 doi:10.20944/preprints201905.0160.v1

SIAMESE NEURAL NETWORK BASED APPEARANCE MODEL FOR MULTI-TARGET TRACKING

Mohib Ullah

Department of Computer Science (IDI), Norwegian University of Science and Technology, Norway.

ABSTRACT

An appearance model plays a crucial rule in multi-target tracking. In traditional approaches, the two steps of appearance modeling i.e visual representation and statistically similarity measure are modeled separately. Visual representation is achieved either through hand-crafted features or deep features and statically similarity is measure through a cross entropy loss function. A loss function based on cross-entropy (KL-divergence, mutual information) find closely related probability distribution for the targets. However, if the targets have similar visual representation, it ends up mixing the targets. To tackle this problem, we come up with a synergetic appearance model named Single Shot Appearance Model based on Siamese neural network. The network is trained with a contrastive loss function for finding the similarity between different targets in a single shot. The input to the network is two target patches and based on their similarity, a contrastive score is output by the network. The proposed model is evaluated on accumulative dissimilarity metric on three datasets. Quantitatively, promising results are achieved against three baseline methods.

Index Terms— Siamese neural network, appearance model, contrastive loss, cross entropy.

1. INTRODUCTION

One of the primary tasks of computer vision is to empower the computers with a vision system similar or even more sophisticated than the humans. Such a capability allow the computers to analyses a visual scene for a variety of tasks including but not limited to tracking [1, 2], anomaly detection [3–6], object detection [7, 8], segmentation [9–11], motion analysis [12, 13], emotion classification [14, 15], and crowd analysis [16, 17]. Designing a robust appearance model plays a crucial rule in aforementioned applications, especially multi-target tracking. For example, the well adopted strategy for multi-target tracking is tracking-by-detection where the tracking is divided into two discrete steps i.e. detection and association. Usually, target detection is achieved through a discriminative [18] or a generative model [19]. While the association is a combinatorial optimization problem and is usually handled globally [20, 21] or locally [22, 23] depending upon the underlying application constraints. The appearance model is mainly used in association to differentiate between the targets. In nutshell, target appearance model consist of two main steps:

- Visual representation
- Similarity metric

Traditionally, hand-engineered features are used for the visual representation of the targets. In a nutshell, they are classified into two broad categories i.e. the local key-point features and the global or region features. The local feature extracts only sparse features from an image or image patch [24–26]. However, global features [27–29] model the whole image patch as a representative descriptor of an image patch. In Fig. 1, the visual representation of a target patch through two feature descriptor is given.

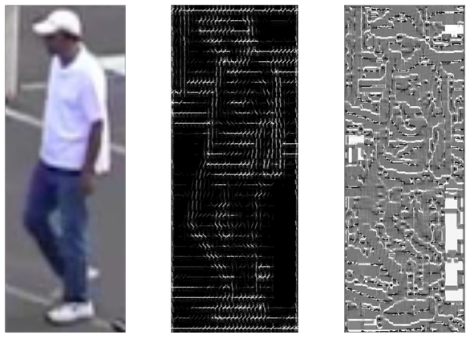

Fig. 1: Visual representation of feature descriptors. A) shows the RGB input of a target patch. B). HOG [27] descriptor. C). LBP [29] descriptor

Based on the level of abstraction, global features can further be classified into zero-order, first-order, and up-to-second [30]. For example, color histogram [31] and raw pixel template [32] are considered to be zero-order descriptor. Level set formulation [33] and gradient descriptor [27] are treated as the first order while region covariance matrix [34] is associated with the second order descriptor. More recently [35], deep features are explored for the visual representation of the targets. Either deep features or the hand-engineered features,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 May 2019 doi:10.20944/preprints201905.0160.v1

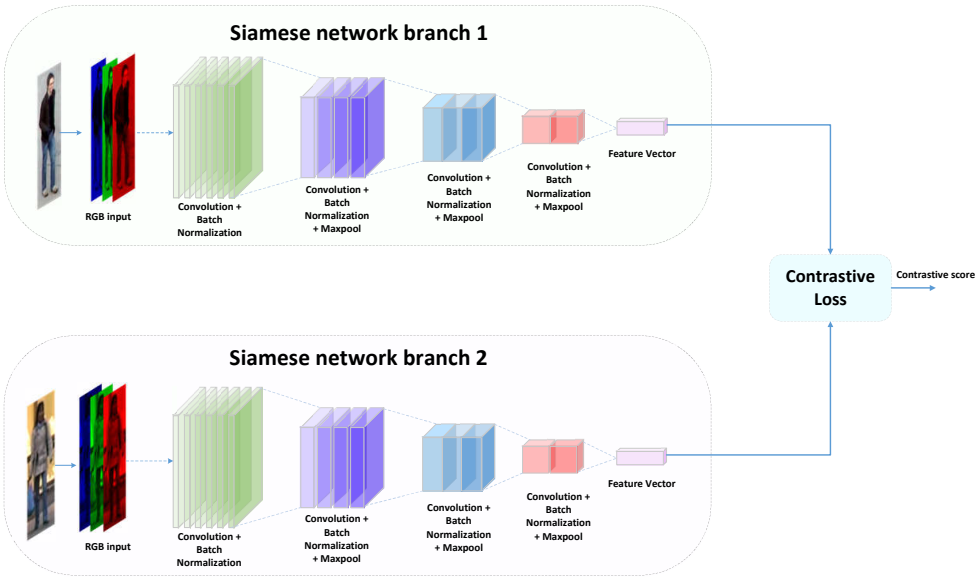

Fig. 2: A pair of targets patch is inserted in each network. The network extracts discriminative features and gives it to the contrastive loss module. The module calculates the contrast between the two patches and output a dissimilarity score.

the aim is to represent the target in the feature space in such a way that it could easily be differentiated from others. In the second step of the appearance model, the similarity between two target patches is calculated through a pre-define metric. The usual adopted metrics are Kullback Leibler divergence [36], Jensen Shannon divergence [37], normalized cross correlation [38], Bhattacharya distance [39], mutual information [40], to name of few. In nutshell, the main characteristic of the existing methods is it's two-step procedure i.e. they perform the visual representation and the statistical similarity analysis independently. The focus of this paper is to introduce an appearance model where the visual representation and the similarity measure are performed in a single synergetic framework which we refer as single shot appearance model. The rest of the paper is organized in the following order. The overview of the proposed method is given in section 2. Network architecture, the contrastive loss function is briefly explained in section 3. Experiments are conducted in section 5 and section 6 concludes the paper.

2. PROPOSED APPROACH

The block diagram of the proposed model is given in Figure **??**. Essentially, the network consists of a Siamese network [41] with a contrastive loss function. Siamese neural network is a special kind of neural network that consists of two parallel Convolutional Neural Networks (CNN). The architecture of both the networks are similar and they share common weight. Essentially, one network is the mirrored version of the other. Each network is taking an image patch corresponding to the target of interest. The target patches are generated by the target detector or manually extracted from the image. The networks extract the discriminative feature through it's linear and non-linear layers (convolution, activation, pooling) from the patches and used as the visual representation of the target. Both the visual representation is given to a contrastive loss function module which outputs the similarity score between the two patches. Compared to the classical CNN, where the network learned to classify the inputs into different categories, the Siamese network gives the dissimilarity score between the corresponding inputs. In the next section 3, a brief description of the architecture and loss function is given.

3. NETWORK ARCHITECTURE

Classically, a CNN consists of convolution, activation, and pooling layers. In any architecture [42–45], these layers are organized in a special order. The depth of the network corresponds to the number of layers in the network. The general rule of thumb is, the deeper the network, the better is feature representation. In Fig.2, the generic structure of the CNN is given. Technically, any architecture of CNN can be used as the Siamese network branch i.e. as a building block of the overall Siamese neural network. However, a well established approach in the research community is to use an already trained CNN network. One of the primary reasons for this is, designing a network from scratch is easy but training is very expensive in terms of computation. Moreover, intrinsically, CNNs are very data hungry. Usually, a pre-trained network gives a good performance because it has been trained on millions of images of Imagenet dataset [46]. In our work,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 May 2019 doi:10.20944/preprints201905.0160.v1

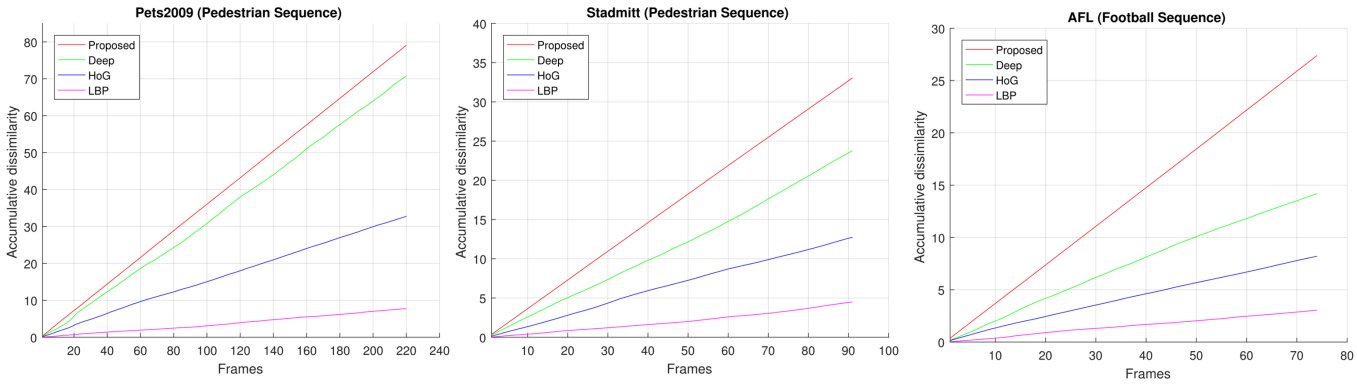

Fig. 3: Quantitative results on three datasets

we used state-of-the-art inception model of CNN [44] as the branches of our Siamese neural network. Originally, inception model [44] is trained for the classification task of 1000 classes. However, we are interested only in the feature extraction capability of the network. So, we truncated the work and removed the fully connected layers responsible for classification and inserted our contrastive lost module. The contrastive loss [47] module compare the visual representation of both the target patches and gives a contrastive score which shows the disparity in the distribution of both target patches. The brief overview of the contrastive lost module is given in the following:

3.1. Contrastive Loss Function

The aim of the loss function is to train and evaluate the network based on the given criteria. For classification, the loss function is defined in such a way to classify an image or image patch into a pre-define class categories. However, the aim of our Siamese network is to differentiate between two target patches rather than classifying them into different categories. Therefore, a classification loss function based on cross-entropy [36] is not suitable as we are not interested in the class probability but rather want to get a dissimilarity score based on the visual representation of the target patches. Therefore, a contrastive loss function is the most suitable for our task. Theoretically, the contrastive loss function assesses how the network is differentiating the given target patches.

Mathematically, the contrastive loss function could be written as:

$$L(D_w, I, m) = (1-I)\frac{1}{2}(D_w)^2 + I\frac{1}{2}\{max(0, m - D_w)\}^2 \quad (1)$$

where I is an indicator variable and set to zero if the inputs are the same (similar targets) and 1 otherwise. m is an empirical parameter and can be seen as a non-negative margin value. It penalizes the fact that if input pairs are very different from the margin, it will not contribute to the overall loss. It is plausible as our goal is to penalized the pairs that are different but the network see them similar. Similarly, D_w is the Euclidean distance between the visual representation of the two target patches.

$$D_w = \sqrt{\{F_{T1} - F_{T2}\}^2} \quad (2)$$

In equation 2, F_{T1} and F_{T2} are the visual representation of target T_1 and T_2.

4. APPLICATION IN TRACKING

One of the important components of multi-target tracking is the association model which establish correspondence between the targets in the consecutive frame. Target association is difficult as the appearance of targets changes temporally. In case of human, it's even more tricky because human body goes under pose changes and articulation. However, the changes in the appearance is temporally continuous and gradual. Therefore, in tracking, it is usually assumed that the appearance remains approximately the same in two consecutive frames. Based on this assumption, ideally a target that exist in two consecutive frames at $t-1$ and t should have a high similarly score compared to the other targets. Consequently, an appearance model that gives optimal scores (high for similar, low for different) for the target patches would yield a high quality tracking performance. In order to validate our claim, we compared the similarity score of our model against three baseline methods that incorporate HoG [27], LBP [29] and deep features [44] with Jensen Shannon divergence [37]. The application of our proposed model is not limited to tracking but potentially could be used for any visual recognition task like image indexing and retrieval.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 May 2019 doi:10.20944/preprints201905.0160.v1

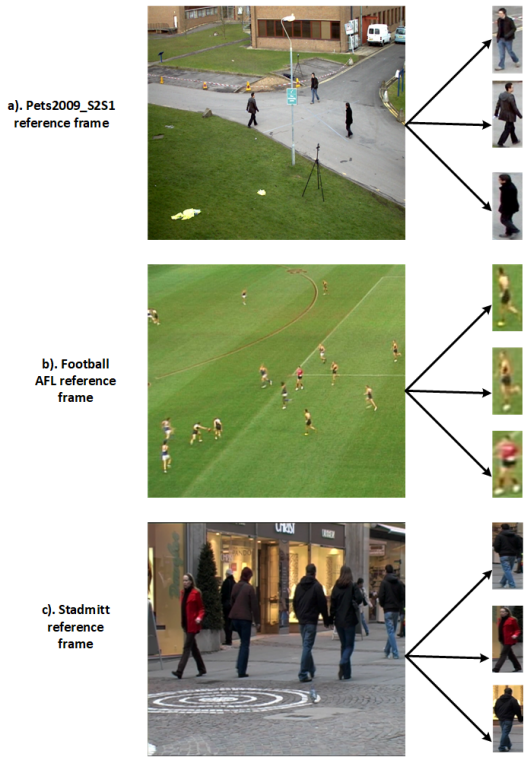

Fig. 4: Reference frames from the three datasets [48–50] and the corresponding target patches

5. EXPERIMENT

The proposed network is implemented in Matlab with Matconvnet toolbox on a core i7 system with 8 GB RAM. To evaluate the network, we have chosen three datasets [48–50] that are commonly used for pedestrian tracking. The sample frames and corresponding target patches are given in Fig. **??**. From each datasets, target patches can be generated by a target detector. However, we have manually annotated the targets patches to effectively evaluate our network and reduce the effect of false positive/negative. Moreover, we used Jensen Shannon divergence [37] as the baseline similarity metric and evaluated two well-known hand-crafted ([27], [29]) and one deep features [44]. The proposed network used similar deep features [44] but with a contrastive loss function. For each dataset, first a reference target is selected and it's contrastive score is calculated against all the targets. The same procedure is followed for each target in every frames. Based on the number of frames used in the simulations, the average score is calculated by normalizing it with the number of frames in the dataset. The number of targets in each dataset is also different. But we have considered only those frames where the number of target stays constant for the effective evaluation of the methods. The quantitative results show the propose network achieved a better contrastive score compared to the baseline method with a good margin.

Datasets	**Frame No.**	**LBP**	**HoG**	**Deep**	**Proposed**
Pets2009	220	0.035	0.148	0.321	0.359
AFL	74	0.04	0.110	0.191	0.370
Stadmitt	91	0.049	0.140	0.261	0.363

Table 1: Normalized average accumulative contrastive score on 3 dataset. The proposed approach gives the best score. The second best score is given by a deep feature based method. HoG [27] and LBP [25] based method gives third and forth best results, respectively.

6. CONCLUSION

We proposed a single shot appearance model for multi-target tracking. It is based on a Siamese neural network with a contrastive loss function. The input to the network is a pair of patches corresponding to targets of interest. The network output a dissimilarity score in a single shot. The quantitative results show that the proposed network gives better visual representation and a better disparity in the feature space between different targets. The model has potential applications in multi-target tracking, image indexing retrieval. In future, we will incorporate the proposed appearance model in the multi-target tracking framework.

7. REFERENCES

[1] Mohib Ullah and Faouzi Alaya Cheikh, "A directed sparse graphical model for multi-target tracking," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1816–1823.

[2] ShiJie Sun, Naveed Akhtar, HuanSheng Song, Ajmal Mian, and Mubarak Shah, "Deep affinity network for multiple object tracking," *arXiv preprint arXiv:1810.11780*, 2018.

[3] Habib Ullah, Mohib Ullah, and Nicola Conci, "Real-time anomaly detection in dense crowded scenes," in *Video Surveillance and Transportation Imaging Applications 2014*. International Society for Optics and Photonics, 2014, vol. 9026, p. 902608.

[4] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, Zahra Moayed, and Reinhard Klette, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," *Computer Vision and Image Understanding*, vol. 172, pp. 88–97, 2018.

[5] Habib Ullah, Ahmed B Altamimi, Muhammad Uzair, and Mohib Ullah, "Anomalous entities detection and localization in pedestrian flows," *Neurocomputing*, vol. 290, pp. 74–86, 2018.

[6] Habib Ullah, Mohib Ullah, Hina Afridi, Nicola Conci, and Francesco GB De Natale, "Traffic accident detection through a hydrodynamic lens," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 2470–2474.

[7] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[9] Habib Ullah and Nicola Conci, "Structured learning for crowd motion segmentation," in *2013 IEEE International Conference on Image Processing*. IEEE, 2013, pp. 824–828.

[10] Paolo Rota, Habib Ullah, Nicola Conci, Nicu Sebe, and Francesco GB De Natale, "Particles cross-influence for entity grouping," in *21st European Signal Processing Conference (EUSIPCO 2013)*. IEEE, 2013, pp. 1–5.

[11] Habib Ullah, Mohib Ullah, and Muhammad Uzair, "A hybrid social influence model for pedestrian motion segmentation," *Neural Computing and Applications*, pp. 1–17, 2018.

[12] Habib Ullah, Mohib Ullah, and Nicola Conci, "Dominant motion analysis in regular and irregular crowd scenes," in *European Conference on Computer Vision (ECCV) workshop on Human Behavior Understanding*. Springer, 2014, pp. 62–72.

[13] Habib Ullah, *Crowd Motion Analysis: Segmentation, Anomaly Detection, and Behavior Classification*, Ph.D. thesis, University of Trento, 2015.

[14] Huanghao Feng, Hosein M Golshan, and Mohammad H Mahoor, "A wavelet-based approach to emotion classification using eda signals," *Expert Systems with Applications*, vol. 112, pp. 77–86, 2018.

[15] Habib Ullah, Muhammad Uzair, Arif Mahmood, Mohib Ullah, Sultan Daud Khan, and Faouzi Alaya Cheikh, "Internal emotion classification using eeg signal with sparse discriminative ensemble," *IEEE Access*, vol. 7, pp. 40144–40153, 2019.

[16] Habib Ullah and Nicola Conci, "Crowd motion segmentation and anomaly detection via multi-label optimization," in *ICPR workshop on pattern recognition and crowd analysis*, 2012.

[17] Habib Ullah, Muhammad Uzair, Mohib Ullah, Asif Khan, Ayaz Ahmad, and Wilayat Khan, "Density independent hydrodynamics model for crowd coherency detection," *Neurocomputing*, vol. 242, pp. 28–39, 2017.

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, , no. 6, pp. 1137–1149, 2017.

[19] Krystian Mikolajczyk, Bastian Leibe, and Bernt Schiele, "Multiple object class detection with a generative model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, vol. 1, pp. 26–36.

[20] Anton Milan, Stefan Roth, and Konrad Schindler, "Continuous energy minimization for multitarget tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 58–72, 2014.

[21] Mohib Ullah and Faouzi Alaya Cheikh, "Deep feature based end-to-end transportation network for multi-target tracking," in *IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 3738–3742.

[22] Wangmeng Zuo, Xiaohe Wu, Liang Lin, Lei Zhang, and Ming-Hsuan Yang, "Learning support correlation filters for visual tracking," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[23] Mohib Ullah, Faouzi Alaya Cheikh, and Ali Shariq Imran, "Hog based real-time multi-target tracking in bayesian framework," in *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2016, pp. 416–422.

[24] David G Lowe, "Distinctive image features from scale-invariant keypoints," *Springer International Journal of Computer Vision (IJCV)*, vol. 60, no. 2, pp. 91–110, 2004.

[25] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua, "Brief: Binary robust independent elementary features," in *Springer European Conference on Computer Vision (ECCV)*, 2010, pp. 778–792.

[26] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski, "Orb: An efficient alternative to sift or surf," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 2564–2571.

[27] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, vol. 1, pp. 886–893.

[28] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer, "Real-time tracking of non-rigid objects using mean shift," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2000, vol. 2, pp. 142–149.

[29] Timo Ojala, Matti Pietikainen, and Topi Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[30] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, Xiaowei Zhao, and Tae-Kyun Kim, "Multiple object tracking: A literature review," *arXiv preprint arXiv:1409.7618*, 2014.

[31] Dennis Mitzel and Bastian Leibe, "Real-time multi-person tracking with detector assisted structure propagation," in *IEEE International Conference on Computer Vision (ICCV Workshops)*, 2011, pp. 974–981.

[32] Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg, "Who are you with and where are you going?," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1345–1352.

[33] Dennis Mitzel, Esther Horbert, Andreas Ess, and Bastian Leibe, "Multi-person tracking with sparse detection and continuous segmentation," in *European Conference on Computer Vision (ECCV)*. Springer, 2010, pp. 397–410.

[34] Fatih Porikli, Oncel Tuzel, and Peter Meer, "Covariance tracking using model update based on lie algebra," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 728–735.

[35] Mohib Ullah, Ahmed Kedir Mohammed, Faouzi Alaya Cheikh, and Zhaohui Wang, "A hierarchical feature model for

multi-target tracking," in *IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 2612–2616.

[36] Jianhua Lin, "Divergence measures based on the shannon entropy," *IEEE transactions on information theory*, vol. 37, no. 1, pp. 145–151, 1991.

[37] Bent Fuglede and Flemming Topsoe, "Jensen-shannon divergence and hilbert space embedding," in *IEEE International Symposium on Information Theory (ISIT)*, 2004, p. 31.

[38] Kai Briechle and Uwe D Hanebeck, "Template matching using fast normalized cross correlation," in *Optical Pattern Recognition XII*. International Society for Optics and Photonics, 2001, vol. 4387, pp. 95–103.

[39] Thomas Kailath, "The divergence and bhattacharyya distance measures in signal selection," *IEEE transactions on communication technology*, vol. 15, no. 1, pp. 52–60, 1967.

[40] Hanchuan Peng, Fuhui Long, and Chris Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[41] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, 2015, vol. 2.

[42] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[43] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *Springer European conference on computer vision*, 2014, pp. 818–833.

[44] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[45] Mohib Ullah, Ahmed Mohammed, and Faouzi Alaya Cheikh, "Pednet: A spatio-temporal deep convolutional neural network for pedestrian segmentation," *Journal of Imaging*, vol. 4, no. 9, pp. 107, 2018.

[46] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)*. IEEE, 2009, pp. 248–255.

[47] Ce Qi and Fei Su, "Contrastive-center loss for deep neural networks," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 2851–2855.

[48] James Ferryman and Ali Shahrokni, "Pets2009: Dataset and challenge," in *Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2009, pp. 1–6.

[49] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele, "Monocular 3d pose estimation and tracking by detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 623–630.

[50] Anton Milan, Rikke Gade, Anthony Dick, Thomas B Moeslund, and Ian Reid, "Improving global multi-target tracking with local updates," in *European Conference on Computer Vision*. Springer, 2014, pp. 174–190.