Person Re-Identification Across Aerial and Ground-Based Cameras by Deep Feature Fusion

Arne Schumann and Jürgen Metzler

Fraunhofer IOSB, Fraunhoferstr. 1, 76131 Karlsruhe, Germany

ABSTRACT

Person re-identification is the task of correctly matching visual appearances of the same person in image or video data while distinguishing appearances of different persons. The traditional setup for re-identification is a network of fixed cameras. However, in recent years mobile aerial cameras mounted on unmanned aerial vehicles (UAV) have become increasingly useful for security and surveillance tasks. Aerial data has many characteristics different from typical camera network data. Thus, re-identification approaches designed for a camera network scenario can be expected to suffer a drop in accuracy when applied to aerial data.

In this work, we investigate the suitability of features, which were shown to give robust results for re-identification in camera networks, for the task of re-identifying persons between a camera network and a mobile aerial camera. Specifically, we apply hand-crafted region covariance features and features extracted by convolutional neural networks which were learned on separate data. We evaluate their suitability for this new and as yet unexplored scenario. We investigate common fusion methods to combine the hand-crafted and learned features and propose our own deep fusion approach which is already applied during training of the deep network.

We evaluate features and fusion methods on our own dataset. The dataset consists of fourteen people moving through a scene recorded by four fixed ground-based cameras and one mobile camera mounted on a small UAV. We discuss strengths and weaknesses of the features in the new scenario and show that our fusion approach successfully leverages the strengths of each feature and outperforms all single features significantly.

Keywords: person re-identification, aerial, camera network, covariance descriptors, deep learning, fusion, retrieval

1. INTRODUCTION

Person re-identification is an important task for automatic understanding of how people move through large facilities such as airports, public events, shopping centers, etc. It is indispensable in establishing consistent person labels across camera networks or even within the same camera. Particularly at public events conventional ground-based cameras are increasingly often supported by cameras attached to UAVs. A mobile aerial perspective greatly improves the situational awareness for security personnel. For the task of automatic person re-identification, however, this scenario introduces new challenges, such as new viewing angles which significantly differ from those of ground-based cameras.

In order to investigate these challenges, we recorded a new dataset which contains images of multiple persons from an aerial camera, as well as a small camera network. Some impressions of the data can be seen in Figure 1. The left side shows examples of person images from the camera network and the right side images of the same persons recorded by the UAV. Note the different aspect ratios of the aerial images and the different positions of body parts within the bounding boxes.

Our approach to person re-identification in this challenging scenario relies on leveraging two types of image features. Due to their recent successes in the field (e.g. [1, 2]), we rely on features learned by convolutional neural networks (CNNs). These features are trained using large amounts of training data. Our dataset is not large enough to split into training and test parts. We thus train our CNN features on separate (camera network)

Further author information: (Send correspondence to Arne Schumann)
Arne Schumann: e-mail: arne.schumann@iosb.fraunhofer.de
Jürgen Metzler: e-mail: juergen.metzler@iosb.fraunhofer.de

(a) (b)

Figure 1: Conventional, camera network bounding boxes (a) compared to bounding boxes of the same individuals occurring in an aerial view (b). All boxes are normalized to a uniform width. The aerial boxes show a much greater variety of aspect ratios, which lead to distortions when the image is scaled to a uniform size prior to re-identification. Furthermore, the extreme angles can result in very different positions of body parts within the aerial images, e.g. head in a top-view is located 'inside' the torso box, instead of above it.

data and transfer them to the aerial scenario. The advantage of learning a feature from large amounts of data lies in its direct and automatic adaptation to the re-identification task. A drawback, however, is the possibility of overfitting to dataset bias which can lead to a decreased transferability of the resulting features. Our second feature type are hand-crafted features. Specifically, we choose covariance descriptors [3], because they are well suited to combine low level image information into a reliable representation for person re-identification [4]. The advantages of hand-crafted features lie in the fact that such features do not require training data and thus cannot overfit to any data biases. Depending solely on the expert knowledge used to design them, their performance might vary less strongly on different data sources. Our main approach aims at combining the strengths of hand-crafted and trained features. We investigate conventional fusion methods, such as early and late fusion. We then propose our own, deep fusion method which includes the information of the hand-crafted features into the training process of the CNN feature. Our experiments show that the deep fusion approaches are superior to conventional fusion methods, strongly outperform all original features by at least 4.5% in mean average precision and successfully combine the strengths of the individual features.

The remainder of this work is organized as follows: In Section 2 we give a brief overview of related approaches. Section 3 outlines the features we use and our main approaches to deep feature fusion. We evaluate our approach on our own aerial dataset in Section 4 and summarize our findings in Section 5.

2. RELATED WORK

In recent years, deep learning methods have increasingly been applied to person re-identification. Notable works include the early approach of Li et al. [5] which uses a filter pairing architecture to match persons. A special neighborhood matching layer was introduced by Ahmed et al. [6]. More recently, Xiao et al. [2] use domain guided dropout to learn multiple domain specific feature representations in one network. In [1] Cheng et al. use a triplet loss and a simple body part segmentation to learn a robust feature embedding. All approaches achieve state-of-the-art performance on public datasets and demonstrate the high accuracy of deeply learned features for person re-identification.

Covariance descriptors were originally introduced for detection and classification tasks in [3]. They were then successfully applied to tracking [7] and have been shown to be well suited for person tracking in aerial images [8, 9]. They also yield high accuracy when used for appearance-based person re-identification [10, 11]. Their high robustness remains competitive, as has been demonstrated in recent approaches such as Gaussians of Local Descriptors (GOLD) [12] and hierarchical gaussian descriptors [13].

Fusion of multiple features is a common practice for person re-identification. Liu et al. [14] have developed an approach to determine the importance of various features for re-identification of a person and combine the features accordingly. Similar in intention, Zheng et al. [15] determine the importance of a feature to the current re-identification query and use it for an adaptive late fusion at score level. Kawai et al. [16] fuse gait and color features to improve re-identification. A recent work by Eisenbach et al. [17] combines multiple features by score fusion. Attributes have also been successfully combined with hand-crafted features [18, 19]. Wu et al. [20] use a fully connected layer to merge hand-crafted feature information into the training process of CNNs. We also investigate this method as one of our deep fusion approaches.

To our knowledge, person re-identification in a setting which combines aerial cameras and a ground based camera network has not yet been studied. Oreifej et al. [21] have proposed an approach based on histograms, HOG features and prior person alignment to re-identify a person within individual aerial recordings. An approach by Layne et al. [22] investigates discriminative re-identification models in scenes recorded by a very low flying UAV. In previous works [23, 24] we have studied re-identification within aerial recordings as well. However, none of these approaches specifically focus on the challenges of matching persons between aerial and ground based views.

3. METHODOLOGY

We base our fusion approach on two recent feature categories which have proven to be very reliable for person re-identification: Hand-crafted covariance descriptors and CNN features which are learned from training data.

3.1 Covariance Descriptors

Covariance descriptors have been shown to be well suited for person re-identification [4, 11] and were first introduced in [3]. A covariance descriptor represents an image region by a covariance matrix of image features. It proposes a natural way of fusing several features which might be correlated with each other, where diagonal entries of the covariance matrix represent the variance of each feature and the off-diagonal entries represent the correlations between the features.

Let $\mathbf{R}_1 = \{(x,y)\,|\,x' \leq x < x'',\ y' \leq y < y''\}$ be a rectangular image region with the width w and the height h. First, for each pixel inside $\mathbf{R}_1$ a feature vector is calculated. One commonly used feature vector in the literature that is also used in our work is e.g. given by

$$\begin{aligned} f(x,y) = \big(\, & x,\ y,\ R(x,y),\ G(x,y),\ B(x,y), \\ & I_x(x,y),\ I_y(x,y),\ I_{xx}(x,y),\ I_{yy}(x,y), \\ & \sqrt{I_x^2(x,y) + I_y^2(x,y)},\ \arctan\frac{I_x(x,y)}{I_x(x,y)}\,\big)^T\,, \end{aligned} \tag{1}$$

where x and y are image coordinates of $\mathbf{R}_1$. $R(x,y)$ $G(x,y)$ $B(x,y)$ are the RGB color values and $I(x,y)$ is the intensity.

The covariance descriptor $\mathbf{\Sigma}_{\mathbf{R}_1}$ for the region $\mathbf{R}_1$ is then given by

$$\mathbf{\Sigma}_{\mathbf{R}_1} = \frac{1}{wh-1} \sum_{(x,y)\in\mathbf{R}_1} \left(f(x,y) - \mu_{\mathbf{R}_1}\right)\left(f(x,y) - \mu_{\mathbf{R}_1}\right)^T\,, \tag{2}$$

where $\mu_{\mathbf{R}_1}$ is the mean feature vector given by

$$\mu_{\mathbf{R}_1} = \frac{1}{wh} \sum_{(x,y)\in\mathbf{R}_1} f(x,y)\,. \tag{3}$$

The geodesic distance between two covariance descriptors $d(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$ is given by

$$\begin{aligned} d(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) &= \sqrt{\langle \log_{\mathbf{\Sigma}_1}(\mathbf{\Sigma}_2) \mid \log_{\mathbf{\Sigma}_1}(\mathbf{\Sigma}_2)\rangle_{\mathbf{\Sigma}_1}} \\ &= \sqrt{\langle \log_{\mathbf{\Sigma}_2}(\mathbf{\Sigma}_1) \mid \log_{\mathbf{\Sigma}_2}(\mathbf{\Sigma}_1)\rangle_{\mathbf{\Sigma}_2}}\,, \end{aligned} \tag{4}$$

Layer	Size, Stride	Output Dim.	# Channels
Input		$3 \times 160 \times 64$	
Conv 1-4	3×3, 1	$32 \times 160 \times 64$	
Pool	2×2, 2	$32 \times 80 \times 32$	
Inception 1a		$256 \times 80 \times 32$	64
Inception 1b	stride 2	$384 \times 40 \times 16$	64
Inception 2a		$512 \times 40 \times 16$	128
Inception 2b	stride 2	$768 \times 20 \times 8$	128
Inception 3a		$1024 \times 20 \times 8$	256
Inception 3b	stride 1	$1536 \times 20 \times 8$	256
Inception 4a		$1024 \times 20 \times 8$	256
Inception 4b	stride 2	$1536 \times 10 \times 4$	256
FC Feat.		256	
FC Loss		#person IDs	

Table 1: The architecture of our CNN.

where

$$\log_{\mathbf{\Sigma}_1}(\mathbf{\Sigma}_2) = \mathbf{\Sigma}_1^{\frac{1}{2}} \log\left(\mathbf{\Sigma}_1^{-\frac{1}{2}} \mathbf{\Sigma}_2 \mathbf{\Sigma}_1^{-\frac{1}{2}}\right) \mathbf{\Sigma}_1^{\frac{1}{2}} . \tag{5}$$

We use this distance to rank the persons in our gallery with respect to the probe image. In order to allow for a direct comparison to the CNN features, we resize the person images to a size of 160×64 pixels.

A more comprehensive description of the calculation of covariance descriptors is given in [3] and details about the Riemannian manifold of the descriptors can be found in [25].

3.2 CNN Features

The CNN architecture we employ for re-identification consists mainly of inception blocks [26] and resembles that of [2]. Details of the architecture are given in Table 1. We first normalize our input images to a size of 160×64 pixels, the input dimensions of our network. Initially, we employ four convolutional layers to extract some basic image features. We then follow this by four inception blocks. Each block consists of two inception layers. The first such layer is a regular inception layer with two 3×3 convolutions, instead of one 5×5. The second layer reduces the dimensions of our feature maps by half. This is achieved by increasing the stride of the final pooling or convolutional layers to 2 [27]. Due to the relatively small size of our input images, we are limited in the number of times we can reduce feature map dimensions. Inception block 3 does thus employ a stride of 1. The final fully-connected layer in our network is of dimensionality 256 and serves as our CNN feature for re-identification.

We train this model in the Market-1501 training set [28] for all our experiments. We employ a simple softmax-loss for person ID classification to train the network. Batch normalization [29] allows us to start with a high learning rate of 0.1, we multiply this with 0.9 each 2,000 iterations and train for a total of 50,000 iterations. Our training batchsize is 50 and training on an NVIDIA Titan X GPU takes approximately 8 hours.

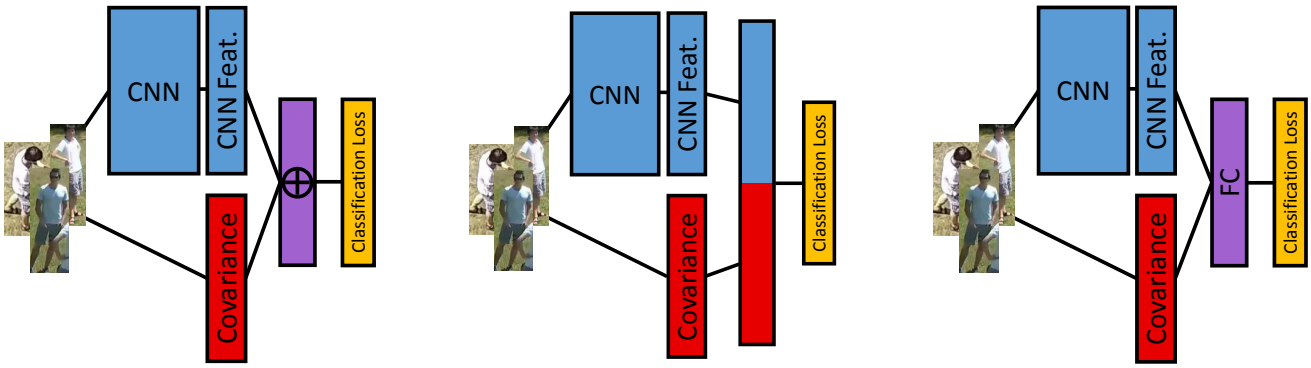
3.3 Conventional Feature Fusion

The general person re-identification pipeline consists of three stages: 1) feature computation, 2) feature comparison by some distance metric, and 3) ranking according to the computed distances. There are three conventional methods which can be employed when the information from two different features is to be fused:

Feature Fusion: After the first stage (i.e. feature computation), feature vectors can be directly combined by concatenating them. This method is sometimes also referred to as early fusion. The distance measure employed in the next step thus has to be suitable for both types of features.

Score* **Fusion:** This fusion method combines the individual distances computed for the two features (i.e.

*Note that we use the word *score* synonymously with *distance* in this work.



(a) (b) (c)

Figure 2: We use three different architectures for deep feature fusion: (a) fusion by elementwise addition layer, (b) fusion by concatenation layer, and (c) fusion by fully-connected layer.

after stage 2). It has the advantage that an individual distance function can be used for each of the features (e.g. geodesic distance for covariance features and cosine distance for deep features). However, care needs to be taken that the resulting distances have similar value ranges. Otherwise one of the features will dominate the fused result. In order to cope with such differences, we compare multiple score normalization methods in our evaluation: normalization to unit length, min-max normalization and normalization by sigmoid function. We apply normalization over all gallery scores obtained for any single query. Score fusion is sometimes also referred to as late fusion.

Decision Fusion: Finally, a fusion step is possible after the third stage of generating a ranking for each individual feature. Two such rankings can be fused by simply considering and averaging their ranks. This has similarity to score fusion but uses a coarser information (i.e. rank instead of score).

3.4 Deep Fusion

In order to achieve a more meaningful and accurate combination of our features, we propose to include the information contained in the hand-crafted feature into the training process of the CNN. If the classification loss can consider this information during training, the task to be learned changes from general person re-identification to learning to compensate only in those cases where the hand-crafted feature does not already perform well. We suggest that this might be a simpler task to learn and our results in Section 4.4 confirm this. We investigate three architectures to include information from hand-crafted features into the CNN, which are depicted in Figure 2.

Deep-add: Our first architecture combines the final feature layer of the CNN with the precomputed hand-crafted feature by a simple element-wise addition. This method adds no parameters to the net. However, the CNN feature layer is restricted to match the dimension of the hand-crafted feature in this case. Note that this method performs a similar operation at train time as score fusion does at test time. This can be easily seen by considering two person images and their corresponding features (CNN_1, HC_1) and (CNN_2, HC_2), where CNN_i and HC_i represent the CNN and hand-crafted feature vectors for person i, respectively:

$$(CNN_1 + HC_1) - (CNN_2 + HC_2) = (CNN_1 - CNN_2) + (HC_1 - HC_2). \quad (6)$$

In the above, the leftmost term corresponds to an elementwise addition prior to a comparison of the persons (i.e. as performed in the network) and the rightmost term corresponds to score fusion.

This idea has parallels to the ResNet [30] architectures with the identity connection being replaced by the hand-crafted feature. Similar to ResNets, our hand-crafted feature path does not add any parameters to the net and lets the parallel block of the network focus on learning an offset to the information provided through the parameter-free path.

Deep-concat: The second approach consists of an architecture which uses a concatenation layer to combine the CNN feature layer with the hand-crafted feature. The dimensions of the two features need not match for

Figure 3: Our dataset consists of a mobile aerial camera (three possible views shown left) and a small ground-based camera-network of four cameras (right). Resolutions and color distortions vary strongly between cameras. The locations of the ground based cameras are marked in the aerial view (best viewed in color).

this approach and the resulting feature will usually have a much larger dimension. Note that this architecture performs the same operation as conventional feature fusion but with the added benefit of giving the network the opportunity to consider the hand-crafted feature's information at train time.

Deep-fc: Finally, we also investigate feature fusion in the learning process by combining the features through a fully-connected layer. This method has no direct equivalent among the conventional fusion methods. It is also the only method that actually increases the number of parameters in the network. A similar approach has previously been described in [20].

All our fusion methods are trained on the same data as our original CNN and use the same training parameters and basic architecture.

4. EVALUATION

Our baseline features and our fusion methods are evaluated on a self-recorded dataset which features aerial and ground-based cameras. Mean average precision (mAP) is used for evaluation of our retrieval results, because this metric best captures overall performance in cases where multiple positive matches are contained in the gallery set.

4.1 Dataset

In order to evaluate our approach to person re-identification between ground-based and aerial cameras we recorded our own dataset which consists of 45 minutes of video from an outdoor scene. A moving quadrocopter was used to record from an aerial perspective with many changes in position and viewing angle. Additionally, a small network of four ground-based cameras was used to record the scene. An impression of the different cameras is given in Figure 3. The cameras vary strongly in viewing angle, color characteristics and resolution (see Table 2).

During the recording, 14 persons entered and exited the fields-of-view of the various cameras frequently. For our evaluations we sampled the recordings at every 100th frame and annotated all occurring persons. We generated an individual probe and gallery set for each of the five cameras. In order to balance our dataset, we limit the number of probe images to at most 10 for each person ID and camera. The number of gallery images for each person and camera is limited to at most 100. Probe and gallery images were chosen at random from the annotations and the sets are non-overlapping. These individual splits into probe and gallery set for each camera allow us to evaluate in detail the performance of our approach across any combination of probe- and gallery-camera. In total, our dataset contains 1217 probe images and 4244 gallery images. A plot of the variation

Camera	Resolution
UAV	3840×2860
G1	1280×720
G2	1920×1080
G3	1280×720
G_HR	4032×2192

Table 2: Resolutions of the five cameras.

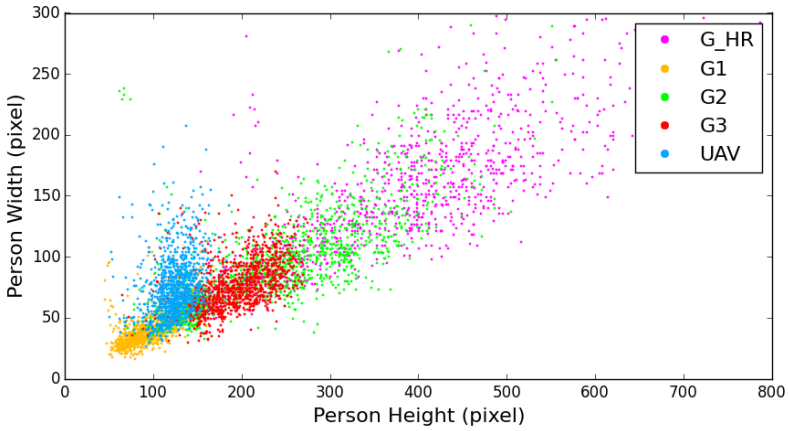


Figure 4: Distribution of person dimensions in our dataset. Note that the aerial camera (UAV) contains markedly different aspect ratios than the ground-based cameras.

in bounding box dimensions and aspect ratio of our annotations is given in Figure 4. Note that the aerial camera has a markedly different distribution of aspect ratios compared to the ground-based cameras.

4.2 Single-Feature Re-Identification

We establish baseline results by first evaluating region covariance and deep features separately on our dataset. In order to better judge the performance of the features, we report the overall mAP as well as five mAP scores for the following settings:

- *Air*: the mAP within the aerial camera (i.e. probe set and gallery set come from the UAV camera),
- *Ground*: the averaged mAP from within each of the ground cameras (i.e. the average over four ground camera values where probe and gallery set come from the same ground camera),
- *Gr.* $\rightarrow$ *Gr.*: the averaged mAP for all ground cameras where probe and gallery set come from different cameras (i.e. an average over 12 mAP scores),
- *Air* $\rightarrow$ *Gr.*: the averaged mAP for the aerial camera probe set and any ground camera gallery set (i.e. an average over 4 mAP scores), and
- *Gr.* $\rightarrow$ *Air*: the averaged mAP for any ground camera probe set and the aerial camera gallery set (i.e. an average over 4 mAP scores).

4.2.1 Region Covariance Features

We evaluated six single-color-space covariance descriptors on our dataset. We considered the HSV, Lab, RGB, YUV, XYZ and YCrCb color spaces. Results are given in Table 3. The color descriptors yield comparable performances with the exception of the HSV color space descriptor, which is less accurate. We also experimented with additional gradient information by adding

$$I_x(x,y),\ I_y(x,y),\ I_{xx}(x,y),\ I_{yy}(x,y),\ \sqrt{I_x^2(x,y)+I_y^2(x,y)},\ \arctan\frac{I_x(x,y)}{I_x(x,y)}$$

as basic features, where $I(x,y)$ is the intensity. However, our results show that gradient information does not help to improve re-identification accuracy on our dataset. A likely reason for this is that the added gradient channels predominantly capture the person contours, which can strongly vary across different images of one person.

Feature	All	Air	Ground	Gr. → Gr.	Air → Gr.	Gr. → Air
Covariance HSV	20.7	28.9	38.8	16.0	17.1	18.2
Covariance Lab+Gradient	26.9	30.4	40.8	24.0	23.4	24.2
Covariance Lab	30.1	31.4	45.2	27.6	26.0	26.5
Covariance RGB+Gradient	27.5	31.4	43.1	24.2	23.6	24.6
Covariance RGB	**30.7**	32.5	47.6	**27.8**	26.0	**26.7**
Covariance YUV	**30.7**	32.5	47.8	27.7	26.0	26.6
Covariance XYZ	30.2	32.6	47.4	27.2	25.8	25.7
Covariance YCrCb	**30.7**	32.5	47.6	**27.8**	26.0	**26.7**
Covariance MCS	29.6	**42.2**	**58.2**	21.7	**27.9**	23.2
Deep Feature	**43.6**	37.2	**69.7**	**43.4**	**31.4**	**31.8**

Table 3: mAP scores for various covariance descriptors and our CNN feature. Note that gradient information does not help with re-identification accuracy in this case. Our combined MCS descriptor performs slightly worse than some individual color space descriptors on average but has a notably better performance for re-identification within the same camera. The CNN feature outperforms the covariance descriptors in all settings, except for re-identification within the aerial camera.

In [13] it has been shown that covariance descriptors calculated on different color spaces have complementary properties and can improve re-identification accuracies. Thus, we construct a combined descriptor, which contains the information from all six color spaces (MultiColorSpace, MCS). This descriptor performs slightly less accurate on average but has a notably higher accuracy for within-camera re-identification. Since this descriptor contains more information than any of the single-color-spaces descriptors and has a dimensionality of 210, which is comparable to that of our CNN feature (256), we use this descriptor for our feature fusion experiments.

The results in Table 3 show some notable common trends for the various camera settings. For all features the average re-identification accuracy within a ground camera is significantly higher than that within the aerial camera. This is likely due to the higher variety of viewing angles within this camera. Re-identification across the various ground cameras often achieves the highest accuracy compared to cross camera settings where the aerial camera is involved. For cross camera re-identification including the aerial view, a notably higher performance is achieved, if the probe images originate from the aerial camera. This can be attributed to the higher quality of probe images available from that camera compared to some of the lower resolution and color distorted ground cameras.

We also performed the covariance feature experiments listed in Table 3 using euclidean or cosine distance instead of the more accurate geodesic distance described in Section 3.1. While the geodesic distance (Equation 4) achieved an average mAP of 28.56% across all covariance descriptors, the euclidean and cosine distances result in average mAP values of 27.56% and 28.47%, respectively. Thus, for our feature fusion experiments we apply cosine distance, because it performs nearly as well as the geodesic distance and is also well suited for comparing CNN features.

4.2.2 CNN Features

The CNN features generated by the network described in Section 3.2 show similar trends as the region covariance features. Again, the average re-identification accuracy within a ground camera is significantly higher than that within the aerial camera. An additional reason for this trend is likely the significant difference in angles between the networks's original training data and the aerial camera. The re-identification across ground cameras, i.e. the setting closest to the deep feature's training data, achieves highest cross-camera accuracy. Finally, the aerial re-identification is again more reliable for the aerial probe setting. The overall performance of the trained deep features is, as expected, better than that of the hand-crafted region covariance features. However, re-identification accuracy within the aerial camera is significantly lower than that of the MCS descriptor. Our feature fusion aims at combining these complementary strengths of the two features.

Feature	All	Air	Ground	Gr. → Gr.	Air → Gr.	Gr. → Air
MCS+CNN feat. fusion (cosine)	44.5	40.6	70.6	**43.5**	34.1	**32.9**
MCS+CNN feat. fusion (eucl.)	44.1	40.6	70.4	42.9	34.1	32.1
MCS+CNN score fusion	31.5	43.2	61.1	23.7	29.3	24.2
MCS+CNN score fusion (unit)	43.5	**44.6**	**72.5**	40.4	**35.6**	31.5
MCS+CNN score fusion (minmax)	**44.8**	43.0	**72.5**	43.3	34.5	32.7
MCS+CNN score fusion (sigmoid)	43.8	43.6	70.5	42.4	33.4	31.8
MCS+CNN decision fusion	39.1	42.6	68.0	34.4	33.3	29.5

Table 4: mAP results of conventional fusion methods applied to our CNN and MCS features. Decision fusion clearly performs least accurate. The best results are achieved by score fusion with a prior min-max normalization of the score matrices.

Feature	All	Air	Ground	Gr. → Gr.	Air → Gr.	Gr. → Air
MCS+CNN deep-add	45.8	43.0	71.4	43.7	35.5	37.7
MCS+CNN deep-concat	**47.8**	45.2	**72.5**	**46.6**	35.8	**39.0**
MCS+CNN deep-fc	44.7	**45.6**	66.5	42.2	**38.3**	36.9

Table 5: Deep fusion results (mAP). Including the MCS descriptor into the training process of the CNN can significantly boost accuracy. The highest improvement can be gained by concatenating the CNN feature with the MCS at the last layer.

4.3 Conventional Feature Fusion

We first investigate the conventional fusion methods outlined in Section 3.3 for the combination of the MCS and CNN features. The results can be seen in Table 4. After feature fusion (i.e. concatenation), we apply either cosine or euclidean distance to rank our galleries. Again, we observe that cosine distance performs better. This further confirms our earlier observation in Section 4.2.1.

Direct score fusion performs comparably poorly, because the feature values, and thus their distances, of the two features have significantly different value ranges. The results are thus dominated by the MCS descriptor and can only be slightly improved by the CNN features. However, if we normalize both feature's distances before fusion, a significant boost in accuracy can be achieved. We applied normalization to unit length (unit), min-max normalization (minmax) or normalization by sigmoid function to the distance matrices. All normalization methods achieve a similar increase in accuracy. The highest accuracy among all fusion methods is achieved by a combination of score fusion and min-max normalization. The overall mAP of 44.8% is a 1.2% improvement over that of the CNN feature. The accuracy within the aerial camera is notably improved by 5.8% compared to the CNN feature and even outperforms that of the MCS feature by 0.8%. In all cross-camera settings the accuracy of the fused feature outperforms that of any of the individual features.

During decision fusion the more nuanced information of individual scores is lost. It is thus not surprising that decision fusion performs with the least accuracy among the three methods.

4.4 Deep Fusion

In Table 5 we show the re-identification accuracies achieved by our proposed deep fusion methods (see Section 3.4). Of necessity, the dimension of the CNN feature layer has to be adapted to 210 in the case of the deep-add architecture. In the two other deep fusion cases we keep it at 256 dimensions in order to match the original CNN architecture. We chose the additional fully-connected layer in the deep-fc to have 256 dimensions in order to keep the increase in network parameters low.

All methods clearly outperform any of the conventional fusion methods and achieve a significant boost in accuracy compared to the original CNN features. This shows that the inclusion of the MCS features in the training process significantly aids the CNN. Particularly the accuracy on the aerial camera, which was the weak point of the CNN features, can be well compensated by all methods. But also in the cross-camera settings the resulting accuracy is notably improved compared to the individual features. Interestingly, the best performance

is achieved by the concatenation architecture. The deep-fc architecture performs far less accurate, even though it contains additional parameters. The likely reason for this difference is the increased dimensionality of the resulting features in the case of deep-concat (i.e. 210+256).

Overall, we were able to significantly improve on the accuracy of our best individual feature (43.6%) by 4.2% through our proposed deep fusion. Analysis of the different evaluation settings shows that we gain the most accuracy in those settings that involve the aerial camera. This result clearly outperforms our earlier work [24] which was conducted solely on the *Air* setting and achieved a mAP of 42.3%.

5. CONCLUSION

We presented three simple, yet effective methods of leveraging information from hand-crafted features in the training process of a CNN. We were able to show that this fusion of the information learned by the CNN and that contained in the hand-crafted feature significantly improved re-identification accuracy on our own dataset. Our analysis of this improvement of overall 4.2% mAP shows that the deep fusion approach is able to compensate for the weaknesses of the individual features and to leverage their complementary information. We could also show that any of the deep fusion approaches outperform any conventional feature fusion method for person re-identification.

REFERENCES

[1] Cheng, D., Gong, Y., Zhou, S., Wang, J., and Zheng, N., "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (2016).

[2] Xiao, T., Li, H., Ouyang, W., and Wang, X., "Learning deep feature representations with domain guided dropout for person re-identification," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (2016).

[3] Tuzel, O., Porikli, F., and Meer, P., "Region covariance: A fast descriptor for detection and classification," in [*Proceedings of the 9th European Conference on Computer Vision (ECCV)*], (2006).

[4] Bak, S. and Brémond, F., "Re-identification by covariance descriptors," *Part of the series Advances in Computer Vision and Pattern Recognition* (2014).

[5] Li, W., Zhao, R., Xiao, T., and Wang, X., "Deepreid: Deep filter pairing neural network for person re-identification," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (2014).

[6] Ahmed, E., Jones, M., and Marks, T. K., "An improved deep learning architecture for person re-identification," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (2015).

[7] Porikli, F., Tuzel, O., , and Meer, P., "Covariance tracking using model update based on means on riemannian manifolds," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (2006).

[8] Metzler, J. and Willersinn, D., "Robust tracking of people in crowds with covariance descriptors," in [*Proceedings of SPIE 7341, Visual Information Processing XVIII*], (2009).

[9] Hübner, Y., Metzler, J., Dürr, B., Jäger, U., and Willersinn, D., "Assessment and optimization of methods for tracking people in riot control scenarios," in [*Proceedings of SPIE 7114, Electro-Optical Remote Sensing, Photonic Technologies, and Applications II*], (2008).

[10] Bak, S., Corvee, E., Brémond, F., and M., T., "Boosted human reidentification using riemannian manifolds," *Journal of Image and Vision Computing* **30(6-7)** (2011).

[11] Metzler, J., "Two-stage appearance-based re-identification of humans in low-resolution videos," in [*Proceedings of IEEE International Workshop on Information Forensics and Security (WIFS)*], (2012).

[12] Serra, G., Grana, C., Manfredi, M., and Cucchiara, R., "Gold: Gaussians of local descriptors for image representation," in [*Computer Vision and Image Understanding*], **134** (2015).

(a) Air-to-ground re-identification.

(b) Air-to-ground re-identification.

(c) Ground-to-air re-identification.

Figure 5: Qualitative results of our approach (top 15 results for each query). The figures show responses of our MSC feature (each first row), the CNN feature (each second row) and our final fused result using deep-concat (each third row) to the same query. In cases (a) and (b), the fusion approach successfully leverages the information from the stronger feature. In case (c) both features result in many errors but the fused result can eliminate most of them.

[13] Matsukawa, T., Okabe, T., Suzuki, E., and Sato, Y., "Hierarchical gaussian descriptor for person re-identification," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (2016).

[14] Liu, C., Gong, S., Loy, C. C., and Lin, X., "Person re-identification: What features are important?," in [*Proceedings of the European Conference on Computer Vision (ECCV)*], (2012).

[15] Zheng, L., Wang, S., Tian, L., He, F., Liu, Z., and Tian, Q., "Query-adaptive late fusion for image search and person re-identification," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (2015).

[16] Kawai, R., Makihara, Y., Hua, C., Iwama, H., and Yagi, Y., "Person re-identification using view-dependent score-level fusion of gait and color features," in [*Proceedings of the International Conference on Pattern Recognition (ICPR)*], (2012).

[17] Eisenbach, M., Kolarow, A., Vorndran, A., Niebling, J., and Gross, H.-M., "Evaluation of multi feature fusion at score-level for appearance-based person re-identification," in [*Proceedings of the Interational Joint Conference on Neural Networks (IJCNN)*], (2015).

[18] Layne, R., Hospedales, T. M., Gong, S., and Mary, Q., "Person re-identification by attributes.," in [*Proceedings of the British Machine Vision Conference (BMVC)*], (2012).

[19] Schumann, A. and Stiefelhagen, R., "Transferring attributes for person re-identification," in [*Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*], (2015).

[20] Wu, S., Chen, Y.-C., Li, X., Wu, A.-C., You, J.-J., and Zheng, W.-S., "An enhanced deep feature representation for person re-identification," in [*Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*], (2016).

[21] Oreifej, O., Mehran, R., and Shah, M., "Human identity recognition in aerial images," in [*Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR)*], (2010).

[22] Layne, R., Hospedales, T. M., and Gong, S., "Investigating open-world person re-identification using a drone," in [*Proceedings of the European Conference on Computer Vision (ECCV)*], (2014).

[23] Schumann, A. and Schuchert, T., "Person re-identification in uav videos using relevance feedback," in [*Proceedings of SPIE IS&T Electronic Imaging*], (2015).

[24] Schumann, A. and Schuchert, T., "Deep person re-identification in aerial images," in [*Proceedings of SPIE, Security+Defence*], (2016).

[25] Pennec, X., "Intrinsic statistics on riemannian manifolds - basic tools for geometric measurements," *Mathematical Imaging and Vision* **25** (2006).

[26] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., "Going deeper with convolutions," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (2015).

[27] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., "Rethinking the inception architecture for computer vision," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (2016).

[28] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q., "Scalable person re-identification: A benchmark," in [*Proceedings of the IEEE International Conference on Computer Vision (CVPR)*], (2015).

[29] Ioffe, S. and Szegedy, C., "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in [*Proceedings of the 32nd International Conference on Machine Learning (ICML)*], (2015).

[30] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (2016).