 2017.

[14] S. Liao, Y. Hu, X. Zhu and S.Z. Li, “Person Re-Identification by Local Maximal Occurrence Representation and Metric Learning”, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2197-2206, 2015.

[15] D. Chen, Z. Yuan, G. Hua, N. Zheng and J. Wang, “Similarity Learning on an Explicit Polynomial Kernel Feature Map for Person Re-Identification”, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1565-1573, 2015.

[16] D. Chen, Z. Yuan, B. Chen and N. Zheng, "Similarity Learning with Spatial Constraints for Person Re-Identification", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1268-1277, 2016.

[17] L. Zhang, T. Xiang and S. Gong, "Learning a Discriminative Null Space for Person Re-Identification", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1239-1248, 2016.

[18] Y. Zhang, B. Li, H. Lu, A. Irie and X. Ruan, "Sample-Specific SVM Learning for Person Re-Identification", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1278-1287, 2016.

[19] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu and Q. Tian, "Query-Adaptive Late Fusion for Image Search and Person Re- Identification", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1741-1750, 2015.

[20] C. Su, F. Yang, S. Zhang, Q. Tian, L.S. Davis and W. Gao, "Multi Task Learning with Low Rank Attribute Embedding for Person Re- Identification", *Proceedings of IEEE Conference on Computer Vision*, pp. 3739-3747, 2015.

[21] N. Martinel, A. Das, C. Micheloni and A. K. Roy Chowdhury, "Temporal Model Adaptation for Person Re-Identification", *Proceedings of European Conference on Computer Vision*, pp. 858-877, 2016.

[22] L. Zheng, Y. Yang and A.G. Hauptmann, "Person Re-Identification: Past, Present and Future", *Proceedings of European Conference on Computer Vision*, pp. 668-687, 2016.

[23] D. Gray and H. Tao, "Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features", *Proceedings of European Conference on Computer Vision*, pp. 262-275, 2008.

[24] A. Mignon and F. Pcca, "A New Approach for Distance Learning from Sparse Pairwise Constraints", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1197-1206, 2012.

[25] R. Zhao, W. Ouyang and X. Wang, "Unsupervised Salience Learning for Person Re-Identification", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3586-3593, 2013.

[26] M. Koestinger, M. Hirzer, P. Wohlhart, P.M. Roth and H. Bischof, "Large Scale Metric Learning from Equivalence Constraints", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2288-2295, 2012.

[27] R. Zhao, W. Ouyang and X. Wang, "Learning Mid-Level Filters for Person Re-Identification", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 144-151, 2014.

[28] A. Dehghan, and M. Shah, "Visual Tracking: An Experimental Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 36, No. 7, pp. 1442-1468, 2014.

[29] A. Karpathy, J. Johnson, and L. Fei Fei, "Visualizing and Understanding Recurrent Networks", *Proceedings of International Conference on Learning Representations*, pp. 1-11, 2015.

[30] M. Gou, S. Karanam, W. Liu, O. Camps and R.J. Radke, "DukeMMTMC4ReID: A Large-Scale Multi-Camera Person Re- Identification Dataset", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 10-19, 2017.

[31] E. Ristani and C. Tomasi, "Features for Multi-Target Multi-Camera Tracking and Re-Identification", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6036-6046, 2018.

[32] Y. Chen, X. Zhu and S. Gong, "Person Re-Identification by Deep Learning Multi-Scale Representations", *Proceedings of IEEE Conference on Computer Vision*, pp. 2590-2600, 2017.

[33] A.S. Hare and P.H. Torr, "Struck: Structured Output Tracking with Kernels", *Proceedings of IEEE Conference on Computer Vision*, pp. 1-8, 2011.

[34] D. Gordon, A. Farhadi and D. Fox, "Re3: Real-Time Recurrent Regression Networks for Visual Tracking of Generic Objects", *IEEE Robotics and Automation Letters*, Vol. 3, No. 2, pp. 788-795, 2018.