TU Delft

Delft University of Technology

RGB-Depth cross-modal person re-identification

Hafner, Frank M.; Bhuiyan, Amran; Kooij, Julian F.P.; Granger, Eric

DOI
10.1109/AVSS.2019.8909838

Publication date
2019

Document Version
Final published version

Published in
Proceedings of the 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS 2019)

Citation (APA)
Hafner, F. M., Bhuiyan, A., Kooij, J. F. P., & Granger, E. (2019). RGB-Depth cross-modal person re-identification. In *Proceedings of the 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS 2019)* IEEE . https://doi.org/10.1109/AVSS.2019.8909838

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

https://www.openaccess.nl/en/you-share-we-take-care

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.

RGB-Depth Cross-Modal Person Re-identification

Frank M. Hafner[a], Amran Bhuiyan[b], Julian F. P. Kooij[c], Eric Granger[b]

[a] *ZF Friedrichshafen AG, Friedrichshafen, Germany*

[b] *Laboratoire d'imagerie, de vision et d'intelligence artificielle, École de technologie supérieure, Montreal, Canada*

[c] *Intelligent Vehicles Group, Delft University of Technology, Netherlands*

frank.hafner@zf.com, amran.apece@gmail.com, j.f.p.kooij@tudelft.nl, eric.granger@etsmtl.ca

Abstract

Person re-identification is a key challenge for surveillance across multiple sensors. Prompted by the advent of powerful deep learning models for visual recognition, and inexpensive RGBD cameras and sensor-rich mobile robotic platforms, e.g. self-driving vehicles, we investigate the relatively unexplored problem of cross-modal re-identification of persons between RGB (color) and depth images. The considerable divergence in data distributions across different sensor modalities introduces additional challenges to the typical difficulties like distinct viewpoints, occlusions, and pose and illumination variation. While some work has investigated re-identification across RGB and infrared, we take inspiration from successes in transfer learning from RGB to depth in object detection tasks. Our main contribution is a novel cross-modal distillation network for robust person re-identification, which learns a shared feature representation space of person's appearance in both RGB and depth images. The proposed network was compared to conventional and deep learning approaches proposed for other cross-domain re-identification tasks. Results obtained on the public BIWI and RobotPKU datasets indicate that the proposed method can significantly outperform the state-of-the-art approaches by up to 10.5% mAP, demonstrating the benefit of the proposed distillation paradigm.

1. Introduction

Person re-identification is an important function in many monitoring and surveillance applications, such as multi-camera target tracking, pedestrian tracking in autonomous driving, access control in biometrics, search and retrieval in video surveillance, and forensics [5, 8], and, as such, has gained much attention in recent years. Given the query image of an individual captured using a network of distributed cameras, person re-identification seeks to recognize that same individual over time within a gallery of previously-captured images [6]. This task remains a challenging problem in real world applications due to low resolution images, occlusions, miss-alignments, background clutter, motion blur, and variations in pose, scale and illumination.

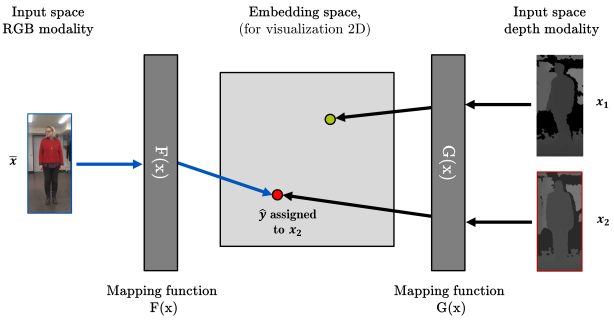


Figure 1. Cross-modal re-identification creates a shared embedding for multiple modalities, each with their own mapping function. Here, the embedding functions are defined as $F(x)$ for RGB and $G(x)$ for depth, respectively.

This paper focuses on deep neural networks for cross-modal person re-identification that allow sensing between RGB and depth modalities. Although some methods have been proposed for cross-modal re-identification between RGB and infrared images [10, 11, 12, 13], almost no research addressing RGB and depth images exists [16, 17]. However, sensing across RGB and depth modalities is important in many real-world scenarios. This is the case, for example, with video surveillance systems that must recognize individuals in poorly illuminated environments [14]. Another use case are autonomous self-driving vehicles, which require tracking pedestrians around their vicinity, where some regions are covered by lidar sensors, and others by RGB cameras. Besides these practical applications, research in cross-modal re-identification can also help legal interpretation of depth-based images concerning privacy data protection (e.g. within GDPR). While it is clear that person data from a RGB camera is highly sensible concerning data privacy, it is still unclear how much private

Copyright Notice: 978-1-5386-9294-3/18/$31.00 ©2019 IEEE

Authorized licensed use limited to: TU Delft Library. Downloaded on February 02,2022 at 13:36:09 UTC from IEEE Xplore. Restrictions apply.

information can be extracted from depth images.

In this paper, a new cross-modal distillation network is proposed for robust person re-identification across RGB and depth sensors. The task is addressed by creating a common embedding of images from both the depth and RGB modalities, as visualized in Figure‘1. The proposed method exploits a two-step optimization process. In the first step a network is optimized based on data from the first modality, and then, in the second step, the embeddings and weights of this first neural network provide guidance to optimize a second network for the other modality. The optimization is based on the final embedding layer of the networks to guarantee an embedding in a common feature space for both modalities. The idea behind this approach is to enable a transfer of learned structural representations from the depth modality to the RGB modality, and, therefore, enforce similar feature embeddings for the modalities.

This paper presents the following contributions: *(i)* A cross-modal deep neural network is adopted to transfer an embedding representation from one modality to the other by exploiting the intrinsic relation between depth and RGB. *(ii)* We intuitively and experimentally show that an ideal deep feature distillation for the task needs to take place from depth to RGB. *(iii)* An extensive experimental validation is conducted to show the performance of state-of-the-art methods on cross-modal person re-identification between RGB and depth. On this basis the advantages of the proposed method are shown on multiple RGB-D based benchmark re-identification datasets.

2. Related Work

The area of person re-identification has received much attention in recent years [8]. This section provides a summary of the state-of-the-art conventional, deep learning and cross-modal techniques as they relate to our research.

Conventional Methods. Conventional approaches for person re-identification from a single modality can be categorized into two main groups – direct methods with handcrafted descriptors or learned features and metric learning based approach. Direct methods for re-identification are mainly devoted to the search of the most discriminant features to design a powerful descriptor (or signature) for each individual regardless of the scene [20, 22, 21]. In contrast, in metric learning methods, a dataset of different labeled individuals is used to jointly learn the features and the metric space to compare them, in order to guarantee a high re-identification rate [22].

Deep Learning Methods. The idea of using a deep learning architecture for person re-identification stems from Siamese CNN with either two or three branches for pairwise verification loss [25] or triplet loss [26, 27] respectively, by proposing new layers [1] or by fusing features from different body parts with a multi-scale CNN structure [2, 3]. Another trend of using deep learning architecture is *transfer learning* [4, 25, 29], for when the distribution of the training data from the source domain is different from that of the target domain. The most common deep transfer learning strategy for re-identification [4] is to pre-train a base network on a large scale or combination of different datasets as source dataset, and transfer learned representation to the target dataset. However, these transfer learning methods depend on the assumption that the tasks are the same and in a single modality and, hence unsuitable when the source and target domains are heterogeneous.

Cross-Modal Methods. While the progress in re-identifying persons in single modalities was significant, only few works [10, 11, 12, 13, 15, 16] investigated the task of cross-modal person re-identification. Recently, several models have been proposed for cross-modal person re-identification between RGB and infrared images [10, 11, 12, 13]. To embed the RGB and IR modalities in a common feature space, the authors in [10, 11, 12] analyze several neural networks architectures: zero padding and one-stream networks [10], and two streams network [11, 12] with different losses. Additionally, problem has been addressed in adversarial way in [13]. There are a few works in the literature that consider a *multi-modal* person re-identification scenario [9, 15] by fusing the RGB and the depth information in order to extract robust discriminative features. In [15], a depth-shape descriptor called eigen-depth is proposed to extract describing features from the depth domain. [16] used the same features to perform cross-modal re-identification between depth and RGB. To the best of our knowledge this is the only existing work on the topic of cross-modal person re-identification between RGB and depth.

In contrast to the approaches described above for cross-modal re-identification, we propose to employ the cross-modal distillation idea by means of a deep transfer learning technique. The idea of the method is inspired by the work on supervision transfer of Gupta et al. [7]. However, supervision transfer [7] and our approach aim at different problems with different focuses of method design: supervision transfer solves the problem of limited data availability for object detection problems with a transfer scheme from RGB to depth. Our method is using the distillation paradigm to transfer knowledge from one modality to a second modality to solve the re-identification task across the two modalities. Therefore, contrary to Gupta et al. [7], the task has to be solved across modalities in the same feature space and is not considered a pre-training procedure as in [7]. Additionally, in Gupta et al. the direction of transfer is defined as from RGB to depth. In contrast, in this work the ideal direction of transfer is investigated in detail and a transfer from depth to RGB is shown to be superior for the application.

Authorized licensed use limited to: TU Delft Library. Downloaded on February 02,2022 at 13:36:09 UTC from IEEE Xplore. Restrictions apply.

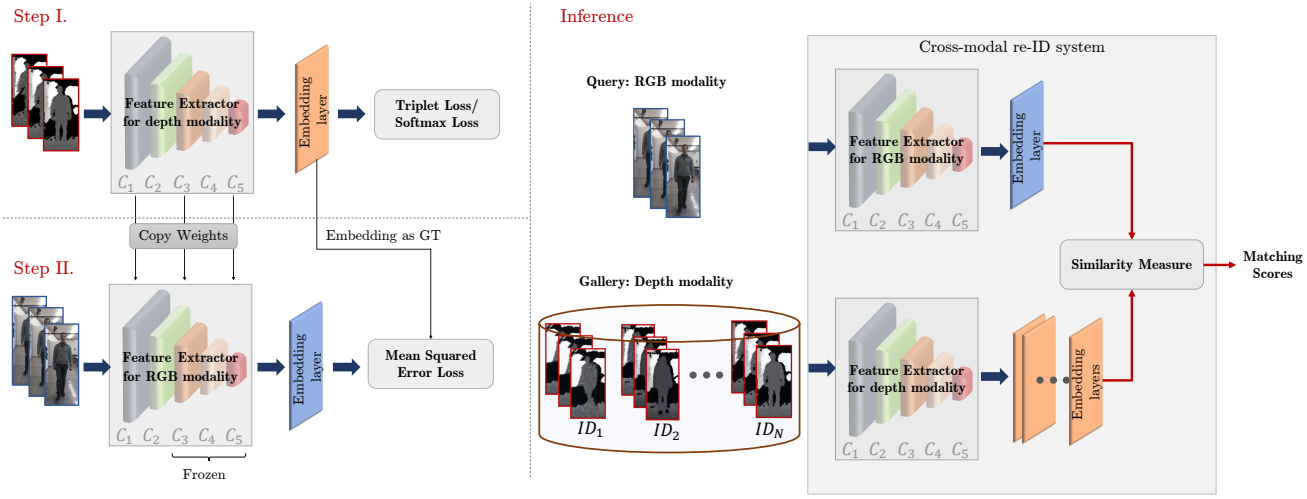


Figure 2. Two step training scheme and inference for the proposed cross-modal distillation network. Step I involves training of a CNN for single-modal re-identification. In step II, the knowledge from the first modality is transferred to the second one. During inference, query and gallery images from different modalities are evaluated to produce feature embeddings and matching scores for cross-modal re-identification. As an example, this figure shows a transfer from depth to RGB, and inference using RGB as query and depth as gallery. The modalities can be interchanged in both cases.

3. Deep Cross-Modal Neural Networks

In this section successful deep neural networks and objective functions for single-modal person re-identification are presented. Based on these key ingredients the new cross-modal distillation network is introduced.

3.1. Methods for Single-Modal Re-Identification: The task of single-modal re-identification is the standard problem within re-identification. For this work a selection of successful feature extraction networks and loss functions will be employed. For feature extraction, our work uses residual neural networks with 50 layers (ResNet50) [24] which are pre-trained on ImageNet. The ResNet architecture was shown to be effective for several person re-identification applications [28, 27]. Furthermore, we consider two possible loss functions, triplet loss and softmax loss, which both have been successfully applied in single-modal person re-identification [26, 8]. We will now shortly discuss both losses in more detail.

Using the *triplet loss* results in a metric learning approach which directly optimizes an embedding layer in euclidean space. During training, this loss compares the relative distances of three training samples, namely an anchor image x_a, a positive image sample x_p from the same individual as x_a, and a negative sample x_n from a different individual. Given an anchor image x_a, this loss assures that the embedding of an image taken from the same class x_p is closer to the anchor's embedding than that of a negative image belonging to another class y_n by at least a margin m in distance metric d. In the following, $F(x)$ denotes the mapping to the embedded space, which is in our case a deep neural network. The triplet loss is therefore defined as:

$$L_{tri} = \sum_{i=1}^{T} \left[d(F(x_{a(i)}), F(x_{p(i)})) - d(F(x_{a(i)}), F(x_{n(i)})) + m \right] . \tag{1}$$

where indices $a(i)$, $p(i)$ and $n(i)$ stand for anchor, positive and negative, of the i-th triplet, and T for the number of triplets used per batch.

For the second considered loss, the *softmax loss*, the embedding is learned indirectly by first treating re-identification on the training set as a classification problem, where all individuals in the training set are considered a different class. Afterwards, the layer of the neural network prior to the softmax loss is used as the embedding. This enables that the network can be applied on test data, which can contain new individuals not present in the training data. Therefore, the softmax loss to optimize the embedding can be expressed as:

$$L_{soft} = -\frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{e^{W_{(y_i)} F(x_i) + b}}{\sum_{j=1}^{C} e^{W_{(j)} F(x_i) + b}} \right) , \tag{2}$$

where N is the batch size, $W_{(j)}$ are the weights leading to the j-th node of the ultimate softmax layer of the network, b is a bias. The amount of classes is defined as C.

3.2 A Cross-Modal Distillation Network: This section explains our novel cross-modal training and inference. The training of the network is divided into two steps to exploit the relationship between depth and RGB and visualized in

Authorized licensed use limited to: TU Delft Library. Downloaded on February 02,2022 at 13:36:09 UTC from IEEE Xplore. Restrictions apply.

Figure 2. In step I of the training of the cross-modal distillation network, a neural network F is trained for sensing in a first modality. In this work the training will be done with Resnet50 and triplet and softmax loss as presented in section 3.1. The networks are optimized by means of an early-stopping criteria based on the mAP in the validation set. Afterwards, the network is frozen as F_{fr}, with corresponding weights $W_{F,fr.}$.

The obtained neural network feature extractor for the first modality is deployed as the baseline network for the training of a feature extractor for the second modality. For training step 2, a network with the same architecture as the corresponding network in step I is initialized. Similarly to [7], the weights of the converged model from step I, $W_{F,fr.}$, are copied to network G which is dedicated to the second modality. Additionally, the weights of the network are frozen from a mid-level convolutional layer up to the final feature embedding. This retains the high-level mapping from the first network. At the same time, the target embedding can still learn meaningful low-level features for the task in the target modality. With this approach, we restrict the learning in the second modality and, hence, force the network to learn to extract similar features in the second modality as in the first modality.

For the actual transfer of knowledge we make use of paired images X_{m1} from modality 1 and X_{m2} from modality 2. The aim is to optimize G in such a way that the embeddings of images from the second modality X_{m2} with label y are close to the embeddings of images from the first modality X_{m1} with the same label y. This is realized by exploiting image pairs $x_{m1,i}$ and $x_{m2,i}$ from the two modalities, which are considered coupled as they are taken at the exactly same time step. Hence, the embedding of $x_{m1,i}$ is obtained with a forward propagation through the frozen network $F_{fr.}$ and is taken as the groundtruth for the embedding of $x_{m2,i}$ with the, at this stage, trainable network G. Since during inference mode the embeddings will be compared based on Euclidean distance, we aim to minimize this metric between the two embeddings. Hence, we make use of the mean squared error (MSE) loss between the embeddings of paired images $F_{fr.}(x_{m1,i})$ and $G(x_{m2,i})$:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\|F_{fr.}(x_{m1,i}) - G(x_{m2,i})\|^2 \qquad (3)$$

where N is the batch size for training. Early-stopping criteria for network training is the loss in validation data.

During inference, the two resulting networks, $F_{fr.}$ and $G_{fr.}$, are evaluated in the corresponding modalities to provide feature embeddings for input images. Similarity between these representations is measured using Euclidean distance. No coupled images are needed for inference.

4. Experimental Methodology

In this section we present the experimental methodology used to validate the proposed approach. Therefore, two RGB-D person re-identification datasets will be presented. As these datasets were originally not designed for cross-modal person re-identification it is important to discuss their intrinsic properties.

The considered datasets are BIWI RGBD-ID [18] and RobotPKU [19] datasets. These datasets were selected because they provide high-resolution depth and RGB images, a decent amount of instances and a large amount of images per instance in different poses. Both datasets were recorded with a Microsoft Kinect camera. The BIWI RGBD-ID dataset targets long-term people re-identification from RGB-D cameras [18]. As in [16] same person with different clothing is considered as a separate instance. Overall, it is comprised of 78 individuals with 22,038 images in depth and RGB. RGB and depth images are provided coupled with no visible difference in capturing time. The dataset consists of 90 persons with 16,512 images in total. The images are provided in a coupled manner. Nevertheless, by visual inspection it is apparent, that there is a slight time difference, in the order of a fraction of a second, between the images captured in depth and RGB. For training and inference images of both datasets in both modalities are resized to 256×128.

For the performance evaluation with the BIWI dataset, the same partitions into training, validation and testing subsets were adopted as in [16], which means 32 individuals were chosen for training, 8 instances for validation and 38 individuals for testing. For the RobotPKU dataset, the division will be videos from 40 individuals for training, 10 for validation, and 40 for testing. This follows the division of [19]. For quantitative evaluation, the average rank 1, 5 and 10 accuracy performance measure is reported along with the mean average precision (mAP). For the reporting of the rank accuracy, a single-gallery shot setting is used, where a random selection of the gallery (G) images is repeated 10 times. For evaluation the exactly same corresponding image in the parallel modality is excluded. To obtain statistically reliable results a 3-fold cross-validation process is used.

5. Results and Discussion

An extensive series of experiments has been considered to validate the proposed cross-modal distillation network. In this section, the results for optimization with the single modalities (i.e., step I. in Fig 2) are first shown to establish a baseline for the individual modalities. Hence, we first investigate how different choices for losses affect the performance on single-modal re-identification, and compare the relative difficulty of the modalities and dataset. Then, the distillation step (step II.) of the proposed method is per-

Authorized licensed use limited to: TU Delft Library. Downloaded on February 02,2022 at 13:36:09 UTC from IEEE Xplore. Restrictions apply.

Table 1. Average test set accuracy of the proposed method (Step I) for different modalities on BIWI dataset.

Modality	Loss	rank-1 (%)	rank-5 (%)	rank-10 (%)	mAP (%)
RGB	Triplet	92.14 ± 1.86	99.71 ± 0.24	99.95 ± 0.08	93.44 ± 1.46
	Softmax	**94.75 ± 0.74**	**99.75 ± 0.19**	**99.96 ± 0.03**	**95.68 ± 0.60**
Depth	Triplet	54.23 ± 1.75	**91.48 ± 0.56**	**99.15 ± 0.18**	55.31 ± 1.71
	Softmax	**59.84 ± 0.66**	90.54 ± 0.81	97.80 ± 0.19	**61.44 ± 0.54**

Table 2. Average test set accuracy of the proposed method (Step I) for different modalities on RobotPKU dataset.

Modality	Loss	rank-1 (%)	rank-5 (%)	rank-10 (%)	mAP (%)
RGB	Triplet	**89.04 ± 3.91**	**99.17 ± 0.33**	**99.46 ± 0.10**	**90.63 ± 3.41**
	Softmax	84.52 ± 0.24	97.91 ± 0.35	99.12 ± 0.23	87.11 ± 0.22
Depth	Triplet	n/a	n/a	n/a	n/a
	Softmax	**44.50 ± 1.02**	**75.83 ± 1.29**	**87.56 ± 0.87**	**44.50 ± 1.02**

formed and evaluated (section 5.2). Here, the ideal direction of transfer is investigated. Finally, the state-of-the-art of the cross-modal person re-identification task between RGB and depth is defined (section 5.3).

5.1. Single-Modal Re-identification Performance: For performance evaluation with individual modalities (RGB and depth separately), the single-modal case, results have been obtained on BIWI and RobotPKU datasets. The representative feature extractor Resnet50 has been optimized with triplet loss, equation (1), and softmax loss, equation (2). For triplet loss an embedding size of 128 and a training batch of 64 with 16 instances á 4 images was used. Batch hard mining was chosen for triplet choice. These parameters were proposed by [26]. To enable a fair comparison also for softmax loss a embedding size of 128 will be chosen. The neural networks trained with softmax were optimized with stochastic gradient descent with Nesterov momentum. Those trained with triplet loss were optimized with the ADAM optimizer. The margin for triplet loss (see formula 1) was set to 0.5.

Table 1 shows the average accuracy of the networks for single-modal re-identification for individual (RGB and depth) modalities on BIWI data. Results show that the networks optimized using RGB modality alone, can reach a high level of accuracy. The best model, optimized with softmax loss provides an average mAP of 95.68%. The performance of networks optimized with triplet loss and softmax loss lead to comparable performance. As expected, the overall accuracy for the networks optimized using depth modality alone is much lower compared to the accuracy achieved for the same task with RGB. The highest accuracy (mAP = 61.44%) is achieved when optimized with triplet loss.

Table 2 shows the average accuracy for single-modal re-identification for individual (RGB and depth) modalities on RobotPKU data. Again, the RGB modality allows to achieve high level of accuracy. For instance, the network trained with triplet loss yields the highest level of accuracy (mAP of 90.63%). In the depth modality, the network trained with softmax loss achieves an average mAP of 44.50%. Due to the inherent complexity of the re-identification task in the depth images, the network trained with triplet loss did not converge to produce meaningful embedding layers. The effect can be explained by the higher level of noise in RobotPKU images in contrast to the BIWI dataset.

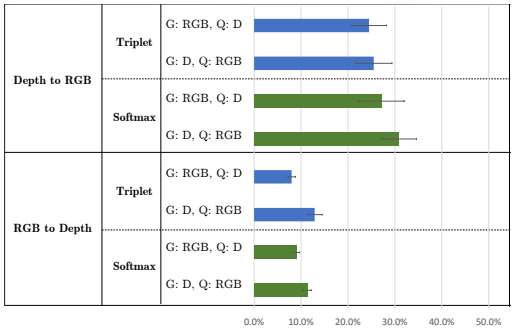


Figure 3. Average mAP accuracy of various cross-modal distillation networks on the BIWI dataset. For all combinations we report varying query (Q) and gallery (G) modalities. The first column indicates the direction of the transfer for the cross-modal distillation.

The difference in performance for sensing in RGB and depth in both datasets provides insight into the complexity of the individual tasks. Given results for both datasets, it is comparably easy to solely sense in RGB as visual cues like color features can be effectively exploited for the re-identification. In depth, color features are not present and the features based on a persons shape are less descriptive and lead to a lower accuracy. Nevertheless, it was also shown that in depth descriptive features can be extracted.

5.2. Performance for Cross-Modal Distillation: In this section the experiments with the cross-modal distillation method as presented in section 3 will be introduced. As baseline or step I the results from section 5 will be considered. In this section experiments are presented to gain insight on step II (distillation), and, in particular, on the advantages of transferring knowledge based on the depth or RGB modality.

Figure 3 presents the average mAP accuracy of the cross-modal distillation networks trained on the BIWI dataset in the cross-modal tasks with varying population of query and gallery between RGB and depth. The top two networks train the baseline network in depth (step I.), and then transfer to RGB (step II.). The bottom two networks train the baseline network in RGB (step I.), and then transfer to depth (step II.). The different colors indicate results with triplet (blue) and softmax (green) loss functions.

Results indicate that the accuracy obtained for when transferring from RGB to depth are significantly lower than from depth to RGB. Using depth images to populate a reference gallery, and RGB images as query achieves an mAP accuracy of about 31% using networks optimized with softmax loss. The best mAP accuracy for the same task and transferring from RGB to depth is about 13%. An expla-

Authorized licensed use limited to: TU Delft Library. Downloaded on February 02,2022 at 13:36:09 UTC from IEEE Xplore. Restrictions apply.

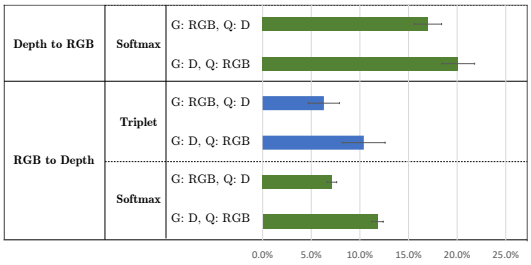


Figure 4. Average mAP accuracy of various cross-modal distillation networks on the RobotPKU dataset. For all combinations we report varying query (Q) and gallery (G) modalities. The first column indicates the direction of the transfer for the cross-modal distillation. As no baseline for depth with triplet was successfully trained (see table 2), no results reported.

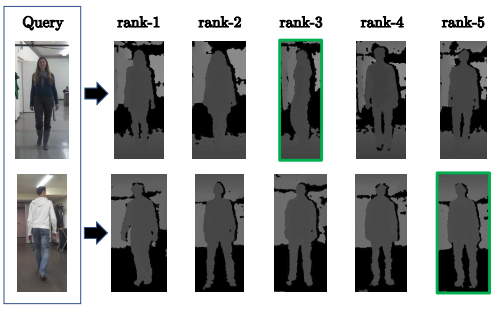


Figure 5. Example of qualitative results for the proposed architecture on BIWI dataset. The green box denotes the correct match. Gallery (G) and Query (Q) varied for the modalities.

nation for this behavior is that the general shape information of a person that is captured in depth data can, to a certain degree, be recovered in the RGB images. In contrast, the additional descriptive information which is inherent in RGB, like color information cannot be found in depth images. This will be further analyzed in section 5.

The performance obtained for models trained with the two losses is only slightly differing (see Table 1). The overall best performance is obtained with a baseline in networks trained with softmax loss with an average mAP of 30.1% with RGB as gallery (G) and depth (D) as query and 27.1% for depth as gallery and RGB as query.

Another finding is the significant difference in performance when alternating the modality used as gallery and query between RGB and depth. Our results suggest that a higher level of performance can be achieved in all networks when the gallery consists of RGB images. This is due to the fact, that if RGB images are in the gallery the probability of meaningful embeddings for the images is higher than for depth in gallery (see table 1 and 2). As the performance indicators are more influenced by meaningful embeddings in the gallery, we see this effect.

Figure 5 shows an example of results for the cross-modal distillation network on BIWI dataset, where the query image is RGB and the gallery image is depth. This figure highlights the complexity of the task, which is very difficult to solve for humans.

Figure 4 presents the average mAP accuracy of the cross-modal distillation networks trained on the RobotPKU dataset in the cross-modal tasks. The results on RobotPKU data mirror the findings from the BIWI dataset. Again, the transfer from depth to RGB significantly outperforms the transfer from RGB to depth. The difference of the best networks in mAP is 11%/7.5% for varying query and gallery population. The best overall network is Resnet50 trained with softmax and a transfer from depth to RGB.

In summary, to obtain the better results with the cross-modal distillation network, the transfer of knowledge should occur from depth to RGB. As shown in section 5 (tables 1 and 2) in the single-modal task a much higher performance was obtained in the RGB modality. Hence, the performance in the single-modal task of the baseline network is not critical to performance for cross-modal distillation. Results suggest that the success of the distillation step is more dependent on the features learned from the modalities. Hence, the features learned in the depth modality were transferable to the RGB modality, while features learned in the RGB modality where not transferable to the depth modality. This gives an indication on the relation between the depth and RGB modality where depth can, to a certain degree, be considered a subset of RGB. Despite the indirect and direct nature of the loss functions, the results indicate that networks with a baseline trained with softmax loss and networks with a baseline network in triplet loss obtain similar results. This shows the robustness of the method itself.

5.3. Comparison with State-of-the-Art Methods: In this section the results from section 5 are taken into a broader scope and are compared to existing methods for cross-modal person re-identification. As deep learning based methods, one-stream network and zero-padding network as of [10] are analyzed. As conventional approaches the WHOS feature extractor [30] and the LOMO feature extractor [22] will be investigated. The same features will be extracted for both modalities and are are compared on basis of Euclidean distance and the additional metric learning step Cross-view Quadratic Discriminant Analysis (XQDA). Additionally, the matching of Eigen-depth and HOG/SLTP features as reported by [16] is included in Table 3 for the BIWI dataset.

Table 3 presents the average accuracy of state-of-the-art and proposed networks for different scenarios on the BIWI dataset. The deep learning based one-stream architecture is outperforming all methods based on hand-crafted features by at least 4%/7% for varying query and gallery in mAP accuracy with a Resnet50 structure. The cross-modal distillation network enables an additional improvement compared to the one-stream network by 11%/7%.

In table 4 the results for the RobotPKU dataset are

Authorized licensed use limited to: TU Delft Library. Downloaded on February 02,2022 at 13:36:09 UTC from IEEE Xplore. Restrictions apply.

Table 3. Average accuracy of state-of-the-art and proposed networks for different scenarios on the BIWI dataset. For results from [16] no detailed information on the evaluation procedure was given.

Approach	Query-RGB, Gallery-Depth				Query-Depth, Gallery-RGB			
	rank-1 (%)	rank-5 (%)	rank-10 (%)	mAP (%)	rank-1 (%)	rank-5 (%)	rank-10 (%)	mAP (%)
WHOS, Euclidean[30]	3.2	16.6	31.5	3.7	5.1	18.7	32.6	5.6
WHOS, XQDA[30]	8.4	31.7	50.2	7.9	11.6	34.1	51.4	12.1
LOMO, Euclidean[22]	2.8	16.4	32.5	4.8	3.3	15.6	29.8	5.6
LOMO, XQDA[22]	13.7	43.2	61.7	12.9	16.3	44.8	62.8	15.9
Eigen-depth HOG/SLTP, CCA [16]	8.4	26.3	41.6	-	6.6	27.6	45.0	-
Eigen-depth HOG/SLTP, LSSCDL [16]	9.5	27.1	46.1	-	7.4	29.5	50.3	-
Eigen-depth HOG/SLTP, Corr. Dict. [16]	12.1	28.4	44.5	-	11.3	30.3	48.2	-
Zero-padding network[10]	5.86 ± 2.18	25.85 ± 6.35	47.13 ± 8.06	7.28 ± 4.03	10.34 ± 2.68	38.91 ± 6.45	62.84 ± 11.48	9.77 ± 3.80
One-stream network[10]	15.68 ± 0.77	50.29 ± 1.18	75.65 ± 0.46	16.86 ± 0.87	19.82 ± 0.33	55.74 ± 0.83	78.92 ± 1.07	23.75 ± 0.30
Cross-modal distillation network (ours)	$\mathbf{26.85 \pm 1.75}$	$\mathbf{65.88 \pm 2.32}$	$\mathbf{84.13 \pm 3.14}$	$\mathbf{27.34 \pm 1.66}$	$\mathbf{29.23 \pm 2.31}$	$\mathbf{70.50 \pm 2.29}$	$\mathbf{88.13 \pm 0.91}$	$\mathbf{30.48 \pm 1.99}$

Table 4. Average accuracy of state-of-the-art and proposed architecture for different scenarios on the RobotPKU dataset.

Approach	Query-RGB, Gallery-Depth				Query-Depth, Gallery-RGB			
	rank-1 (%)	rank-5 (%)	rank-10 (%)	mAP (%)	rank-1 (%)	rank-5 (%)	rank-10 (%)	mAP (%)
WHOS, Euclidean[30]	3.8	16.3	29.5	3.9	3.5	16.1	31.2	5.4
WHOS, XQDA[30]	10.0	31.8	49.8	8.2	9.8	31.0	48.0	9.8
LOMO, Euclidean[22]	3.6	15.0	28.0	3.9	3.7	15.3	28.7	4.9
LOMO, XQDA[22]	12.9	36.4	56.1	10.1	12.3	37.4	56.1	12.3
Zero-padding network[10]	7.76 ± 0.85	29.04 ± 2.57	47.79 ± 3.34	7.67 ± 0.59	6.57 ± 0.64	26.80 ± 2.14	45.62 ± 2.78	8.31 ± 0.56
One-stream network[10]	11.92 ± 0.63	38.13 ± 1.01	57.34 ± 2.14	11.42 ± 0.52	12.48 ± 1.01	38.51 ± 1.51	56.77 ± 0.85	14.19 ± 1.37
Cross-modal distillation network (ours)	$\mathbf{17.50 \pm 2.23}$	$\mathbf{51.92 \pm 3.62}$	$\mathbf{72.73 \pm 3.21}$	$\mathbf{17.10 \pm 1.85}$	$\mathbf{19.45 \pm 2.01}$	$\mathbf{54.28 \pm 3.08}$	$\mathbf{74.44 \pm 2.27}$	$\mathbf{19.82 \pm 2.10}$

shown. The one-stream network with Resnet50 structure again outperforms the conventional feature extractors in average mAP. The performance increase of the cross-modal distillation network above that of the one-stream network in mAP is at 5%/5%. Overall results show that the cross-modal distillation network can significantly improve accuracy compared to state-of-the art methods for both BIWI and RobotPKU datasets.

5.4. Analysis of Neural Network Activations: In this section an explanation of the superior performance of the distillation network will be given, by analyzing deconvolution images of relevant deep learning methods. Figure 6 shows deconvolution images for different networks on two images from RGB (a. and c.) and depth (b. and d.) from the BIWI RGBD-ID dataset. The guided backpropagation algorithm was used for visualization of the activations for the networks [23]. The architectures which are shown are separate training for the single-modality task (as in section 3.1), the one-stream network, as the second in the state-of-the-art table [10], and our cross-modal distillation method.

The images show that the activations for the different networks are varying considerably. When optimized for the single modalities, the networks in the RGB modality are activated by features inside the torso region of a person, like the color of the same. The network sensing in the depth modality is activated by the outer structure of the torso. For the one-stream network in the RGB modality the network is mostly activated by colors of torso and upper legs, while in the depth modality a cluttered outer structure of the torso is captured. For the RGB modality in the cross-modal distillation network a very different activation map can be observed (images (a) and (c)). Instead of being activated by color features, we see that the network is mostly activated the structure of the torso for those images. Therefore, the knowledge from depth, which is a descriptiveness of the problem with structural details, was transferred to the RGB modality. This finding underlines that the transfer of knowledge between the modalities was successful. As the describing features for the images are similar, the task of embedding to a common feature space is facilitated. This explains the better performance in cross-modal person re-identification as found in section 5.3.

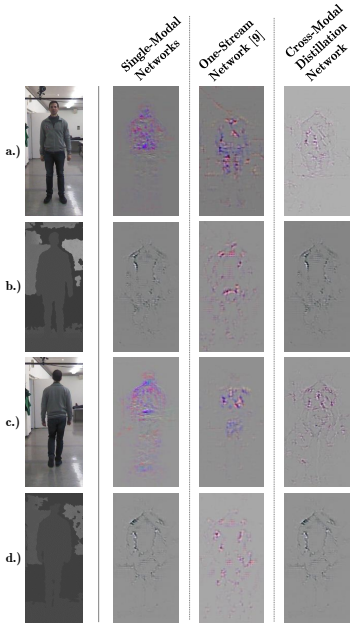


Figure 6. Comparison of deconvolution images for different networks on BIWI data. Visualization is performed with guided backpropagation [23]. Activation maps of cross-modal distillation network in RGB highly differing to the other techniques.

Authorized licensed use limited to: TU Delft Library. Downloaded on February 02,2022 at 13:36:09 UTC from IEEE Xplore. Restrictions apply.

6. Conclusions

In this paper, a new deep neural networks is proposed for cross-modal person re-identification that allow sensing between RGB and depth modalities. Its two-step approach enables the network to exploit the relation between these two relevant modalities, and thereby provide a high level of performance. Experimental results on two benchmark public datasets indicate that our proposed network can outperform related state-of-the-art methods for cross-modal re-identification by up to 10.5% in mAP. Results also show that features which are descriptive in the depth modality can successfully be extracted in the RGB modality for person re-identification. This implies that information captured in depth is to some extent retrievable in the RGB modality. Following this, we were able to show that the depth modality can be seen as subset of the RGB modality.

References

[1] Ahmed, E., Jones, M., Marks, T.K. "An improved deep learning architecture for person re-identification". CVPR 2015. 2

[2] Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N. "Person re-identification by multi-channel parts-based CNN with improved triplet loss function". CVPR 2016. 2

[3] Li, D., Chen, X., Zhang, Z., Huang, K. "Learning deep context-aware features over body and latent parts for person re-identification". CVPR 2017. 2

[4] Geng, M., Wang, Y., Xiang, T., Tian, Y. "Deep transfer learning for person re-identification," CoRR abs/1611.05244, 2016. 2

[5] Gong, S., Cristani, M., Yan, S., Loy, C. C. "Person re-identification". Springer Science & Business Media, 2014. 1

[6] Shoubiao T., Feng Z., Li L., Jungong H., Ling S., "Dense Invariant Feature-Based Support Vector Ranking for Cross-Camera Person Reidentification". TCSVT, 2018. 1

[7] Gupta, S., Judy H., Malik, J. "Cross modal distillation for supervision transfer". CVPR 2016. 2, 4

[8] Zheng, L., Yi, Y., Alexander G. H. "Person re-identification: Past, present and future," ArXiv:1610.02984 2016. 1, 2, 3

[9] Enzweiler, M., Gavrila, D. M. "A Multi-Level Mixture-of-Experts Framework for Pedestrian Classification". TIP, 2011. 2

[10] Wu, A., Zheng, W. S., Yu, H. X., Gong, S., Lai, J. "RGB-infrared cross-modality person re-identification". ICCV 2017. 1, 2, 6, 7

[11] Ye, M., Lan, X., Li, J., Yuen, P. C. "Hierarchical Discriminative Learning for Visible Thermal Person Re-Identification". AAAI 2018. 1, 2

[12] Ye, M., Wang, Z., Lan, X., Yuen, P. C "Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking". IJCAI 2018. 1, 2

[13] Dai, P., Ji, R., Wang, H., Wu, Q., Huang, Y. "Cross-Modality Person Re-Identification with Generative Adversarial Training". IJCAI 2018. 1, 2

[14] Sudhakar, P., Anitha Sheela, K., Satyanarayana, M. "Imaging Lidar system for night vision and surveillance applications". ICACCS, 2017. 1

[15] Wu, A., Wei-Shi Z., Jian-Huang L. "Robust depth-based person re-identification". TIP, 2017. 2

[16] Zhuo, J., Zhu, J., Lai, J. and Xie, X. "Person Re-identification on Heterogeneous Camera Network". CCF CCCV, 2017. 1, 2, 4, 6, 7

[17] Hafner, F., Bhuiyan, A., Kooij, J. F. P., Granger, E. "A Cross-Modal Distillation Network for Person Re-identification in RGB-Depth". CoRR abs/1810.11641, 2018. 1

[18] Munaro, M., Fossati, A., Basso, A., Menegatti, E., Van Gool, L. "One-shot person re-identification with a consumer depth camera". Person Re-Identification, Springer, 2014. 4

[19] Liu, H., Liang H., Liqian M.. "Online RGB-D person re-identification based on metric model update". CAAI Trans. Intelligence Technology 2.1 (2017): 48-55 4

[20] Bhuiyan, A., Perina, A., Murino, V. "Person re-identification by discriminatively selecting parts and features". ECCV 2014. 2

[21] Panda, R., Bhuiyan, A, Murino, V., Roy-Chowdhury, A. K. "Unsupervised Adaptive Re-identification in Open World Dynamic Camera Networks". CVPR 2017. 2

[22] Liao, S., Hu, Y., Zhu, X., Li, S. Z. "Person re-identification by local maximal occurrence representation and metric learning". CVPR 2015. 2, 6, 7

[23] Springenberg, J. T., Dosovitskiy, A., Brox, T., Riedmiller, M. "Striving for simplicity: The all convolutional net". ICLR 2015. 7

[24] He, K., Zhang, X., Ren, S., Sun, J. "Deep residual learning for image recognition". CVPR 2016. 3

[25] Xiao, T., Li, H., Ouyang, W., Wang, X. "Learning deep feature representations with domain guided dropout for person re-identification". CVPR 2016. 2

[26] Hermans, A., Lucas B., Bastian L. "In defense of the triplet loss for person re-identification". CoRR abs/1703.07737, 2017. 2, 3, 5

[27] Ristani, E., Carlo T. "Features for Multi-Target Multi-Camera Tracking and Re-Identification". CoRR abs/1803.10859, 2018. 2, 3

[28] Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., Tian, Q. "Person Re-identification in the Wild". CVPR 2017. 3

[29] Li, Y. J., Yang, F. E., Liu, Y. C., Yeh, Y. Y., Du, X., Wang, Y. C. F. "Adaptation and Re-Identification Network: An Unsupervised Deep Transfer Learning Approach to Person Re-Identification". CoRR abs/1804.09347, 2018. 2

[30] Lisanti, G., Masi, I., Bagdanov, A. D., Del Bimbo, A., "Person re-identification by iterative re-weighted sparse ranking". TPAMI, 2015. 6, 7

Authorized licensed use limited to: TU Delft Library. Downloaded on February 02,2022 at 13:36:09 UTC from IEEE Xplore. Restrictions apply.