Aalborg Universitet

Attention in Multimodal Neural Networks for Person Re-identification

Lejbølle, Aske Rasch; Krogh, Benjamin; Nasrollahi, Kamal; Moeslund, Thomas B.

Attention in Multimodal Neural Networks for Person Re-identification

Aske R. Lejbølle[1]   Benjamin Krogh[2]   Kamal Nasrollahi[1]   Thomas B. Moeslund[1]

[1]Visual Analysis of People (VAP), Aalborg University
Rendsburggade 14, 9000 Aalborg
{asrl,kn,tbm}@create.aau.dk

[2]BlipTrack Veovo
Haekken 2, 9310 Vodskov
benjamin.krogh@veovo.com

Abstract

In spite of increasing interest from the research community, person re-identification remains an unsolved problem. Correctly deciding on a true match by comparing images of a person, captured by several cameras, requires extraction of discriminative features to counter challenges such as changes in lighting, viewpoint and occlusion. Besides devising novel feature descriptors, the setup can be changed to capture persons from an overhead viewpoint rather than a horizontal. Furthermore, additional modalities can be considered that are not affected by similar environmental changes as RGB images. In this work, we present a Multimodal ATtention network (MAT) based on RGB and depth modalities. We combine a Convolution Neural Network with an attention module to extract local and discriminative features that are fused with globally extracted features. Attention is based on correlation between the two modalities and we finally also fuse RGB and depth features to generate a joint multilevel RGB-D feature. Experiments conducted on three datasets captured from an overhead view show the importance of attention, increasing accuracies by 3.43%, 2.01% and 2.13% on OPR, DPI-T and TVPR, respectively.

1. Introduction

Person re-identification (re-id) is the task of matching person descriptors extracted from images captured across, typically, non-overlapping cameras and persists as a hot topic within the research community [38]. This is not only due to major challenges, including changes in lighting, viewpoint and occlusion between cameras but also the potential usage within applications such as forensics or long-term tracking of pedestrians [5, 23].

A person re-id system, typically, consists of tracking, features extraction and feature matching using simple distance metrics, for example, Euclidean distance or more sophisticated ones such as Keep It Simple and Straight-forward MEtric (KISSME) based on Bayesian theory and Mahalanobis distance [11]. Variations such as Cross-view Quadratic Discriminant Analysis (XQDA) [17] additionally considers subspace learning while Support Vector Machines (SVM) [31] aims at maximizing distance between features of non-similar pairs. For a comprehensive overview of metrics applied in person re-id, please see [9]. Most often, researchers focus on either feature extraction or matching using supervised learning, although, following recent developments of deep learning, Convolution Neural Networks (CNN) have been proposed also in the case of person re-id [1, 3, 26, 28, 30, 32]. These networks are able to learn both discriminative features and a classifier simultaneously by training in an end-to-end fashion.

Due to more focus on CNN in re-id, more data has become a necessity to properly train the networks. As a result, larger datasets in recent years have emerged [16, 36, 37], not only allowing proper use of CNN but also increasing the realism of re-id evaluation. Common for these datasets is the viewpoint which is mostly horizontal, allowing occlusions between persons or persons and objects. Another option is to place the camera overhead, resulting in a vertical viewpoint, an option only considered by few [7, 13, 18]. This both has the potential of decreasing the probability of occlusions and improve privacy preservation. Examples of the two different viewpoints are shown in Figure 1.

By changing the viewpoint, less color and texture information might be available and it is therefore crucial to extract features that represent the most important parts of a persons appearance. One way is to learn part-specific CNN models by splitting the image into local regions and feed those to separate CNN streams [26, 30]. Even though, these models learn local feature responses, they still consider regions that are not relevant to the feature descriptor, decreasing invariance to lighting, background clutter, etc.

Another way is to apply an *attention mechanism*, originally introduced and applied in Neural Machine Translation problems (NMT) [2], which can be used to consider only certain local parts of an image. Within computer vision, this
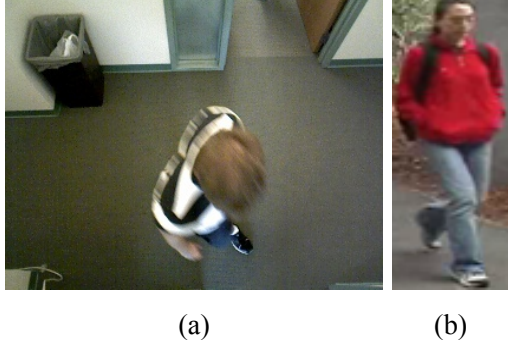
(a)   (b)

Figure 1. Examples of images captured from (a): an overhead viewpoint [7] and (b): a horizontal viewpoint [6].

method has been applied with great success to both image captioning [29], action recognition [24] and, more recently, person re-id [7, 19, 33].

Attention works by calculating a set of positive weights defined as a 2D attention map. Attention maps are then used to summarize features extracted from a CNN. Two types of attention are often considered, soft attention where attention weights are calculated based on a differentiable deterministic mechanism which can easily be trained along with the rest a neural network, and hard attention where weights are calculated by a stochastic process.

Besides capturing local information, additional modalities can be considered to extract different heuristics. Based on extracting features from images captured in an overhead view, it makes sense to include depth information as an additional modality. To that end, previous work on multimodal person re-id has shown RGB and depth based features to complement each other well [10, 13, 18].

In this paper we apply soft attention to person re-id, considering images captured from an overhead view. Instead of only applying attention using color or depth information, we consider a multimodal approach by calculating attention weights based on fusing RGB and depth features, both extracted using pre-trained CNN. As a result, attended regions in the RGB image are based on the representation in depth domain which better captures information around regions with significant change in depth. Vice versa, attended regions in the depth domain are based on the RGB representation to better capture depth information in areas with discriminative color information. To extract features from different discriminative regions, we learn attention maps at multiple layers of the CNN and fuse locally summarized features. Additionally, local features are fused with global feature descriptors to capture information at different abstraction levels as previously proposed with success [14, 28].

Finally, we also learn a joint feature representation by fusing RGB and depth features in the late layers of the network to produce a multilevel RGB-D based feature descriptor and train the entire network end-to-end. To summarize, our contributions include:

- We implement soft attention in a multimodal CNN by fusing RGB and depth features.

- We analyze the importance of attention in a multimodal context by visualizing calculated attention maps in different scenarios.

The rest of the paper is structured as follows. Related work is presented in Section 2 followed by a description of the proposed methodology in Section 3. Experimental results are presented in Section 4, including an impact analysis of applying attention. Finally, the paper is concluded in Section 5.

2. Related Work

Ever since the first significant results in object recognition [12], CNN have been proposed in person re-id [1, 16]. While these focus on globally extracted features, more recent proposals are based on part-based learning to capture more local information [3, 26, 32]. Ustinova *et al.* [26] propose a Bilinear-CNN by splitting the body into three parts and train part specific CNN that are summarized by bilinear combination of features. Finally, features from the three parts are fused in a fully connected layer. Part specific CNN are also proposed by Cheng *et al.* [3] who split the body into four parts and learn both part specific and global features that are fused in the late layers of the CNN. A different approach is followed by Zhao *et al.* [32] who apply a Regional Proposal Network (RPN) to locate 14 human body joints and extract seven body sub-regions. A CNN is applied to each sub-region to learn part specific features that are afterwards fused in a four layered feature fusion network (FFN). Part localization is also proposed by Li *et al.* [15], but instead of localizing the joints, they apply a Spatial Transformer Network (STN) to localize head-shoulder, upper-body and lower-body regions. Once again, part specific features are learned and later fused with globally extracted features. Common for aforementioned methods is the requirements of a horizontal viewpoint in order to either have a properly division of body parts or localize the joints. In case of an overhead view, this is not possible.

Soft attention in CNN can be related to saliency learning using hand-crafted features which also aims at locating discriminative regions. Little work has been done within this area, most notable are the works of Zhao *et al.* [34, 35]. In [35] they propose salience learning by matching patches within a constrained window between images of persons captured by two different cameras. For each patch, a salient score is calculated using either K-Nearest Neighbors or One-class SVM. Meanwhile, in [34] they propose learning
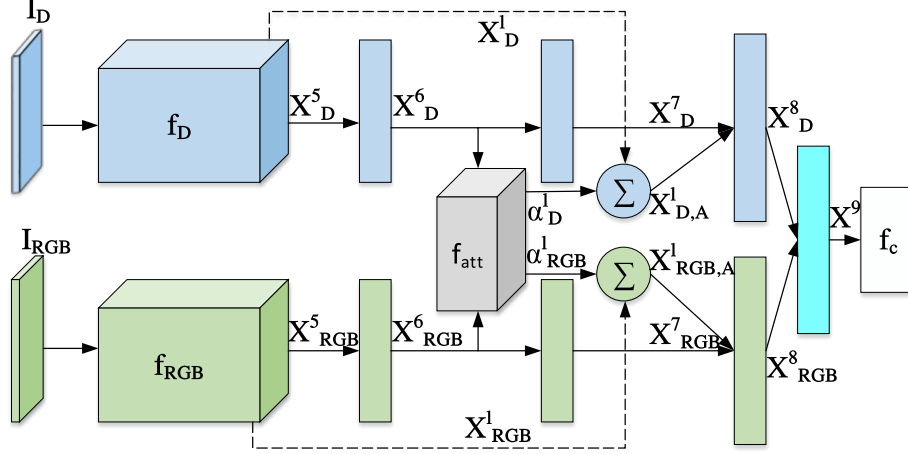
Figure 2. Overview of the Multimodal ATtention network (MAT). An RGB and depth image is encoded by an RGB based encoder, shown by the green stream, and depth based encoder, shown by the blue stream, respectively. Outputs from the last convolution layer are embedded and applied to the attention module which calculates attention maps for each modality. Feature maps from the encoders are summarized using the attention maps and fused with global feature representations at each modality. Finally, features from the two modalities are fused to a multilevel RGB-D based feature descriptor and used for prediction.

discriminative mid-level filters by clustering image patches with coherent appearance and apply SVM. These filters are then used to calculate filter responses of input images prior to feature matching.

Attention has been previously proposed only a few times within person re-id [7, 19, 33]. Liu *et al.* [19] propose a Comparative Attention Network (CAN) which is trained end-to-end by producing and comparing attended regions of positive and negative image pairs, i.e., images of similar and non-similar persons. By combining a CNN with a Long Short-Term Memory (LSTM) network, attention maps are produced at different time steps to capture different local regions by using the same encoded image as input at each time step. The work of Zhao *et al.* [33] is also motivated by attention which is used to model a part-aligned human representation by learning attention weights through end-to-end training using a triplet loss function. Finally Haque *et al.* [7] propose a depth-based recurrent visual attention network by combining a CNN with an LSTM to learn spatiotemporal features. By adding a localization network, discriminative features are extracted from glimpses, i.e., a minor region in the input. The localization network is trained using reinforcement learning to focus on discriminative regions. While [19, 33] apply attention in the RGB domain, [7] apply attention in depth domain. This work, to our knowledge, is the first to apply attention in a multimodal context.

Multimodal fusion of RGB and depth information is rarely considered in person re-id [18, 21, 27]. Liciotti *et al.* [18] propose a combination of hand-crafted RGB and depth features to capture both color, texture and anthropometric information. RGB-D based hand-crafted features are

also proposed by Wu *et al.* [27] who extract a rotation invariant Eigen-depth feature and fuse it with low-level color and texture features [17]. Only two previous proposals fuse RGB and depth features using a CNN [10, 13]. Karianakis *et al.* [10] learn spatiotemporal features from a combined CNN and LSTM. Considering the small sample size issue, they add hard attention to incorporate regularization. Finally, Lejbølle *et al.* [13] propose a multimodal CNN which jointly learns a multimodal feature descriptor based on individually trained RGB and depth CNN. Common in aforementioned work is fusion of features which is simply done by concatenation which does not capture the correlation between features from different modalities. In this work, we use correlation between depth and RGB features to extract local information from the input images and, additionally, exploit the advantage of multimodal feature fusion by learning a joint descriptor based on RGB and depth.

3. Methodology

An overview of the proposed network is shown in Figure 2. RGB and depth images I_{RGB} and I_D are encoded using an RGB based encoder f_{RGB} and depth encoder f_D, respectively, represented by CNN. The outputs from the last convolution layer are embedded in fully connected layers and used as input to the attention model f_{att}. The attention model multiplies features to capture correspondence between modalities, following the idea of multiplicative interaction [20]. Attention weights, α^l, are afterwards calculated separately for the l'th layer of the RGB and depth encoders, and used to summarize feature maps X_{RGB}^l and X_D^l. The summarized features are fused with globally ex-

tracted feature descriptors and the two modality based features are fused to learn a joint feature representation. Finally, a classification module f_c is added for prediction when training the network. In the rest of the paper, we refer to our proposed network as Multimodal ATtention network (MAT).

3.1. Visual Encoder

The input to the MAT is an RGB image I_{RGB} and a corresponding depth image I_D that are separately processed by modality based encoders f_{RGB} and f_D given by,

$$
\begin{aligned}
X^5_{RGB} &= f_{RGB}(I_{RGB}, \theta_{RGB}) \\
X^5_D &= f_D(I_D, \theta_D),
\end{aligned}
\tag{1}
$$

where $X^5_{RGB} \in R^{N \times N \times K}$ and $X^5_D \in R^{N \times N \times K}$ are the outputs from the fifth and last convolution layer denoted by the superscript 5, θ_{RGB} and θ_D are the encoder weights while K represents the number of feature maps of size $N \times N$.

The encoders follow the Caffe variation (CaffeNet) [8] of the AlexNet CNN [12] for better comparison with the related method of [13] which does not consider attention. The CaffeNet consists of five convolution layers, the first and second followed by local response normalization and max pooling. Max pooling is also added after the fifth convolution layer and followed by three fully connected layers, the last one used to calcuate an output score for each class normalized by a softmax function. Rectified Linear Units (ReLU) are used as nonlinear activation while dropout with a probability of 0.5 is added between fully connected layers to increase network generalization [25].

Followeing the baseline architecture of [13], the encoders take as input images of size 227×227 and output feature descriptors $X^5_{RGB} \in R^{13 \times 13 \times 256}$ and $X^5_D \in R^{13 \times 13 \times 256}$. Two fully connected layer afterwards embed features to sparse feature descriptors $X^7_{RGB} \in R^{4096}$ and $X^7_D \in R^{4096}$, representing global information. Different from [13], we do not fuse X^7_{RGB} and X^7_D to a joint RGB-D feature, but first fuse each modality based feature with locally extracted features from the attention model.

3.2. Attention Model

The attention model f_{att} is based on using depth information to calculate attention weights for the RGB features and vice versa. In this subsection, we outline the calculation of RGB attention weights, similar calculations are defined in case of depth by simply exchanging subscripts *RGB* and *D*.

As input to the attention model, we use features $X^6_{RGB} \in R^{4096}$ and $X^6_D \in R^{4096}$ extracted from the first fully connected layer. The attention weights to extract local features

from the output of any given layer of f_{RGB} are then calculated as,

$$
\begin{aligned}
e^l &= f_{att}(X^6_{RGB}, X^6_D, \theta_\alpha), \quad e^l \in R^{N^2} \\
\alpha^l_i &= \frac{\exp(e^l_i)}{\sum_i e^l_i}, \qquad\qquad \alpha^l \in R^{N^2},
\end{aligned}
\tag{2}
$$

where e^l is a vector of unnormalized attention weights of size N^2 and θ_α represents the attention model parameters. To calculate a weighted average of features, attention weights are normalized using a softmax function, resulting in α^l, as originally proposed [2]. Thus, given a feature map of, e.g., size 13×13, we calculate 169 normalized attention weights.

The attention model implements multiplicative interaction to learn relations between RGB and depth features, and calculation of attention weights can therefore also be written as,

$$
e^l = W^l_{att}(X^6_{RGB} \odot X^6_D) + b^l_{att},
\tag{3}
$$

where $\odot$ represents an element-wise multiplication while $W^l_{att} \in R^{M \times N}$ and $b^l_{att} \in R^N$ are the weights and bias of the attention model, respectively, and M is the number of hidden units in X^6_{RGB}.

The normalized attention weights calculated in Equation 2 are then used to calculate the weighted average of features from the l'th layer of f_{RGB} as,

$$
X^l_{RGB,A} = (\hat{X}^l_{RGB})^T \alpha^l, \quad X^l_{RGB,A} \in R^K,
\tag{4}
$$

where $\hat{X}^l_{RGB} \in R^{N^2 \times K}$ is the flattened output from layer l and $X^l_{RGB,A}$ is a feature descriptor containing local information from the input RGB image dependent on features from the depth image. Since the attention maps are used to summarize features across all feature maps, only local regions of interest are considered. In our experiments presented Section 4, we calculate attention maps for the fourth and fifth convolution layers of f_{RGB} and f_D to capture different local information, resulting in calculations of, in total, four attention maps. We observe that consideration of additional low-level information from earlier convolution layers does not improve accuracy. Given the outputs $X^4_{RGB} \in R^{13 \times 13 \times 384}$ and $X^5_{RGB} \in R^{13 \times 13 \times 256}$, we thereby summarize features using attention maps $\alpha^4_{RGB} \in R^{169}$ and $\alpha^5_{RGB} \in R^{169}$ resulting in attention based features $X^4_{RGB,A} \in R^{384}$ and $X^5_{RGB,A} \in R^{256}$.

The attention based features are afterwards fused with X^7_{RGB} by adding a new fully connected layer, to form a modality based multilevel feature $X^8_{RGB} \in R^{4096}$. Finally, multilevel RGB and depth features are fused by a second new fully connected layer resulting in a multimodal feature descriptor X^9_{RGBD} used for prediction.

Prediction is implemented by calculating a probability score of each class given X^9_{RGBD}. A softmax layer is added

to normalize scores and the entire network is trained end-to-end by minimizing the logistic loss function defined as,

$$\min_{\theta_{RGB},\theta_D,\theta_\alpha,\theta_c} -\frac{1}{J}\sum_{i=1}^{J} log(\hat{p}_i) \tag{5}$$
$$\hat{p}_i = f_{MAT}(I_{RGB}, I_D; \theta_{RGB}, \theta_D, \theta_\alpha, \theta_c),$$

where the loss is calculated over a mini-batch of size J and $\hat{p}_i$ represents the normalized score for the i'th image predicted by f_{MAT}.

4. Experiments

This section outlines the experimental results and analysis of the MAT. First, evaluated dataset and corresponding test protocols are described followed by details of training f_{RGB}, f_D and f_{MAT}. Finally, the results are presented with a comparison to state-of-the-art methods and the attention module is analyzed by a visualization of calculated attention maps.

4.1. Datasets and Protocols

When evaluating the MAT, we only consider datasets collected from an overhead viewpoint. Three RGB-D based datasets, to our knowledge, have been proposed for overhead person re-id, including the Depth-based Person Identification from Top (DPI-T) [7], Top View Person Re-identification (TVPR) [18] and Overhead Person Re-identification (OPR) [13].

DPI-T: Recorded in a hallway, this dataset consists of 12 persons, appearing in an average of 25 sequences in five different appearances, both in the training and test set. A total of 213 sequences are included in the training set, while 249 are used for testing. During test, all sequences from the test set are matched with those in the training set.

TVPR 23 videos are recorded in a hallway, including a total of 100 persons, each appearing twice. The training set consists of persons walking from left to right, while walking from right to left in the test set. At test time, sequences from the test set are compared with those of the training set. Due to missing frames in one of the recorded videos at time of testing, 94 persons are considered in our evaluation. Different from [13] who consider full-frame images, we apply a You Only Look Once (YOLO) detector [22] optimized for person detection, to automatically extract the ROI around persons.

OPR This dataset, recorded in a university canteen, consists of 64 persons captured twice, when entering and leaving the canteen. Similar to protocols in widely used re-id datasets captured from a horizontal view, 10 random train/test splits are performed, each consisting of 32 persons in both training and test set. The final result is then calculated as an average of accuracies across all experiments.

4.2. Implementation details

Before training the MAT, f_{RGB} and f_D are fine-tuned by initializing a Caffe model, pre-trained on the ImageNet dataset. In case of f_D, we encode depth images by applying a JET colormap which has shown to outperform other encoding methods such as surface normals or Horizontal disparity, Height and Angle (HHA) [4]. In addition to also being faster, applying a color map allows us to initialize weights using a pre-trained ImageNet model since each depth value is mapped to a value in the RGB color space ranging from blue (close to the camera) to red (far from camera). Fine-tuning is performed using Stochastic Gradient Descent (SGD) with momentum of $\mu = 0.9$ and a batch-size of 128. The base learning rate is set to $\eta^0 = 0.01$ and reduced by $\eta^i = \eta^{i-1}0.99$ after each epoch. Similar to [13], data augmentation such as cropping and flipping are applied to extend the dataset. To that end, we resize images to 256×256 and draw cropping values from a discrete distribution in range $[0, 29]$. After fine-tuning RGB and depth encoders, we add and initialize the attention module and fusion layers, and similarly train f_{MAT} by SGD. We reduce the base learning rate to $\eta^0 = 0.001$ and train the network using a batch-size of 32. In case of both fine-tuning encoders and training the MAT, training runs for up to 100 epochs which takes 4-5 hours using an Nvidia GTX 1080 GPU.

At test time, we extract features X_{RGBD}^9 from the last fully connected layer and use Euclidean distance to match features extracted in different camera views. Results are ranked according to the distance, intuitively, having the match with the shortest as the most similar. Since all datasets contain several images of each person, we apply a multi-shot approach and pool features extracted from all images of each person. Pooling is implemented by calculating average features which has previously shown superior to, e.g., maximizing when combining CNN features and a Euclidean distance metric [26, 36]. Although, in case of TVPR, we observe feature maximization to perform better and therefore provide results on that dataset using maximized features.

4.3. Experimental Results

We present results as Cumulative Matching Characteristic (CMC) curves that are produced by calculating a cumulative score for each *rank-i* indicating the number of persons having their true match within the i most similar in the ranked list.

The CMC curves produced from results on DPI-T, TVPR and OPR are shown in Figure 3, along with results without the use of attention, similar to the proposed method of [13][1].

---

[1]In the original study, the authors identified a minor error in the input of OPR after publication, hence, results differ from those reported in [13].
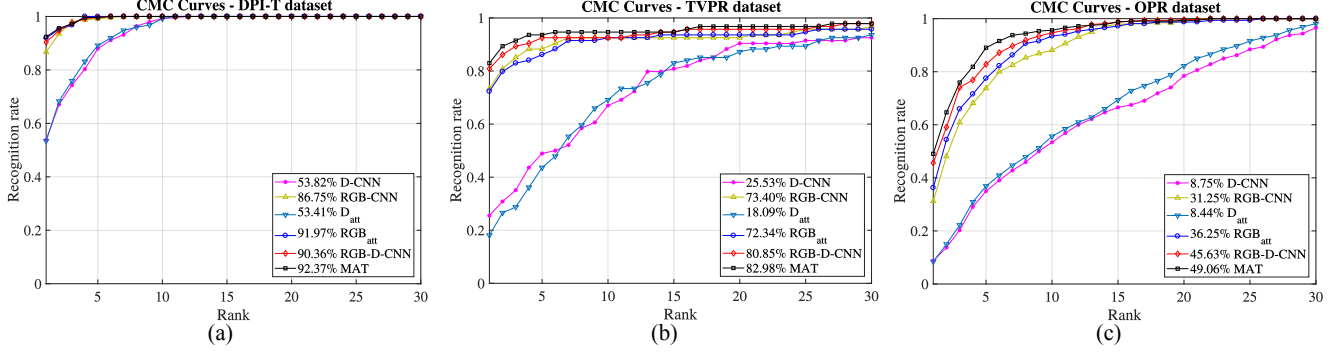
Figure 3. Experimental results on (a): DPI-T (p=249), (b): TVPR (p=94) and (c): OPR (p=32) for our multilevel attention-based RGB and depth features (D_{att} and RGB_{att}) along with MAT, and D-CNN, RGB-CNN and RGB-D-CNN proposed in [13].

RGB_{att} and D_{att} represent the attention-based multilevel color and depth features X_{RGB}^8 and X_D^8, while D-CNN, RGB-CNN and RGB-D-CNN represent the baseline depth, color and joint models, respectively.

From Figure 3 it is clear that addition of attention-based features increases the rank-1 accuracy, even though, the accuracy is already high. Additionally, fusion of RGB and depth features outperform the use of RGB or depth individually. This is the case for DPI-T where the MAT increases the rank-1 accuracy by 2.01% and 0.4% compared to RGB-D-CNN and RGB_{att}, respectively. It is also worth noticing the increase of 5.22% when comparing RGB_{att} and RGB-CNN which shows the effect of using attention maps to extract local features and fuse those with global features. Similarly on TVPR, the rank-1 accuracy is increased by 2.13% and 10.64% compared to RGB-D-CNN and RGB_{att}, respectively. Comparing RGB_{att} and D_{att} to RGB-CNN and D-CNN, respectively, the use of attention does not seem to have a positive impact which could be due to misalignment issues from the detection, leading to missing information. This will be further analyzed in Subsection 4.4. Nonetheless, fusing the attention-based features results in a higher accuracy when comparing MAT and RGB-D-CNN. Finally on OPR, the MAT increases rank-1 accuracy by 3.43% and 12.81% compared to RGB-D-CNN and RGB_{att}, respectively. Similar to DPI-T, fusing local and global information increases rank-1 accuracy by 5.00% when comparing RGB_{att} and RGB-CNN.

4.4. Analysis of Attention

To identify the contribution from the attention model, we visualize examples of attention maps α_{RGB}^4, α_{RGB}^5, α_D^4 and α_D^5 for all evaluated datasets. The visualizations are shown in Figure 4. We show examples of persons having their true match as most similar ((a), (c) and (e)) and persons having their true match outside top-10 ((b), (d) and (f)). In case of TVPR and OPR, we randomly sample an image from each view and calculate attention maps. Since

DPI-T consists of multiple sequences of each person, we randomly sample images from the most similar sequences between views.

Generally, attention maps differs between the datasets. In case of DPI-T, attention is mostly focusing on parts of the floor, although, attended regions also include parts of the person. This is most notable in case of α_{RGB}^4 in Figure 4 (a) where attention is mostly centered around the person and edges of the images. The pattern of α_{RGB}^5 is more random, almost only capturing features from the floor. This behavior could be due to the encoding of depth images resulting in larger gradient changes in the floor compared to the persons, causing the floor to have a higher impact on the RGB based attention maps. Meanwhile, attention maps α_D^4 and α_D^5 focus on minor local regions centered around the floor. Considering full-frame images, combined with uniform colors of the scene, depth based attention maps are more affected by the colors of the floor, causing local features to almost not contain any information from the persons. This results in addition of noisy information, decreasing accuracy which is also clear when comparing D_{att} and D-CNN in Figure 3 (a). This indicates the importance of extracting the ROI around the persons to remove as much background information as possible. In order to identify contributing regions, calculated attention maps before and after training the MAT should be compared. This will be considered in future work.

The attention maps for TVPR are less random but more similar across persons. In case of both Figure 4 (c) and (d), α_{RGB}^4 and α_D^5 capture local information from the bottom right part of the images while α_{RGB}^5 and α_D^4 capture information in the center right part of the images. This cause images with misaligned detections to capture local features from the floor, in some cases, negatively affecting accuracy, as also shown in Figure 3 (b) when comparing D_{att} and RGB_{att} to D-CNN and RGB-CNN. A reason for attention maps to be concentrated at the edges of the images could be the low resolution of depth information which results in
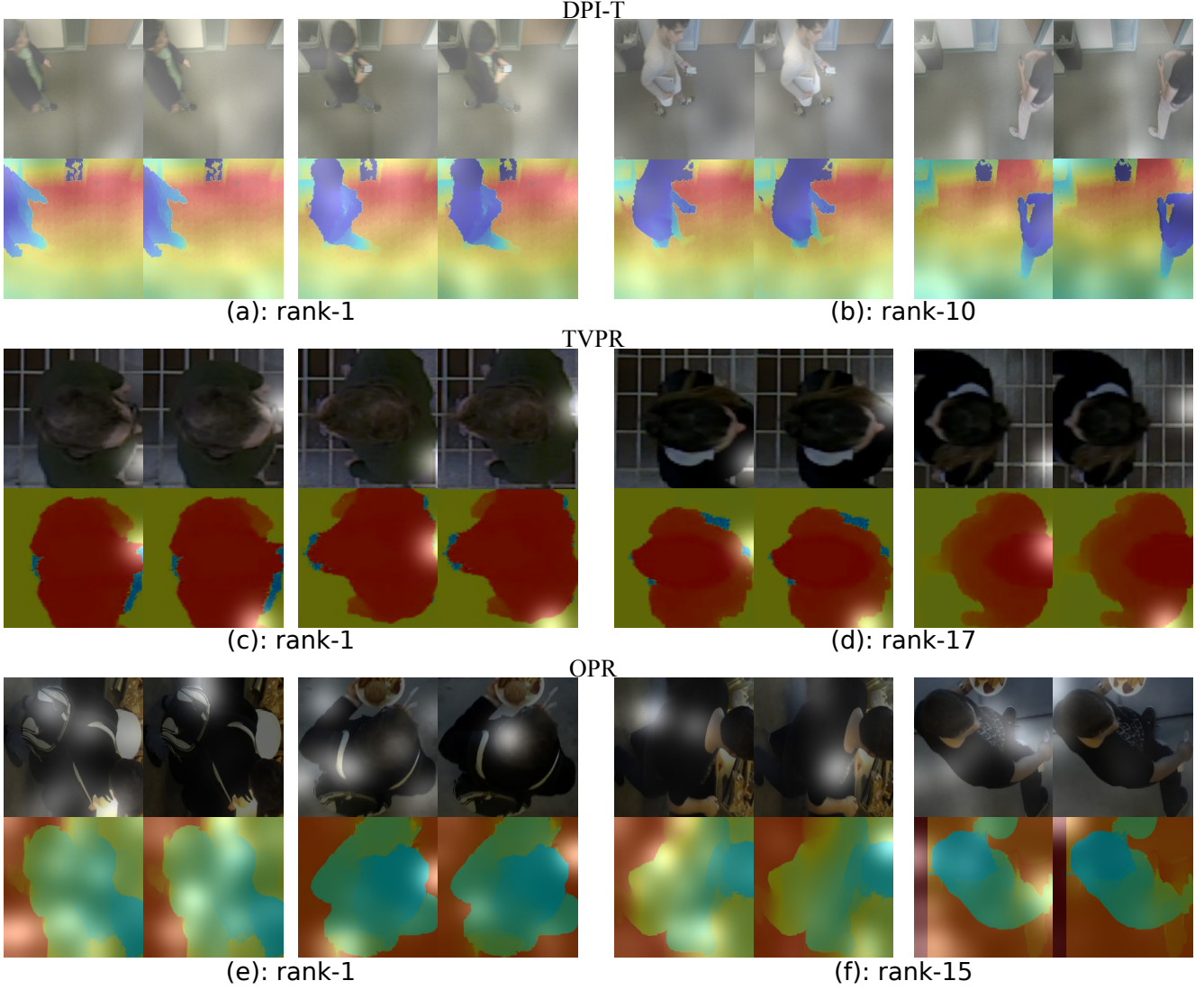
Figure 4. Examples of calculated attention maps in case of DPI-T, TVPR and OPR. Each sub-figure consists of four attention maps from the same person in two different views. The four attention maps are organized as follows; top-left corner: α^4_{RGB}, bottom-left corner: α^4_D, top-right corner: α^5_{RGB}, bottom-right corner: α^5_D. Brighter areas means higher attention weights.

useful gradient information only at the edge of the person. Although, in most cases, discriminative local information is extracted leading to a higher accuracy when fused with complementary global features.

Attention maps calculated in case of OPR are more centered around useful information. Comparing RGB based attention maps α^4_{RGB} and α^5_{RGB}, both mostly focus on the clothing which, typically, provide more discriminative information compared to, for example, the hair. Nevertheless, they still focus on different parts of the image, while α^4_{RGB} focus on multiple local regions with corresponding gradient changes in the depth image, α^5_{RGB}, focus on a single region. Additionally, the impact of fusing RGB and depth is shown by the attended regions, mostly centered near regions with

larger gradient changes, for example, at the shoulder. This has a positive impact since these areas can be assumed to contain more useful information, considering the overhead view. α^4_D and α^5_D are more view dependent, focusing on several regions in the first view, while only focusing on a couple regions in the second. Like DPI-T, this could be due to a more diverse background in the first view. They both capture information around regions with larger gradients, indicating that the attention model learns to calculate depth based attention maps that capture regions with useful color information while still preserving gradient information. A few failure cases exist as seen in the second (right) view of Figure 4 (f). Here, a large gradient change in the left part of the depth image greatly affects the calculation of attention

maps, causing attended regions to be centered around this edge. This is most likely a product of the depth calculations in [13] and should simply be removed in future evaluations.

4.5. Comparison to State-of-the-art

We compare our results with state-of-the-art for the three evaluated datasets. Due to the novelty of these datasets, only few results previously have been presented, including the 4D Recurrent Attention Mechanism (4D RAM) [7] and recurrent network with temporal attention (Depth ReID) [10] in case of DPI-T, and TVDH [18] in case of TVPR. Finally, the results of RGB-D-CNN$_{avg}$ (RGB-D-CNN) presented in [13] are compared. The comparisons are summarized in Table 1-3, in all tables, "–" indicate non present results.

Method/Rank	r = 1	r = 5	r = 10	r = 20
4D RAM [7]	55.60	–	–	–
Depth ReID [10]	77.50	96.00	–	–
RGB-D-CNN [13]	90.36	**99.60**	**100**	**100**
MAT (ours)	**92.37**	**99.60**	**100**	**100**

Table 1. Comparison between MAT and state-of-the-art systems on the DPI-T dataset (p=249). Best results are in bold.

Method/Rank	r = 1	r = 5	r = 10	r = 20
TVDH* [18]	75.50	87.50	89.20	91.90
RGB-D-CNN [13]	63.83	89.36	93.62	**97.87**
RGB-D-CNN†	80.85	92.55	92.55	95.74
MAT (ours)	**82.98**	**93.62**	**94.68**	96.81

Table 2. Comparison between MAT and state-of-the-art systems on the TVPR dataset (p=94). Best results are in bold. (*Results are estimated from the CMC curve, †Reproduced by training and testing on images from detection).

Method/Rank	r = 1	r = 5	r = 10	r = 20
RGB-D-CNN [13]	45.63	82.81	94.69	**99.69**
MAT (ours)	**49.06**	**89.06**	**95.62**	99.38

Table 3. Comparison between MAT and state-of-the-art systems on the OPR dataset (p=32). Best results are in bold.

Comparisons in Table 1 show the MAT to outperform previously proposed methods. While 4D RAM and Depth ReID only consider depth information, RGB-D-CNN also considers color, showing the importance of fusing color and depth information. As also mentioned in Subsection 4.3 the MAT still increases accuracy, indicating the importance of including local discriminative features.

In case of TVPR shown in Table 2, the MAT outperforms both RGB-D-CNN [13] and TVDH [18], increasing the rank-1 accuracy by 2.13% and 7.48%, respectively. Additionally, we note the importance of eliminating background noise which is shown by an increased rank-1 accu-

racy of 17.02% when comparing the original RGB-D-CNN results of [13] which considers full-frame images, and our evaluation using a similar system.

Finally, we compare the rank-1 through rank-20 accuracies, also depicted in Figure 3, for the OPR dataset. Besides the rank-1 increase of 3.43%, the rank-5 accuracy is also greatly increased by 6.25% which is important to note, considering an image retrieval context where often the top-k most similar images are inspected by a person.

5. Conclusion

In this paper, we have proposed a Multimodal ATtention network (MAT) which implements an attention model with a multimodal CNN to calculate attention maps that capture local discriminative features from RGB and depth images. Attention maps are calculated by fusing RGB and depth information, resulting in attention maps that are calculated in a multimodal fashion. In total, four attention maps are calculated to extract local features from the fourth and fifth convolution layers of an RGB and depth CNN, respectively. Local RGB and depth based features are separately fused with global feature descriptors resulting in modality dependent multilevel features. Finally, multilevel RGB and depth features are fused to a multilevel RGB-D feature descriptor which better captures the correlation between RGB and depth information while including information at different abstraction levels. Evaluations on three overhead based datasets DPI-T, TVPR and OPR show the importance of fusing local and global information by increasing the rank-1 accuracy by 2.01%, 2.13% and 3.43%, respectively, compared to a similar network not considering attention.

To further increase accuracy, a more novel CNN should be considered while also the addition of an LSTM layer can be used to extend the network by additionally capture temporal information. By adding an LSTM, different attention modules can be considered, either spatial, temporal, or spatiotemporal.

Acknowledgement

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *Proc. CVPR*, pages 3908–3916, 2015.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proc. CVPR*, pages 1335–1344, 2016.

[4] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard. Multimodal deep learning for robust rgb-d object recognition. In *Proc. IROS*, pages 681–687, 2015.

[5] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales. The re-identification challenge. In *Person re-identification*, pages 1–20. Springer, 2014.

[6] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proc. ECCV*, pages 262–275, 2008.

[7] A. Haque, A. Alahi, and L. Fei-Fei. Recurrent attention models for depth-based person identification. In *Proc. CVPR*, pages 1229–1238, 2016.

[8] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proc. ACMMM*, pages 675–678, 2014.

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

[10] N. Karianakis, Z. Liu, Y. Chen, and S. Soatto. Person depth reid: Robust person re-identification with commodity depth sensors. *arXiv preprint arXiv:1705.09882*, 2017.

[11] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Proc. CVPR*, pages 2288–2295, 2012.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, pages 1097–1105, 2012.

[13] A. R. Lejbølle, K. Nasrollahi, B. Krogh, and T. B. Moeslund. Multimodal neural network for overhead person re-identification. *Proc. BIOSIG*, pages 25–34, 2017.

[14] A. R. Lejbølle, K. Nasrollahi, and T. B. Moeslund. Enhancing person re-identification by late fusion of low-, mid- and high-level features. *IET Biometrics*, 7(2):125–135, 2018.

[15] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proc. CVPR*, pages 384–393, 2017.

[16] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proc. CVPR*, pages 152–159, 2014.

[17] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proc. CVPR*, pages 2197–2206, 2015.

[18] D. Liciotti, M. Paolanti, E. Frontoni, A. Mancini, and P. Zingaretti. Person re-identification dataset with rgb-d camera in a top-view configuration. In *Video Analytics. Face and Facial Expression Recognition and Audience Measurement*, pages 1–11. Springer, 2016.

[19] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 26(7):3492–3506, 2017.

[20] R. Memisevic. Learning to relate images: Mapping units, complex cells and simultaneous eigenspaces. *arXiv preprint arXiv:1110.0107*, 2011.

[21] F. Pala, R. Satta, G. Fumera, and F. Roli. Multimodal person reidentification using rgb-d cameras. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(4):788–799, 2016.

[22] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In *Proc. CVPR*, pages 6517–6525, 2017.

[23] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Proc. ECCV*, pages 17–35, 2016.

[24] S. Sharma, R. Kiros, and R. Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015.

[25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[26] E. Ustinova, Y. Ganin, and V. Lempitsky. Multi-region bilinear convolutional neural networks for person re-identification. In *Proc. AVSS*, pages 1–6, 2017.

[27] A. Wu, W.-S. Zheng, and J.-H. Lai. Robust depth-based person re-identification. *IEEE Transactions on Image Processing*, 26(6):2588–2603, 2017.

[28] S. Wu, Y.-C. Chen, and W.-S. Zheng. An enhanced deep feature representation for person re-identification. In *Proc. WACV*, pages 1–8, 2016.

[29] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. ICML*, pages 2048–2057, 2015.

[30] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *Proc. ICPR*, pages 34–39, 2014.

[31] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan. Sample-specific svm learning for person re-identification. In *Proc. CVPR*, pages 1278–1287, 2016.

[32] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proc. CVPR*, pages 1077–1085, 2017.

[33] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In *Proc. ICCV*, pages 3219–3228, 2017.

[34] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *Proc. CVPR*, pages 144–151, 2014.

[35] R. Zhao, W. Oyang, and X. Wang. Person re-identification by saliency learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(2):356–370, 2017.

[36] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *Proc. ECCV*, pages 868–884, 2016.

[37] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proc. ICCV*, pages 1116–1124, 2015.

[38] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.