People Orientation Recognition by Mixtures of Wrapped Distributions on Random Trees

Davide Baltieri, Roberto Vezzani, and Rita Cucchiara

DIEF - University of Modena and Reggio Emilia
Via Vignolese 905, 41125 - Modena, Italy
{davide.baltieri,roberto.vezzani,rita.cucchiara}@unimore.it
http://imagelab.ing.unimore.it

Abstract. The recognition of people orientation in single images is still an open issue in several real cases, when the image resolution is poor, body parts cannot be distinguished and localized or motion cannot be exploited. However, the estimation of a person orientation, even an approximated one, could be very useful to improve people tracking and re-identification systems, or to provide a coarse alignment of body models on the input images. In these situations, holistic features seem to be more effective and faster than model based 3D reconstructions. In this paper we propose to describe the people appearance with multi-level HoG feature sets and to classify their orientation using an array of Extremely Randomized Trees classifiers trained on quantized directions. The outputs of the classifiers are then integrated into a global continuous probability density function using a Mixture of Approximated Wrapped Gaussian distributions. Experiments on the TUD Multiview Pedestrians, the Sarc3D, and the 3DPeS datasets confirm the efficacy of the method and the improvement with respect to state of the art approaches.

Keywords: Orientation recognition, Mixtures of Wrapped Distributions, Random Trees.

1 Introduction

Computer vision is devoting great efforts in automatic detection and analysis of people in images and videos, for recognizing interactions, social behaviors, gestures and actions. To this aims, several applications require the recognition of the person pose, both in term of his/her intrinsic posture and in terms of orientation with respect to the camera. If the image resolution is not sufficient for precise 3D scene and people reconstruction and if motion information is not available, *people orientation recognition* is still an open problem.

In this paper we address the recognition of people orientation in images acquired by surveillance cameras, exploiting the visual appearance only, extracted with a people detection technique. We assume that people are in typical pedestrian postures, walking, running, gazing, or chatting upstanding.

Many effective results have been reached if the image resolution is high enough to allow the detection of body parts, in order to reconstruct a 3D model of

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part V, LNCS 7576, pp. 270–283, 2012.

the body [1] or 2D structural models [2], but often in surveillance videos these details are not available. Instead, whenever the monitored person can be correctly detected and tracked along time, its orientation can be inferred by the trajectory on the ground plane [3]. Nerveless, this approach too cannot be applied to detect the orientation when (see Fig. 1):

- People are still, or are in a social setting;
- People suddenly change direction;
- People are in cluttered, crowded areas where the trajectory cannot be computed;
- People have an orientation different to the motion direction.

We explore a general solution for recognizing the people orientation by looking at the appearance only, discarding any motion information or 2D and 3D body models. Very few attempts have been presented in the recent past, with limited results due to the intrinsic challenging aspects. Moreover, having humans an articulated body, it is not possible to define and measure a precise angular direction of the people orientation, but a discretized value (corresponding to one of the main four or eight directions) is usually satisfactory. Indeed, even under this simplification, current results are limited to about 50% of accuracy [4].

1.1 Our Proposal

In this paper we define a new and very general approach which exploits state of the art descriptors and detectors, but ensembles them in a unique angle-oriented classifier. Since the body orientation is mainly related to shape and edges, the best features are straightforwardly related to luminance gradients, without the influence of colors. Histograms of Oriented Gradients (HoG) features [5] are adopted. The main orientations are singularly recognized with an array of Extremely Randomized Trees classifiers [6], which proved to be very fast and powerful in this case. Moreover, the main novelty is the integration of the detectors responses in a single probability density function generated as a Mixture of Wrapped Distributions, and in particular as a Mixture of Approximated Wrapped Gaussian (MoAWG) weighted by the detector outputs. The maximum

Fig. 1. From left to right: people talking without motion; people changing direction because of an obstacle; women walking in a crowd; a goalkeeper moving laterally but with frontal orientation

of this probability density function is the answer of the orientation problem that we further quantize in the main directions, for filtering errors and noise and for making an easy comparison with ground truth data.

As a matter of fact, the estimation of people orientation is an intrinsically continuous problem. The discretized classes are not well separated and sometimes even overlapped (due to the torsion movements of the body). The MoAWG acts as "interpolation" of the outputs of different trained classifiers: each binary classifier is trained to select a positive region of the feature space, which is not related and not imperatively disjoint with the others. The final step (AWG) integrates different contributions and thus improves the classification when the orientation is quite ambiguous.

The resulting approach is fast, simple and very effective, it can be generalized for a different number of main directions, and it can be adopted in many other problems where main directions must be recognized. In Section 3 we report several tests and comparisons on different publicly available datasets, such as the TUD Multiview Pedestrians dataset [1], SARC3D [7] and 3DPES [8], both available on OpenVisor[1], and video sequences from PETS. Comparisons with previous methods demonstrated a very satisfactory improvements of about 18% with respect to the state of the art, achieveing up to 65% of accuracy on the TUD Multiview Pedestrians dataset using eight directions and more than 80% of accuracy using four directions.

1.2 Related Works

As introduced, the problem of people orientation has been deeply studied for 2D and 3D body modeling, in particular for the detection of the position and the orientation of body parts, such as the head, the arms, the torso, and the legs [2,9]. More specific works have been proposed to capture the head/face orientation only when details such as eyes, nose or even the eyeglasses are detectable [10–13], or when motion tracklet can be grouped for detecting orientation changes [14]. In [15], the head pose is estimated using a combination of Random Forests on Gabor features and an additional LDA step as node test for each tree.

Very few works focus on people orientation recognition using the holistic appearance only: some of them have been defined for gaze orientation in social interaction analysis [16], searching only for the main four orientations using covariance matrix descriptors. However, the more strictly related works concern people orientation recognition using appearance only and feature classification. In the recent work of Adriluka *et al.* [1] an initial 2D body orientation estimation has been proposed which uses HoG detectors followed by a suitable Support Vector Machine. A later work by Chen *et al.* [4], specifically designed for surveillance, suggests the adoption of similar features for a multi-class classifier trained on the main eight orientations. Results are very promising in the four main orientations (front, back, right and left) but are limited in the diagonal ones. Recently, the original work has been improved by the same authors [17]

[1] `http://www.openvisor.org`

integrating an additional step based on the head position and orientation, when available. In [18] the skeleton pose of a walking person is estimated using HoG descriptors and Random Forests as classifiers using an exemplar based approach. However this system introduces a pre-alignment step based on 3D information and complex training data, which is not required in our proposal.

2 System Overview

The proposed system aims at estimating the orientation of a person with respect to the camera point of view. The input is a single detection I_k of a person, obtained cropping an image or a video frame on the bounding box provided by a generic appearance based people detector [5, 19]. Since motion or background information are neglected, the person silhouette is not available. The orientation $\theta_k \in (-\pi, \pi]$ of the person is the angle between the main direction of the person and the horizontal axis of the image plane. The precise value of θ_k is ambiguous and impossible to measure since head, shoulders and legs could be differently oriented. Thus, a discrete set of directions instead of a continuous range of values is more appropriated. Let us define the set $\boldsymbol{C}$ of N discrete orientations sampled from the interval $(-\pi, \pi]$ as

$$\begin{aligned} \boldsymbol{C} &= \left\{c^i\right\}, i \in (0, N-1), \\ c^i &= \left(\left(\tfrac{2\pi i}{N} + \pi\right) \mod 2\pi\right) - \pi. \end{aligned} \tag{1}$$

In our experiments we set $N = 8$, obtaining the eight main directions depicted in Fig. 2. For each class c^i, we trained a specific binary classifier on a set of HoG descriptors [5]; a classification score ψ^i instead of a boolean response is also required. A first orientation estimation $\bar{c_k}$ of the image I_k could be directly obtained from the outputs $\Psi_k = \left\{\psi_k^1, \ldots, \psi_k^N\right\}$ of the classifiers:

$$\bar{c_k} = c^j, j = \underset{i}{\operatorname{argmax}}\, \psi_k^i. \tag{2}$$

Using Eq.2 the estimated orientation does not take into account the output of all the classifiers, but the winner one only. In particular, due to the ambiguity of the human direction above described, more than one classifier could positively react and a more precise estimation of the main orientation could be obtained by combining the results of all the classifiers. To this aim, we propose to estimate the continuous distribution $p(\theta|I_k)$ as a function of the classifier outputs. The person orientation $\bar{\theta_k}$ and its corresponding discretized class $c(\bar{\theta_k})$ is now computed maximizing the previous distribution:

$$\bar{\theta_k} = \underset{\theta \in (-\pi, \pi]}{\operatorname{argmax}}\, p(\theta|I_k). \tag{3}$$

Algorithms 1 and 2 report a pseudo-code description of the classification steps, while the following subsections will detail the set of classifiers and the integration of their outputs using a circular statistic approach.

(a) (b)

Fig. 2. The eight directions recognized by the proposed system and the corresponding color labels

2.1 Discrete Orientation Classifiers

For each detected person, a 2268-dimensional feature vector is computed based on the HoG descriptor [5]. The color image cropped around the person is firstly converted into a single channel image; the first direction of a Principal Component Analysis space reduction is selected. With respect to the luminance channel, the PCA-based image channel preserves and even enhances the edge gradients. Thus, a multi-level HoG feature vector is computed, dividing the input image into blocks at three different levels: the first level contains 8x24 non-overlapping blocks, the second level 4x12 blocks and the last level 2x6 blocks. At each level the image is down-sampled with a scale factor of 0.5. An histogram of oriented gradients quantized in 9 wrapped bins is computed on each of the 252 blocks and normalized over 2x2 sets of blocks. During the histogram computation, the tri-linear interpolation described in [5] is preserved. The 2268-dimensional feature vector ϕ_k is obtained concatenating the 9 histogram values of the 252 blocks computed over I_k; ϕ_k acts as appearance descriptor of the images and it is sent to the array of classifiers.

Due to the very high dimensionality of the input feature vector, we adopted the Extremely Randomized Trees classifiers introduced by Geurts *et. al.* [6]. The Extremely Randomized Trees are similar to Random Trees but instead of using bagging selection they keep the same input training set to train all the trees. For binary classification problems only, Random Trees allow the estimation of a fuzzy-predicted class label, i.e., a confidence value of the binary classification result. In our case, given a feature vector ϕ_k, each of the N classifiers provides a value $\{\psi^i, i = 1, \ldots, N\}$ calculated as the proportion of decision trees that classified the input to the winner class. A discrete label of the image orientation could be generated using Eq. 2.

2.2 Output Filtering by Circular Statistics

Instead of directly using the outputs of the N trained classifiers to generate a discrete class label, we integrated the results in a continuous probabilistic

Algorithm 1. Discrete Orientation Classifiers

Require: $N, \Gamma = \{\Gamma^1, \ldots, \Gamma^N\}$, set of trained classifiers

function MULTILEVELHOG(I)
  $\{B_j^1\}, j = 1 \ldots 192 \leftarrow$ SPLIT($I, 8, 24$) ▷ Level 1
  $\phi_j^1 = HoG(B_j^1)$

  $I \Leftarrow$ RESIZE($I, 0.5$)
  $\{B_j^2\}, j = 1 \ldots 48 \leftarrow$ SPLIT($I, 4, 12$) ▷ Level 2
  $\phi_j^2 = HoG(B_j^2)$

  $I \Leftarrow$ RESIZE($I, 0.5$)
  $\{B_j^3\}, j = 1 \ldots 12 \leftarrow$ SPLIT($I, 2, 6$) ▷ Level 3
  $\phi_j^3 = HoG(B_j^3)$

  $\phi = [\phi_1^1 \ldots \phi_{128}^1 | \phi_1^2 \ldots \phi_{48}^2 | \phi_1^3 \ldots \phi_{12}^3]$
  Normalize(ϕ)
  return ϕ
end function

function FINDORIENTATIONS(I)
  $\{I_k\} \leftarrow$ PEOPLEDETECTOR(I)
  for $k = 1 \rightarrow K$ **do**
    $\phi_k \leftarrow$ MULTILEVELHOG(I_k)
    for $i = 1 \rightarrow N$ **do**
      $\psi^i \leftarrow \Gamma(\phi_k)$
    end for
    $\bar{c_k} \leftarrow \text{argmax}_i \, \psi_k^i$.
  end for
end function

distribution $p(\theta|I)$. The reason of this step is mainly due to the overlapping of the orientation classes, which leads to have more than one high response from the set of discrete-orientation classifiers.

The terms ψ^i are used as weights of a mixture of wrapped distributions, each centered on the N selected orientations θ_i. Directional statistics has been widely studied in the past and Wrapped Gaussians or the most general von Mises distributions are the widest adopted models [20] to manage periodic data such as angles [21]. The probability density function of a wrapped normal distribution is

$$\mathcal{WN}(\theta|\theta_0, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \sum_{k=-\infty}^{+\infty} e^{-\frac{(\theta-\theta_0+2k\pi)^2}{2\sigma^2}}, \tag{4}$$

where μ and σ^2 are the corresponding mean and variance. A very interesting approximated version of the Wrapped Gaussian has been presented by Bahlmann in [22] to deal with semi-periodic multivariate data in handwritten character recognition and successively used in [3] for trajectory description and clustering.

The corresponding probability density function is

$$\mathcal{AWN}(\theta|\theta_0,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{((\theta-\theta_0) \mod 2\pi)^2}{2\sigma^2}}. \quad (5)$$

A mixture of $\mathcal{AWN}$ is obtained as a weighted sum of $\mathcal{AWN}$ probability density functions:

$$MoA\mathcal{WN}(\theta|\boldsymbol{w},\boldsymbol{\theta}_0,\boldsymbol{\sigma}) = \sum_{i=1}^{N} w_i \cdot \mathcal{AWN}(\theta|\theta_{0,i},\sigma_i) \quad (6)$$

The required function for the orientation estimation is thus obtained using a Mixture of Approximated Wrapped Gaussian as in Eq. 6:

$$p(\theta|I_k) = MoA\mathcal{WN}(\theta|\psi_k,\boldsymbol{C},\sigma) \quad (7)$$

where the variance σ was set to a fixed value for all the components. The σ parameter of Eq. 7 depends on the number of adopted classifiers: if σ is 0 the AWG step is disabled. Increasing the σ value includes in the final response the contributions of more neighbor classifiers.

The person direction is estimated using Eq. 3 through Mean-Shift optimization with starting seeds on the c^i values. Fig. 3 shows all the steps of the proposed method and in particular the final filtering step obtained with the Mixture of Approximated Wrapped Gaussians which is also described in Algorithm 2.

3 Experimental Results

The proposed approach has been extensively tested on three public datasets, namely the TUD Multiview Pedestrians dataset [1], 3DPeS [8] and SARC3D [7]. In the following, quantitative results are presented and compared to the state of the art; finally some qualitative outcomes on video sequences from PETS and 3DPeS are shown.

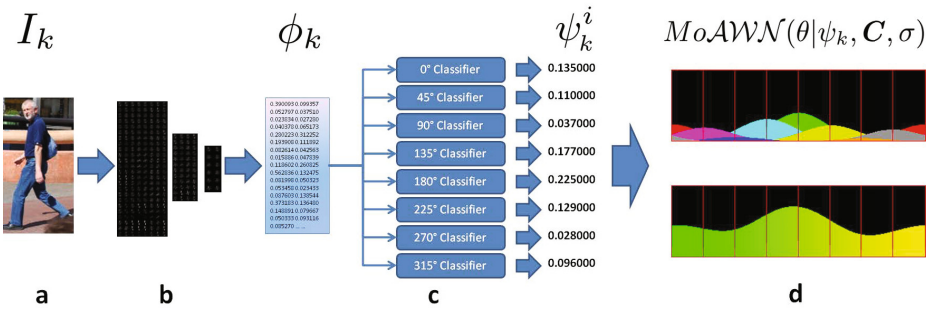


Fig. 3. A schema of the proposed method, (a) Input image, (b) Multi-Level HoG, (c) Array of classifiers, (d) the Mixture of Approximated Wrapped Gaussians

Algorithm 2. Orientation estimation with a Mixture of Approximated Wrapped Gaussians

Require: $N, \Gamma = \{\Gamma^1, \ldots, \Gamma^N\}$, set of trained classifiers

function FINDORIENTATIONS2(I)
  $\{I_k\} \leftarrow$ PEOPLEDETECTOR(I)
  for $k = 1 \rightarrow K$ **do**
    $\phi_k \leftarrow$ MULTILEVELHOG(I_k)
    for $i = 1 \rightarrow N$ **do**
      $\psi^i \leftarrow \Gamma^i(\phi_k)$
    end for
    $p(\theta|I_k) \leftarrow MoA\mathcal{WN}(\theta|\psi_k, \boldsymbol{C}, \sigma)$
    $\theta_k \leftarrow \text{argmax}_{\theta \in [-\pi \ldots \pi]}\, p(\theta|I_k)$ ▷ Mean Shift Maximization
    if Continuous Output **then**
      return θ_k
    else
      $i = \left(\theta_k + 2\pi \cdot \frac{N}{2\pi}\right) \mod N$
      return c_i
    end if
  end for
end function

3.1 Datasets and Metrics

The TUD Multiview Pedestrian dataset contains 5288 snapshot of pedestrians, fully annotated with bounding boxes, orientation classes and skeletons. We randomly selected 20% of the images from the provided training set to train the classifiers, then we use the same split originally proposed for validation and testing: 248 snapshots for validation and parameters estimation, and 300 images for testing.

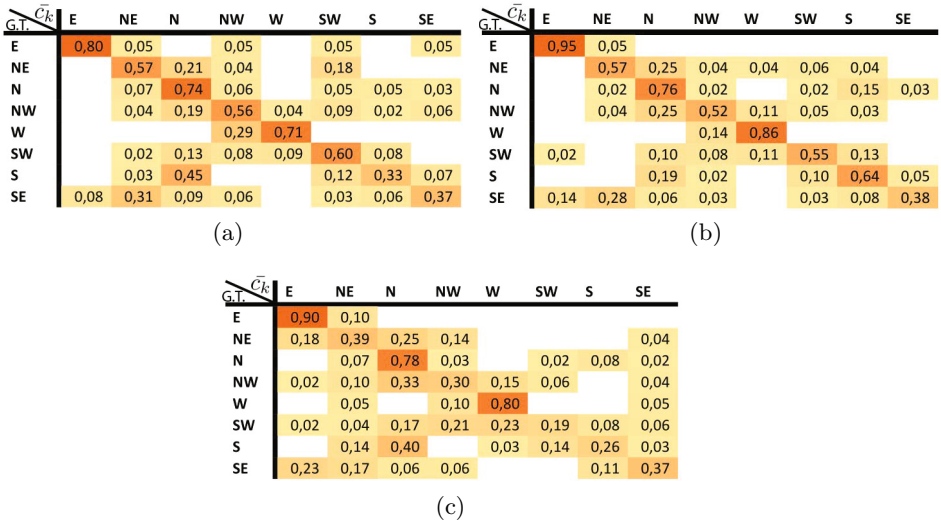
For the 3DPeS dataset, people snapshots were randomly selected from the provided videos and manually annotated. We obtained 1012 snapshots, 360 were used for training, 652 for testing.

Sarc3D provides 200 snapshots of 50 people taken from 4 predefined points of view. We used all the provided images for testing only, exploiting the same classifiers and parameters learned from 3DPeS, since images have similar characteristics (image resolution and camera point of view).

Test results are presented in terms of classification confusion matrices, where each row contains the ground-truth label whilst each column indicates the predicted one. We also report classification accuracy for each class, and two final measures: "Accuracy 8", where we consider exact hits only, and "Accuracy 4" where adjacent classes are also considered correct.

3.2 Performance Evaluation

The first extensive experiments were carried out on the TUD Multiview Pedestrian dataset. Fig. 4 shows the results of our method without (Fig. 4(a)) and with (Fig. 4(b)) the Mixture of Approximated Wrapped Gaussians filtering step. Predictably,



G.T. \ $\bar{C}_k$	E	NE	N	NW	W	SW	S	SE
E	0,80	0,05		0,05		0,05		0,05
NE		0,57	0,21	0,04		0,18		
N		0,07	0,74	0,06		0,05	0,05	0,03
NW		0,04	0,19	0,56	0,04	0,09	0,02	0,06
W				0,29	0,71			
SW		0,02	0,13	0,08	0,09	0,60	0,08	
S		0,03	0,45			0,12	0,33	0,07
SE	0,08	0,31	0,09	0,06		0,03	0,06	0,37

(a)

G.T. \ $\bar{C}_k$	E	NE	N	NW	W	SW	S	SE
E	0,95	0,05						
NE		0,57	0,25	0,04	0,04	0,06	0,04	
N		0,02	0,76	0,02		0,02	0,15	0,03
NW		0,04	0,25	0,52	0,11	0,05	0,03	
W				0,14	0,86			
SW	0,02		0,10	0,08	0,11	0,55	0,13	
S			0,19	0,02		0,10	0,64	0,05
SE	0,14	0,28	0,06	0,03		0,03	0,08	0,38

(b)

G.T. \ $\bar{C}_k$	E	NE	N	NW	W	SW	S	SE
E	0,90	0,10						
NE	0,18	0,39	0,25	0,14				0,04
N		0,07	0,78	0,03		0,02	0,08	0,02
NW	0,02	0,10	0,33	0,30	0,15	0,06		0,04
W		0,05		0,10	0,80			0,05
SW	0,02	0,04	0,17	0,21	0,23	0,19	0,08	0,06
S		0,14	0,40		0,03	0,14	0,26	0,03
SE	0,23	0,17	0,06	0,06			0,11	0,37

(c)

Fig. 4. Confusion matrices on the TUD Multiview Pedestrian Dataset: (a) without MoAWG filtering, (b) with MoAWG filtering, (c) using only the 4 main classifiers (E,N,W,S) and the MoAWG step. Each row contains the ground-truth label whilst each column indicates the predicted one.

the intermediate directions are more difficult to recognize; additionally opposite and specular directions are difficult to disambiguate. Directly using the outputs of the classifiers the average accuracy is around 58%. However, the MoAWG step usually improves the classification of ambiguous cases, reaching an average overall accuracy of 65%. In order to highlight the contribution of the MoAWG step, we performed an additional test. We reduced the training set using four main directions instead of eight, generating an array of 4 classifiers only (E,N,W,S). The continuous p.d.f. of Equation 7 is obtained as a Mixture of four components, but the final label is quantized in 8 classes, recovering the intermediate directions. The corresponding confusion matrix is reported in fig. 4(c).

Fig. 5(a) shows the confusion matrix of our method on the 3DPeS dataset, while Fig. 5(b) shows the system results on the Sarc3D dataset. The average accuracy of the system is still around the 60% in both cases. The Sarc3D dataset contains people oriented in 4 main directions only, but we used the array of 8 classifiers trained on the 3DPeS dataset.

Fig. 6 summarizes the results obtained on the three datasets, both with the MoAWG step and without it (indicated as No AWG in the table). As reported, the MoAWG step always improves the classification performance (from 4% on 3DPeS, to 25% on Sarc3D). Additionally accuracy 8 and accuracy 4 scores are reported. A visual example of the classification output on the TUD dataset is depicted in Fig. 7. Each image has been recolored following the rules of Fig. 2. The top half color of each image represents the ground-truth value, while the bottom half shows our results.

G.T. \ $\bar{C}_k$	E	NE	N	NW	W	SW	S	SE
E	0,86	0,12						0,02
NE	0,31	0,60	0,07			0,01		0,01
N	0,02	0,05	0,69	0,15	0,01	0,01	0,06	0,01
NW			0,08	0,51	0,29	0,04	0,04	0,04
W	0,15			0,15	0,70			
SW	0,04	0,01		0,02	0,29	0,56	0,08	
S	0,01	0,02	0,22	0,04		0,11	0,47	0,13
SE	0,32		0,05	0,05			0,05	0,53

(a)

G.T. \ $\bar{C}_k$	E	NE	N	NW	W	SW	S	SE
E	0,52	0,46				0,02		
NE								
N		0,24	0,74				0,02	
NW								
W	0,02			0,06	0,44	0,44	0,04	
SW								
S		0,02	0,26			0,06	0,66	
SE								

(b)

Fig. 5. Confusion matrices on the (a) 3DPeS and (b) Sarc3D datasets

	E	NE	N	NW	W	SW	S	SE	Accuracy 8	Accuracy 4
TUD	0,95	0,57	0,76	0,52	0,86	0,55	0,64	0,38	0,65	0,83
3DPeS	0,86	0,6	0,69	0,51	0,7	0,56	0,47	0,53	0,61	0,89
Sarc3D	0,52		0,74		0,44		0,66		0,59	0,87
TUD - No AWG	0,8	0,57	0,74	0,56	0,71	0,6	0,33	0,37	0,58	0,76
3DPeS - No AWG	0,82	0,68	0,69	0,66	0,54	0,58	0,38	0,42	0,59	0,87
Sarc3D - No AWG	0,3		0,78		0,26		0,54		0,47	0,9

Fig. 6. Performance summary on the three datasets

Using the same method, qualitative results on an excerpt of the "PETS2009 - Flow Analysis and Event Recognition" video sequence and on a video from 3DPeS are shown in Fig. 8 and Fig. 9 respectively. In this last case, the ground-truth was generated from the person trajectory on the ground plane and thus a precise orientation angle is used in the evaluation.

3.3 Comparative Evaluation

The system results have been compared against two state of the art techniques [1, 4] and other alternative solutions exploiting different classifiers and features on the TUD dataset. In particular, we evaluated the classification accuracy of the system where variations of Support Vector Machines and Random Forest classifiers are adopted instead of the Extremely Randomized Trees, and Covariance Descriptors [23] replace the HoG-based feature vector (Table 10). The first row (HoG - ERT -AWG) reports the performances of our complete method which outperforms all the other solutions:

- Eight Randomized Forests on HoG features with and without the MoAWG step (HoG - RT - AWG and HoG - RT - NoAWG);
- Eight SVMs in regression mode on HoG features with and without the MoAWG step (HoG - Multi SVMr - AWG and HoG - Multi SVMr - NoAWG);
- A single multi-class SVM used in Classification mode on HoG features (HoG - Single SVMC - No AWG); the response vector ψ is not available and thus the AWG step is not applicable;

Fig. 7. Qualitative results on some snapshots from the TUD dataset

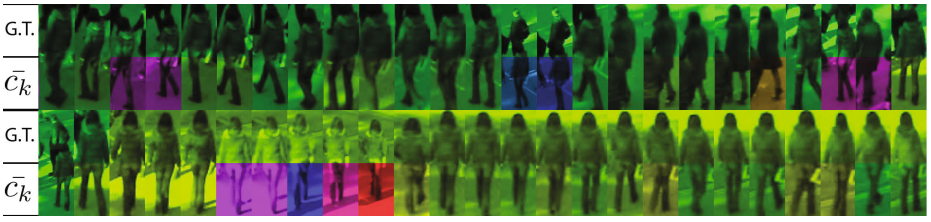


Fig. 8. Qualitative results on a excerpt from PETS2009; a woman dressed in red is initially walking from left to right and then away from the camera

- A single Multi-class Random Trees classifier on HoG features (HoG - Single RT - No AWG); similarly to the SVM based solution, the AWG step is not allowed in this case;
- Just four Extremely Randomized Forests on HoG features with and without the MoAWG step (HoG - 4 ERT - AWG and HoG - 4 ERT - NoAWG); The AWG step allows to recover the intermediate directions.
- Extremely Randomized Trees on classic Covariance descriptors [23] (COV - ERT - AWG and COV - ERT - NoAWG);
- Eight SVMs in regression mode on classic Covariance descriptors [23] (COV - SVM - AWG and COV - SVM - NoAWG);
- Extremely Randomized Trees and SVMr on modified Covariance descriptors, which embed information on mutual correlation between gradients in different image cells (COV2 - ERT - AWG, COV2 - ERT - NoAWG, COV2 - SVM - AWG and COV2 - SVM - NoAWG);

In all experiments the following parameters where used: for the Extremely Randomized Trees and Random Forests Classifiers the number of trees was set to 50, the maximum depth for each tree was set to 20. For the SVMs, ν-SVM classifiers were used, with a standard RBF kernel. The parameters γ was set to 0.000407,

Fig. 9. Qualitative results on images from a person tracked in 3DPeS

	E	NE	N	NW	W	SW	S	SE	Overall
HoG - ERT - AWG	0,95	0,57	0,76	0,52	0,86	0,55	0,64	0,36	0,65
HoG - ERT - NoAWG	0,8	0,57	0,74	0,56	0,71	0,6	0,33	0,37	0,58
HoG - RT - AWG	0,73	0,57	0,58	0,45	0,55	0,68	0,5	0,32	0,54
HoG - RT - NoAWG	0,56	0,42	0,51	0,45	0,55	0,65	0,28	0,28	0,46
HoG - Multi SVMr - AWG	0,5	0,26	0,47	0,41	0,3	0,6	0,18	0,44	0,39
HoG - Multi SVMr - NoAWG	0,43	0,15	0,38	0,31	0,15	0,16	0,6	0,33	0,31
HoG - Single SVMC - No AWG	0,75	0,35	0,65	0,52	0,7	0,68	0,65	0,44	0,59
HoG - Single RT - No AWG	0,87	0,58	0,77	0,54	0,6	0,49	0,2	0,23	0,53
HoG - 4 ERT - AWG	0,9	0,39	0,78	0,31	0,81	0,19	0,26	0,37	0,5
HoG - 4 ERT - NoAWG	1	0	0,88	0	0,9	0	0,42	0	0,4
COV - ERT - AWG	0,3	0,15	0,7	0,29	0,2	0,21	0,26	0,28	0,3
COV - ERT - No AWG	0,33	0,1	0,53	0,29	0,3	0,18	0,28	0,28	0,28
COV - SVM - AWG	0,3	0,15	0,38	0,06	0,4	0,1	0,29	0,12	0,23
COV - SVM - NoAWG	0,2	0,1	0,32	0,11	0,45	0,06	0,33	0,15	0,21
COV2 - ERT - AWG	0,16	0,15	0,28	0,11	0,15	0,16	0,72	0,25	0,25
COV2 - ERT - NoAWG	0,16	0,15	0,28	0,11	0,15	0,16	0,72	0,25	0,25
COV2 - SVM - AWG	0,36	0,15	0,41	0,18	0,2	0,05	0,23	0,12	0,21
COV2 - SVM - NoAWG	0,33	0,05	0,31	0,18	0,15	0,05	0,26	0,12	0,18

Fig. 10. Comparison of the proposed method with alternative solutions exploiting different classifiers and features

ν set to 0.5. For the MoAWG, σ was set to 0.75. Lastly, Fig. 11 compares our method against the one presented by Chen *et al.* in [4], which uses a similar feature vector and a sparse representation technique, and against the methods presented by Andriluka *et al.* in [1], which exploit banks of viewpoint specific part based detectors (linear SVMs) trained on the 8 orientation classes. In the first row, the results of our proposal are reported as reference. The second row shows the performance of [4], while the other rows contains three variants of the method by Andriluka *et al.* reported in [1]. In the first case (referenced as Max in the table), the orientation is estimated as the maximum over the outputs of 8 specific detectors; in the second case (SVM), eight additional SVMs are trained on top of the viewpoint specific detectors using training images from one class as positive samples and the remaining ones as negatives samples. Finally, in the

	E	NE	N	NW	W	SW	S	SE	Overall
HoG - ERT - AWG	0,95	0,57	0,76	0,52	0,86	0,55	0,64	0,36	0,65
Chen et al. [4]	0,65	0,37	0,71	0,53	0,7	0,59	0,41	0,36	0,55
Andriluka et al.- Max [3]	0,54	0,35	0,46	0,23	0,38	0,08	0,4	0,08	0,31
Andriluka et al. - SVM [3]	0,73	0,13	0,49	0,12	0,56	0,44	0,7	0,16	0,42
Andriluka et al.- SVM-adj [3]	0,71	0,22	0,29	0,18	0,85	0,18	0,5	0,29	0,35

Fig. 11. Comparison of the proposed method with the state-of-the-art

last case (SVM-Adj) orientations are grouped into triplets of adjacent directions and 8 linear SVMs are trained on such groups.

The method proposed in this paper outperforms all the previous solutions; in particular, the improvement with respect to the one presented by Chen *et al.* in [4] is around the 18%.

4 Conclusions

We proposed a new method for estimating the orientation of a person on single images based on appearance features only. A three level HoG feature set is extracted from each people detection. The feature vector is provided as input to an array of binary classifiers trained on a set of discrete orientations. Instead of assuming as output the orientation related to the winner classifier, a continuous probability density function of all the orientations is generated with a circular statistic approach. A mixture of approximated wrapped Gaussians is generated, where each component is centered on one of the discrete trained orientations and its weight is proportional to the corresponding classifier output. Several experiments show that a perfect solution to the problem is still missing, but the proposed method outperforms state of the arts techniques.

References

1. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition, pp. 623–630 (2010)
2. Ferrari, V., Marín-Jiménez, M., Zisserman, A.: 2D Human Pose Estimation in TV Shows. In: Cremers, D., Rosenhahn, B., Yuille, A.L., Schmidt, F.R. (eds.) Visual Motion Analysis. LNCS, vol. 5604, pp. 128–147. Springer, Heidelberg (2009)
3. Calderara, S., Prati, A., Cucchiara, R.: Mixtures of von mises distributions for people trajectory shape analysis. IEEE Trans. Circuits Syst. Video Technol. 21, 457–471 (2011)
4. Chen, C., Heili, A., Odobez, J.: Combined estimation of location and body pose in surveillance video. In: Proc. of IEEE Conf. on Advanced Video and Signal-Based Surveillance, pp. 5–10 (2011)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893. IEEE Computer Society, Washington, DC (2005)

6. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine Learning 63, 3–42 (2006)
7. Baltieri, D., Vezzani, R., Cucchiara, R.: Sarc3d: a new 3d body model for people tracking and re-identification. In: Proc. of IEEE Int. Conf. on Image Anal. and Process., Ravenna, Italy, pp. 197–206 (2011)
8. Baltieri, D., Vezzani, R., Cucchiara, R.: 3dpes: 3d people dataset for surveillance and forensics. In: Proc. of the 1st International ACM Workshop on Multimedia access to 3D Human Objects, Scottsdale, Arizona, USA, pp. 59–64 (2011)
9. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Pose search: Retrieving people using their pose. In: Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition, pp. 1–8 (2009)
10. Lanz, O., Brunelli, R.: Dynamic head location and pose from video. In: IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems, pp. 47–52 (2006)
11. Canton-Ferrer, C., Casas, J.R., Pardàs, M.: In: Head Orientation Estimation Using Particle Filtering in Multiview Scenarios, pp. 317–327. Springer, Heidelberg (2008)
12. Gourier, N., Hall, D., Crowley, J.L.: Estimating Face Orientation from Robust Detection of Salient Facial Features. In: Proceedings of Pointing 2004, International Workshop on Visual Observation of Deictic Gestures, ICPR (2004)
13. Setthawong, P., Vannija, V.: Improving the estimation of head pose orientation: By using eyeglasses as a key feature. In: Proc. of Int. Conf. on Information Technology and Multimedia (ICIM 2011), pp. 1–6 (2011)
14. Ozturk, O., Yamasaki, T., Aizawa, K.: Estimating human body and head orientation change to detect visual attention direction. In: Proc. IEEE Int. Conf. Comput. Vision, ACCV 2010, pp. 410–419. Springer, Heidelberg (2011)
15. Huang, C., Ding, X., Fang, C.: Head pose estimation based on random forests for multiclass classification. In: 2010 20th International Conference on Pattern Recognition (ICPR), pp. 934–937 (2010)
16. Cristani, M., Bazzani, L., Pagetti, G., Fossati, A., Tosato, D., Del Bue, A., Menegaz, G., Murino, V.: Social interaction discovery by statistical analysis of f-formations. In: British Machine Vision Conference, BMVC (2011)
17. Chen, C., Heili, A., Odobez, J.M.: A joint estimation of head and body orientation cues in surveillance video. In: Proc. IEEE Int. Conf. Comput. Vision Workshops, pp. 860–867 (2011)
18. Rogez, G., Rihan, J., Ramalingam, S., Orrite, C., Torr, P.: Randomized trees for human pose detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8 (2008)
19. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition, pp. 304–311 (2009)
20. Mardia, K.V.: Statistics of directional data. Journal of the Royal Statistical Society. Series B (Methodological) 37, 349–393 (1975)
21. Agiomyrgiannakis, Y., Stylianou, Y.: Wrapped gaussian mixture models for modeling and high-rate quantization of phase data of speech. IEEE Trans. on Audio Speech And Language Processing 17, 775–786 (2009)
22. Bahlmann, C.: Directional features in online handwriting recognition. Pattern Recognition 39, 115–125 (2006)
23. Tuzel, O., Porikli, F., Meer, P.: Pedestrian detection via classification on riemannian manifolds. IEEE Trans. Pattern Anal. Mach. Intell. 30, 1713–1727 (2008)