Neurocomputing 272 (2018) 647–659

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Appearance based pedestrians' head pose and body orientation estimation using deep learning

Mudassar Raza, Zonghai Chen*, Saeed-Ur Rehman, Peng Wang, Peng Bao

Department of Automation, University of Science and Technology of China (USTC), Hefei, 230027, PR China

ARTICLE INFO

Article history:
Received 4 October 2016
Revised 6 June 2017
Accepted 13 July 2017
Available online 21 July 2017

Communicated by Guan Ziyu

Keywords:
Convolutional neural network (CNN)
Full-body orientation
Head-pose
Pedestrians
Proposed training dataset

ABSTRACT

Pedestrian orientation recognition, including head and body directions, is a demanding task in human activity-recognition scenarios. While moving in one direction, a pedestrian may be focusing his visual attention in another direction. The analysis of such orientation estimation via computer-vision applications is sometimes desirable for automated pedestrian intention and behavior analysis. This paper highlights appearance-based pedestrian head-pose and full-body orientation prediction by employing a deep-learning mechanism. A supervised deep convolutional neural-network model is presented as a deep-learning building block for classification. Two separate datasets are prepared for head-pose and full-body orientation estimation. The proposed model is subsequently trained separately on the two prepared datasets with eight orientation bins. Testing of the proposed model is performed with publicly available datasets, as well as self-taken real-time image sequences. The experiments reveal mean accuracies of 0.91 for head-pose estimation and 0.92 for full-body orientation estimation. The performance results illustrate that the proposed approach effectively classifies head-poses and body orientations simultaneously in different setups. The comparison with existing state-of-the-art approaches demonstrates the effectiveness of the presented approach.

1. Introduction

Automated human activity recognition (HAR) based on visual equipment is becoming the epicenter of current research. Head and body poses provide hints about human behaviors and intentions. Orientation analysis, a co-area of HAR, is used to estimate an object's direction of motion and attention. Some of the many potential areas of application include robotic surveillance, implementing security of no-pass-through areas, intelligent driver assistance through the observation of pedestrians' movements while crossing roads, observing people watching advertisement boards and marketing stands to gain automated information and determine current trends in people's interests. Human body orientation provides hints about a person's behavior; for example, it can indicate where the person is going and in which direction he is looking. Head-pose and the direction of a pedestrian's movement are in some cases less associated. A pedestrian moving in one direction can direct his visual attention in another direction. Additionally, the main challenge in the automated pedestrian-orientation detection of both head and body is a dealing with very-low-resolution images. Estimation techniques based on facial and body features are not suitable in such scenarios. Additionally, feature extraction for varying appearances, such as different clothing, becomes very difficult.

Hand-crafted features perform well when there is not enough data available to extract features. Apart from hand-crafted features [1–5], deep learning solves the abovementioned problems by automatic feature extraction in an end-to-end learning mechanism. CNN-based deep learning has proven to be very competitive for a variety of classification tasks, such as character recognition [6], object detection [7–9], object recognition (ImageNet) [10], pedestrian detection [11] and traffic-sign classification [12].

The proposed approach follows CNN as a deep-learning tool to classify head and body appearances. The CNN-based approach overcomes various performance issues. The main contributions of this manuscript are as follows:

i. CNN is used as a building block to represent pedestrian head-pose and body-orientation classes with low-resolution images.
ii. The proposed system is an appearance-based full-body-orientation estimation and head-pose estimation approach and it is applicable to both still images and image sequences.
iii. CNN requires a huge number of images for the learning step. Therefore, two separate big datasets for head-pose and body orientation are prepared to employ deep learning.

* Corresponding author.
E-mail address: chenzh@ustc.edu.cn (Z. Chen).

http://dx.doi.org/10.1016/j.neucom.2017.07.029
0925-2312/

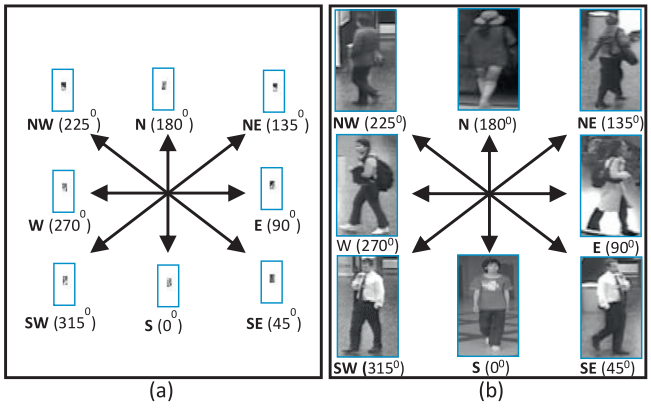


Fig. 1. Eight different representations of pedestrians' in terms of (a) head-pose and (b) full-body orientation.

iv. Only grayscale images from 2D cameras are considered as input to the proposed model. Additionally, the time dimension is not taken into account while processing video sequences; however, computation times for CNN predictions are provided.
v. Promising classification results are achieved, which are compared to current state-of-the-art approaches.

The manuscript is organized as follows: Sections 1 and 2 consist of the introduction and literature review, respectively. Section 3 describes the materials and method for representing the proposed system. This section also highlights the architecture and configuration of the proposed convolutional neural network. Section 4 discusses the proposed big datasets for orientation classification and head-pose estimation, which will be helpful for deep learning. Section 5 describes experiments and their results. Section 6 presents the manuscript's conclusions.

2. Related work

Few approaches in the literature consider both head-pose and full-body orientation at the same time; instead, researchers utilize either head-pose estimation or body-orientation extraction. The discussion below highlights the current state-of-the-art approaches found in the literature for determining pedestrian direction of attention and direction of motion.

Various authors characterize pedestrian head or full-body direction with numerous types of orientation-bin configurations, such as frontal bins (deal with frontal views only) [13], three bins (left, right and front/back) [14], four bins (left, right, front and back) [4,14–18], and eight orientation bins (0°–315°, each bin has an edge distinction of 45° with its neighboring bin, as shown in Fig. 1) [19–22].

Several authors also consider only the upper body for orientation analysis [23–25]. To cope with the challenges of classifying pedestrian direction, most methodologies use traditional hand-crafted features, such as local binary patterns (LBP) [26], mean energy features [27], silhouette features [28], scale-invariant feature transform (SIFT) features [29], histogram of oriented gradients (HOG) [14,16,30–32], discrete cosine transform HOG (DCT-HOG) [16], aggregated channel features (ACF) [24] and sparse representations (SR) [33]. The classifiers used in combination with these features are support vector machines (SVM) [1,17,33,34], random forest [16,25,32], extremely randomized trees [35], census transform (CT), R-transforms [27], principal component analysis (PCA) [30], linear sum [31] and ExtRaTree [28,36]. Apart from the abovementioned methods, some other methods such as probabilistic methods [15,32], geometrical shape-based methods [18], part-based methods [26,37], template matching [38], appearance-based methods [19] tracking-based methods (the Kalman-filter approach [39] and the particle-filter approach [33,40,41]) have also gained the attention of researchers for orientation analysis. Different methodologies use diverse image-acquisition devices such as stereo-vision cameras, 3D cameras, RGB-D cameras, monocular cameras and overlapping cameras. The images from these sources are in different formats; therefore, different techniques are used to process them. A brief description of some recent approaches is given in the following paragraphs.

Rehder et al. [26] utilize part-based classification to classify head-orientation, and discrete orientation classification is used over local binary pattern (LBP) features. A syntactic post-processing part-based dictionary algorithm [37] is used in aware vehicle systems to reduce the number of accidents between humans and vehicles. The framework is constrained to training ranges with all pedestrians dressed uniformly. Gandhi, T. and M.M. Trivedi [34] introduce a single-image-based pedestrian safety system that predicts the facing directions of pedestrians. The authors utilize SVM to predict pedestrian orientation, which aids to lower the odds of a crash between the vehicle and walker. A decision tree consolidated with SVM [17] is likewise utilized to estimate pedestrian direction. Vishwakarma et al. [27] convert human images to mean energy-silhouette images. The R-transforms are then applied to extract direction information and multi-class SVM is applied thereafter for further characterization. Tao and Klette [16] consider eight bins of pedestrian direction. The method uses discrete cosine transform HOG (DCT-HOG) features in combination with random decision forest as a classifier. Regression with supervised learning [42] is used to solve the problem of orientation estimation on images acquired from a 3D range camera. In another methodology by Liem and Gavrila et al. [43], pedestrian orientation is estimated by finding the gap between a learned texture model and the original 3D shape. Tangent space with multi-class LogitBoost [44] is applied to single images in the Diamler Chrysler dataset. Fitte-Duval et al. [24] compare their work based on multiscale variants of ACF with the multi-level HoG-feature approach and claim ACF:GM+HoG+LBP (MACF) as their best strategy. Chen et al. [33] highlight multi-level HOG features along with a sparse representation (SR) approach with SVM-, SVM-adj- and MultiSVM-based methods. The input tracks are provided by a particle-filtering-based tracker. Ardiyanto and Miura [25] implement various features and classifiers for orientation analysis and find the block importance feature model of partial least squares with random forest (BIFMS-PLS-RF) classifier to be the superior classifier. Baltieri et al. [35] experiment on distinctive variants of their proposed approach, named mixture of approximated wrapped Gaussian (MoAWG), and claim the HoG-based extremely randomized trees and mixture of approximated wrapped Gaussian (MoAWG: HoG - ERT -AWG) method as their best approach. Schulz et al. [45] use head detection and single-frame-based pose estimation with modified CT as a classifier. In another work by Schulz et al. [41], instead of using a single-frame approach, the authors use a particle-filtering tracking mechanism over time for head-pose estimation.

Recently, significant progress has been made in the area of learning with automatic feature extraction by employing deep-learning strategies. Through an extensive internet search, we are unable to find any appropriate deep-learning approach in our domain. Several deep-learning approaches have gained attention for pedestrian body-pose estimation [46–52]; these approaches, however, do not cover pedestrian body-orientations. Although some head-pose estimation methods are found that employ deep-learning strategies [13,53,54], those methods are based on face images with good resolution and visibility and only consider frontal face images. B. Ahn et al. [13] use deep neural network (DNN) for head-orientation classification. However, these researchers in-

Table 1
Mathematical notations used in this manuscript.

Notation	Description	Notation	Description
$\otimes$	Convolution operation	V'	Number of kernels in a bank of filter for a layer L
I	Input vectors	V	Number of color channels
L	Layer number	$r_L,\ c_L,\ v,\ v', i, j$	Indexing variables
$L+1$	Next layer	s	Stride
R	Total number of rows	I_i	i^{th} feature map
R_L	Total number of rows of input vectors at layer L	k	Height and width of non-overlapping region for pooling operation
R_F	Total number of rows of kernel vectors at layer L	P_i	Result of*max*-pooling operation
C	Total number of columns	ρ	Dropout rate
C_L	Total number of columns of input vectors at layer L	φ	Softmax probability
C_F	Total number of columns of kernel vectors at layer L	F_i	Corresponding kernel vector of i^{th} neuron
F	Bank of kernels/ Filters	N	Total number of neurons/ Classes
V'	kernel	η	Learning Rate
ε	Cross-entropy loss	j	Batch number
TP	True positives	FP	False positives
TN	True negatives	FN	False negatives
Prc.	Precision	TPR	True positive rates
FPR	False positive rates	M	Confusion matrix
t	Total number of classes		

vestigate frontal face images only with good resolution. Bao and Ye [54] presented a head-pose estimation method that employs a CNN-based approach, but their scope is limited to frontal images with good visibility and yaw and pitch movements only. J. Choi et al. [55] employ lightweight CNN for a human body-orientation problem and claim 81.58% accuracy on high-resolution 3D human pose images from the human3.6 M dataset [56]. Another approach related to convolutional neural networks (CNN) and orientation estimation is proposed by Wagner et al. [57], but their work is based on infrared-camera images of animals (not pedestrians), especially deer images, with three orientation views, i.e., left, right, and front/back.

This work focuses on images taken at far distances (low-resolution images), and both the body and face images usually have poor visibility. A more appropriate work related to our domain is presented by Rolf et al. [21], which employs deep-belief networks (DBN) for head-pose extraction. The images are first histogram equalized and then provided to a network model for classification.

Through a review of the available literature, we have come to understand that for pedestrian images the background should be removed correctly. After background removal, hand-crafted features can perform effectively. Abundant research is available on background removal, but there is still room for improvement. Our research is focused on both still images as well as sequences of images of pedestrians that are acquired from distant cameras. Therefore, there is the possibility of complex backgrounds and blurred images (because of the quality of the cameras and the distance of the object from the camera, the image quality is affected). For automatic feature-learning mechanisms (like CNN features), we do not need to remove the background. All that we need to do is provide enough data, from which the deep-learning model learns potential features. These points raise our motivation to select CNN features for our problem.

3. Convolutional neural network (CNN)

The proposed method employs CNN to perform pedestrian orientation and gaze-direction classification. Fig. 1 shows the eight different pedestrian head-orientations (Fig. 1(a)) and body-orientations (Fig. 1(b)) on which we focus in this paper. All the images in the dataset are first normalized to the same format. Images in the training datasets are then divided into training and validation sub-datasets. Afterwards, all images are labeled with their respective orientations to perform supervised learning. These two sub-datasets are then passed to the proposed deep CNN model for training. The testing is then performed, in similar fashion, to check the performance of the system. The test images are again normalized and then passed to the trained classifier for prediction. The block diagram of the proposed system is shown in Fig. 2.

The proposed CNN architecture consists of a hierarchical structure containing several types of layers attached one after another. Each layer that processes the input vectors outputs feature vectors or feature maps. In addition to the input layers, the main layers used in the proposed architecture are convolution, down-sampling, regularization, dropout, fully connected and softmax layers. Table 1 contains the mathematical notations used in this manuscript.

In the *Convolution layer*, the set of input vectors I at layer number L (assume each input is a vector with order 3 tensor having dimension $R_L \times C_L \times V$) is convolved with a bank of kernels F having V' kernels (each kernel is assumed to have dimension $R_F \times C_F \times V$), where, R and C are the row and column sizes of the vector. The term V stands for the color channels. Kernel bank F, as a whole, is an order 4 tensor with size $R_L \times C_L \times V \times V'$. The proposedmodel is designed for grayscale images. Therefore, in this case $V = 1$ for initial image inputs. The expanded form of the convolution (Eq. 1) step is expressed in as:

$$c_{r_{L+1},c_{L+1},v'} = \sum_{r_L=0}^{r_L<R_L} \sum_{c_L=0}^{c_L<C_L} \sum_{v=0}^{v<V} F_{r_L,c_L,v,v'} \otimes I_{r_{L+1}+r_L,c_{L+1}+c_L,v} \tag{1}$$

where, r_L, c_L, v and v' are indexing variables, $L+1$ represents the next layer and $\otimes$ represents the convolution operation. The convolution of an order 3 tensor input, as mentioned above, with an order 4 set of tensor kernels produces output of size $(R_L - R_F + 1) \times (C_L - C_F + 1) \times V'$. If it is desired for the input and output images to have the same dimensions, a hyper-parameter named*zero-padding* (adding additional rows or columns containing zero values) is normally used. Padding $\lfloor \frac{R_L-1}{2} \rfloor$ rows above and $\lfloor \frac{R_L}{2} \rfloor$ rows below all input vectors results in the same number of rows for all feature vectors. Similarly, padding $\lfloor \frac{C_L-1}{2} \rfloor$ columns above and $\lfloor \frac{C_L}{2} \rfloor$ columns below all input vectors results the same number of columns for all feature vectors.

Another hyper-parameter named *stride* ((s) indicates that while sliding the kernel over the input vector, s cells at a time are skipped. This results in output of a reduced size.

The *Pooling layer* down-samples the input vectors [57] by reducing the spatial resolution. The proposed architecture employs

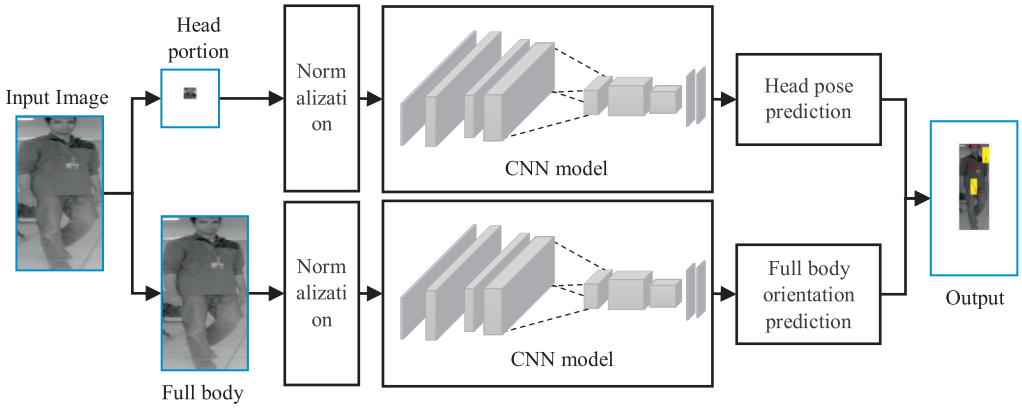


Fig. 2. Block diagram of the proposed orientation-classification scheme through CNN. The final step of the training phase is a trained CNN classifier, which is used in the testing phase for predictions.

max-*pooling*, which is mathematically expressed as

$$P_i = \max_{k\times k} I_i \tag{2}$$

where I_i is the ith feature map, which is divided into $k \times k$ non-overlapping regions. P_i represents the pooling result of size $k \times k$, which contains the maximum values over the corresponding regions.

The *Rectified Linear Units (ReLUs) layer* [10] enhances the non-linear characteristic of the decision function. For input I, the activation function in terms of ReLU is represented as

$$f(I) = \max(0, I) \tag{3}$$

The *fully connected layer* is usually prone to overfitting. The *Dropout layer* is also used to avoid overfitting. The nodes are dropped out with probability $1 - \rho$ (where ρ is the dropout rate) during each training stage. The *Softmax layer* is used as a linear classifier in the last layer of CNN and is mathematically expressed as

$$\varphi(I_i) = \frac{e^{F_i I_i}}{\sum_{j=1}^{N} e^{F_j I_i}} \tag{4}$$

where I_i is the input vector and F_i is the corresponding kernel vector of the ith neuron. N is the total number of neurons corresponding to N classes. The output $\varphi(I_i)$ is the probability of the ith input class. The predicted class contains the highest-probability value.

3.1. Proposed CNN model architecture

The layers described in the previous sub-section are combined one after another to form a hierarchy of feature-extraction operations along with a fully connected neural-network classifier. The network parameters (padding, stride, number of filters, filter-bank sizes, dropout rate) initially follow ImageNet classification values [10]. Afterwards, considering the limited hardware resources, these values are updated for the proposed model after extensive experimentation and fine-tuning. The detailed block diagram of the proposed CNN architecture is shown in Fig. 3.

Circles represent operations of the layers, while labels in the circles indicate the type of operation performed. Rectangular and cubic boxes represent input/output vectors and the corresponding labels indicate the dimensions of the vectors, i.e., rows, columns and depth. Five convolutional layers (C1–C5) are used in total. To maintain the same resolution of input and output vectors, zero padding is applied to all convolutional layers (except layers C5 and FC). The input image size is set to $64 \times 64 \times 1$, where 1 represents a single channel of the image out of the R, G, and B channels. Layer C1 is employed as the first layer of the CNN. This layer convolves the input image with 25 kernels of size 11×11 to form 25 feature maps. Layer C2 consists of 50 feature maps with 9×9 filter size. The output size of the feature maps is $32 \times 32 \times 50$. Layer C3 uses a filter bank containing 75 kernels, each of size 7×7. Layer C4 convolves 75 input vectors with 100 kernels, each of size 5×5. Padding is used in layers C1–C4 to maintain the resolution of the input vectors and output feature maps. Padding, however, is not utilized in layer C5. Layer C5 employs a filter bank of 500 kernels, each of size 3×3. *Max-pooling* is used after all convolutional layers except C5. Pooling layers P2 to P4 employ a 2×2 pooling operation with zero padding while P1 uses 3×3 pooling with single padding on all sides of each input vector. Stride 2 is used on all pooling layers to down sample the input vectors by a factor of 2.

To avoid overfitting, dropout layer D is applied before the fully connected layer. Layers R1 and R2 use ReLU transforms to introduce non-linearity to the linear mappings. The Softmax layer is applied last to perform prediction.

Supervised learning is performed with stochastic gradient descent as an optimization method. The input samples are divided into small batches, which allows them to be processed in parallel on GPU. This speeds up the training process. After computing each batch, weights are updated as follows:

$$V'_{j+1} = V'_j - \eta \frac{\partial \varepsilon}{\partial V'} \tag{5}$$

where j is the batch number, V' is the kernel, η is the learning rate, and ε is the cross-entropy loss. The term $\frac{\partial \varepsilon}{\partial V'}$ represents the gradient, which is averaged over j.

Most of the time, a pedestrian is moving in one direction, while his visual focus is in another other direction. Therefore, this paper is focused on classifying two attributes of pedestrians, i.e., pedestrian orientation classification to predict the direction of motion and head-pose estimation to predict the direction of attention. The proposed CNN architecture is trained separately for the two attributes. Hence, the pedestrian's whole-body image is used as an input for full-body orientation classification, while the upper portion is used as input for classifying the visual-attention direction (see Fig. 2).

4. Training dataset preparation

A large dataset is required to perform deep-learning tasks. To the best of our knowledge, there is no such large dataset available for pedestrian orientation classification and head-pose estimation. Therefore, images are collected from various available person re-identification datasets. The targeted datasets are VIPeR [58], CAVIAR4REID [59], ETHZ [60], iLIDS-VID [61], QMUL under Ground *Re*-IDentification (GRID) [62], CAMPUS-Human dataset for human re-identification [63] and Market-1501 [64]. The selected full-body

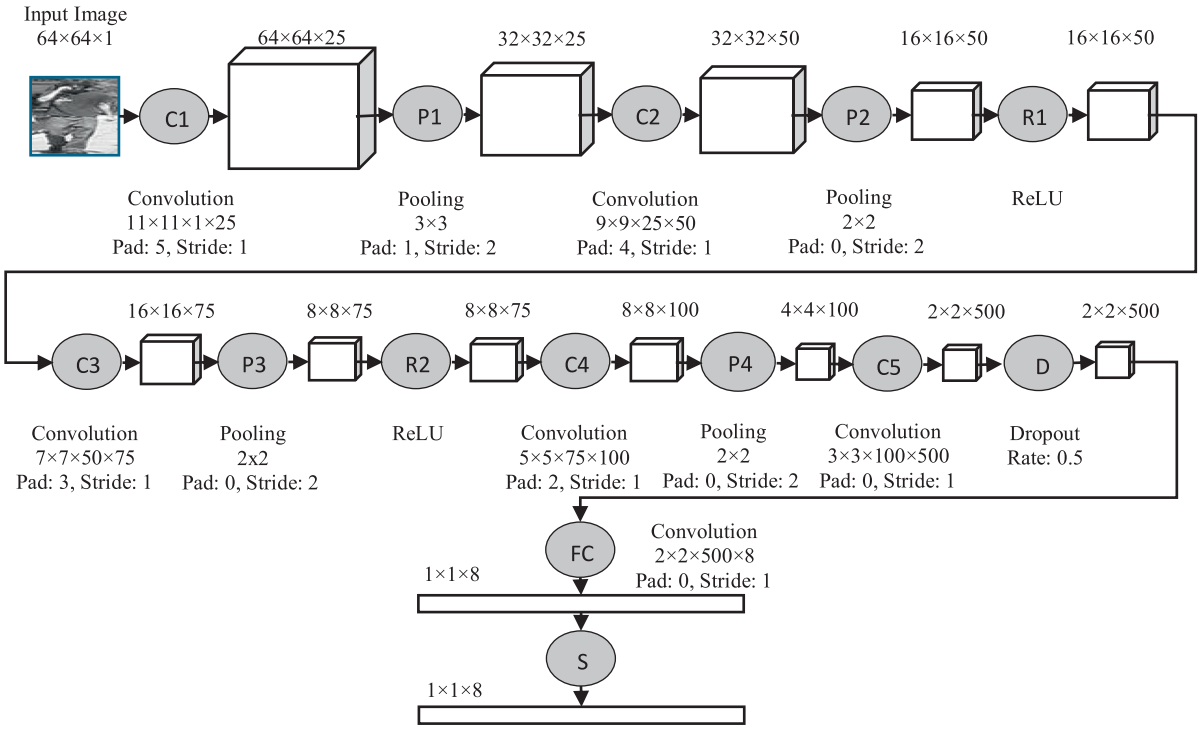


Fig. 3. Layered architecture of proposed CNN, where, C1 - C5 are convolutional layers, P1 – P4 represent pooling (max) layers, R1 – R2 symbolize ReLU layers, D is the dropout layer, FC is the fully connected layer and S represents the Softmax layer.

Table 2
Number of unique images in the proposed dataset.

Orientation/Head-pose angle	Number of full body images (BDBO)	Total Number of Head Images (BDHP)
0°	4438	4214
45°	4330	4204
90°	4112	4058
135°	4624	4266
180°	4648	4318
225°	4624	4446
270°	4112	4046
315°	4330	4282
Total	**35,218**	**33,834**

images are annotated and labeled with pedestrians' orientations. The dataset for head-pose is prepared by cropping the upper-body portion containing the head. The images are separated according to orientation, annotated and labeled well thereafter. The two datasets are then named "big dataset for body-orientation (BDBO)" and "big dataset for head-pose (BDHP)". The total numbers of unique images collected are 17,609 for BDBO and 16,917 for BDHP.

4.1. Dataset augmentation

To increase the size of a dataset, augmentation is applied. Images are mirrored first and then added to the opposite class; for example, a body image oriented at 90° is flipped and then added to the class of images with orientation of 270°. This technique doubles the size of the desired dataset (total BDBO images = 35,218 and total BDHP images = 33,834). Additionally, each RGB channel of each image is treated as a separate image, which increases the size of the dataset by up to threefold. Table 2 shows the total number of unique images per class for both datasets. These abovementioned datasets contain the different qualities of images taken in various indoor and outdoor environments. The main aim of collecting these images is to prepare a challenging dataset for deep learning.

5. Results and discussions

The experiments are performed on a computer with Intel core i5 2.3 Ghz CPU with Nvidia Geforce GTX 750Ti GPU having 2GB memory and computing capability 5.0. Training on the proposed datasets is executed on mini-batches, each of size 100, to decrease GPU memory usage. The learning rate is selected as 0.001. The weight decay and momentum values are set as 0.0005 and 0.9, respectively. The total number of epochs selected is 50, although the network becomes stable after 30 epochs most of the times during training experiments.

We first fine-tune the proposed model and perform learning. Afterwards, we assess our model based on different approaches.

5.1. Fine tuning for finalizing the networkdesign

The performance of the CNN network depends on various factors such as augmenting the dataset (increasing or decreasing the sample sizes), the number of layers, values of network parameters, depth of the filters, and sequences of the layers. To the best of our insight, there is no principal method available to define a suitable CNN model for a specific problem. The traditional scheme is to study the similar approaches with good results and develop an improved model through experiments. In this work, many experiments are preformed to finalize the proposed model. These experiments are based on varying the number of CNN layers, dataset augmentation and changing the hyper parameters (padding, stride, number of filters and their depths) values. The values for the learning rates, weight decay, momentum, and epochs (as mentioned above) are kept unchanged for all experiments. Moreover, we design several CNN based models with a different number of layers. Fig. 4 illustrates the effects on validation error (VE) of some of the major experimentswith these models during the learning stage.

Train and val errors in Fig. 4 represent the highest-scoring predictions of training and VEs, respectively, while train-5 and val-5 represent top 5 scoring predictions. The most important among

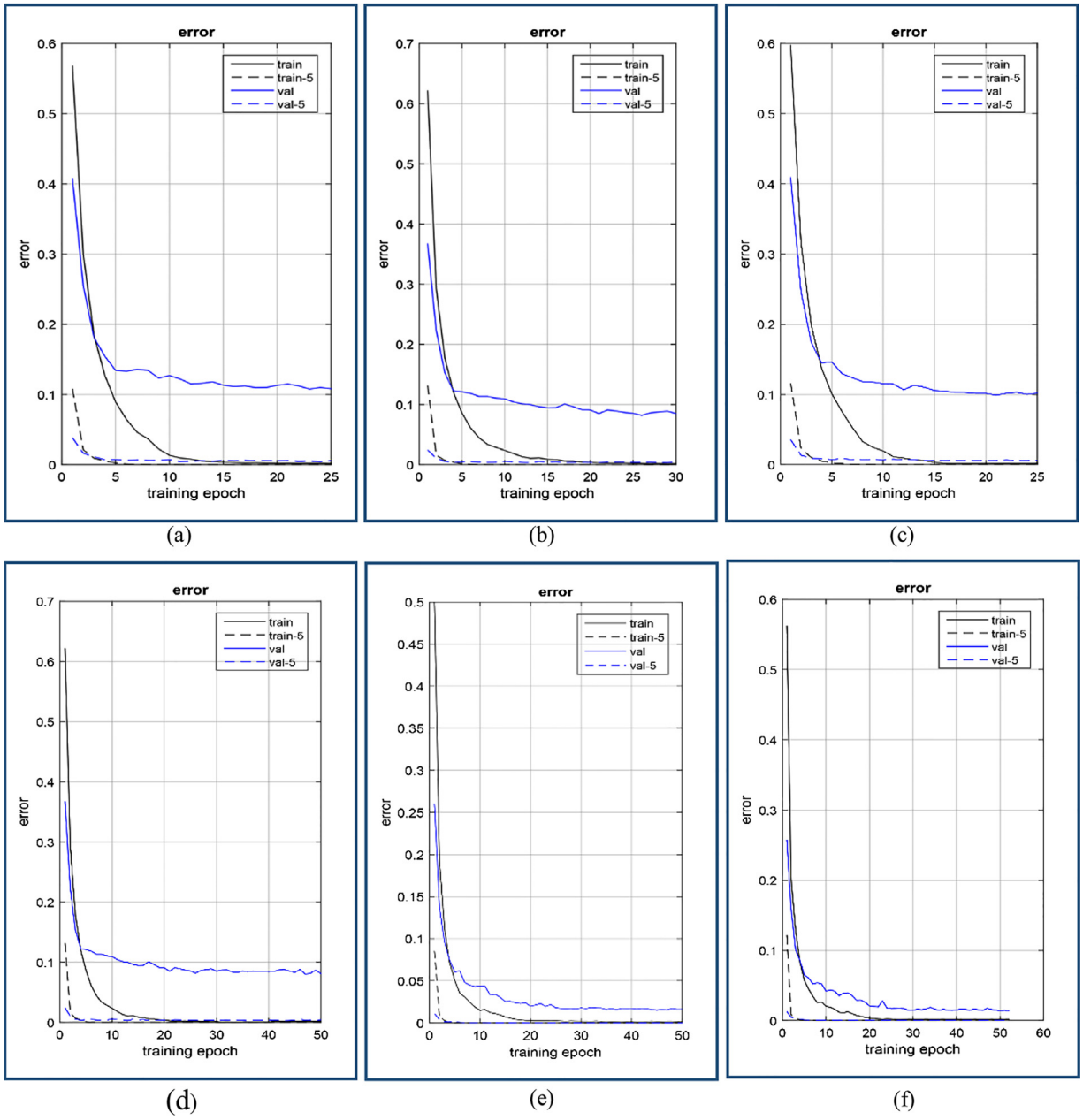


Fig. 4. Fine-tuning of the proposed CNN model: (a) training with a twelve-layer model, (b) training with a thirteen-layer model, (c) training with a fifteen-layer model, (d) training with an augmented dataset (BDBO), (e) training for full-body orientation with a fourteen-layer model with BDBO dataset, (f) training for head-pose with a fourteen-layer model with BDHP dataset.

them is VE. When the CNN network is trained with twelve layers on BDBO as a deep-learning training dataset, VE is reduced to 0.107 (Fig. 4(a)). By using thirteen layers, VE is further reduced to 0.080 (Fig. 4(b)). With fifteen layers, VE is increased to 0.099 (Fig. 4(c)). Furthermore, after applying dataset augmentation, VE decreases to 0.075 (Fig. 4(d)). Fig. 4(e) represents the minimum VE achieved (VE = 0.015) by the final selected CNN model, which has fourteen layers, including a dropout layer (with dropout rate $\rho = 0.5$). The finalized model with fourteen layers is trained with BDHP as well for head-pose classification (VE = 0.017); see Fig. 4(f).

Table 3 illustrates the training results for some of the experiments with various CNN models and their architectures (number of layers and the sequence of fundamental layers). These models are first trained on the non-augmented images of the BDBO dataset and then these models are again trained on the augmented images of the BDBO dataset. It is observed that by increasing the number of images (augmentation) VE is reduced. The models are also designed with a lower number of layers. The training with these lower number of layers results in higher validation errors. The addition of more layers results in lower VE. Furthermore, models with more than fourteen layers also resulted in relatively higher VEs. The least VE is achieved for the fourteen-layered model.

Moreover, different experiments are also performed by adjusting the values of different hyper parameters (padding, stride, number of filters and their depths). The illustration of all these experiments is difficult to mention here.

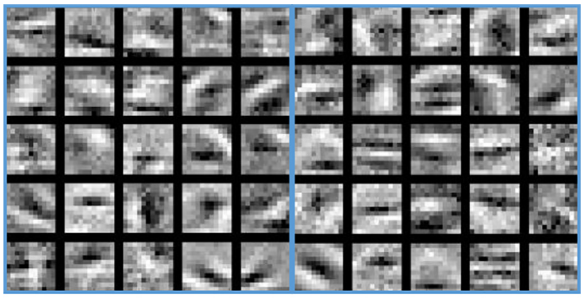
Table 4 provide an example of the effect on validation error with varying depth of the filters at layer C1 of the proposed CNN model. The results reveal that the model with filter depth 25 shows the least VE. Through the variety of experiments, the final fine-tuned model (Fig. 3) is selected for further performance evaluations.

Table 3
VEs with different CNN models on BDBO dataset.

Total Layers	VE (Without augmented dataset)	VE (With augmented dataset)	Network Layers sequence
6	0.376241	0.295711	C1+P1+C2+P2+FC+S
8	0.253431	0.191643	C1+P1+R1+C2+P2+D1+FC+S
10	0.175425	0.091607	C1+P1+C2+C3+P2+C4+R1+D1+FC+S
12	0.189953	0.107169	C1+C2+P1+R1+C3+C4+P2+C5+R2+D1+FC+S
13	0.113270	0.080001	C1+P1+C2+P2+R1+C3+P3+R2+C4+P4+C5+FC+S
14	0.064901	0.015152	C1+P1+C2+P2+R1+C3+P3+R2+C4+P4+C5+D1+FC+S
15	0.099782	0.075388	C1+P1+C2+C3+P2+R1+C4+P3+C5+P4+C6+R2+D+FC+S
16	0.143788	0.080302	C1+C2+P1+C3+C4+P2+R1+C5+P3+C6+P4+C7+R2+D+FC+S

Table 4
Effect on validation error with varying depth of the filters at layer C1 of the proposed CNN model.

Filter depth at C1	VE (With augmented dataset)
$5(11 \times 11 \times 1 \times 5)$	0.024696
$15(11 \times 111 \times 15)$	0.019847
$25\ (11 \times 11 \times 1 \times 25)$	0.015152



(a) (b)

Fig. 5. Learned filters in the first layer of CNN (a) head-pose estimation, (b) full-body orientation estimation.

Fig. 5 presents a visualization of the first-layer filters learned during the training of the proposed CNN model with BDBO (Fig. 5(a)) and BDHP (Fig. 5(b)), which shows that most of the filters achieved standard edges. The increase in depth of the filters in the next layers makes them difficult to visualize.

An illustration of some feature maps (at layers C1, C2, R1, C3, R2, and C4) for a sample image is shown in Fig. 6. In the deeper layers, these feature maps extract more comprehensive features.

5.2. Testing protocol

Three types of datasets are used: static-image dataset (TUD Multiview pedestrian dataset [65]), image-sequence dataset (CAVIAR [66] data) and self-taken real-time captured streams. In all three types, the bounding boxes of pedestrians' images are selected and, after normalization, their full-body images are passed on to the trained full-body CNN classifier. Afterwards, head portion is cropped, normalized and passed to the trained head-pose CNN classifier for prediction. The results are then compared with those of state-of-the-art approaches found in the literature.

The results are presented in terms of confusion matrices. The row labels of the matrices represent ground-truth values, whereas the column labels indicate predicted values. The performance, based on confusion matrix M with a total of $t = 8$ classes, is quantified in terms of accuracy, precision (*Prc.*), true-positive rates (*TPR*) (or Recall) and false-positive rate (*FPR*), which are represented as follows:

$$Accuracy = \frac{Sum\ of\ correct\ hits}{Sum\ of\ all\ classifications} = \frac{\sum_{i=1}^{t} M(i,i)}{\sum_{i=1}^{t}\sum_{j=1}^{t} M(i,j)} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$TPR = Recall = \frac{TP}{TP + FN} \quad (8)$$

$$FPR = \frac{FP}{FP + TN} \quad (9)$$

where *TP* represents the sum of all true positives, *FP* represents the sum of all false positives, TN is the sum of true negatives and FN symbolizes the sum of false negatives.

Two types of accuracy are considered for performance evaluation. 'Accuracy1' or 'Acc1' is considered the exact accuracy and lies on the diagonal of exact hits of the confusion matrix. 'Accuracy2' or 'Acc2' adds actual accuracies to adjacent classes' accuracies along the ground-truth row. The adjacent labels are also considered true for orientation classification.

It should be noted that the results based on existing methods are based on either head-pose estimation or full-body orientation estimation for a particular dataset. The proposed approach presents both head-pose and full-body orientation estimation. Therefore, full-body orientation results are compared with existing schemes on the TUD-Multiview pedestrian dataset [65], while head-pose estimation results are compared with existing schemes on CAVIAR sequences. However, along with these comparisons, we present our results for both head and body-orientations on all datasets.

5.3. Experiment onthe TUD-Multiview pedestrian dataset

The proposed model is assessed with the TUD-Multiview pedestrian dataset for benchmarking with existing works. The TUD-Multiview dataset [65] contains 4732 full-body pedestrian images, annotated in terms of orientation, with 400–479 images per class. The dataset also contains separate subsets of 248 validation and 248 test images. The performance of the proposed model is evaluated by comparison with state-of-the-art approaches. The confusion matrices shown in Fig. 7 correspond to the results of head-pose and full-body orientation extraction obtained on the TUD-Multiview dataset. The exact hit values along diagonals show the acceptable true-positives rates.

Table 5 represents a class-wise performance comparison of full-body orientation estimation in terms of true-positive rates (TPRs) and false-positive rates (FPRs).The approaches selected for comparison are MACF [24], sparse representation (SR) [33], BIFMS-PLS-RF [25] and MoAWG: HoG-ERT-AWG [35]. The proposed approach achieves higher values for precision and TPR and lower values for FPR. Therefore, the use of CNN improves the performance of orientation prediction and the proposed approach completely outperforms all the solutions presented previously. This is because deep supervised learning involves automatic feature extraction, which is proved to be more robust than traditional hand-crafted features.

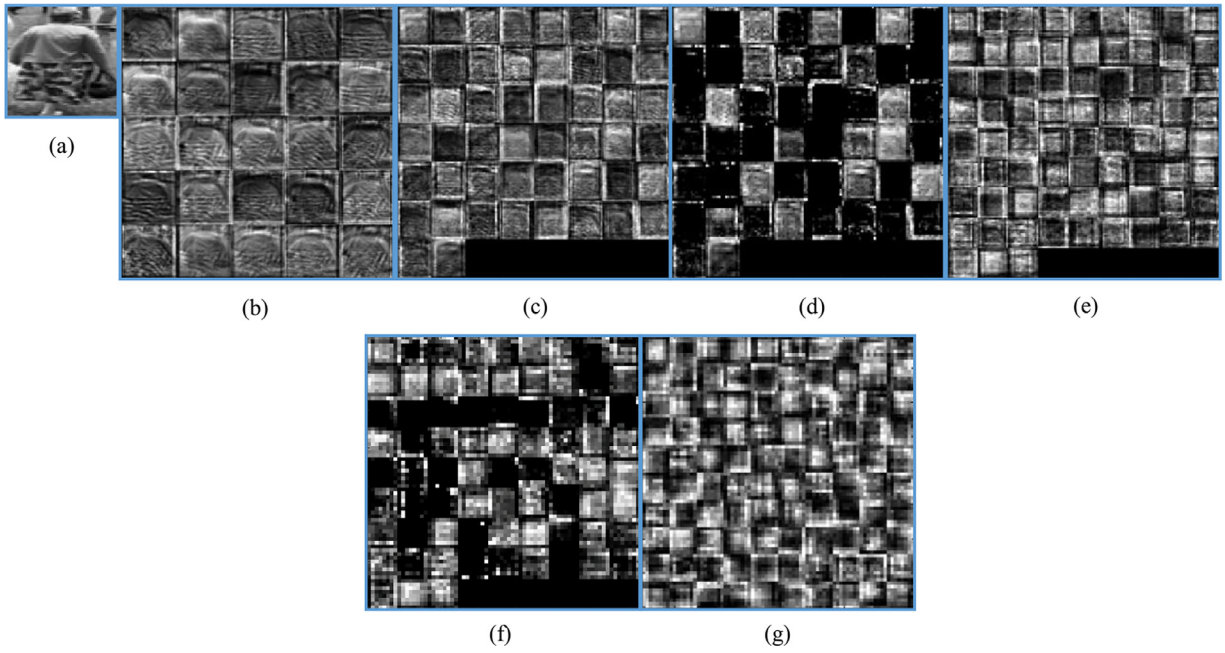


Fig. 6. An Illustration of some feature maps of an input image formed in different layers: (a) Input image, (b) Feature maps at C1, (c) Feature maps at C2, (d) Feature maps at R1, (e) Feature maps at C3, (e) Feature maps at R2, and (g) Feature maps at C4.

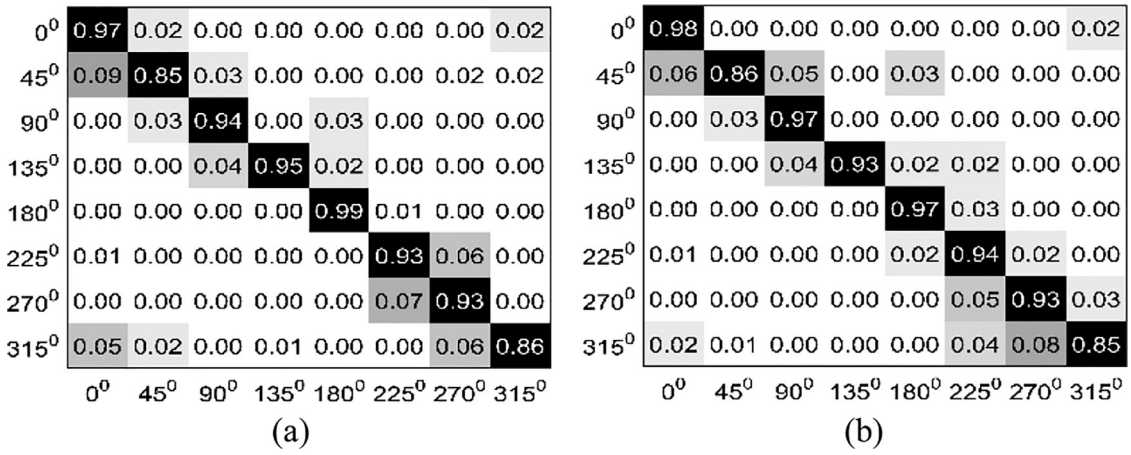


Fig. 7. Confusion matrices for the proposed CNN model on the TUD-Multiview dataset: (a) head-pose and (b) full-body orientation.

Table 5
Class-wise performance comparison of full-body orientation estimation on the TUD-Multiview dataset.

Class	MACF			SR			BIFM-PLS-RF			MoAWG			CNN (Proposed)		
	Prc.	TPR	FPR	Prc.	TPR	FPR	Prc.	TPR	FPR	Prc.	TPR	FPR	Prc.	TPR	FPR
0°	0.32	0.32	0.099	0.71	0.40	0.024	0.58	0.75	0.077	0.59	0.64	0.061	0.90	0.98	0.014
45°	0.48	0.38	0.059	0.55	0.36	0.042	0.64	0.57	0.045	0.83	0.38	0.011	0.96	0.86	0.006
90°	0.39	0.40	0.088	0.62	0.65	0.057	0.74	0.79	0.039	0.85	0.95	0.022	0.92	0.97	0.012
135°	0.39	0.29	0.066	0.45	0.37	0.065	0.63	0.57	0.049	0.59	0.57	0.055	1.00	0.93	0.000
180°	0.36	0.60	0.157	0.42	0.71	0.139	0.63	0.69	0.059	0.47	0.76	0.121	0.92	0.97	0.011
225°	0.36	0.32	0.083	0.57	0.53	0.057	0.68	0.52	0.034	0.61	0.52	0.047	0.88	0.94	0.019
270°	0.53	0.44	0.056	0.66	0.70	0.052	0.60	0.77	0.073	0.77	0.86	0.037	0.89	0.93	0.015
315°	0.46	0.45	0.074	0.49	0.59	0.089	0.80	0.56	0.020	0.68	0.55	0.037	0.95	0.85	0.006

Fig. 8 shows the two accuracies for the methods represented in Table 5. The proposed model obtained an accuracy1 value of 93% and an accuracy2 value of 99%. The improved overall accuracy rates prove the effectiveness of the proposed model.

Class-wise performance for head-pose estimation on the TUD-Multiview dataset is shown in Table 6. In this instance, accuracy1 is found to be 0.92, while accuracy2 is calculated as 0.99. The performance results of head-pose prediction seem closer to the performance results for full-body orientation prediction.

5.4. Experiment on CAVIAR dataset

We evaluate the proposed approach on the CAVIAR dataset. CAVIAR [66] provides labeled and annotated pedestrian images

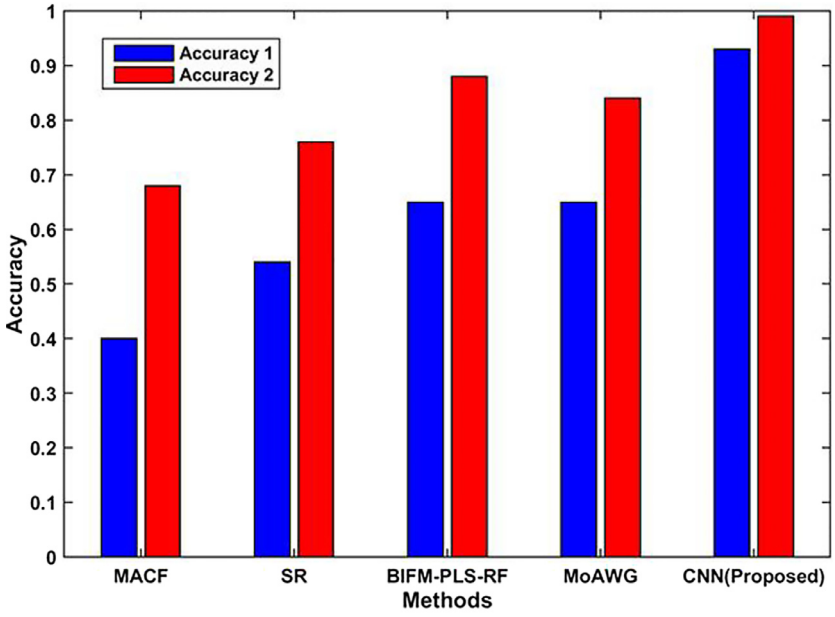


Fig. 8. Comparison of accuracy1 and accuracy2 for full-body orientation on the TUD-Multiview dataset.

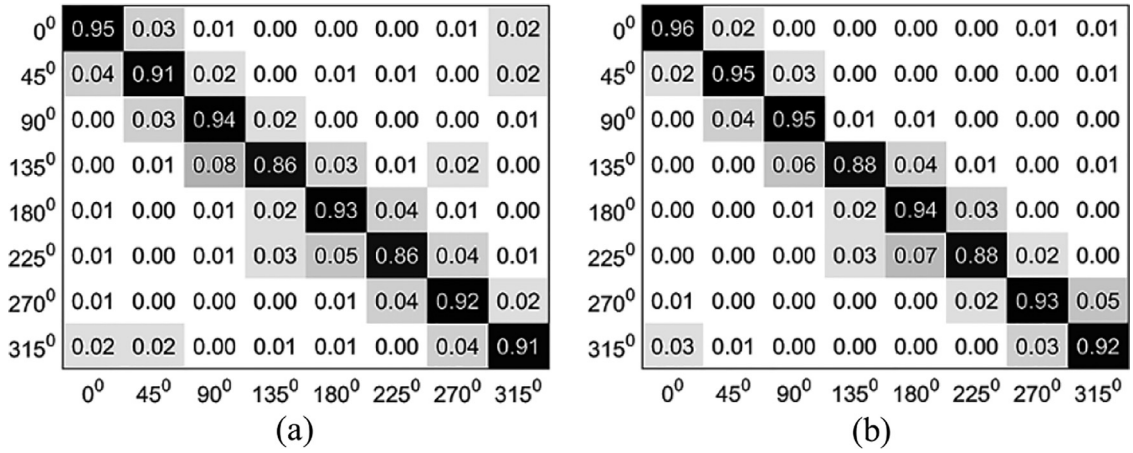


Fig. 9. Confusion matrices for the proposed CNN model on CAVIAR sequences: (a) head-pose, (b) full-body orientation.

Table 6
Class-wise performance for head-pose estimation on the TUD-Multiview dataset.

Class	Acc1	Acc2	Prc.	TPR	FPR
0°	0.92	0.99	0.86	0.97	0.021
45°			0.93	0.85	0.009
90°			0.93	0.94	0.010
135°			0.99	0.95	0.002
180°			0.96	0.99	0.006
225°			0.91	0.93	0.013
270°			0.87	0.93	0.019
315°			0.96	0.86	0.005

with head pan-angle estimates. The dataset contains image sequences of pedestrians walking in a shopping mall. Four CAVIAR sequences are selected for the experiment: ShopAssistant1cor, ThreePastShop1cor, OneShopOneWait2cor and ThreePastShop2cor. The pedestrians' labeled bounding boxes are extracted from the provided ground truths. Confusion matrices for the proposed CNN model on CAVIAR sequences are shown in Fig. 9.

The class-wise performance comparison of head-pose estimation is given in Table 7. The selected methods for evaluation are the single-frame (SF) approach [41,45], Head-tracking (HT) approach [41] and Deep-belief network (DBN) approach [21]. The CNN-based approach shows high performance in terms of precision, TPR and FPR. This performance makes the proposed methodology highly competitive compared to other approaches.

The results in terms of the two accuracies can be seen in Fig. 10. The actual hit rate (accuracy1) is 0.91, which is much better than those of other methods.

Class-wise performances for full-body orientation estimation on CAVIAR sequences are shown in Table 8. Here, accuracy1 for head-pose estimation is calculated as 0.93, while accuracy2 is calculated as 0.99, indicating the better performance rate of the proposed CNN model.

5.5. Experiment on real-time video sequences

The validity of the proposed approach is further evaluated by using real-time video sequences captured from a webcam placed in different indoor and outdoor environments. The total number of acquired video sequences is 50, taken in diverse scenarios, containing 4027 frames (each with a resolution of 640×480). Approximately 2000 pedestrians' full-body images are extracted (and re-

Table 7
Class-wise performance comparison of head-pose estimation on CAVIAR sequences.

Class	SF			HT			DBN			CNN (Proposed)		
	Prc.	TPR	FPR	Prc.	TPR	FPR	Prc.	TPR	FPR	Prc.	TPR	FPR
0°	0.37	0.62	0.154	0.37	0.78	0.187	0.75	0.95	0.044	0.92	0.94	0.011
45°	0.40	0.34	0.073	0.36	0.32	0.080	0.88	0.80	0.016	0.91	0.91	0.012
90°	0.50	0.16	0.022	0.32	0.09	0.028	0.85	0.91	0.022	0.89	0.94	0.016
135°	0.37	0.37	0.088	0.51	0.66	0.092	0.91	0.50	0.007	0.93	0.86	0.009
180°	0.42	0.29	0.058	0.45	0.18	0.031	0.54	0.91	0.113	0.89	0.93	0.017
225°	0.32	0.41	0.124	0.53	0.51	0.064	0.78	0.25	0.010	0.90	0.86	0.014
270°	0.32	0.23	0.069	0.63	0.28	0.023	0.77	0.91	0.039	0.90	0.92	0.015
315°	0.38	0.55	0.129	0.40	0.67	0.140	0.88	0.90	0.017	0.93	0.91	0.009

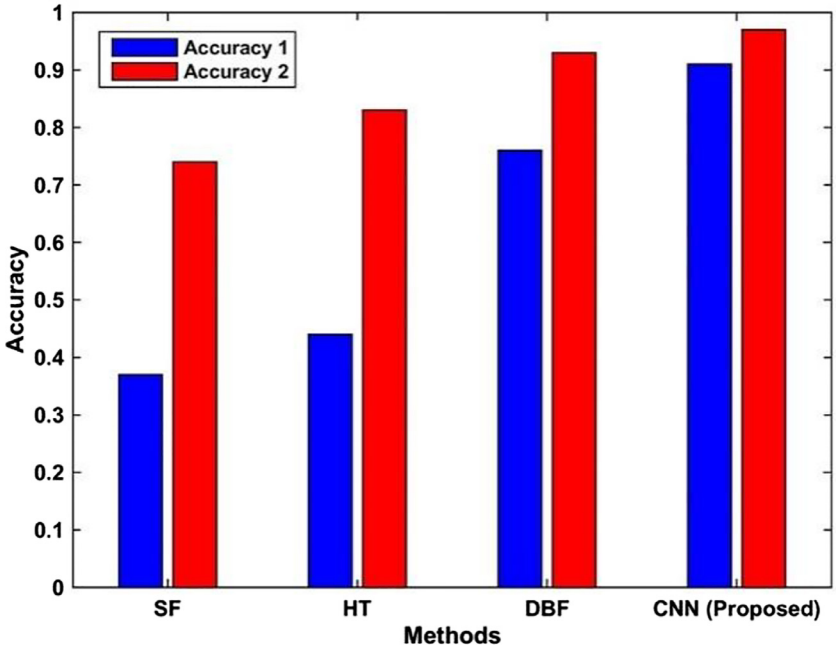


Fig. 10. Comparison of accuracy1 and accuracy2 for head-pose estimation on CAVIAR sequences.

Table 8
Class-wise performance for full-body orientation estimation on CAVIAR sequences.

Class	Acc1	Acc2	Prc.	TPR	FPR
0°	0.93	0.99	0.95	0.96	0.008
45°			0.93	0.95	0.011
90°			0.91	0.95	0.013
135°			0.94	0.88	0.008
180°			0.88	0.94	0.018
225°			0.94	0.88	0.008
270°			0.94	0.93	0.008
315°			0.92	0.92	0.012

sized to 64×64 resolution) and selected for the experiment. These full-body images and images of head portions (cropped from full-body images and resized to 64×64 resolution) are then supplied to the CNN classifiers for predictions.

Fig. 11 shows the confusion matrices of the proposed approach (Fig. 11(a) for head-pose and Fig. 11(b) for full-body orientation).

The latest work for head and body orientation estimation on real-time sequences (to the best of our knowledge) is performed by Shinya et al. [40]. Their accuracies for head-pose are estimated as Accuracy1: 0.50 and Accuracy2: 0.92. The accuracies for full-body orientation are estimated as Accuracy1: 0.53 and Accuracy2: 0.93. (Note: The accuracies are calculated from confusion matrices given in [40]). Comparing with our approach, we achieve Accuracy1: 0.88 and Accuracy2: 0.98 for head-pose, and Accuracy1: 0.90 and Accuracy2: 0.98 for full-body orientation. Due to the unavailability of the video sequences used in the work of Shinya et al. [40], direct comparison is not possible. However, as the most recent work and state-of-the-art method, we believe it is important to mention it in this work and, therefore, we provide undirected comparison. Table 9 depicts the class-wise performance comparison of proposed approach on real-time self-taken sequences. The CNN predictor achieves high recall, which is highly desirable. The proposed approach yields promising results, with both recall and precision being above 82% in all classes with an average computation time of 34.48 predictions per second (0.029 s per prediction). The proposed approach significantly outperforms the existing approach.

Fig. 12 depicts classified images for head and body orientation on the selected datasets used for experiments. The upper arrows represent the head-pose angles, while the lower arrows represent the directions in which the pedestrians are moving (see Fig. 12(a), TUD-Multiview dataset; Fig. 12(b), CAVIAR dataset; and Fig. 12(c), real-time self-taken image sequences).

Finally, mean accuracy results of head-pose and full-body orientation on public and self-taken datasets are summarized in Table 10. The results indicate acceptable accuracy rates, which proves the robustness of the proposed approach.

The relatively lower performance of head-pose extraction, compared to full-body orientation estimation, is because low-

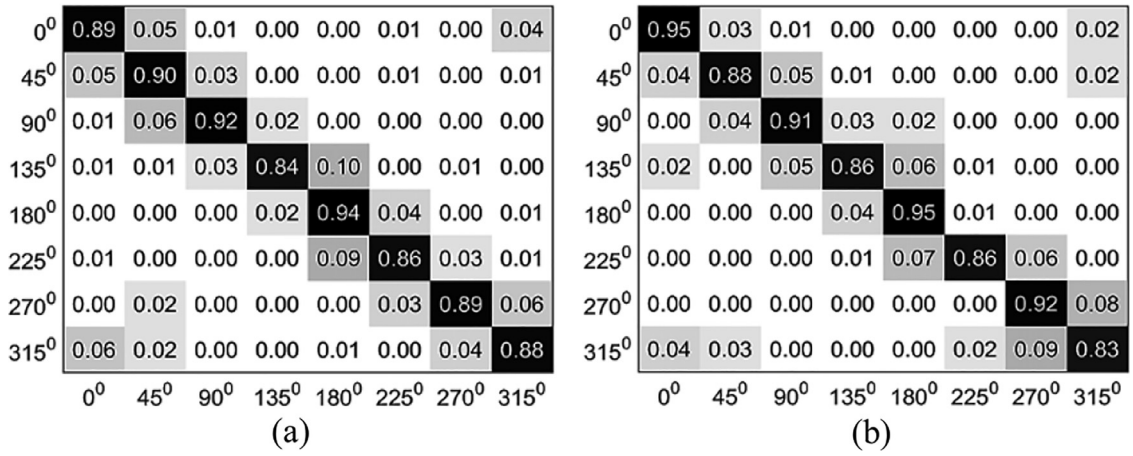


Fig. 11. Confusion matrices for the proposed CNN model on real-time self-taken sequences: (a) head-pose, (b) full-body orientation.

Table 9
Class-wise performance for head-pose and full-body orientation estimation using the proposed CNN approach on real-time self-taken sequences.

Class	Head					Body				
	Acc1	Acc2	Prc.	TPR	FPR	Acc1	Acc2	Prc.	TPR	FPR
0°	0.89	0.99	0.87	0.89	0.019	0.90	0.98	0.91	0.95	0.014
45°			0.86	0.89	0.021			0.91	0.88	0.013
90°			0.93	0.92	0.009			0.89	0.91	0.015
135°			0.96	0.83	0.005			0.90	0.86	0.013
180°			0.82	0.94	0.030			0.86	0.95	0.022
225°			0.90	0.86	0.013			0.96	0.86	0.005
270°			0.92	0.89	0.012			0.87	0.92	0.021
315°			0.87	0.88	0.018			0.87	0.83	0.017

Fig. 12. Several classified images from different datasets; the upper arrow in each image represents the direction of head-pose while the lower arrow indicates the orientation of the person; (a) TUD-Multiview pedestrian dataset, (b) CAVIAR dataset, (c) self-taken image sequences.

Table 10
Mean accuracies of the proposed CNN approach.

	Head-pose orientation accuracy		Full-body orientation accuracy	
	Acc1	Acc2	Acc1	Acc2
Mean accuracy	0.91	0.98	0.92	0.99

resolution and low-illumination distant images of pedestrians' heads are sometimes not completely visible through cameras, or even to human eyes. This makes it difficult to distinguish front and back views of the human head. Fig. 13 shows some difficult cases of head-poses. In addition, we believe that these results can be further improved by increasing the size of the training set.

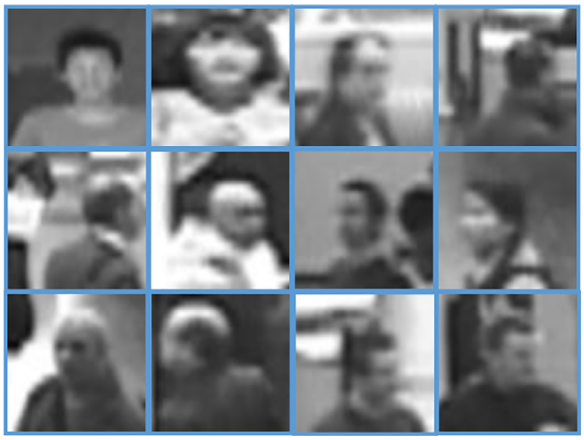


Fig. 13. Several difficult cases of head-poses.

6. Conclusions

Head-pose and the direction of a pedestrian's movement are occasionally very loosely correlated. Pedestrians who are following a path in one direction can have orient their heads in other directions. Therefore, we propose a CNN-based approach for classifying pedestrian head-pose and full-body orientation. The proposed CNN model is trained separately with the two proposed training datasets for both head and body orientations. Bounding-box images of pedestrians are provided to a trained CNN model for classification of automatically extracted features. The use of appearance-based classification makes this approach applicable for both still images as well as image sequences. The trained CNN classifier is tested with different publicly available datasets, as well as self-taken image sequences. The results are compared with those of the available state-of-the-art approaches. The proposed deep-learning approach outperforms other classifiers by a prominent margin. The average value of accuracy 1 is found to be 0.91 for head-pose estimation and 0.92 for full-body orientation estimation. This acceptable accuracy proves the robustness of the proposed approach.

The proposed CNN model can be further fine-tuned for more accurate results. Additionally, the proposed approach is applied for only eight orientation classes in this study and can be extended to more orientations for better predictions in the future. In addition, the proposed approach will be useful in improving people tracking for enhanced path prediction, behavior analysis, risk assessment and human activity recognition.

Acknowledgments

The authors would like to thank National Natural Science Foundation of China (61375079), and the Chinese Academy of Science-The World Academy of Sciences (CAS-TWAS) President's Fellowship.

References

[1] M. Guo, Y. Zhao, J. Xiang, C. Zhang, Z. Chen, Review of object detection methods based on SVM, Control Decis 2 (2014) 001.
[2] M. Lin, Z. Chen, Salient region detection via low-level features and high-level priors, in: Proceedings of the IEEE International Conference on Digital Signal Processing (DSP), IEEE, 2015, pp. 971–975.
[3] M. Lin, C. Zhang, Z. Chen, Global feature integration based salient region detection, Neurocomputing 159 (2015) 1–8.
[4] J. Lallemand, A. Ronge, M. Szczot, S. Ilic, Pedestrian orientation estimation, in: German Conference on Pattern Recognition, Springer, Münster Germany, 2014, pp. 476–487.
[5] M. Lin, Z. Chen, C. Zhang, Fused motion and color features for object tracking, Int. J. Inf. Process. Manag 5 (2014) 95.
[6] D. Bouchain, Character recognition using convolutional neural networks, in: Seminar Statistical Learning Theory Institute for Neural Information Processing, University of Ulm, Germany, 2006.
[7] G. Cheng, P. Zhou, J. Han, Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images, IEEE Trans. Geosci. Remote Sens 54 (2016) 7405–7415.
[8] G. Cheng, J. Han, A survey on object detection in optical remote sensing images, ISPRS J. Photogramm. Remote Sens 117 (2016) 11–28.
[9] X. Chu, W. Ouyang, X. Wang, CRF-CNN: Modeling structured information in human pose estimation, Adv. Neural Inf. Process. Syst (2016) pp. 316–324.
[10] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2012) pp. 1097–1105.
[11] M. Szarvas, A. Yoshizawa, M. Yamamoto, J. Ogata, Pedestrian detection with convolutional neural networks, in: Proceedings of the Intelligent Vehicles Symposium, IEEE, 2005, pp. 224–229.
[12] P. Sermanet, Y. LeCun, Traffic sign recognition with multi-scale convolutional networks, in: Proceedings of the International Joint Conference on Neural Networks, IEEE, 2011, pp. 2809–2813.
[13] B. Ahn, J. Park, I.S. Kweon, Real-time head orientation from a monocular camera using deep neural network, in: Proceedings of the Asian Conference on Computer Vision, Springer, 2014, pp. 82–96.
[14] A. Gepperth, M.G. Ortiz, B. Heisele, Real-time pedestrian detection and pose classification on a GPU, in: Proceedings of the 16th International IEEE Conference on Intelligent Transportation Systems (ITSC), IEEE, 2013, pp. 348–353.
[15] M. Enzweiler, D.M. Gavrila, Integrated pedestrian classification and orientation estimation, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 982–989.
[16] J. Tao, R. Klette, Part-based RDF for direction classification of pedestrians, and a benchmark, in: Proceedings of the Computer Vision-ACCV 2014 Workshops, Springer, 2014, pp. 418–432.
[17] G. Santoshi, S. Mishra, Pedestrian with direction detection using the combination of decision tree learning and SVM, in: Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India (CSI), 1, Springer, 2015, pp. 249–255.
[18] D. Tosato, M. Spera, M. Cristani, V. Murino, Characterizing humans on riemannian manifolds, IEEE Trans. Pattern Anal. Mach. Intell 35 (2013) 1972–1984.
[19] H. Liu, L. Ma, Online person orientation estimation based on classifier update, in: IEEE International Conference on Image Processing (ICIP), IEEE, 2015, pp. 1568–1572.
[20] S. Yano, Y. Gu, S. Kamijo, Estimation of pedestrian pose and orientation using on-board camera with histograms of oriented gradients features, Int. J. Intell. Transp. Syst. Res (2014) 1–10.
[21] R.H. Baxter, M.J. Leach, S.S. Mukherjee, N.M. Robertson, An adaptive motion model for person tracking with instantaneous head-pose features, IEEE Signal Process. Lett 22 (2015) 578–582.
[22] W. Liu, Y. Zhang, S. Tang, J. Tang, R. Hong, J. Li, Accurate estimation of human body orientation from RGB-D sensors, IEEE Trans. Cybern 43 (2013) 1442–1452.
[23] M. Hayashi, K. Oshima, M. Tanabiki, Y. Aoki, Upper body pose estimation for team sports videos using a poselet-regressor of spine pose and body orientation classifiers conditioned by the spine angle prior, Inf. Media Technol 10 (2015) 531–547.
[24] L. Fitte-Duval, A.A. Mekonnen, F. Lerasle, Upper body detection and feature set evaluation for body pose classification, in: Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP'15), Berlin (Germany), 2015.
[25] I. Ardiyanto, J. Miura, Partial least squares-based human upper body orientation estimation with combined detection and tracking, Image Vis. Comput 32 (2014) 904–915.
[26] E. Rehder, H. Kloeden, C. Stiller, Head detection and orientation estimation for pedestrian safety, in: Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), IEEE, 2014, pp. 2292–2297.
[27] D. Vishwakarma, A. Dhiman, R. Maheshwari, R. Kapoor, Human motion analysis by fusion of silhouette orientation and shape features, Procedia Comput. Sci 57 (2015) 438–447.
[28] S. Piérard, M. Van Droogenbroeck, Estimation of human orientation based on silhouettes and machine learning principles, in: Proceedings of the International Conference on Pattern Recognition Applications and Methods (ICPRAM), 2012.
[29] O. Ozturk, T. Yamasaki, K. Aizawa, Tracking of humans and estimation of body/head orientation from top-view single camera for visual focus of attention analysis, in: Proceedings of the IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, IEEE, 2009, pp. 1020–1027.
[30] R. Furuhashi, K. Yamada, Estimation of street crossing intention from a pedestrian's posture on a sidewalk using multiple image frames, in: Proceedings of the First Asian Conference on Pattern Recognition, IEEE, 2011, pp. 17–21.
[31] K. Goto, K. Kidono, Y. Kimura, T. Naito, Pedestrian detection and direction estimation by cascade detector with multi-classifiers utilizing feature interaction descriptor, in: Proceedings of the Intelligent Vehicles Symposium (IV), IEEE, 2011, pp. 224–229.
[32] F. Flohr, M. Dumitru-Guzu, J.F. Kooij, D.M. Gavrila, Joint probabilistic pedestrian head and body orientation estimation, in: Proceedings of the Intelligent Vehicles Symposium, IEEE, 2014, pp. 617–622.

[33] C. Chen, A. Heili, J.-M. Odobez, Combined estimation of location and body pose in surveillance video, in: Proceedings of the 8th International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2011, pp. 5–10.
[34] T. Gandhi, M.M. Trivedi, Image based estimation of pedestrian orientation for improving path prediction, in: Proceedings of the Intelligent Vehicles Symposium, IEEE, 2008, pp. 506–511.
[35] D. Baltieri, R. Vezzani, R. Cucchiara, People orientation recognition by mixtures of wrapped distributions on random trees, in: Proceedings of the European Conference on Computer Vision, Springer, 2012, pp. 270–283.
[36] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn 63 (2006) 3–42.
[37] J.P. Wachs, M. Kölsch, D. Goshorn, Human posture recognition for intelligent vehicles, J. Real-Time Image Process 5 (2010) 231–244.
[38] L. Chen, G. Panin, A. Knoll, Human body orientation estimation in multiview scenarios, in: International Symposium on Visual Computing: Advances in Visual Computing, Crete, Greece, 2012, pp. 499–508.
[39] J.K. Lee, E.J. Park, Minimum-order Kalman filter with vector selector for accurate estimation of human body orientation, IEEE Trans. Robot 25 (2009) 1196–1201.
[40] S. Yano, Y. Gu, S. Kamijo, Estimation of pedestrian pose and orientation using on-board camera with histograms of oriented gradients features, Int. J. Intell. Transp. Syst. Res 14 (2016) 75–84.
[41] A. Schulz, R. Stiefelhagen, Video-based pedestrian head pose estimation for risk assessment, in: Proceedings of the 15th International IEEE Conference on Intelligent Transportation Systems, IEEE, 2012, pp. 1771–1776.
[42] S. Piérard, D. Leroy, J.-F. Hansen, M. Van Droogenbroeck, Estimation of human orientation in images captured with a range camera, in: International Conference on Advanced Concepts for Intelligent Vision Systems, Springer, Berlin, Heidelberg, 2011, pp. 519–530.
[43] M.C. Liem, D.M. Gavrila, Person appearance modeling and orientation estimation using spherical harmonics, in: Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE, 2013, pp. 1–6.
[44] W. Lianshi, X. Limin, L. Dayong, Unified pedestrian detection and orientation recognition using single-frame, Int. Inf, Inst (Tokyo) Inf 15 (2012) 3811.
[45] A. Schulz, N. Damer, M. Fischer, R. Stiefelhagen, Combined head localization and head pose estimation for video–based advanced driver assistance systems, in: Proceedings of the Joint Pattern Recognition Symposium, Springer, 2011, pp. 51–60.
[46] W. Ouyang, X. Chu, X. Wang, Multi-source deep learning for human pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2329–2336.
[47] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, R. Moore, Real-time human pose recognition in parts from single depth images, Commun. ACM 56 (2013) 116–124.
[48] X. Chen, A.L. Yuille, Articulated pose estimation by a graphical model with image dependent pairwise relations, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 1736–1744.
[49] A. Jain, J. Tompson, Y. LeCun, C. Bregler, Modeep: A deep learning framework using motion features for human pose estimation, in: Proceedings of the Asian Conference on Computer Vision, Springer, 2014, pp. 302–315.
[50] S. Li, Z.-Q. Liu, A.B. Chan, Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 482–489.
[51] A. Jain, J. Tompson, M. Andriluka, G.W. Taylor, C. Bregler, Learning human pose estimation features with convolutional networks, Cornel University Library (2013). arXiv preprint arXiv: 1312.7302 (accessed 02.09.2016)
[52] E. Brau, H. Jiang, 3D human pose estimation via deep learning from 2D annotations, 3D vision (3DV), in: Proceedings of the Fourth International Conference on 3D vision (3DV), IEEE, 2016, pp. 582–591.
[53] L. Beyer, A. Hermans, B. Leibe, Biternion nets: continuous head pose regression from discrete training labels, in: Proceedings of the German Conference on Pattern Recognition, Springer, 2015, pp. 157–168.
[54] J. Bao, M. Ye, Head pose estimation based on robust convolutional neural network, Cybern. Inf. Technol 16 (2016) 133–145.
[55] J. Choi, B.-J. Lee, B.-T. Zhang, Human body orientation estimation using convolutional neural network, Cornel University Library (2016). arXiv preprint arXiv: 1609.01984 (accessed 02.09.2016).
[56] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3. 6m: large scale datasets and predictive methods for 3d human sensing in natural environments, IEEE Trans. Pattern Anal. Mach. Intell 36 (2014) 1325–1339.
[57] R. Wagner, M. Thom, M. Gabb, M. Limmer, R. Schweiger, A. Rothermel, Convolutional neural networks for night-time animal orientation estimation, in: Proceedings of the Intelligent Vehicles Symposium (IV), IEEE, 2013, pp. 316–321.
[58] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: Proceedings of the Computer Vision–ECCV 2008, Springer, 2008, pp. 262–275.
[59] D.S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, V. Murino, Custom pictorial structures for *Re*-identification, in: BMVC, 2011, p. 6.
[60] A. Ess, B. Leibe, K. Schindler, L.V. Gool, A mobile vision system for robust multi-person tracking, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.
[61] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in: Proceedings of the Computer Vision–ECCV 2014, Springer, 2014, pp. 688–703.
[62] C.C. Loy, T. Xiang, S. Gong, Time-delayed correlation analysis for multi-camera activity understanding, Int. J. Comput. Vis 90 (2010) 106–129.
[63] Y. Xu, L. Lin, W.-S. Zheng, X. Liu, Human re-identification by matching compositional template with cluster sampling, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3152–3159.
[64] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.
[65] M. Andriluka, S. Roth, B. Schiele, Monocular 3d pose estimation and tracking by detection, in: Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 623–630.
[66] CAVIAR, CAVIAR project/IST 2001 37540 http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/).

Mudassar Raza is a Ph.D. Scholar at University of Science and Technology of China (USTC), China under CAS-TWAS fellowship. He has more than seven years of experience of teaching undergraduate classes at COMSATS Institute of Information Technology, Pakistan. His interests include are Deep Learning, pattern recognition, and parallel & distributed computing.

Zonghai Chen, born in December 1963, a native of Tongcheng County, Anhui Province, is a professor and doctoral supervisor at the Department of Automation, University of Science and Technology of China (USTC). Prof. Chen is also a recipient of special allowances from the State Council. Chen obtained his bachelor's degree from the Department of Management & Systems Science of USTC in 1988, and his master's degree in Control Theory and Control Engineering from USTC in 1991. After completing his postgraduate studies in 1991, Chen joined the faculty of USTC, dedicating himself to research and teaching in the field of control science and engineering. Prof. Chen's main research area covers modeling and control of complex systems, intelligent science and technology.

Saeed-Ur Rehman has received his MS degree from Mohammad Ali Jinnah University Islamabad Pakistan. Currently he is a Ph.D. Scholar under CAS-TWAS Fellowship; in the University of Science and Technology of China (USTC), Hefei, China. His Research Interests include computer vision, deep learning and use of machine learning and its applications.

Peng Wang received his BS degree from University of Science and Technology of China (USTC) in 2010. He continued to be a Ph.D. candidate ever since 2010 in USTC and got his PHD degree in 2015. His currently work as a postdoctor in the Automation Department. His current research is focused on uncertain information process in mobile robot navigation, interval analysis, and deep/reinforcement learning.

Peng Bao received his BS degree from University of Science and Technology of China (USTC) in 2014. He continued to be a Ph.D. candidate ever since 2014 in USTC. His current research is focusing on uncertain information process in mobile robot navigation, environment sensing and representation, and deep learning.