Intelligent Environments 2018
I. Chatzigiannakis et al. (Eds.)

doi:10.3233/978-1-61499-874-7-139

A Framework for Human Recognition and Counting in Restricted Area for Video Surveillance

Alessandro Moro[a,1] Jun Wakabayashi[a] Tetsuro Toda[a] Kazunori Umeda[a]
[a]*Chuo University*

Abstract. We introduce a multi-process framework for human counting and recognition that exploits the combination of multiple deep neural networks. Deep networks have advanced the state of the art in many fields and play an essential role in computer vision for detection and recognition. However, very deep networks are still slow at inference time, and they require a substantial amount of hardware to perform complex operations. Real-time recognition from video source is still an issue due to complexity of scenario and the amount of data to process. In this paper, we propose an approach that combines multiple neural networks, that is fast and accurate.

Keywords. Video Surveillance, Human Recognition, Deep Learning

1. Introduction

Detection and recognition of people in restricted areas is an important task to prevent or persecute theft of data and objects. Offices and laboratories are locations which can contain valuable documents or devices, and it is not difficult for an unauthorized person to enter in non-secured areas and act freely. In the recent years new technologies allow to control restricted areas, and video surveillance technologies are frequently used for control and detection. Even if accuracy of non-invasive systems is increasing, the problem of recognition is still a hot research topic. This paper addresses the problem of recognizing authorized humans in a restricted area, while preserving real-time processing and high accuracy.

Human detection problem has been widely studied in the past decades. A successful algorithm based on histograms was introduced by Dalal et al [**7**]. More recently, with the advent of Deep Learning technologies, Convolutional Neural Network (CNN) algorithms had improved the accuracy rate in particular for the important task of pedestrian detection [**3**].

The problem of human recognition has been explored by analysis of locomotion properties [**24**], or additional invasive hardware [**16**]. More recently, computer vision algorithms have shown successful result in the recognition of human from faces [**23**]. However, a minimum resolution and sufficient angle of observation is required.

[1] Corresponding Author, Alessandro Moro, Department of Precision Mechanic, Chuo University; E-mail: moro@sensor.mech.chuo-u.ac.jp.

The main contribution of this work is a human recognition system which combines multiple neural networks to discern visitors from accredited people. The system shows high accuracy rate and is able to analyze real-time stream data.

This paper is organized as follow: Section 2 briefly reviews the related work in the area. The system overview and technical details are discussed in Section 3. Finally, the experimental results are demonstrated in 4. Section 5 concludes the paper.

2. Related Work

Accuracy of human and object detection increased significantly (i.e. [**11**, **26**]). In particular the advent of parallel computing offered the possibility to compute complex data and extract information at pixel level. While semantic segmentation is gaining accuracy and importance [**14**], collecting a relative sufficient amount of categories may be expensive [**22**]. Even if instance segmentation has become more common after the work of Hariharan et al. [**13**], and performance increased considerably [**2**], expensive hardware and high computation time are limitations. For specific task, such as human detection and recognition, bounding box segmentation reached high level of performance [**25**], and recently Cao et al. shown a very high speed algorithm for human parts detection [**4**].

Convolutional and recurrent operations became important building blocks for many research applications. For long-range dependency modeling, recurrent operations [**6**, **15**], are the dominant solutions. For image data or temporally local data, accurate results are obtained by convolutional operations [**10**, **21**]. Convolutional network shown high accuracy in face recognition [**28**].

While the reliability of recognition by image analysis is increasing as shown in [**28**], proper facial alignment is an important requirement for accurate results. Face detection has been deeply studied as shown in the milestone work of Viola and Jones [**32**]. More recently, the problem of face alignment has successfully addressed by a cascade of convolutional networks [**35**].

3. Proposed Recognition Method

The baseline of our propose system follows the conventional information retrieve systems. We chose our system by following these considerations: an observed scene contains important information for a limited period of time, and it is necessary to collect the maximum number of frames from the observable period. The whole system is shown in Fig. 1. We propose a distributed system which performs real-time operations and offline operations to take benefit of both high speed, lower accuracy algorithms, and higher precision but slower algorithms.

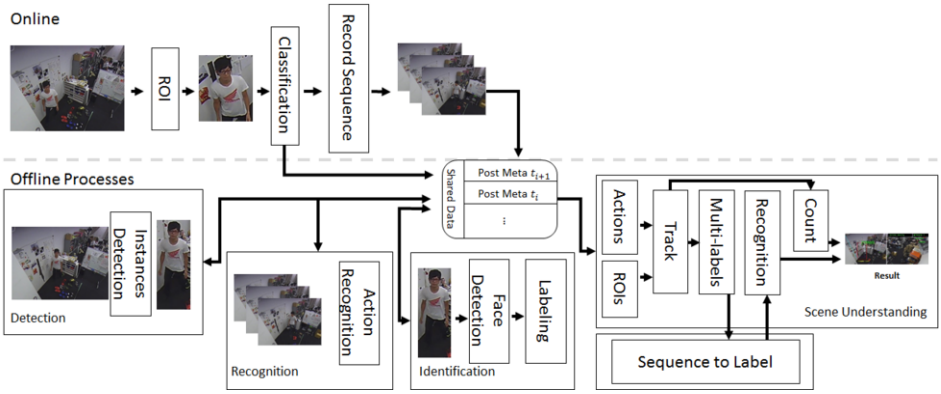

Figure 1. The proposed system pipeline. The data is analyzed temporally from left to right, in real-time (online), or as soon as available (offline).

3.1. Human Detection

Common location for video surveillance cameras allow a clear observation of a human target for a limited number of frames. In order to analyze the largest number of data, we use a fast preprocessing of each captured frames. As shown in [**3**], CNN can be used to perform a pedestrian/human detection with high accuracy rate and low computation time. We opted for a background subtraction algorithm with a Disjoin-set data-structure, to partition nearest blobs and to estimate Regions of Interest (ROIs) in real-time. Each ROI is converted in grayscale and rescaled to a 32x32 pixel size. For each ROI, we perform a dual class (*human/other*) classification. A 3 layers CNN connected by 3 max-pooling layers and a dense layer, all with Rectification Linear Unit (ReLU) activation function are used. An example of input source and topology of the network is shown in Fig.2. Data augmentation has been used to artificially enlarge the training set and reduce the over fitting. We used several image transformation (similar to [30]), to increase our relative small training set to 33000 images from the observed scenario. The classes are equally balanced, batch size of 64, and learning rate of 0.001 for 25 epochs.

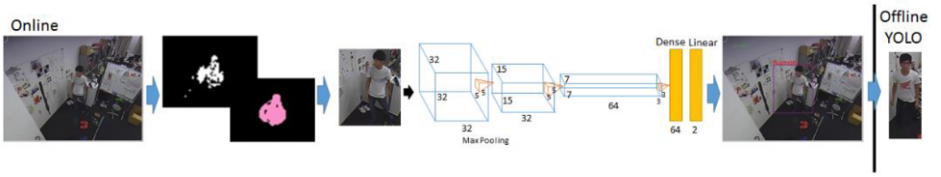

Figure 2. Detected human in the scene.

In our scenario, the video source is acquired at 30Hz. The online preprocessing works at a computation time of about 20ms, which guarantees a real-time acquisition. It is important to avoid loss of frames due to the small number of frames in which valuable information can be extracted.

Since shadow and human persistence in the same position alter the size of ROIs, we use an accurate and fast, but not real-time, algorithm described in [**25**] for an offline analysis. The algorithm is used to refine the regions detected and perform instance segmentation of humans in each frame where a ROI containing a human has been detected.

3.2. Face Detection and Labeling

We rely on a face recognition algorithm as primary method to recognize the people in the scene detected in 3.1. A face may result in partially visible or totally occluded. We consider that even if the face is not completely visible, the region of the head can still contain considerable amount of information. We use a joint detection algorithm [**35**] to align the head when visible and a cascade of CNN [**28**] for recognition (Fig. 3).

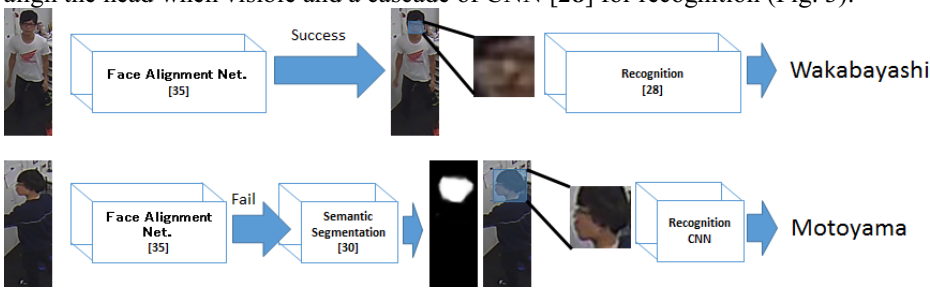

Figure 3. Flow of face alignment and recognition. Detection of the face succeeds and recognition (Top). In case of failure, a head segmentation is performed and a CNN used for recognition (bottom).

We trained the recognition on images of 128x128 pixels without margin, due to the low resolution of the source. We used a set of 3500 images for the training divided in seven different classes. Each class represented a person and about 500 images for each person is used.

We use a semantic segmentation algorithm (Teichmann et al. [**30**]) to find the areas expected to contain side faces and head in the selected ROIs. The network uses a VGG-16 architecture [**27**], and it is refined with 200 manually segmented images from the observed environment. In order to recognize the extracted region, we used a simplified class recognition of popular architecture [**1**]. The architecture is shown in Fig. 4.

Input
Conv + ReLU Kernel: 3x3, channel: 64, padding: 1
Max Pooling (kernel: 2x2, stride: 2)
2 x (Conv + ReLU) Kernel: 3x3, channel: 64, padding: 1
Max Pooling (kernel: 2x2, stride: 2)
Fully Connected + ReLU channel: 512
Dropout (rate: 0.5)
Fully Connected + softmax channel: 3
Linear (channel: # classes)

Figure 4. Convolutional Neural Network Architecture.

We trained the model with a set of about 500 images for each individual for a total of seven people. We performed data augmentation to obtain a training set of 30000 images. A 90/10 training test set shown an accuracy of 98.5% and loss of 0.098.

3.3. Action Recognition

Long short-term memory (LSTM) networks has shown great results in the recognition of sequences, where early inputs remain in memory instead of being forgotten [**15**]. The recognition of an action gives an additional information on the persistence of an object in the memory of the system.

Many highly accurate architectures have been proposed [**34**]. In our scenario, we consider a simple dictionary for the possible actions: *{enter, leave, walk around}*. For this reason, a simple and relatively fast architecture to compute is used (Fig. 5). The architecture has the advantage of use a single stream of images, requiring limited memory and minimal preprocessing.

Deep features from every five frames are extracted using a pre-trained Inception V3 [**29**]. We used the last pool layer of the InceptionV3 network as feature source for the LSTM network.

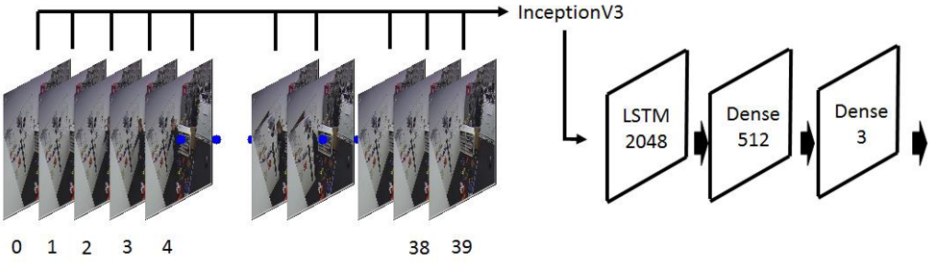

Figure 5. Action recognition from an observed period of time.

We trained the action recognition with a set of 600 sequences of 40 frames for each sequence. We used a set of 200 sequences for each type of action (enter, leave, walk around). Each frame is resized at resolution of 300x300 pixels. With the described architecture an accuracy of 84% is obtained.

3.4. Tracking multiple views

Each human instance detected at time *t* is associated with previously detected object. We opted for a Siamese network solution similar to the idea described in [**19**]. Instead of computing the Euclidean distance as suggested in [**12**], we observed that we could obtain good performances by training a linear regression over the estimated composition of the convolution networks. While other tracking algorithms such as [**8**] can perform well in sequential images, we consider that ROIs can be paired by different space-time information, since a human can be detected in a different position after long time or from different video sources. We associate a tracker for each detected human, and for each new frame, we compare the tracker ROIs' human image with current frame ROIs (Fig. 6). Two convolution layers are used to generate the features, and combined to estimate the similarity between the images.

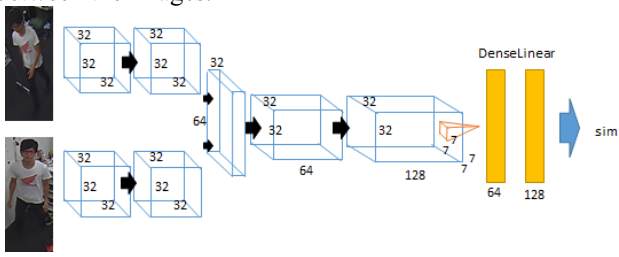

Figure 6. Head detection in unaligned faces.

Our goal is to pair new detected objects with previously tracked objects by estimating the similarity between images. We calculate the similarity between all the tracked objects and new observed objects (1).

$$O_i = \begin{cases} created & if\ |E| < |N| \\ removed & if\ |N| = \emptyset \wedge O_i\ leave \\ j & if\ j \neq \emptyset \wedge \underset{j \in N}{\operatorname{argmax}}\left(sim\left(x_i, x_j\right)\right) \end{cases} \quad (1)$$

where $sim\left(x_i, x_j\right) \in [0,1]$ is the similarity function obtained by the CNN networks, E is the set of existing tracked objects from previous frame, N is the set of new ROIs from the current frame, and O_i the object to track where $O_i \subseteq E$. We trained our model with a set of 60000 different pairs of images from different sequences equally balanced between similar and different. We calculated a RMSE of 0.091. A tracker is not immediately removed if the human tracked disappeared. We chose for a temporal threshold of 1 second to remove lost tracked objects.

The same subject can be visible from multiple views and generate multiple tracked objects. By following the idea of Varga et al. [**20**], we performed a projective transformation to align the human centroids (Fig. 7).

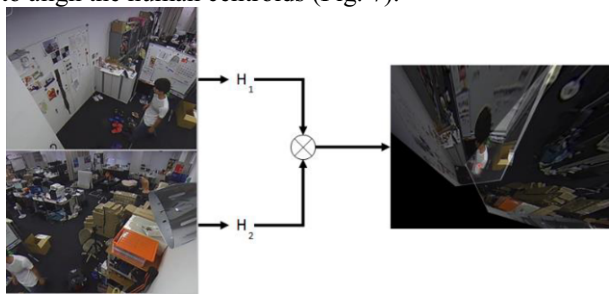

Figure 7. Projective transformation around the common observed area. Body center of mass is coincident.

Two tracked objects are merged together if they share the same centroid projected inside the overlapping areas.

3.5. Sequence Recognition

Sequence of multiple labels of an observed tracked human may contain spurious values. We consider that a failure in the proper recognition is a possible event. However, if multiple observations of the same person are performed, it is possible that the correct label is associated. We propose the use of variable length sequence classification, inspired by the problem of learning from multiple labels [**16**], and by the model described in [**18**].

We consider a tracked object history as a sequence of assigned labels and likelihood associated. Let $\hat{x}_i$ be the input sequence for the i-th training sample, and C_i the associated classes. We define an encoding function φ so that each new instance of a class is a normalized index of the cardinality of C_i (2).

$$\theta_i = \varphi(\hat{x}_i, |C_i|) \quad (2)$$

Our goal is to estimate a set of parameters $\lambda \in \Lambda$ in the class of models M so that we can predict **y** and the likelihood associated $\mathcal{L}$ for a test input $\hat{\boldsymbol{x}}$, where y has the highest probability to be a member of the set *C*.

$$\bar{\lambda} = \underset{\lambda}{\operatorname{argmax}} \prod_i p(y \in C_i | \theta_i, \lambda) \quad (3)$$

Since our training set is limited, we consider the adoption of shown in Gated Recurrent Unit (GRU) [**18**]. These recurrent neural network shown the ability to

converge to the solution with a smaller training set compared to LSTM. The full sequence recognition is shown in Fig. 8.

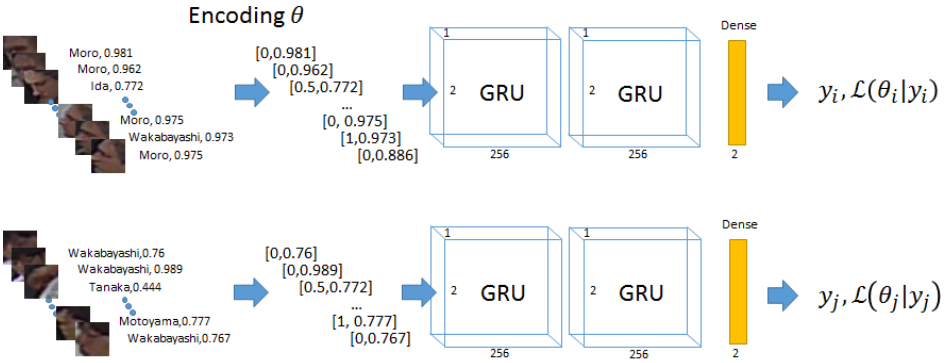

Figure 8. Multiple labels are combined in a sequence to recognize the correct name.

Even if promising results are given by RNN, as shown in [**35**], the network lack robustness in learning optimal label orders. In order to increase the robustness, we perform data augmentation by randomly shuffle of the training sequences. We generated a set of 20000 synthetic sequences and calculated a loss function of 0.02.

3.6. Scene Understanding

The observed scene is described by analyzing of available meta-data collected by the online and offline processes. We use a queue concept to elaborate the content of each frame. Even if partially completed meta-data can be analyzed, we prefer to elaborate only the completed frames meta-data. This solution guarantee a sequential analysis, and an easier evaluation and interpretation. The flow of the data analysis, recognition and counting of human in the environment is shown in Fig. 9.

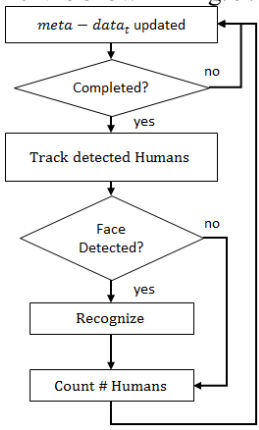

Figure 9. Flow chart of the analysis of the collected data. Only completed data is elaborated. The recognition is update only when new biometric information exist.

Since the detection and recognition from the side of the face is prone to error, we used a weighted value for the recognized sequences:

$$\bar{y} = w_1 y_1 + w_2 y_2$$
$$\bar{\mathcal{L}} = w_1 \mathcal{L}_1 + w_2 \mathcal{L}_2 \quad (4)$$

where $y_1, \mathcal{L}_1$ are the label and likelihood associated to the face recognition based on [2], and $y_2, \mathcal{L}_2$ with the face recognition from side. We consider a person recognized if the likelihood is greater than 0.5.

Recently, a promising counting algorithm based on the direct scene analysis has been described in [**5**]. In indoor environment, however, a person can be invisible to the camera (too far, occlusion) and a single frame analysis is prone to error. We counted the people by detection. Each active tracker is counted as single instance of human (Fig. 10).

Figure 10. An example of described scene. The action and # people in the scene is shown on the top. On the right the recognized human from the sequence. In the region of interest tracked (red), the current associated label (green) with relative likelihood, and similarity (yellow).

4. Experiments

The network models have been implemented in Tensorflow r1.5, Keras 2.1.4, and Cognitive Network Toolkit (CNTK) v.2.4. All models were trained only once and used for all result throughout the paper. We performed a specialized training for our networks with the exception of the offline human detection based on YOLO dataset. We used a GPU GeForce GTX 1080 Ti, and a GeForce GTX 980 for our experiments. Computation time for each frame is measured as follow: Human Detection based on [25] about 300ms, face detection and recognition about 200ms each. Semantic segmentation about 300ms, and 200ms for the action recognition. The proposed algorithm elaborates all the captured frames without frame loss. Offline analysis returns an accurate result with the output that has an incremental delay directly proportional to the number of frames collected. A normal observed sequence has an output delay of about 5 to 10 seconds.

To evaluate our proposed framework, we consider a laboratory as restricted areas. We used a set of 20 videos which do not belong of the training set for the evaluation. We calculated the correctness of the face recognition algorithm and the recognition with multiple labels on a total of 4317 frames where at least a person has been detected.

Table 1. The following table summarize the result of the framework parts used to describe the scene.

Recognition Method	True Positive (TP)	False Positive (FP)	False Negative (FN)
Face Detection [35]	57.46%	42.54%	0%
Face Recognition [28]	81.7%	18.3%	0%
Multi-Label (proposed)	90.63%	5.96%	3.41%
Count (proposed)	94.14%	3.81%	2.05%

Results in Table 1 show that the proposed algorithm increased the recognition performance by combining sequential information and recognition of human from images captured from side face.

We compare our proposed algorithm with the result obtained by using only a face recognition [28]. Since biometric information are available only when the face is clearly visible, we included the detection rate obtained with an angled image (Fig. 11). We experienced that the success in the recognition is strongly influenced by the correct detection of at least one face.

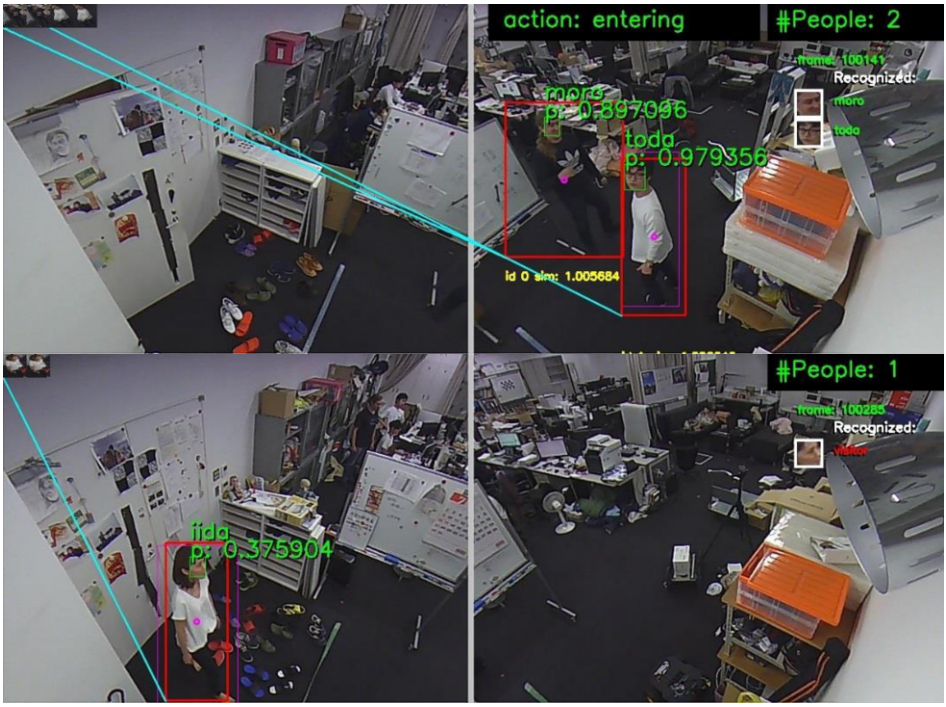

Figure 11. Example of successful detection, recognition, and counting of the observed scene. On the bottom, the subject is not in the training model, and she is correctly labelled as visitor.

5. Conclusion and Future Work

In this work, we studied the problem of recognizing and counting the number of occupant in a restricted area. We proposed a framework for analysis of an observed environment which combines the advantages of online and offline algorithms. We described a multi-label recognition to increase the robustness of recognized people in the scene. This framework is based on multiple neural networks which summarize the flow of each process activity at each frame. The proposed framework has the advantage that can be easily extended and distributed on multiple machines.

References

[1] Y. Abouelnaga et al., CIFAR-10: KNN-based Ensemble of Classifier, International Conference on Computational Science and Computational Intelligence (2016)
[2] A. Arnab and P. Torr, Pixelwise Instance Segmentation with a Dynamically Instantiated Network, CVPR(2017)
[3] A. Angelova et al., Real-Time Pedestrian Detection With Deep Network Cascades, BMVC(2015).
[4] Z. Cao et al., Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, CVPR(2017)
[5] P. Chattopadhyay et al., Counting Everyday Objects in Everyday Scenes, CVPR(2017).
[6] Cho et al., Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, EMNLP(2014).
[7] N. Dalal and B. Triggs, Histograms of Oriented Gradients for Human Detection, CVPR(2005).
[8] M. Danelljan et al., Accurate scale estimation for robust visual tracking, BMVC(2014).
[9] K.-I. Funahashi and Y. Nakamura, Approximation of dynamic systems by continuous time recurrent neural networks, Neural Netw.(1993) vol.6, 801-806.
[10] K. Fukushima and S. Miyake, Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition, Competition and cooperation in neural nets (1982).
[11] R. Girshick et al., Rich feature hierarchies for accurate object detection and semantic segmentation, CVPR (2014).
[12] R. Hadsell et al., Dimensionality Reduction by Learning an Invariant Mapping, CVPR(2006)
[13] B. Hariharan et al., Simultaneous detection and segmentation, ECCV (2014), 297-312.
[14] K. He et al., Mask R-CNN, ICCV(2017).
[15] S. Hochreiter and J. Schmidhuber, Long short-term memory, Neural computation (1997).
[16] R. Jin and Z. Ghahramani, Learning with Multiple Labels, NIPS(2002).
[17] N. Karimian et al., Human recognition from photoplethysmography (PPG) based on non-fiducial features, Acoustics, Speech and Signal Processing (2017).
[18] S.-Y. Kim et al., Multi-Label Learning with the RNNs for Fashion Search, ICLR(2017)
[19] G. Kock et al., Siamese Neural Network for One-shot Image Recognition, International Conference on Machine Learning(2015).
[20] A. Krizbeysky et al., Imagenet classification with deep convolutional neural network, Proc. Adv. Neural Inf. Process Syst. (2012), 1097-1105.
[21] Y. LeCun et al., Backpropagation applied to handwritten zip code recognition, Neural computation (1989).
[22] T.-Y. Lin et al., Microsoft COCO: Common objects in context, ECCV(2014).
[23] O. M. Parkhi et al., Deep Face Recognition, BMVC(2015).
[24] C. Prakash. et al., *A framework for human recognition using a multimodel Gait analysis approach*, Computing, Communication and Automatica (ICCCA) 2016.
[25] Redmon et al., YOLO9000: Better, Faster, Stronger, CVPR(2017)
[26] S. Ren et al. Faster R-CNN: Towards real-time object detection with region proposal networks, NIPS(2015)
[27] K. Simonyan et al., Very Deep Convolutional Networks for Large-Scale Image Recognition, ICLR(2015)
[28] F. Schroff et al., *FaceNet: A Unified Embedding for Face Recognition and Clustering*, CVPR (2015), 815-823.
[29] C. Szegedy et al., Rethinking the Inception Architecture for Computer Vision, CVPR(2015).
[30] M. Teichmann et al., MultiNet: Real-time Joint Semantic Reasoning for Autonomous Driving, arXiv:1612.07695(2016)
[31] D. Varga et al., A multi-view pedestrian tracking method in an uncalibrated camera network, ICCVW(2015)
[32] P. Viola and M. J. Jones, Robust real-time face detection, International Journal of Computer Vision **57** (2004), 137-154
[33] J. Wang et al., Cnn-rnn: A unified framework for multi-label image classification, CVPR(2016), 2285-2294.
[34] A. Ullah et al., Action Recognition in Video Sequences using Deep Bi-Directional LSTM with CNN Features, IEEE Access 6: 1155-1166 (2018)
[35] K. Zhang et al., *Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks*, IEEE Signal Processing Letters (2016), vol. 23, no. 10, 1499-1503.