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/

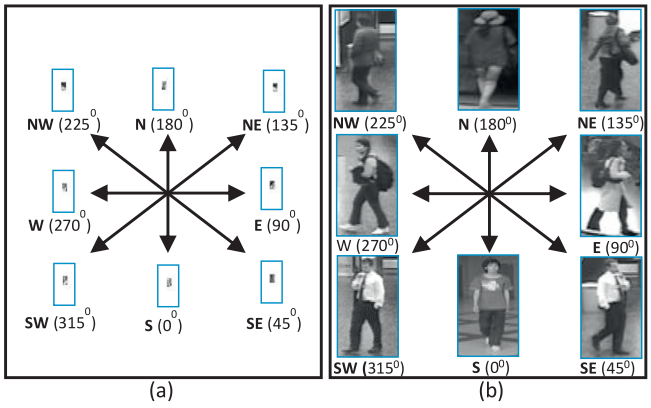


**Fig. 1.** Eight different representations of pedestrians' in terms of (a) head-pose and (b) full-body orientation.

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-