MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Body Part Alignment and Temporal Attention for Video-Based Person Re-Identification

Jones, M.J.; Rambhatla, S.

TR2019-108 September 25, 2019

Abstract

We present a novel deep neural network for video-based person re-identification that is designed to address two of the major issues that make this problem difficult. The first is dealing with misalignment between cropped images of people. For this we take advantage of the OpenPose network [2] to localize different body parts so that corresponding regions of feature maps can be compared. The second is dealing with bad frames in a video sequence. These are typically frames in which the person is occluded, poorly localized or badly blurred. For this we design a temporal attention network that analyzes feature maps of multiple frames to assign different weights to each frame. This allows more useful frames to receive more weight when creating an aggregated feature vector representing an entire sequence. Our resulting deep network improves over state-of-theart results on all three standard test sets for video-based person re-id (PRID2011 [8], iLIDS-VID [21] and MARS [27]).

British Machine Vision Conference (BMVC)

Body Part Alignment and Temporal Attention for Video-Based Person Re-Identification

Sai Saketh Rambhatla
rssaketh@umd.edu
University of Maryland
College Park, MD, USA
Michael Jones
mjones@merl.com
Mitsubishi Electric Research Labs
Cambridge, MA, USA

Abstract

We present a novel deep neural network for video-based person re-identification that is designed to address two of the major issues that make this problem difficult. The first is dealing with misalignment between cropped images of people. For this we take advantage of the OpenPose network [2] to localize different body parts so that corresponding regions of feature maps can be compared. The second is dealing with bad frames in a video sequence. These are typically frames in which the person is occluded, poorly localized or badly blurred. For this we design a temporal attention network that analyzes feature maps of multiple frames to assign different weights to each frame. This allows more useful frames to receive more weight when creating an aggregated feature vector representing an entire sequence. Our resulting deep network improves over state-of-the-art results on all three standard test sets for video-based person re-id (PRID2011 [8], iLIDS-VID [21] and MARS [27]).

1 Introduction

Person re-identification is the problem of identifying a person from full-body images or videos. The problem is usually formulated as a comparison between two images or two videos each cropped around a person. This problem typically arises in multi-camera tracking problems, but is also useful for finding different instances of a person in surveillance video from a single camera. For practical applications, the problem of person re-identification most often arises in video applications (as opposed to applications in which only still images are available). For this reason, we focus on the video-based person re-id problem. When comparing two videos of a person, a fundamental problem is how to take advantage of the information in multiple video frames. We solve this problem using *temporal attention pooling (TAP)* that assigns a different weight to each frame to compute an aggregated feature vector representing the identity of the person over the whole video sequence. Whereas some frames have rich information about the subject, some frames may not be helpful due to occlusion or change in illumination, etc. This issue is illustrated in Figure 1 (b) where large occlusions occur in the frames in the middle and hence would be detrimental for re-identification. Our approach deals with this problem by learning a TAP network that assigns higher weights to more informative frames.

One of the major difficulties for person re-id is misalignment. People can have a large variety of poses because of articulated limbs as well as camera viewpoints which often result

(a) (b)

Figure 1: (a) Illustration of alignment problem between two person image being compared for person re-id. Same colored ovals correspond to the same regions in the image but not to corresponding body parts. (b) Illustration of a sequence with occluded human body parts. Features extracted from frames with occlusion are corrupted and detrimental to accuracy.

in particular body parts for cropped person images being in very different image locations. See Figure 1 (a) for an illustration of this problem. Recent work on person re-id [5, 18] has tried to address this issue to some degree, but the alignment is usually imprecise. In our work, we use the OpenPose network [2] to get more precise alignment of person images.

Our main contributions can be summarized as follows: (1) We propose a novel architecture for precise spatial alignment using the OpenPose network that solves the registration problem. (2) We propose a temporal aggregation pooling (TAP) network to aggregate the information across frames to form a rich video descriptor. (3) Our deep network improves the state-of-the-art results on all three standard datasets.

2 Related Work

Most of the work on person re-identification has focused on comparing a single image to a single image. Within this body of work, some recent methods have achieved superior results through better alignment of pedestrian images. The work of Su et al. [19], Zheng et al. [29], Zhao et al. [26], and Zhang et. al [25] all fall into this category. Of these, Zhao et al.'s Spindle Net is the closest to our work in that they use the forerunner of OpenPose [2], a Convolution Pose Machine (CPM) [22], to segment different body parts in each pedestrian image. Unlike ours, their method models each body part as a rectangular region which is explicitely cropped from the input image.

The problem of video-based person re-id has so far received less attention than that of single-frame person re-id. Some papers that have addressed the video-based person re-id problem have focused on how to combine the information from multiple frames. For example, Liu et al. [13], Zhou et al. [31] and McLaughlin et al. [15] all fall into this category. A recent paper by Chen et al. [3] compares two person sequences by splitting each sequence into short sub-sequences called snippets and trains a deep network to compare and aggregate two sets of snippets. Unlike our method that can use pre-computed gallery features to compute similarity at inference time, Chen et al.'s method requires a forward pass through their network to compare each probe sequence to each gallery sequence, greatly increasing the computational overhead.

There has been relatively little work on video-based person re-identification that focuses both on careful spatial alignment of body parts as well as temporal attention. One paper along these lines is [10] which uses a spatiotemporal attention network to learn a number of spatial attention detectors (CNNs) and then aggregates the resulting receptive fields over time to yield a single feature vector. Their work is similar in spirit to ours but uses a different spatial attention network and a different temporal attention network. The spatial attention detectors

used in [10] automatically discover salient image regions useful for re-identification. The ability of the network to look for discriminative regions on its own is analogous to the *bottom-up* attention in the human visual cortex. Studies [1, 16] suggest that the human visual cortex has a separate, independent pathway for *top-down* attention. Bottom-up attention arises due to the *"pop-out"* effect in the visual scene while top-down attention is due to a conscious effort (analogous to a pre-defined detector). Our work presents a simple yet efficient way to give the network the ability to perform top-down attention on the input image. As suggested by [16], both the described methods operate independently of each other and hence [10] can be used in conjunction to further improve performance. Another recent work by Liu et al. [14] proposed to refine feature maps, which are corrupted by occlusion, blur etc. by using a modified version of a GRU (called RRU) and then use a 3D convolution module to extract spatio-temporal features for re-identification. It is to be noted that the motivation for TAP is very similar, where we down-weight the "importance" of an occluded frame, Liu et al. [14] utilize RRU to recover missing parts and supress noisy features using temporal data. While the focus of this work is alignment, Liu et al. focus on the interplay between spatial and temporal features to improve person re-identification.

Another recent work by Suh et al. [20] uses the OpenPose network to find body parts, but uses a different method of incorporating the part maps into a part-aligned representation than us. Their method is very memory intensive (due to outer product in bilinear pooling) and requires some form of dimensionality reduction of the features to be feasible. Also, Suh et al. handle multiple frames in the simplest way by simply averaging their descriptor over frames. Song et al. [18] deals with misalignment by pooling over three fixed regions within each pedestrian bounding box which are determined by analyzing the positions of different parts detected in the training data using OpenPose. Their method focuses more on temporal processing by estimating the quality of the regions in each frame and weighting the feature vector for each region in each frame according to its quality estimate. Similarly, Dai et al. [5] uses a weaker alignment model than we do (they use only scale and translation for aligning pedestrian images) but a strong temporal model for handling multiple frames (consisting of a bi-directional LSTM network). Xu et al. [24] also present a method for video-based person re-identification that mainly focuses on temporal pooling using a recurrent network but does not use precise body part alignment.

In our work, we take advantage of the OpenPose deep network [2] to align pedestrian images and also use a temporal attention pooling network (a fully convolutional network) to estimate non-uniform weights for different frames in a pedestrian video sequence. This allows us to intelligently combine well-aligned feature maps across many frames to yield a distinctive feature vector that leads to improved state-of-the-art accuracy on the three main video person re-id test sets: PRID2011, i-LIDS-VID and MARS.

3 Method

We propose a new deep learning alignment method to better handle the alignment issue in video-based person re-id. Given an input sequence of video frames, the pose of the person in each frame is extracted (Section 3.1) and the corresponding attention maps are combined to form composite heat maps (Section 3.2) for regions of the human body. These heatmaps are combined with appearance features to better align the features of corresponding parts and then are concatenated to form the per frame person descriptors. The per frame person descriptors are then combined using a novel temporal attention module (Section 3.3) to form a descriptor for the entire video. A cross entropy loss and metric loss are then applied on top

of the fully connected (fc) layer to train the network in an end-to-end fashion.

3.1 Pose Estimation

For surveillance applications, the resolution of people in a video frame is usually low, which implies that instead of using a biometric such as faces, it is necessary to use full-body appearance for person re-identification. However, misalignment makes it difficult to compare the full body appearance of two people in different images. To alleviate this issue, we utilize the state-of-the-art pose estimation network OpenPose [2] to extract the pose of the person in the frame and leverage it to get the full-body appearance of the person of interest. The OpenPose network has 7 stages, where each stage refines the location of the joint predicted by the previous stage. At the head of the network is a VGGNet [17] to extract features from the image which is then sent to the first stage of OpenPose. The output of OpenPose is 19 heatmaps (18 joints and 1 background), giving the probability of occurrence of a particular body part such as the nose, neck, right shoulder, etc. and 38 part affinity field maps (19 in x-direction and 19 in y-direction) corresponding to edges between part nodes.

3.2 Pose Alignment

Given a video sequence $\mathcal{V} = \{I_1, I_2, ..., I_t\}$ consisting of t frames, where $\mathcal{V} \in \mathcal{R}^{H\times W\times 3\times t}$ and $I_t \in \mathcal{R}^{H\times W\times 3}$, an appearance network $\mathcal{A}$ extracts appearance features from each frame. We use the ResNet-50 [6] architecture as the appearance network. We remove the fc and the final average pooling layers and use the remaining network as the feature extractor. Let $f_a = \{f_a^1, f_a^2, ..., f_a^t\}$ be the appearance features for the video sequence $\mathcal{V}$, where $f_a^t \in \mathcal{R}^{h\times w\times C}$ and $f_a \in \mathcal{R}^{h\times w\times C\times t}$ i.e.

$$f_a = \mathcal{A}(\mathcal{V}). \tag{1}$$

Let $f_p = \{f_p^1, f_p^2, ...f_p^t\}$ be the joint information from the OpenPose network, $f_p^t \in \mathcal{R}^{h\times w\times F}$, in our case $F = 57$ (19 joint maps and 38 part affinity maps). Let $f_p^t = \{f_{p,1}^t, f_{p,2}^t, ..., f_{p,F}^t\}$ be the individual joint heat and part affinity field maps and let $\mathcal{S}_u$, $\mathcal{S}_m$, $\mathcal{S}_l$ be sets containing indices of heatmaps (and part affinity field maps) of upper, middle and lower body parts respectively. We group the heatmaps of {eyes, ears, nose and neck} and the part affinity field maps corresponding to the edges {R-eye $\leftrightarrow$ R-ear, L-eye $\leftrightarrow$ L-ear, L-eye $\leftrightarrow$ nose, R-eye $\leftrightarrow$ nose, L-eye $\leftrightarrow$ nose, nose $\leftrightarrow$ neck }, to form one of the composite parts. Similarly we group heat maps of {elbows, shoulders, wrists} and {hip, knees, ankles} with part affinity maps of {R-shoulder $\leftrightarrow$ R-elbow, R-elbow $\leftrightarrow$ R-wrist, neck $\leftrightarrow$ R-hip L-shoulder $\leftrightarrow$ L-elbow, L-elbow $\leftrightarrow$ L-wrist, neck $\leftrightarrow$ L-hip } and {R-hip $\leftrightarrow$ R-knee, R-knee $\leftrightarrow$ R-ankle, L-hip $\leftrightarrow$ L-knee, L-knee $\leftrightarrow$ L-ankle }, respectively to form the other composite parts. Since the receptive field of VGGNet is large, these composite parts have the receptive field to look at the corresponding upper, torso and lower body of the human in the image. This partition is done by leveraging the prior information about humans. The composite part maps, each one corresponding to upper, middle and lower body, are formed by pooling over the indices of joint maps in each of the sets $\mathcal{S}_u$, $\mathcal{S}_m$, $\mathcal{S}_l$, i.e.,

$$r_p^t = \frac{1}{|S_r|}\sum_{k\in\mathcal{S}_r} f_{p,k}^t \tag{2}$$

where $r \in \{u, m, l\}$ is a region of the body. The formation of such composite heatmaps, gives the network the top down attention previously discussed. By combining these maps with appearance features, the network is directed to focus on specific human body parts. The composite part maps $u_p^t, m_p^t, l_p^t \in \mathcal{R}^{h\times w\times 1}$ are shown in Figure 2.

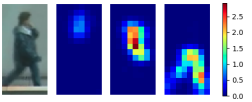

Figure 2: Composite parts of upper, middle and lower body for an example person.

Each of the composite part maps reflect the confidence of finding the head, torso and legs of the person in the current frame and are then converted into a probability distribution by applying a softmax over the $h \times w$ spatial locations.

$$\bar{r}_p^t[k,n] = \frac{e^{r_p^t}[k,n]}{\sum_{j=1}^{w}\sum_{i=1}^{h} e^{r_p^t}[i,j]} \tag{3}$$

where $r \in \{u,m,l\}$, $k \in \{0,1,\ldots,h-1\}$, and $n \in \{0,1,\ldots,w-1\}$. These composite part maps are then multiplied with appearance feature maps f_{app}^t from the appearance net to form the appearance feature map of the corresponding composite part. The composite part maps are a kind of mask and multiplying by appearance feature maps retains only the parts of the feature maps corresponding to each composite body part.

$$f_{a,r}^t = f_a^t * \bar{r}_p^t \tag{4}$$

where $r \in \{u,m,l\}$. It will be motivated in Section 3.3 that the features $f_{a,r}^t$ carry different information for each t. So we compute a set of weights $w_i, i \in \{1,2,....t\}$ using the TAP described in the next section. The appearance feature is then spatially pooled and concatenated with the appearance features of the composite parts to form the final frame descriptor. These operations ensure that the corresponding appearance features of the body parts of different/same individuals are compared, thereby effectively achieving alignment. More precisely,

$$\bar{f}_a^t = avgpool(f_a^t), \qquad \bar{f}_{a,r}^t = avgpool(f_{a,r}^t) \tag{5}$$

$$f_{frame}^t = [\bar{f}_a^t; \bar{f}_{a,u}^t; \bar{f}_{a,m}^t; \bar{f}_{a,l}^t] \tag{6}$$

and r takes values in $\{u,m,l\}$. Here f_a^t, $f_{a,u}^t$, $f_{a,m}^t$, $f_{a,l}^t \in \mathcal{R}^C$ where $C = 2048$. Finally, the video descriptor can be formed by taking the weighted average of per frame descriptors.

$$f_{video} = \sum_{i=1}^{t} w^i * f_{frame}^i \tag{7}$$

In the next Section 3.3, we propose a novel way to compute the weights of the summation in the above equation. The entire pipeline described above can be seen in Figure 3. We refer to parts 1,2 and 3 in the figure as AlignNet (the pose and appearance nets and composite part maps without temporal attention pooling).

3.3 Temporal Attention Pooling

We observe that, for images with very poor quality or occluded parts, the composite parts do not look as expected and hence do not contribute much information for classification of the identity in question. So a naive average of the frame features to form the video descriptor would be sub-optimal. Our observation is supported by previous work [10, 13], in which it was shown that mere average/max pooling of the features is not effective to aggregate the per frame features. Hence we propose a temporal attention pooling (TAP) module which

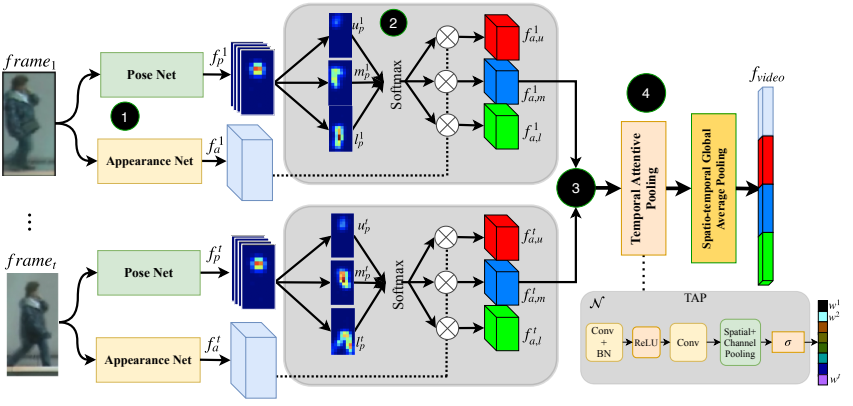

Figure 3: **Pose guided Alignment Network Architecture**. Each video is divided into chunks of t consecutive frames. (1) Each image is passed through an appearance network (ResNet-50) and pose network (OpenPose) to give appearance and pose features. (2) Pose heatmaps corresponding to upper, middle and lower parts are grouped after softmax is applied spatially. These part maps give a probabilistic mask of each composite body part's location in the image. (3) Appearance and part features from all the frames are then combined and (4) finally a temporal attention pooling (TAP) unit computes the weights required to combine the t frames to form the video descriptor.

assigns an "importance" to the feature of a particular frame. These values can be interpreted as a measure of quality of the pose prediction for each frame.

The proposed TAP is a fully convolution network consisting of two 2D 3×3 convolution layers (with ReLU and batch normalization layers after the first convolutional layer) which maintain the spatial dimension and finally pools along the spatial and channel dimensions to get a quality estimate for each frame of the video. We normalize these values to $[0,1]$ using a sigmoid operation to avoid amplifying the activation maps in the aggregation step.

Mathematically, the operations are described as below. From Section 3.2, let $f_{a,u} = [f^1_{a,u}, f^2_{a,u}, ..., f^t_{a,u}]$ and $f_{a,u} \in \mathcal{R}^{h \times w \times C \times t}$. Let $\mathcal{N}$ be the fully convolutional TAP network, just described and shown in Figure 3. This gives,

$$\bar{f}_{a,u} = \mathcal{N}(f_{a,u}). \tag{8}$$

Note that $\bar{f}_{a,u} \in \mathcal{R}^t$. Finally we apply a point-wise sigmoid function, σ, to get the temporal attention values,

$$w = \sigma(\bar{f}_{a,u}). \tag{9}$$

The vector $w(= [w^1, w^2, ..., w^t]) \in \mathcal{R}^t$ gives the attention value for each of the t per frame features. The final video descriptor is then computed using Eq 7.

For frames with occlusion, OpenPose and thus the composite part maps often fail while the global appearance maps still contain useful features. Because of this, we apply separate TAP to the appearance (f_a) and the part features ($f_{a,r}$ where $r \in \{u,m,l\}$). In our experiments to make the training faster and reduce memory consumption, we apply the TAP on one of the composite part features ($f_{a,u}$) and use the obtained weights w for other composite part features. We experimented with learning separate TAP nets for each component but the minor improvement in accuracy did not justify the cost in memory and time.

Figure 4: Example sequences in our synthetic dataset containing 1000 identities each with 2 sequences of 50 frames.

3.4 Loss

The proposed pose alignment network (Section 3.2) and TAP (Section 3.3) can be trained end-to-end. We use the K-way softmax cross entropy loss and a metric loss to train the network. The advantages of using both the losses for training has been discussed in image based re-identification literature and is seldom done in video re-identification. We believe that the discussions hold strong even for video-based re-identification. For the metric loss we use the mining strategy proposed in [7] which is widely used in image-based re-identification. Specifically, we use a margin ranking loss with the formulation as follows

$$\mathcal{L}(x,y) = \max(0, -y * (d_{an} - d_{ap}) + m) \tag{10}$$

where m is the margin. The purpose of this loss is to rank d_{an} (distance between anchor and negative samples) higher than d_{ap} (distance between anchor and positive) for a set of anchor, positive and negative samples. We mine the positive, negative and anchor from a batch instead of the whole dataset as [7] suggests.

4 Experiments

4.1 Datasets

To compare with previous methods, we evaluate our method on three standard video-based person re-id datasets : PRID2011 [8], iLIDS-VID [21], and MARS [27]. PRID2011 contains 400 sequences for 200 people from 2 cameras each. Each sequence has a variable number of frames with 100 frames on average. iLIDS-VID consists of 600 image sequences of 300 subjects from 2 cameras each. The average sequence length is 73 frames. The MARS dataset is the largest video-based person re-id dataset with $1,261$ identities and around $20,000$ video sequences. Each identity is captured by at least 2 cameras and has 13.2 sequences on average. There are $3,248$ distractor sequences in the dataset.

For the PRID2011 and iLIDS-VID datasets, we follow the evaluation protocol from [21]. Half the identities are randomly put into training and the other half into testing. This procedure is repeated 10 times for computing averaged accuracies. During test time all videos from camera 1 are treated as probes and videos from camera 2 are treated as gallery. For the MARS dataset we follow the original splits provided by [27] which predefined 631 identities for training and the remaining identities for testing. We use the predefined query set for evaluation.

4.2 Implementation Details

We divide each video sequence in the training set, into multiple chunks of 15 frames each with 5 frames overlap. All the models in Table 1 (b) for PRID2011 and iLIDS-VID are

first trained on a synthetic dataset with 1000 identities before fine-tuning on the individual dataset. We found this did not help with MARS because of its large size. The synthetic images (used for iLIDS-VID and PRID) of people were generated using Blender from animated 3D parametric models. We will make this synthetic training set available. The number of synthetic images we used is about the same as the number of outside images used for training in [10]. Some example images are shown in Figure 4. Table 1 (a) shows the accuracy of our AlignNet model with and without synthetic data pretraining for a single partition of each dataset.

METHOD	PRID2011	iLIDS-VID
AlignNet - No synthetic pretrain	90.5%	75.7%
AlignNet - Synthetic pretrain	**94.4%**	**80.7%**

(a)

METHOD	PRID2011	iLIDS-VID	MARS
Baseline	85.4%	75.7%	79.7%
AlignNet + AvgPool	**94.4%**	**80.7%**	**82.6%**
AlignNet + TAP	94.4%	**86.0%**	**83.2%**

(b)

Table 1: (a) Effect of pre-training with synthetic data. Both models are initially trained on ImageNet. (b) Component Analysis. CMC-rank1 number is reported on one split of iLIDS-VID and PRID2011 and the entire MARS dataset.

The OpenPose network is initialized with weights trained on MS COCO keypoint dataset. The input image is kept at the original size of 128×64 for PRID2011 and iLIDS-VID. For MARS, we operate on the original image size of 256×128. Batch stochastic gradient descent is used to optimize the network. All the different parts, ResNet-50, OpenPose, TAP and the classifier are trained with separate learning rates. We do not want OpenPose network's weights to change too much, so it is assigned a lower learning rate. The learning rates are dropped whenever training loss saturates and is trained for an average of 300 epochs.

During test time, due to the fully convolutional nature of our model, a variable number of frames can be used as input. Although we could feed the entire testing video sequence to our model at once, because of GPU memory contraints, we break the test sequence into chunks of 15 frames each without overlap, extract features and then average the features to form the video descriptor. For a particular query, we find its similarity with all the gallery sequences and rank them in decreasing order of similarity for evaluation. We use the Euclidean distance between query and gallery video descriptor as a measure of dissimilarity.

Re-identification performance is usually reported using CMC plots. For MARS, since there are multiple instances of the sequence in the gallery, we use CMC and mAP, while only CMC for PRID2011 and iLIDS-VID.

4.3 Ablation Studies

We conducted an ablation study to show the effectiveness of each component of our architecture. The detailed analysis of this study is in the supplemental material. The results are summarized in Table 1 (b) (using a single partition for PRID2011 and iLIDS-VID) and show that adding alignment to our network leads to large improvements across all three datasets. TAP leads to significant improvement on iLIDS-VID which has many frames with occlusion and poor contrast while the other two datasets which are "cleaner" do not benefit as much from the temporal weights.

4.4 Comparison with the state-of-the-art methods

Table 2 reports the performance of our approach with other state-of-the-art methods. Our method improves over the previous best approach on the PRID2011 dataset by almost 3%.

For iLIDS-VID and MARS, our method is better than all previous approaches except the recent snippet method [3] which is much more computationally expensive at inference time.

We also look at the effect of two post-processing methods that can be applied to any identification method. Re-ranking [30] is commonly used in the person re-identification community to boost the accuracy. It mainly leads to significant improvements especially for mAP for datasets with multiple sequences of probe identity in the gallery set (unlike PRID2011 and iLIDS-VID) so we report our improvement using re-ranking only on the MARS dataset. The other post-processing method we applied is score normalization [23] which is better known in the speech recognition community. Each similarity score is normalized by subtracting the means and dividing by the standard deviations of that score's row and column. The intuition behind this procedure is that a similarity score is relative to how unique a particular gallery (or probe) image is and normalizing by the mean and standard deviation accounts for the image's uniqueness. The final row of Table 2 shows that score normalization improves the CMC-rank1 rate by 0.7% for PRID2011 (where there is little room for improvement) and by 2.9% on iLIDS-VID and 1.7% on MARS (where there is more room for improvement). Thus, we improve over the results of [3] on iLIDS-VID. For MARS, we see an improvement of 11% on mAP which also achieves a new state-of-the-art accuracy.

Method	PRID2011	iLIDS-VID	MARS
SI2DL [32]	76.7	48.7	-
mvRMLLC+Alignment [4]	66.8	69.1	-
AMOC + EpicFlow [12]	82.0	65.5	-
RNN [15]	70	58	-
IDE [28] + XQDA[11]	-	-	65.3 (47.6)
end AMOC + EpicFlow [12]	83.7	68.7	68.3 (52.9)
MARS [27]	77.3	53.0	68.3(49.3)
SeeForest [31]	79.4	55.2	70.6 (50.7)
QAN [13]	90.3	68.0	-
PAM-LOMO+KISSME [9]	92.5	79.5	-
ASTPN [24]	77.0	62.0	44.0 (-)
RQEN [18]	92.4	76.1	73.7 (51.7)
TRL [5]	87.8	57.7	80.5 (69.1)
SpaAtn+TemAtn [10]	93.2	80.2	82.3 (65.8)
PAB [20]	-	-	83.0 (72.2)
Snippet [3]	93.0	85.4	**86.3** (76.1)
STMP [14]	92.7	84.3	84.4 (72.7)
Ours	96.0	83.0	83.2 (71.8)
Ours+Score Normalization	**96.7**	**85.9**	84.9 (73.6)
Ours + Re-ranking[30]	-	-	85.4 (**82.8**)

Table 2: Comparison with state-of-the-art Methods. For PRID2011 and iLIDS only CMC-rank1 number is reported, for MARS, mAP (in parentheses) along with CMC-rank1 is reported.

5 Conclusion

Evidence from recent studies shows that alignment is a key factor in improving re-identification performance. In this work, we propose a pose-guided alignment network, which mimics the top-down attention of the human visual cortex. We use a state-of-the-art pose estimation network to find key points of humans, and use the keypoints to generate composite parts corresponding to upper, middle and lower body. We then leverage these composite part maps as attention to provide the network with aligned features for comparison. This method solves

the problem of aligning corresponding image patches across frames, which occur due to large changes in pose. Finally, we propose a novel temporal attention pooling module which assigns an "importance" value based on image quality and visibility of the part. This module gives the network a chance to penalize features of frames which do not contain rich information for better re-identification. The module is motivated from self-attention. We evaluate the effectiveness of each proposed unit and demonstrate improvements over the state of the art using our approach for video-based person re-identification.

References

[1] Farhan Baluch and Laurent Itti. Mechanisms of top-down attention. *Trends in Neurosciences*, 34(4):210 – 224, 2011. ISSN 0166-2236. doi: https://doi.org/10.1016/j.tins.2011.02.003. URL http://www.sciencedirect.com/science/article/pii/S0166223611000191.

[2] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[3] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] J. Chen, Y. Wang, and Y. Y. Tang. Person re-identification by exploiting spatio-temporal cues and multi-view metric learning. *IEEE Signal Processing Letters*, 23(7):998–1002, July 2016. ISSN 1070-9908. doi: 10.1109/LSP.2016.2574323.

[5] J. Dai, P. Zhang, H. Lu, and H. Wang. Video person re-identification by temporal residual learning. *IEEE Trans. on Image Processing*, 2018.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. doi: 10.1109/CVPR.2016.90.

[7] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017. URL http://arxiv.org/abs/1703.07737.

[8] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In Anders Heyden and Fredrik Kahl, editors, *Image Analysis*, pages 91–102, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-21227-7.

[9] F. M. Khan and F. Brémond. Multi-shot person re-identification using part appearance mixture. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 605–614, March 2017. doi: 10.1109/WACV.2017.73.

[10] S. Li, S. Bak, P. Carr, and X. Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, 2018.

[11] S. Liao, Y. Hu, Xiangyu Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2197–2206, June 2015. doi: 10.1109/CVPR.2015.7298832.

[12] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, and J. Feng. Video-based person re-identification with accumulative motion context. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2788–2802, Oct 2018. ISSN 1051-8215. doi: 10.1109/TCSVT.2017.2715499.

[13] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. In *CVPR*, 2017.

[14] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li, editors. *Proceedings of the Thirty Third AAAI Conference on Artificial Intelligence, January 27- Febuary 1, 2019, Honolulu, Hawaii USA*, 2019. AAAI Press.

[15] N. McLaughlin, J.M. del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, 2016.

[16] Yair Pinto, Andries R. van der Leij, Ilja G. Sligte, Victor A. F. Lamme, and H. Steven Scholte. Bottom-up and top-down attention are independent. *Journal of Vision*, 13(3): 16, 2013. doi: 10.1167/13.3.16. URL +http://dx.doi.org/10.1167/13.3.16.

[17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[18] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai. Region-based quality estimation network for large-scale person re-identification. In *AAAI*, 2018.

[19] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017.

[20] Y. Suh, J. Wang, S. Tang, T. Mei, and K.M. Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018.

[21] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV 2014*, pages 688–703, 2014. ISBN 978-3-319-10593-2.

[22] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[23] D. Wu, B. Li, and H. Jiang. Normalization and transformation techniques for robust speaker recognition. In France Mihelic and Janez Zibert, editors, *Speech Recognition*, chapter 17. IntechOpen, Rijeka, 2008.

[24] 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. In *ICCV*, 2017.

[25] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun. Alignedreid: Surpassing human-level performance in person re-identification. In *arXiv:1711.08184v2*, 2018.

[26] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, and S. Yi. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017.

[27] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, pages 868–884, 2016.

[28] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Krishna Chandraker, and Qi Tian. Person re-identification in the wild. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3346–3355, 2017.

[29] Z. Zheng, L. Zheng, and Y. Yang. Pedestrian alignment network for large-scale person re-identification. *IEEE Trans. on Circuits and Systems for Video Technology*, 2017.

[30] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3652–3661. IEEE, 2017.

[31] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tian. See the forest for the trees: Joint spatial and temporal recurrent neual networks for video-based person re-identification. In *CVPR*, 2017.

[32] X. Zhu, X. Jing, X. You, X. Zhang, and T. Zhang. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. *IEEE Transactions on Image Processing*, 27(11):5683–5695, Nov 2018. ISSN 1057-7149. doi: 10.1109/TIP.2018.2861366.