A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 25

Multimodal Neural Network for Overhead Person Re-identification

Aske R. Lejbølle[1], Kamal Nasrollahi[2], Benjamin Krogh[3], Thomas B. Moeslund[4]

Abstract: Person re-identification is a topic which has potential to be used for applications within forensics, flow analysis and queue monitoring. It is the process of matching persons across two or more camera views, most often by extracting colour and texture based hand-crafted features, to identify similar persons. Because of challenges regarding changes in lighting between views, occlusion or even privacy issues, more focus has turned to overhead and depth based camera solutions. Therefore, we have developed a system, based on a Convolutional Neural Network (CNN) which is trained using both depth and RGB modalities to provide a fused feature. By training on a locally collected dataset, we achieve a rank-1 accuracy of 74.69%, increased by 16.00% compared to using a single modality. Furthermore, tests on two similar publicly available benchmark datasets of TVPR and DPI-T show accuracies of 77.66% and 90.36%, respectively, outperforming state-of-the-art results by 3.60% and 5.20%, respectively.

Keywords: Multimodal; Person Re-identification; Convolutional Neural Networks; Feature Fusion

1 Introduction

Person re-identification (re-id) i.e. identifications of persons across two or more cameras, is a topic with increasing interest due to potential usage in forensics, analysis of pedestrian flow in urban areas or monitoring of queue times in, for example, an airport. Meanwhile, it is also a topic still in research due to challenges that include changes in lighting, view and pose between camera views. To cope with these challenges, focus often lies in extracting robust hand-crafted feature descriptors from each person that are matched between views. For this purpose, soft biometrics are considered, such as colour and texture of the clothing, either represented as histograms [Li15] or transformed to sparse descriptors [LSF15]. To further improve accuracy of correct matches, supervised learning algorithms are applied that learn to separate similar feature pairs from dissimilar ones [Ch16, ZXG16]. More recently, deep learning has drawn increasing interest from the research community with Convolution Neural Networks (CNN) outperforming hand-crafted feature descriptors, as they are able to learn more expressive features [AJM15, WCZ16].

Besides aforementioned challenges, privacy preservation is often related to person re-id as a potentially large amount of data needs to be stored. Other than representing images as

[1] Visual Analysis of People (VAP), Aalborg University, Rendsburggade 14, 9000 Aalborg, Denmark, asrl@create.aau.dk
[2] Visual Analysis of People (VAP), Aalborg University, Rendsburggade 14, 9000 Aalborg, Denmark, kn@create.aau.dk
[3] BLIP Systems A/S, Vester Hassing, Haekken 2, 9310 Vodskov, Denmark, bbk@blipsystems.com
[4] Visual Analysis of People (VAP), Aalborg University, Rendsburggade 14, 9000 Aalborg, Denmark, tbm@create.aau.dk

feature descriptors, camera placement can be considered as a means of privacy preservation. Most current benchmark datasets within re-id consider a frontal view [GT08, Li14] while only few consider an overhead view which has the advantage of reducing privacy issues and avoid occlusions between persons or objects and persons in the scene [HAFF16, Li17]. Furthermore, other modalities that are more anonymous can be used, for example depth, from which information is captured using either passive stereo, i.e. a stereo camera or active, for example, a Microsoft Kinect. From depth information, the height and width of the person can be extracted along with different body ratios [Ba12]. Instead of relying on a single modality, combining (fusing) different modalities have shown to improve performance in related applications such as object recognition [Ei15] and object segmentation [Ha16]. Such fusing can be done either at feature level (feature fusion), for example, by concatenation of respective feature descriptors or at decision/score level (late fusion) by fusing the output decisions/scores from different modalities [Ki98].

To consider challenges regarding changes across views and the advantages of fusing different modalities, we propose a novel framework for applying colour and depth (RGB-D) based re-id to images, captured with an overhead view. More specifically, we take advantage of the recent advances within deep learning and train a CNN using information from both RGB and depth modalities to improve accuracy compared to using either modality independently. To that extend, we collect a novel RGB-D based dataset in an uncontrolled environment from a stereo camera placed overhead to avoid occlusions and, at the same time, preserve privacy by not recording faces. Our dataset is collected to resemble real-life situations by having multiple persons within view, while current overhead datasets only consider a single person within view at a time. In summary, the main contributions of our work include:

- We train a CNN using RGB and depth modalities information and show that fusion of these improves accuracy.
- We collect and annotate a novel RGB-D and overhead based dataset which can be used to both evaluate re-id accuracy but also multi-target detection and tracking algorithms in RGB and depth domain.

2 Related Work

While re-id using hand-crafted colour and texture features or CNN's are widely studied, overhead re-id is rarely considered. In addition, only a limited number of articles suggest depth modality for this purpose.

Overhead re-id As most current re-id datasets are collected in outdoor scenes, a frontal view is typically considered. A few systems have been proposed for evaluating datasets with an overhead view [Ar08, AC12]. [AC12] proposes feature extraction using a Histogram of Oriented Gradients (HOG) algorithm combined with a linear Support Vector Machine (SVM) for classification while [Ar08] extracts features based on the colour and texture of the hair. While both datasets are recorded in an indoor environment, they only extract colour information.

Overhead RGB-D-based re-id More RGB-D based datasets for re-id are currently being proposed. While the first considered a frontal view [Mu14], the most recent consider an overhead view [HAFF16, Li17]. [HAFF16] collected a dataset in a hallway and applies a combined CNN and Long-Short-Term-Memory (LSTM) network using depth based image sequences to learn spatio-temporal representations of each person. Meanwhile, [Li17] extracts seven different depth features and two colour features that are feature fused by concatenation. While the former extracts only depth information, the latter considers only hand-crafted features from both modalities.
Multi-modal CNN While the work of [HAFF16] to our knowledge is the only previous proposed neural network using depth information for re-id, multi-modal CNN's have been proposed for related applications [Ei15, Ha16]. [Ei15] trains a CNN for object recognition using both colour and depth images by fusing respective features in late layers of the network to consider both modalities during training. To that extend, [Sa16] shows that feature fusion of colour and depth features in a CNN outperforms similar fusion scheme using other classification methods, such as SVM and Deep Belief Networks (DBN). Meanwhile, [Ha16] proposes a multi-modal encoder-decoder network for semantic segmentation by fusing outputs from each layer in an RGB and depth based encoder, respectively, before passing the output through an RGB-D based decoder. In this case, fusing is applied as an element-wise summation. To our knowledge, no multi-modal neural networks have previously been proposed for re-id. Although, [WCZ16] proposes a fusing scheme similar to that of [Ei15], but instead of fusing different modalities, complementary feature types are fused, i.e. CNN and hand-crafted features. To our knowledge, the system proposed in this paper, is the first to incorporate multiple modalities in a CNN to learn a multi-modal feature representation.

3 Methodology

As we desire to exploit both colour and depth information, along with the potential of CNN's, our aim is to use an architecture which jointly processes the two modalities, RGB and depth, simultaneously. For person re-id, such architecture has not previously been applied, although, in object recognition the work of [Ei15] shows an increase in accuracy compared to using a single modality.

We apply an architecture similar to that of [Ei15], having two CNN streams separately processing an input image while being fused in a later fully connected layer, as shown in Figure 1. The structure of each separate CNN follows the AlexNet architecture (please see [KSH12] for details) and consists of five convolution layers, the first, second and fifth followed by a max pooling and normalization layer. The outputs from the last convolution layers are followed by two fully connected layers, transforming the feature maps to sparse representations for RGB and depth, respectively. The feature representations ($fc7^{RGB}$ and $fc7^{D}$) are concatenated and used as input to a fully connected layer (fc8) which learns a joint RGB-D feature representations based on both colour and depth images. Finally, a softmax layer (fc9) calculates output probabilities for each class, defined as a person ID, which combined with a loss function is used to update the parameters of the network. We

refer to our proposed system as RGB-D-CNN. At test time, the softmax layer is discarded and features are extracted from fc8.

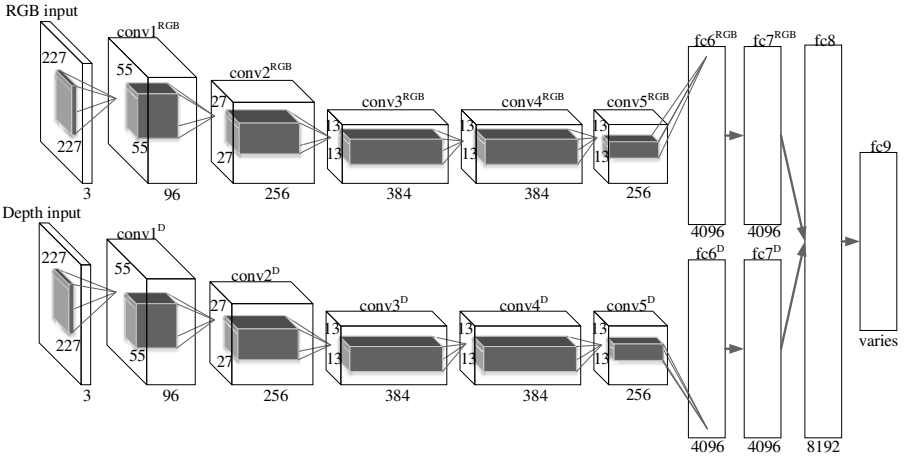


Fig. 1: Overview of the RGB-D based CNN (RGB-D-CNN). Lower part processes a depth image, while the upper part processes a colour image, features from last fully connected layer of the streams are fused in a joint fully connected layer before classification.

Individual training Before training the RGB-D-CNN, CNN models are trained for RGB and depth, respectively. We refer to these models as RGB-CNN and D-CNN. Both follow similar structure as the upper/lower part of the joint CNN, with a softmax layer replacing fc8 and fc9. The model weights are initialized using a pre-trained model of the CaffeNet version [Ji14] of AlexNet trained on the ImageNet dataset. Following the architecture of AlexNet, the input is an image of size 227×227, randomly cropped from an image of size 256×256, to make the network robust to changes in translation. Both colour and depth images are therefore resized accordingly before being processed by the network. In addition, the images are randomly flipped to increase the amount of training data. In case of depth images, [Ei15] shows that applying a jet colourmap enhances the accuracy compared to encoding the images using surface normals [BRF13] or Horizontal disparity, Height and Angle (HHA) encoding [Gu14]. This colour transformation maps each depth value to a colour in RGB colour space from blue(close) over green to red(far). This enables us to initialize the weights using the pre-trained CaffeNet model without additional preprocessing. We therefore perform similar step before training the depth model.

Given sets of parameters and datasets $(W^{RGB}, b^{RGB}, X^{RGB}, Y)$ and (W^{D}, b^{D}, X^{D}, Y) for RGB and depth, respectively, where W and b are the model weights and bias, while (X^{RGB}, X^{D}) are the set of RGB and depth images with corresponding labels Y, we train the models by minimizing a loss function, L, as given in Equation 1:

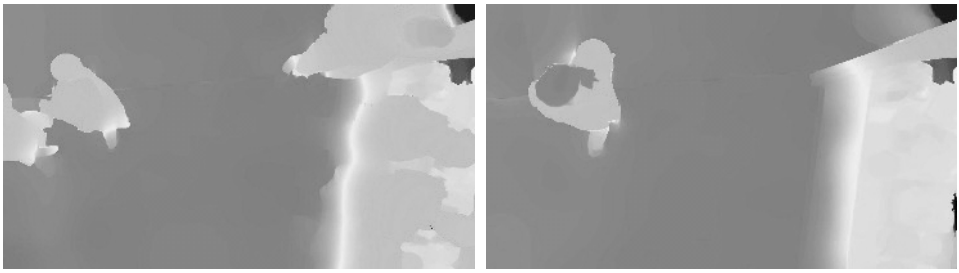
$$\min_{W,b} -\frac{1}{N}\sum_{i=1}^{N} L(W,b,x_i,y_i) \qquad L(W,b,x_i,y_i) = log(\hat{p}_i, y_i) \tag{1}$$

where W, b are the weights and bias of the model currently being trained, $X = \{x_1, .., x_N\}$ is the sample set and $\hat{p}_i$ is the output probability from the softmax layer of the i'th sample given the true label y_i.

Joint training After training RGB-CNN and D-CNN, the model parameters are used to initialize the two CNN streams in RGB-D-CNN. The softmax layers are replaced by a randomly initialized fully connected layer (fc8) and new softmax layer (fc9). By fusing outputs from both fc7^{RGB} and fc7^{D} in fc8, the parameters of the depth stream are updated depending on the input to the RGB stream and vice versa, while the weights and bias of fc8 are updated based on both inputs, resulting in a fused output. [WCZ16] shows how fusion of hand-crafted and CNN features in the late layers of the network affects parameter update of the CNN. Similar proof applies to this context.

4 Experimental Results

Datasets For evaluation we present a novel RGB-D based dataset collected from an overhead view. We refer to the dataset as Overhead Person Re-identification (OPR). The dataset is collected using a calibrated ZED stereo camera from Stereolabs [St17], mainly due to its ability to record depth from a range 0.7m-20m covering both low and high ceilings. In addition, it captures video in resolutions up to 4416×1242 pixels which is much higher than RGB sensors in solutions such as the Microsoft Kinect. The camera is placed in the ceiling at a university canteen (uncontrolled environment) to capture a populated area. From this perspective, persons are captured when approaching (walking from top to bottom), and leaving (walking from bottom to top) the canteen a few minutes later, enabling us to evaluate re-id performance. Data is collected on a single day during a two hour period around midday to capture video when the number of persons in the canteen is increasing and decreasing. As a result, cases of having a large number of persons and only a single person are recorded, examples of captured depth images in both cases are shown in Figure 2 (a) and (b), respectively.



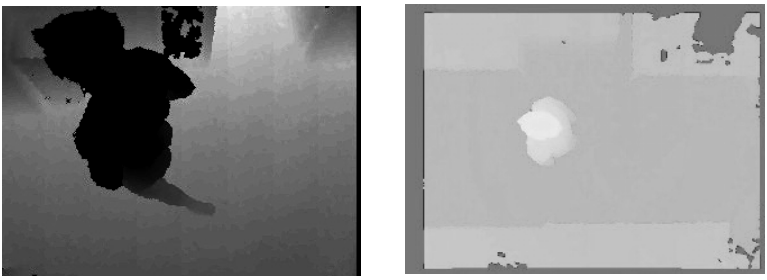
(a) (b)

Fig. 2: Examples of depth images containing (a): multiple persons and (b): containing a single person. Each person is captured when approaching (right side) and leaving (left side) the canteen.

Disparity maps are computed using Semi-Global Block Matching (SGBM) as it has shown as a good compromise between accuracy and processing time [Ka11], followed by filtering

using a Weighted Least Square (WLS) kernel to eliminate noise and make the background more uniform, resulting in more precise depth information. Finally, we manually annotate bounding boxes around persons and use those for our system, the annotations enables us to further test detection, tracking and segmentation algorithms in future work. A total number of 78742 frames with 64 different persons have been annotated for re-id.

To our knowledge, only the datasets of [HAFF16] (DPI-T) and [Li17] (TVPR) have previously been proposed for RGB-D and overhead based re-id. Both are recorded in a hall with only a single person within view at all times. Examples of depth images from these datasets are shown in Figure 3. In addition to evaluating on our own dataset, we apply our system to those of [HAFF16] and [Li17] for comparison with their original results.



(a) (b)

Fig. 3: Examples of depth images from (a): DPI-T and (b): TVPR.

Evaluation protocols Depending on the dataset, different training and testing protocols are followed.

OPR Similar to most RGB-based datasets within re-id, we perform 10 random train and test splits, each set containing 32 persons. After training the CNN models, features from the test set are extracted from the last fully connected layer.

TVPR The training set originally consists of 100 persons walking from left to right while the test set consists of same persons walking from right to left. During test, features from the test set are compared with those from the training set. Although, due to issues at test time regarding one of the video sequences, only 94 persons were considered for training and testing.

DPI-T 12 persons appear in five different sets of clothing in both the training and test, while the number of recordings in each set differs. A total of 213 sequences are used for training while the test set consists of 249 sequences which are all classified by comparing with those of the training set.

When training RGB-CNN and D-CNN, a batch size of 128 is used while a size of 64 is used in case of RGB-D-CNN. Network parameters are updated using Stochastic Gradient Descent (SGD) with momentum is to avoid getting stuck in a local minimum. Hyper parameters are set accordingly to [Ei15] with a momentum of 0.9 and base learning rate of 0.01 which is reduced by multiplying with 0.97 for each epoch. At each epoch, the training set is randomly shuffled for faster convergence [Be12]. We present our results by calculating the *rank-1* to *rank-k* accuracies based on feature matching where *rank-i* indicates a

cumulative percentage of persons having their true match within the i most similar with k indicating the total number of persons. For OPR, the average accuracies over all train/test splits are calculated. Matches are calculated using Euclidean distance between extracted features following a multi-shot approach, i.e. features from all images of each person/sequence are extracted and either maximized or averaged, indicated by subscripts *max* and *avg*.

Figure 4 (a) shows the resulting Cumulative Matching Characteristic (CMC) curves for applying RGB-CNN, D-CNN and RGB-D-CNN to OPR. It is clear that fusing of RGB and depth modalities clearly increases accuracy compared to using a single modality. The best result is achieved by RGB-D-CNN$_{avg}$, increasing accuracy by 16.00% compared to RGB-CNN$_{avg}$. Furthermore, Figure 4 (b) and (c) show the results of our system applied to DPI-T and TVPR, respectively. In case of DPI-T, RGB-D-CNN$_{avg}$ still outperforms RGB-CNN and D-CNN with an increase of 3.61% compared to RGB-CNN$_{avg}$. Finally for TVPR, RGB-CNN provide better results compared to RGB-D-CNN. A reason for this could be the quality of depth information (see Figure 3 (b)) negatively affecting the training of RGB-D-CNN in combination with corresponding colour images. Even though, D-CNN results are slightly worse in case of DPI-T, the level of detail in depth images are higher (see 3 (a)) causing the modality to better complement RGB. The quality of depth information therefore seems important when training an RGB-D CNN. Looking at results across all datasets, averaged features mostly provides the highest accuracies, although, in case of depth features, feature maximization seems better. This could be due to encoding of features as colourized images combined with an overhead view from which the height of each person, and thereby the colour gradient, is important. By averaging features, this information more easily gets lost if the representation changes between images.

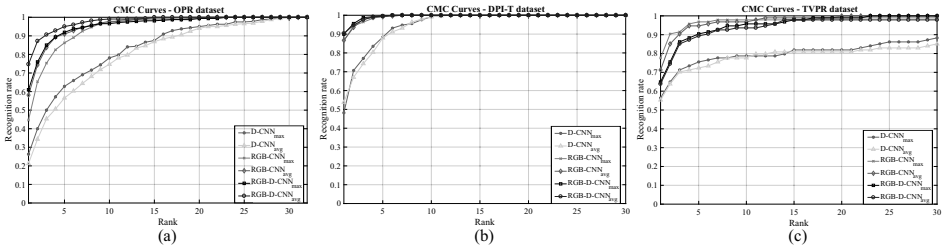


Fig. 4: Results on (a) OPR (p=32), (b) DPI-T (p=249) and (c) TVPR (p=94) for RGB-CNN, D-CNN and RGB-D-CNN, respectively, using maximized (*max*) and averaged (*avg*) features.

Tables 1 summarizes our results on TVPR and DPI-T, compared to their original results. As [HAFF16] only provides a rank-1 accuracy while [Li17] only provides CMC curves, only the rank-1 accuracy is considered. For [Li17], rank-1 is estimated from the CMC curves. *Ours* refers to the best results achieved by our system (RGB-D-CNN$_{avg}$ in case of DPI-T and RGB-CNN$_{avg}$ in case of TVPR). In both cases we outperform original results, for DPI-T by 34.76% by also using RGB. From Figure 4 (b), it is worth noting that our D-CNN alone achieves almost similar accuracies as [HAFF16] who also adds an LSTM layer on top of a similar CNN.
Even though, six persons are missing for the tests on TVPR, our system shows potential to be improved further. For RGB alone, our system outperforms that of [Li17] by ≈5.16%.

	Rank-1 accuracy [%]	
Method	*DPI-T*	*TVPR*
4D RAM [HAFF16]	55.60	–
TVDH [Li17]	–	72.50
Ours	**90.36**	**77.66**

Tab. 1: Comparison of our RGB-D-CNN to original results on DPI-T and TVPR datasets.

Processing time We evaluate processing time for stereo and feature matching on OPR to discuss on the potential of using passive stereo for re-id applications. 20 matching iteration are run using an Intel i7-6700HQ CPU @ 2.60GHz and 16GB of RAM and average timings are provided. Stereo matching matching is performed on images of size 960×540.

While feature matching only takes $4.0e10^{-5}$s, SGBM and WLS are more processing intensive taking 0.136s and 0.103s, respectively. Nonetheless, $\approx$4 FPS is achieved using the CPU. For real-time applications, GPU implementations of SGBM and WLS algorithms could be used speed up the process. No such implementations are available at the moment.

5 Conclusion

In this paper, we have presented an RGB-D based CNN applied to person re-identification. Two CNN models are trained using colour and depth images, respectively, captured from an overhead view and resulting trained parameters are used to initialize a joint RGB-D-CNN model trained using both modalities. To test the system, we collected a novel RGB-D and overhead based dataset which is annotated for evaluation on both re-id accuracy, but also detection and tracking algorithms. By applying our system to our novel and two previously proposed datasets, we have shown that the combination of RGB and depth modalities increase accuracy by 16.0% and 3.6% on our OPR dataset and DPI-T, respectively. In case of TVPR, RGB modality alone achieved higher accuracy than combining modalities due to the quality of depth information. This indicates an importance to capture detailed depth information to proper complement the RGB modality. In addition, our system shows an FPS of 4 using a CPU, with potential of being increased if processing intensive algorithms such as SGBM and WLS are implemented on a GPU. For future work, the system should be evaluated on bounding boxes extracted automatically from a person detector. To increase detection performance, depth information could also be used for this purpose. Furthermore, our proposed system could be extended with an LSTM to handle video rather than averaging or maximizing features extracted from a sequence of images. This would allow for temporal information to be captured as well. Finally, more recently developed neural networks could replace the AlexNet architecture to increase performance and decrease processing time.

Acknowledgement

The work carried out in this paper is supported by Innovation Fund Denmark under Grant 5189-00222B.

References

[AC12] Ahmed, Imran; Carter, John N: A robust person detector for overhead views. In: Proc. ICPR. IEEE, pp. 1483–1486, 2012.

[AJM15] Ahmed, Ejaz; Jones, Michael; Marks, Tim K: An improved deep learning architecture for person re-identification. In: Proc. CVPR. IEEE, pp. 3908–3916, 2015.

[Ar08] Aradhye, Hrishikesh; Fischler, Martin; Bolles, Robert; Myers, Gregory: Headprint-Based Human Recognition. In: Advances in Biometrics: Sensors, Algorithms and Systems. Springer London, chapter 15, pp. 287–306, 2008.

[Ba12] Barbosa, Igor; Cristani, Marco; Del Bue, Alessio; Bazzani, Loris; Murino, Vittorio: Re-identification with rgb-d sensors. In: Computer Vision–ECCV 2012: Workshops and Demonstrations. Springer, pp. 433–442, 2012.

[Be12] Bengio, Yoshua: Practical recommendations for gradient-based training of deep architectures. In: Neural networks: Tricks of the trade, pp. 437–478. Springer, 2012.

[BRF13] Bo, Liefeng; Ren, Xiaofeng; Fox, Dieter: Unsupervised feature learning for RGB-D based object recognition. In: Experimental Robotics. Springer, pp. 387–402, 2013.

[Ch16] Chen, Ying-Cong; Zheng, Wei-Shi; Lai, Jian-Huang; Yuen, Pong: An Asymmetric Distance Model for Cross-view Feature Mapping in Person Re-identification. IEEE Transactions on Circuits and Systems, 2016.

[Ei15] Eitel, Andreas; Springenberg, Jost Tobias; Spinello, Luciano; Riedmiller, Martin; Burgard, Wolfram: Multimodal deep learning for robust rgb-d object recognition. In: Proc. IROS. IEEE, pp. 681–687, 2015.

[GT08] Gray, Douglas; Tao, Hai: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proc. ECCV. Springer, pp. 262–275, 2008.

[Gu14] Gupta, Saurabh; Girshick, Ross; Arbeláez, Pablo; Malik, Jitendra: Learning rich features from RGB-D images for object detection and segmentation. In: Proc. ECCV. Springer, pp. 345–360, 2014.

[Ha16] Hazirbas, Caner; Ma, Lingni; Domokos, Csaba; Cremers, Daniel: Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In: Asian Conference on Computer Vision. Springer, pp. 213–228, 2016.

[HAFF16] Haque, Albert; Alahi, Alexandre; Fei-Fei, Li: Recurrent Attention Models for Depth-Based Person Identification. In: Proc. CVPR. IEEE, pp. 1229–1238, 2016.

[Ji14] Jia, Yangqing; Shelhamer, Evan; Donahue, Jeff; Karayev, Sergey; Long, Jonathan; Girshick, Ross; Guadarrama, Sergio; Darrell, Trevor: Caffe: Convolutional architecture for fast feature embedding. In: Proc. ACMMM. ACM, pp. 675–678, 2014.

[Ka11] Kalarot, Ratheesh; Morris, John; Berry, David; Dunning, James: Analysis of real-time stereo vision algorithms on GPU. In: Proc. IVCNZ. 2011.

[Ki98] Kittler, Josef; Hatef, Mohamad; Duin, Robert PW; Matas, Jiri: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3):226–239, 1998.

[KSH12] Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey E: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. Citeseerx, pp. 1097–1105, 2012.

[Li14] Li, Wei; Zhao, Rui; Xiao, Tong; Wang, Xiaogang: Deepreid: Deep filter pairing neural network for person re-identification. In: Proc. CVPR. IEEE, pp. 152–159, 2014.

[Li15] Liao, Shengcai; Hu, Yang; Zhu, Xiangyu; Li, Stan Z: Person re-identification by local maximal occurrence representation and metric learning. In: Proc. CVPR. IEEE, pp. 2197–2206, 2015.

[Li17] Liciotti, Daniele; Paolanti, Marina; Frontoni, Emanuele; Mancini, Adriano; Zingaretti, Primo: Person Re-identification Dataset with RGB-D Camera in a Top-View Configuration. In: Video Analytics. Face and Facial Expression Recognition and Audience Measurement. Springer, pp. 1–11, 2017.

[LSF15] Li, Sheng; Shao, Ming; Fu, Yun: Cross-view projective dictionary learning for person re-identification. In: Proc. AAAI. AAAI Press, pp. 2155–2161, 2015.

[Mu14] Munaro, Matteo; Basso, Alberto; Fossati, Andrea; Van Gool, Luc; Menegatti, Emanuele: 3D reconstruction of freely moving persons for re-identification with a depth sensor. In: Proc. ICRA. IEEE, pp. 4512–4519, 2014.

[Sa16] Sanchez-Riera, Jordi; Hua, Kai-Lung; Hsiao, Yuan-Sheng; Lim, Tekoing; Hidayati, Shintami C; Cheng, Wen-Huang: A comparative study of data fusion for RGB-D based visual recognition. Pattern Recognition Letters, 73:1–6, 2016.

[St17] Stereolabs: , ZED - Depth Sensing and Camera Tracking. https://www.stereolabs.com/zed/specs/, 2017.

[WCZ16] Wu, Shangxuan; Chen, Ying-Cong; Zheng, Wei-Shi: An enhanced deep feature representation for person re-identification. In: Proc. WACV. IEEE, pp. 1–8, 2016.

[ZXG16] Zhang, Li; Xiang, Tao; Gong, Shaogang: Learning a Discriminative Null Space for Person Re-Identification. In: Proc. CVPR. IEEE, pp. 1239–1248, 2016.