Combination of RGB-D Features for Head and Upper Body Orientation Classification

Laurent Fitte-Duval, Alhayat Ali Mekonnen, Frédéric Lerasle

Combination of RGB-D Features for Head and Upper Body Orientation Classification

Laurent Fitte-Duval, Alhayat Ali Mekonnen and Frédéric Lerasle

LAAS-CNRS, Université de Toulouse, CNRS, UPS, Toulouse, France
{lfittedu, aamekonn, lerasle}@laas.fr

Abstract. In Human-Robot Interaction (HRI), the intention of a person to interact with another agent (robot or human) can be inferred from his/her head and upper body orientation. Furthermore, additional information on the person's overall intention and motion direction can be determined with the knowledge of both orientations. This work presents an exhaustive evaluation of various combinations of RGB and depth image features with different classifiers. These evaluations intend to highlight the best feature representation for the body part orientation to classify, i.e, the person's head or upper body. Our experiments demonstrate that high classification performances can be achieved by combining only three families of RGB and depth features and using a multiclass SVM classifier.

Keywords: Head Pose Estimation, Upper Body Pose Estimation, Multiclass Classification, Feature Combination

1 Introduction

A person's head and body orientations convey important cues about the intention of the person. Whether the person is trying to interact with an intelligent machine or another person, orienting ones head and body towards the agent is a natural way to establish engagement. As a result, automated perception of people's head and body orientation has attracted a lot of attention in computer vision, human-machine interaction (HMI), and robotics disciplines. Possible applications are many: relevant examples include, user's intention characterization in human-robot interaction (HRI) [15], social interaction trends analysis [2], automated sport video analysis [9], human attention understanding for business and perceptual interface, etc. It can also be used to improve people tracking [2], body pose estimation [9], and action recognition [14] functionalities.

Nevertheless, correct estimation of people's head and body orientation is very challenging due to low image resolution, poor lighting conditions, frequent partial occlusions, and articulated body poses. In the past, the majority of approaches relied on RGB cameras [16, 2, 9]. But, their performance has been hampered because of their sensitivity to lighting condition, resolution, and lack of 3D information. With the advent of commercially available consumer RGB-D cameras like the Kinect and Asus Xtion, improved performance has been recorded, primarily as a result of the added depth information and its insensitivity to lighting conditions [5, 7, 13].

RGB-D based head pose estimation has been popularly addressed as a regression problem with approaches that provide continuous head pose angular estimates [5, 19]. But, these approaches require high resolution data and hence work only in very close range ($<$ 2m). For applications entailing further operating ranges, a classification approach with coarse discrete orientation classes is privileged [10] (referred here as orientation classification than pose estimation). This alleviates the need to obtain precise ground truth for head pose, which is difficult, and is reasonably sufficient for user intention understanding applications. On the other hand, to determine body orientation, the trend is to extract discriminant features from a coarsely segmented full person (usually obtained by employing a pedestrian or people detector) and apply a trained classifier [7, 13]. These approaches, however, deteriorate in presence of partial occlusions, for instance, partial occlusions of the legs, which is a common occurrence in close human-machine interaction. Estimating body orientation based on upper body data (pertaining to shoulder orientation) helps alleviate this shortcoming. Similar to head orientation, by using both RGB and depth data and considering discrete orientation classes rather than continuous estimation, overall performance over a wide operating range can be improved [7].

In this work, we investigate head and upper body orientation classification (discrete classes) based on RGB and depth image features, and linear and non-linear classifiers. Our aim is to classify the orientation of a person's head (yaw angle) and body (horizontal shoulder anterior orientation) independently into eight discrete classes.The upper body consideration enables body orientation classification in all ranges (especially in close range where full body based approach severely deteriorates). In both cases, the depth information robustifies performance in close and medium range operation, and the added RGB compensates the deficit in depth data in far range. Our work relies on popular RGB and depth features: local binary patterns (LBP) [18], histogram of oriented gradients (HOG) [3], depth local binary patterns (LBP_D), and histogram of depth difference (HDD) [24]. Additionally, multiscale variants of HOG and HDD features are also considered. For orientation classification, three different multiclass classifiers are considered: Random forest (RF), linear support vector machine (SVM), and sparse based classifier (SBC). This kind of systematic RGB-D feature combinations evaluation for head and upper body orientation classification is lacking in the literature. All evaluations are based on a recently released RGB-D public dataset [13]. The work is presented organized as follows: The rest of this section discusses related work and contributions. Section 2 addresses the adopted head and upper body representation with emphasis on the feature sets considered. It is then followed by a description of the different classifiers used in Section 3. Experiments and comparative results are presented in Section 4 and finally, the paper finishes with concluding remarks in Section 5.

1.1 Related work

The majority of works in head orientation estimation are presented as head pose regression, predicting the continuous 2D or 3D head orientation, e.g., [5,

16]. Though very useful and informative, these approaches obtain acceptable performance in close range. For medium and far range applications, a classification approach with discrete orientation classes is preferred [9, 10]. Depending on the intended application, as is the case in this work, the coarse orientation estimate provided could be sufficient. Another point on head orientation estimation is the data used. RGB data has been extensively used (see survey [16]), but the recent advent of consumer RGB-D sensors have shifted the focus from RGB based approaches to mainly depth based approaches [5, 19]. Furthermore, though demonstrated in close range, improved performance can be obtained by using both RGB and depth image data [8].

On the other hand, in human body orientation classification, the objective is to determine a person's body orientation angle (yaw angle). This is a relatively simplified problem than 3D human pose estimation, and yet conveys invaluable information about the heading direction of a person and his/her intention. Human body orientation classification can be achieved based on either full body [2, 22, 7] or upper body [9, 6] image data analysis. Most works are based on RGB images in video-surveillance contexts, e.g., [2, 22], though recent trends in RGB-D sensors made it possible for more improved full body orientation classification [7, 13]. Full body approaches do not work well in presence of partial occlusions, either due to close human-camera distance or intra-person occlusions when multiple persons are in view. This shortcoming is better alleviated with upper body approaches, e.g., [6, 9]. These works are based on RGB images. As shown in this work, further improvement can be obtained by adding depth information. Our work investigates RGB-D data for both head and upper body orientation classification which, though evident, is lacking in the literature.

A popular paradigm for orientation classification is to first extract relevant features from a bounding box encapsulating a body part (head or upper body), usually provided by a detector, then to utilize a trained multiclass classifier to determine orientation class. Although some approaches have tried a coupled detection-orientation classification paradigm that detects and determines orientation in one go, e.g., [22], the former is preferred as it dissociates the tackled problem. Furthermore, detector performance has improved significantly [4] so more focus can be dedicated to orientation classification. In the literature, relevant features considered include HOG [9, 22, 2], LBP [18], and HDD [7]. The trend seems to pick a specific family of feature set and use it without any systematic feature combination evaluation. For classification, popular choices are multiclass SVM (one-vs-all configuration) [7], sparse representation based classifier [2, 6], and random forest [22, 9]. This non-comparative feature-classifier trends, added with the lack of benchmarking in RGB-D approaches using public dataset makes feature-classifier choices difficult. Even though Liu *et al.* [13] introduced a public RGB-D dataset, called MCG-RGBD dataset, it has not been extensively used by the community yet. In light of these challenges, our work presents evaluation of several RGB and depth feature combinations for head and upper body orientation classification. All experiments are carried on MCG-RGBD to facilitate future benchmarking.

1.2 Contributions

The main contribution of this work is a comprehensive evaluation of several RGB and depth features, and their combinations, on a public dataset for head and upper body pose estimation. It also reports classification results based on three multi-class classifiers, capturing the essence of existing approaches and making future benchmarking easy.

2 Head and upper body representation

We start with the premise that the position of the head and upper body of people in an image are known (in the form of a bounding box). Typically, this can easily be obtained using an upper body detector and a head region segmentation technique [6, 12]. The next steps for orientation classification are relevant feature set extraction and classification, bearing in mind an underlying discrete orientation class representation. This section presents orientation class representation aspects and the adopted heterogeneous feature sets.

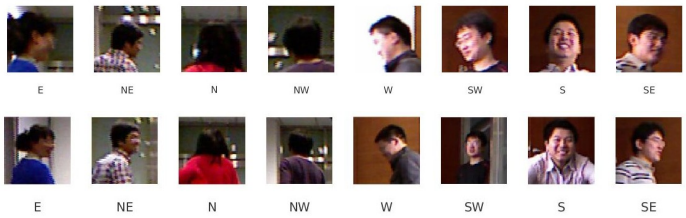


Fig. 1: Illustration of the head (top row) and upper body (bottom row) eight discrete orientation classes.

2.1 Discrete Orientation Classes

The head pose representation is usually defined by its pitch, yaw and roll angles [16]. Considering our problematic, we focus on the yaw angle which is discretized into eight orientation classes equally distant at 45° [9, 20]. Similar to previous works on body orientation classification [1, 2], we also quantize the upper body orientation into the same eight discrete orientations.

These eight orientations (Figure 1) analogically denotes the four cardinal directions with the four intercardinal directions where E, NE, N, NW, W, SW, S, NE corresponds to these directions considered around the yaw rotation axis. In order to determine the actual pose of the body part, a multi-instances classification problem is considered where the classes are the eight possible directions of the considered body part. The predicted direction computed as an output of the classifier will give essential cues that indicate a user's intention.

2.2 Feature Sets

The choice of the features for our work has been inspired from previous work in orientation classification and person detection.

HOG [3] The HOG features is the most widely used feature for person detection and full body pose classification in the literature [1]. The computation of the feature is based on a division of the considered window in cells of equal sizes which will be efficiently associated in order to gather the gradient orientations computed in a histogram. The variations of values in this histogram are characteristics of local shape of the classified object. The final feature is obtained by concatenating all block histogram with a dimensionality function of the number of bins used to divide the gradient orientation range and the number of subdivisions in the windows.

HDD [21, 24] This feature set is extracted by applying similar procedure as in the original HOG feature on the depth data. It tries to compute a histogram of depth difference based on the disparity of depth variations, and extending the orientation space and scaling the depth information in a suitable way to improve its representation.

Multiscale HOG and HDD [2, 22] These feature sets generally compute the features (HOG or HDD) at three scales which are multiples of each other by a factor of two before concatenating the generated feature vectors into a final multilevel feature. M-HOG and M-HDD denote these multiscale variants.

LBP and its depth variant LBP$_D$ [17, 11] LBP is a robust texture descriptor because of its invariance to gray-scale and rotation. It mainly consists to label a pixel after testing a threshold on its neighborhood. The simplicity of this image analysis allows a fast computation in addition to its ability to underline patterns while being immune to contrast changes. In the vein of [11] which proposes a new LBP-based feature for gender recognition, we decide to apply the LBP pattern to depth images in order to enrich our set of features with depth-based texture information.

3 Multiclass Classification

For classification, both head and upper body orientation classification is treated as a multiclass classification problem with as many classes as number of considered discrete orientations. In this work, this results in an eight class multiclass classification problem. The classifiers are trained and tested based on the set of features described in Section 2.2. The three classifier types considered are presented below.

Random Forest (RF) Random forest is an ensemble methods that uses N randomly trained decision trees (separately trained in parallel) of depth D to create a strong classifier. It uses the average of each tree output to define the final classification. There are several variants of random forest learning strategies. In this work, each decision tree is learned using random samples drawn with replacement from the training set. In addition, when splitting a decision tree node during the construction of the tree, the split that is chosen is no longer the best split among all features. Instead, the split that is picked is the best split among a random subset of the features.

SVM One-vs-All Support vector machines (SVMs) are statistical supervised learning methods used for classification and regression [23]. Linear SVM (used here) specifically, uses a hyperplane to define the decision boundary that separates the two classes. To extend it as a multiclass classifier, the one-vs-all strategy is adopted which involves fitting one classifier per class. For each classifier, the class is fitted against all the other classes. In addition to its computational efficiency, it is easily interpretable. Since each class is represented by one and one classifier only, it is possible to gain knowledge about the class by inspecting its corresponding classifier. The final classification label is determined as the one that maximizes the classification score.

Sparse Representation based Classification (SBC) In [2], a sparse representation approach for multi-instances classification with proven efficiency in face recognition was introduced. The objective is to project the feature vector in a base of the considered classes by approximating the feature vector as a linear combination of the training features. The reconstruction weights of this decomposition are subject to a non-negative constraint and obtained using an L_1 regularization. These weights will have non-zero values if they corresponds to the data associated to the same class. Summing all the values of this sparse decomposition, it is possible to calculate the probability of each class. Then the maximal probability gives the output label.

4 Experiments

We carry out an exhaustive evaluation of different feature combinations in order to emphasize a trade-off between feature representation, classification effectiveness, and CPU cost.

4.1 Evaluation Metrics

The classification is evaluated based on standard confusion matrices where the columns corresponded to the predicted classes while each row corresponds to the ground truth classes. Concentrated detections along the diagonal indicate good performance. We can extract the classification accuracy per class considering the exact instances of the class normalized by all the classified instances for this same class. Then we can average the accuracy for all the classes which is our first performance criteria, accuracy 1 ($acc1$). We consider a second criteria, accuracy 2 ($acc2$), where the predictions to one of the two adjacent classes are considered as correct as in [1].

4.2 Dataset

In order to evaluate our approach and make future benchmarking easy, the MCG-RGBD public dataset [13] is used. The dataset contains 10 RGB-D video sequences of 11 people acting indoors in three different scenes (meeting room, corridor, and entrance) with a 640×480 resolution. It includes a wide variety

of situations as walking, standing, jumping, running rotating, etc. The sensor limits the field of the acquisition allowing to observe people between 2.5 and 10 meters. As it focuses on people's global body orientation in a similar way as [1], the yaw angle of each person is provided as ground truth. The dataset contains a total of 4000 images. Head and upper body bounding boxes have been manually annotated for evaluation.

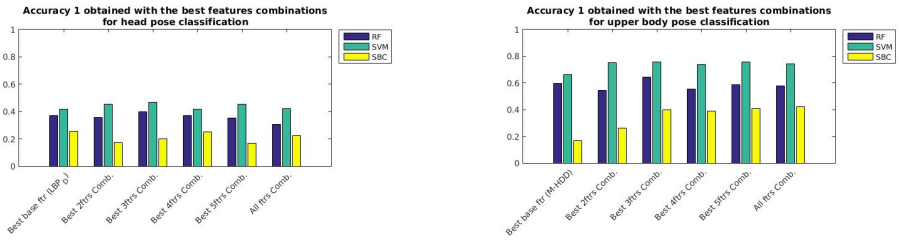
We divide the dataset into training and testing sampled in a 2/3 and 1/3 proportion, respectively. The training set is doubled by adding reflections of each annotated data. It is then filtered to discard occluded samples. In our experiments, we observed that using equal number of samples in each class for training improved overall performance. Hence, the final training set consisted of an equal sample of 116 instances in each class, resulting in a total of 928 samples for training. The final test set consists of 2256 annotated samples.

4.3 Implementation Details

Before generating the different features, we extracted head and upper body windows from the dataset which are normalized to a fixed size of 64×64 pixels. When computing HOG, an 8×8 pixels cell size, a 2×2 cells block size, and a 9 bins gradient orientation quantization (same parameters as in [3]). As in [2], HOG features and the derived HDD features are computed using non-overlapping blocks, meaning a block stride of 8 pixels or more generally equal to the cell length. The multiple scale variants of HOG and HDD features (M-HOG and M-HDD) use respectively one and two additional cells with sizes of 16×16 and 32×32 pixels for head and upper body feature computation. The difference between RGB and depth based features is with the spacing of the bins which extends from 180° in the first case and 360° in the second case. The additional texture information from the LBP is computed on both RGB and depth channels (LBP_D) using an efficient implementation inspired by the works in [17] and integrated in a channel way as in [6]. The six base features presented in Section 2.2 (LBP, LBP_D, HOG, M-HOG, HDD, M-HDD) are all combined in every possible way leading to 63 features sets which are all evaluated. The dimensions of our six base features vary from 336 to 1512 while the dimensions of their combinations vary from 672 to 4668.

Regarding the multiclass classifiers, the random forest (RF) parameters, number of trees and the maximum depth of tree, are obtained by cross validation varying the values from 5 to 60 and from 5 to 30, respectively. The SVM classifier is used with default parameters and a unitary penalty parameter C. The sparse-based classifier (SBC) does not require any mandatory parameters but requires the same necessary parameters than any classification approaches: matrices of training and testing data associated to a ground truth associating a class to each of the considered sample. During our experiments, the sparse-based approach proposed in [2] reveals to be compelling to tune. A tolerance depending to the features dimension has to be set up in the L_1 solver used and computation of its value for the wide variety of feature combination has revealed to be constraining. Relaxing non-negative constraint and using a L_2 norm for the regularization, the approach remains time consuming although slightly faster and

presenting better performances. The L_2 regularization is easily computed using the pseudo-inverse of the feature matrix. The sparse based classifier used for our evaluation differs of the original approach but presents more pertinent results for our evaluation.



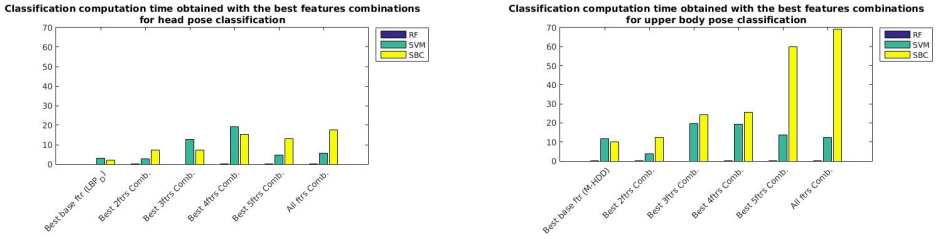
(a) (b)

Fig. 2: Classification performance result, *acc*1, for: (a) head orientation classification, and (b) upper body classification by classifier. See text for description.

4.4 Results

In this section, we will present the main results of our analysis but an extensive comparison of the feature combination is developed in the supplementary material[1]. The results observed allow to compare some approaches of the literature with the new combination of features and classifiers observed. The main works dealing with orientation classification in the literature are using the HOG features associated with SVM [7] or random forest [9] or the multiscale variant of HOG associated with random forest [22] and sparse-based classifier [2]. Fig. 2a and Fig. 2b depict the *acc*1 results obtained for head orientation and upper body orientation classification, respectively. For brevity[1], of all combinations evaluated, we present the best single feature (base ftr), and two (2fts comb.), three (2fts comb.), four (2fts comb.), and five (2fts comb.) features combinations. Additionally, we also present the result obtained with all combined features (all ftrs comb.). The results are reported for each multiclass classifier. The best features for head orientation classification are (considering single, two, three, four, and five combinations): LBP_D, LBP_D+M-HOG, LBP_D+M-HOG+M-HDD, LBP+LBP_D+HOG+M-HOG, and LBP_D+HOG+M-HOG + HDD+M-HDD. For that of upper body, they are: M-HDD, HOG+HDD, LBP_D+M-HOG+HDD, LBP+LBP_D+HOG+HDD, and LBP_D+HOG+M-HOG+HDD+M-HDD. As observed, in the two histograms in Fig. 2, the first two classifiers (SVM and RF) are in the same order of performance whereas the sparse-based approach achieve lower performances. This observation is easily explained by the use of the entirety of data available without selection or optimization of its parameters in this approach used whereas the two first are trained to keep the most pertinent features or parameters during the classification process.

[1] For extensive evaluation results, please refer to the supplemental material at http://homepages.laas.fr/aamekonn/acivs16/supplement.pdf



(a) (b)

Fig. 3: Single instance classification times (time reported in seconds) for: (a) head orientation classification, and (b) upper body classification by classifier. See text for description.

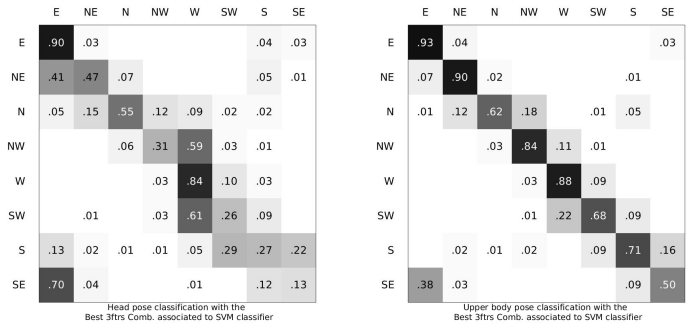
For computation time aspect, Fig. 3 presents the CPU cost of the best feature combinations determined during evaluation on the test set. Regardless of the number of features combined, which might directly affect the CPU cost, the random forest runs in times of the range of the tenth of a second whereas the SVM classifier runs in tens of seconds. This difference is due to the features selection realized during the random forest training whereas the SVM compute an optimal separation hyperplane using all the features available. This gap in scale would be a decision factor when integrating these functionalities, head and upper body orientation classification, in a wider application framework.

Table 1: Comparison of our best results and common approaches in the literature.

Approach	Classifier	Head			Upper body		
		Feature	*acc1*	*acc2*	Feature	*acc1*	*acc2*
Hayashi et al. [9]	RF	HOG	0.32	0.66	HOG	0.21	0.52
Tao et al. [22]	RF	M-HOG	0.32	0.64	M-HOG	0.21	0.51
Fumito et al. [7]	SVM	HOG	0.38	0.74	HOG	0.30	0.69
Chen et al. [2]	SBC	M-HOG	0.29	0.68	M-HOG	0.24	0.63
Ours best base ftr	SVM	LBP_D	0.42	**0.92**	M-HDD	0.66	0.88
Ours best 3ftrs comb.	SVM	LBP_D+M-HOG+M-HDD	**0.47**	**0.92**	LBP_D+M-HOG+HDD	**0.76**	**0.98**

Regarding the classification accuracy and more precisely *acc*1, common approach in the literature are superseded seeing that the best scores obtained using one feature easily exceed them (Table 1). We can notice that in both problem, the best score classifying the orientation with unique features are obtained by depth based feature. The LBP_D feature in the head case and the M-HDD for the upper body. The use of RGB-based features from the literature approaches appears insufficiently informative on our dataset. Globally, the scores obtained combining from two up to six heterogeneous features mixing depth and RGB information and using the SVM classifier features are of the same range (Fig. 2). The maximum *acc*1 scores are obtained combining three features. A score of 47% *acc*1 is obtained mixing LBP_D, M-HOG and M-HDD for head orientation classification, whereas a 76% *acc*1 is achieved mixing LBP_D, M-HOG and HDD for upper body orientation classification. Upper body based classification demonstrates a significantly better accuracy, nearly 30%, than head orientation. This

gap is due to the extension of the analyzed area which reduce the ambiguity between classes. Considering the second performance criteria, $acc2$, scores of 92 and 98% are observed. This means that, on average, there is a 47% confidence to correctly classify a head orientation and there is a 45% possibility that it is misclassified for a neighboring orientation. An average error score of 8% for head orientation and 2% for upper body (considering neighboring classes as correct) on our orientation classifications is an outstanding performance with regards to the literature.



(a) (b)

Fig. 4: Confusion matrix for (a) head and (b) upper body orientation classification obtained using the best approaches – best three features combined with SVM classifier.

To further illustrate the classification performance in each discrete class, Figs. 4a and 4b show the confusion matrices for the best head and upper body orientation classification results respectively. The best results pertain to the three feature combinations highlighted in Table 1. From the confusion matrices, we can observe a matrix presenting some full lines for the head case whereas the upper body one present a sparser structure. On the first case, we have concentrated scores for the lateral classes (W and E) and spread estimations with imprecision on the orientation classification for the frontal and dorsal class (N and S). However, there is a global concentration of the estimations around the diagonal explaining the high score of 92% $acc2$. On the second case, we have a sparse matrix presenting a light cross shape. This low score symmetrically to the high score are due to the front/rear ambiguity. Nevertheless, high scores are reached all along the diagonal leading to the high scores for the two performance criteria. These differences confirm the complexity rising according to the size of the body part considered but an overlap would easily be established if these analysis were realized jointly.

5 Conclusion

In this work, we presented an extensive evaluation of several RGB and depth feature set combinations for head and upper body orientation classification. We

showed the interest of adding the depth information. Using the heterogeneity of this information, we obtain a 47% and a 92% accuracies ($acc1$ and $acc2$ respectively) for head orientation classification. For upper body orientation classification, accuracy scores of 76% and 98% are obtained. Our results also attest that by using a combined feature set composed of a single variant of each LBP, HOG, and HDD features, it is possible to obtain the best classification performance. The preferred variants are LBP_D and M-HOG, with M-HDD for head orientation and with HDD for upper body orientation classification. The best results are indeed obtained by using both RGB and depth based feature sets. In addition, our experimental results indicate that better results are obtained with the SVM classifier. But, this should be considered in light of the intended application, as the improvement obtained using SVM over RF based approach might not justify the incurred CPU cost (which is at least higher by an order of magnitude).

Future prospects include integration of the best trained model in human-robot interaction context and using the classification output as percepts for user's intention estimation. We also believe further improvements on head and upper body orientation estimation can be obtained using probabilistic filtering approaches, possibly with underlying head-should physiological models.

Acknowledgment

This work is funded by the ROMEO2 project (http://www.projetromeo.com/) in the framework of the Structuring Projects of Competitiveness Clusters (PSPC).

References

1. M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*, pages 623–630, June 2010.
2. C. Chen, A. Heili, and J. Odobez. Combined estimation of location and body pose in surveillance video. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS'11)*, pages 5–10, 2011.
3. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 886–893 vol. 1, June 2005.
4. P. Dollar, R. Appel, S. Belongie, and P. Perona. Fast Feature Pyramids for Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1532–1545, 2014. 00127.
5. G. Fanelli, J. Gall, and L. Van Gool. Real time head pose estimation with random regression forests. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*, pages 617–624, June 2011.
6. L. Fitte-Duval, A.A. Mekonnen, and F. Lerasle. Upper Body Detection and Feature Set Evaluation for Body Pose Classification. In *International Conference on Computer Vision Theory and Applications (VISAPP'15)*, pages 439–446, 2015.
7. S. Fumito, D. Daisuke, I. Ichiro, M. Hiroshi, and F. Hironobu. Estimation of Human Orientation using Coaxial RGB-Depth Images. In *International Conference on Computer Vision Theory and Applications (VISAPP'15)*, pages 113–120, 2015.

8. R.-S. Ghiass, O. Arandjelović, and D. Laurendeau. Highly accurate and fully automatic head pose estimation from a low quality consumer-level rgb-d sensor. In *Workshop on Computational Models of Social Interactions: Human-Computer-Media Communication*, pages 25–34, 2015.
9. M. Hayashi, T. Yamamoto, Y. Aoki, K. Ohshima, and M. Tanabiki. Head and Upper Body Pose Estimation in Team Sport Videos. In *IAPR Asian Conference on Pattern Recognition (ACPR'13)*, pages 754–759, November 2013.
10. C. Huang, X. Ding, and C. Fang. Head Pose Estimation Based on Random Forests for Multiclass Classification. In *International Conference on Pattern Recognition (ICPR'10)*, pages 934–937, August 2010.
11. T. Huynh, R. Min, and J.-C. Dugelay. An Efficient LBP-Based Descriptor for Facial Depth Images Applied to Gender Recognition Using RGB-D Face Data. In *ACCV Workshops (ACCVW'12)*, pages 133–145, November 2012.
12. O.H. Jafari, D. Mitzel, and B. Leibe. Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras. In *IEEE International Conference on Robotics and Automation (ICRA'14)*, pages 5636–5643, May 2014.
13. W. Liu, Y. Zhang, S. Tang, J. Tang, R. Hong, and J. Li. Accurate Estimation of Human Body Orientation From RGB-D Sensors. *IEEE Transactions on Cybernetics*, 43(5):1442–1452, October 2013.
14. S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3177–3184, June 2011. 00124.
15. C. Mollaret, A.A. Mekonnen, I. Ferrane, J. Pinquier, and F. Lerasle. Perceiving user's intention-for-interaction: A probabilistic multimodal data fusion scheme. In *IEEE International Conference on Multimedia and Expo (ICME'15)*, pages 1–6, June 2015.
16. E. Murphy-Chutorian and M. M. Trivedi. Head Pose Estimation in Computer Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626, April 2009. 00859.
17. T. Ojala and M. Pietikinen. Gray Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *Lecture Notes in Computer Science*, 1842:404–420, 2000.
18. T. Ojala, M. Pietikinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, January 1996.
19. C. Papazov, T. K. Marks, and M. Jones. Real-time 3d head pose and facial landmark estimation from depth images using triangular surface patch features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*, pages 4722–4730, June 2015.
20. T. Siriteerakul. Advance in Head Pose Estimation from Low Resolution Images: A Review. *International Journal of Computer Science Issues*, 9(3), March 2012.
21. L. Spinello and K.O. Arras. People detection in RGB-D data. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'11)*, pages 3838–3843, September 2011.
22. J. Tao and R. Klette. Integrated Pedestrian and Direction Classification Using a Random Decision Forest. In *IEEE International Conference on Computer Vision Workshops (ICCVW'13)*, pages 230–237, December 2013.
23. V. N. Vapnik. *The Nature of Statistical Learning Theory*. 1999.
24. S. Wu, S. Yu, and W. Chen. An attempt to pedestrian detection in depth images. In *Chinese Conference on Intelligent Visual Surveillance (IVS'11)*, pages 1–3, 2011.