Upper Body Detection and Feature Set Evaluation for Body Pose Classification

Laurent Fitte-Duval, Alhayat Ali Mekonnen and Frédéric Lerasle
CNRS, LAAS, 7, Avenue du Colonel Roche, F-31400 Toulouse, France
Université de Toulouse, UPS, LAAS, F-31400 Toulouse, France

Keywords: Upper Body Detection, Body Pose Classification, Fast Feature Pyramid, Sparse Classification, Aggregated Channel Features, Feature Evaluation

Abstract: This work investigates some visual functionalities required in Human-Robot Interaction (HRI) to evaluate the intention of a person to interact with another agent (robot or human). Analyzing the upper part of the human body which includes the head and the shoulders, we obtain essential cues on the person's intention. We propose a fast and efficient upper body detector and an approach to estimate the upper body pose in 2D images. The upper body detector derived from a state-of-the-art pedestrian detector identifies people using Aggregated Channel Features (ACF) and fast feature pyramid whereas the upper body pose classifier uses a sparse representation technique to recognize their shoulder orientation. The proposed detector exhibits state-of-the-art result on a public dataset in terms of both detection performance and frame rate. We also present an evaluation of different feature set combinations for pose classification using upper body images and report promising results despite the associated challenges.

1 INTRODUCTION

In Human Robot Interaction (HRI), one of the fundamental requirements is the correct detection and localization of human agents in the vicinity of the robot. The robot should be able to perceive the whereabouts of human agents in order to coordinate with them seamlessly. Depending on the exact application, the interaction space can vary from few centimeters to several meters. This proximity, in turn, introduces constraints about the field of view of the cameras mounted on the robot for instance. Usually, in close interaction the majority of cameras mounted on a robot can only see parts of the person, specifically the part above the thigh (figure 1), and any person detection mechanism adopted should take this into consideration.

In HRI, person detection relies on either classical RGB cameras, e.g., (Mekonnen et al., 2011), or RGB-D cameras that can provide 3D data like the Kinect sensor (Jafari et al., 2014). Due to physical, economical, and design constraints classical RGB cameras are predominantly found on robots. Hence, in this paper we will focus on perceptions based on 2D RGB images. The most popular approach for human detection in HRI is using a pedestrian detector that has been trained on full body annotated people dataset, e.g., (Dollár et al., 2012). Unfortunately, these detectors fail to detect people in presence of partial occlusions, specifically partial occlusions of the legs. But, by focusing on the upper part of the body, principally the head and the shoulders, it is possible to identify the presence of humans in the image under these circumstances. The approach, referred as upper body detection, is similar to a pedestrian detection but focuses on a smaller area which is less variable than the complete human body and is less exposed to the problem of occlusion (Li et al., 2009; Zeng and Ma, 2010). Figure 1 illustrates this point: given a typical Human-Robot (H/R) situation depicted on the left, the best pedestrian detector fails to correctly detect the two persons in front of the robot (bottom right) whereas our proposed upper body detector handles it perfectly (top right).

After identifying the human agents, we need to characterize their global behavior and their degree of intentionality to interact with the robot or with another human agent. Generally, the analysis of these cues is related to the head cues (Katzenmaier et al., 2004; Sheikhi and Odobez, 2012). The head cue indicates the direction of the person's visual point of interest and eventually the recipient of the person's

Fitte-Duval L., Mekonnen A. and Lerasle F..
Upper Body Detection and Feature Set Evaluation for Body Pose Classification.
DOI: 10.5220/0005313104390446
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 439-446
ISBN: 978-989-758-090-1

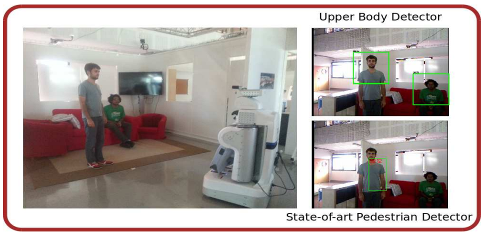


Figure 1: Output of our proposed detector (top right) and state-of-art pedestrian detector (Dollár et al., 2014), bottom right, in an HRI context.

speech if there is a discussion (Bazzani et al., 2013). But the estimation of orientations from both indicators, the head and the body, helps to know its posture or the direction of its movement in addition to its visual point of interest (Chen et al., 2011).

In this vein, we propose to estimate the orientation (body pose) of people using upper body cues. We present an extensive evaluation of different feature sets to identify the best feature set that can capture required discriminative cues for upper body pose classification. Given different scenarios as those pictured in Fig. 2, the upper body pose allows to differentiate a situation where two agents interact amongst each other without paying attention to the robot (left) from a situation where the agent is facing the robot forward for a possible interaction (right).

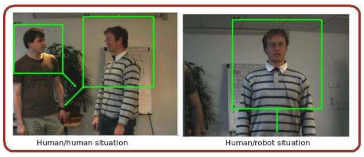


Figure 2: Output of our pose classifier in several interaction scenarios. The protruding arrows indicate pose orientation.

Related Works. Many researchers have investigated 2D upper body detection for human detection, e.g., (Li et al., 2009; Zeng and Ma, 2010; Eichner et al., 2012). The most frequent features used in these works are the Histogram of Oriented Gradient (HOG) (Dalal and Triggs, 2005) features which capture gradient distribution in the image. To date, these are the most discriminant features and the best results are achieved by approaches that use some variants of HOG (Dollár et al., 2012). Some works further improve the detection performance by considering heterogeneous pool of features, for example, combining Local Binary Pattern (LBP) and HOG features (Zeng and Ma, 2010; Hu et al., 2014).

Some recent advances on pedestrian detection focus on feature representation introducing the notion of Integral Channel Features (ICF) (Dollár et al., 2012). This representation takes advantage of the fast computation using integral image and combines different type of heterogeneous features to obtain outperforming detection results in comparison to state-of-the-art detectors based on HOG features. Dollár *et al.* (2014) propose an alternative representation of the channel features called Aggregated Channel Features (ACF) which slightly improves these performances. Both ICF and ACF are quite appealing as they have recorded outstanding performance both in terms of detection and computation time. Combined with a cascade classifiers configuration (Bourdev and Brandt, 2005) and an efficient multiscale representation using approximation of rescaled features for the detection process, this approach produces one of the fastest state-of-the-art pedestrian detector (Dollár et al., 2014).

Upper body pose estimation has also been investigated by various researchers, e.g, (Eichner et al., 2012; Weinrich et al., 2012). These works use different methods to retrieve an articulated model of a good configuration on a single image or on a sequence of images. To avoid the complexity associated with articulated models some works have investigated using data obtained from global pedestrian detection for pose classification in video surveillance context (Andriluka et al., 2010; Chen et al., 2011; Baltieri et al., 2012). These works use the same HOG features computed at several scales with different classifiers similar to those used in the detection process to attribute one of the possible direction to the observed pedestrian, for examples, SVMs (Andriluka et al., 2010), or random trees (Baltieri et al., 2012). The original approach of Chen *et al.* (2012), which is adopted in this work, uses sparse representation technique. This approach has proved robust to occlusion and data corruption.

Contributions. This paper makes two core contributions: (1) it presents an upper body detector based on 2D RGB image using Aggregate Channel Features (ACF) and soft cascade classifier that leads to state-of-the-art results in terms of both detection performance and frame rate; and (2) it presents a detailed comparative evaluation of various feature sets for upper body based pose classification using a sparse representation technique.

This paper is structured as follows: It begins with an overview of the framework in section 2. Sections 3 and 4 detail the approach used for upper body detection and pose classification respectively. All experiments carried out and associated results are presented in section 5. Finally, the paper ends with concluding remarks in section 6.

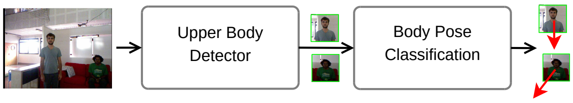


Figure 3: Adopted complete perceptual framework.

2 FRAMEWORK OVERVIEW

Figure 3 depicts a concise summary of the framework adopted in this work that highlights the two core components: upper body detector and body pose classification. As stated in section 1, given a sequence of RGB images, we are interested to detect people in the image stream and correctly classify their body poses. The upper body detector finds people in the image by using a sliding window approach which exhaustively scans the image at all possible locations and scales applying the trained model. This step is computationally intensive and is the main bottleneck that inhibits achieving high frame rates. Taking this into account, we use the ACF feature set combined with soft cascade classifier, which have been proven to lead to superior detection and speed, for detection. This block passes all detected upper bodies to the upper body pose classification block.

The body pose classification block determines the orientation of the provided upper body data. For this we have also adopted a proven technique based on sparse representation. We evaluate a plethora of feature sets including classical and multiscale ACF and HOG features and their combinations but only retain the feature set that leads to the best result for actual applications. Finally, this block provides orientation information for the detected upper bodies.

3 UPPER BODY DETECTOR

We present the key processes used in the detection framework: the feature representation, the classification algorithm structure and the algorithm for multiscale feature representation.

3.1 Aggregated Channel Features

Aggregated Channel Features (ACF) differs from Integral Channel Features (ICF) by using pixel lookups as features instead of sums over rectangular regions (Dollár et al., 2014). A channel C is a representation of an image I where the pixels are obtained applying a feature generation function Ω. So several channels $C = \Omega(I)$ can be defined using different feature transformations. The channel features used in this work are:

Gray and Color. The gray-scale is the simplest color channel of an image as $C = I$. For our detection module, we use the three LUV color channels which have proved informative for person detection (Dollár et al., 2012).

Gradient Magnitude. A non-linear transformation which captures edge strength.

Gradient Oriented Histogram. It consists of a histogram indexed by six gradient orientations (one channel per orientation bin) and weighted by the gradient magnitude. Normalizing the histogram using the gradient magnitude allows to approximate the HOG features.

Local Binary Pattern. We introduce this new channel feature in addition of the previous already used in the work in (Dollár et al., 2014). The Local Binary Pattern (LBP) is one of the best texture descriptor in the literature. We use an efficient implementation of the gray-scale and rotation invariant LBP descriptor inspired by the works of Ojala et al., (2002).

Once, all the channels $C = \Omega(I)$ have been computed, the channels are further divided into blocks. The pixels in each block are summed and the resulting lower resolution channels are smoothed to make up the ACF. Then, the ACF obtained are used to train a soft cascaded boosted tree (section 3.2). Figure 4 illustrates the different channels used in our work and the steps to generate the aggregated channels.

3.2 Soft Cascade

Boosting is a classification method that consists of combining weighted weak classifiers to create a strong classifier. The soft cascade classifier is a boosted classifier variant proposed by Zhang and Viola (2008) and used for pedestrian detection successfully in (Dollár et al., 2014). Unlike a classical cascade which has a predefined number of distinct stages, the soft cascade has a single stage with many weak classifiers. Rejection thresholds that make the “soft” cascades are calculated for each weak classifier in a process called cascade calibration that allows reordering of the “soft” stages to improve the classification accuracy. Its main strength is that it considers the information of each stage in the global process contrary to the classical attentional cascade which trained each stage individually. It also allows optimization of detection rate and execution time.

3.3 Fast Feature Pyramids

A feature pyramid is a multiscale representation of an image where channels $C_s = \Omega(I_s)$ are computed at

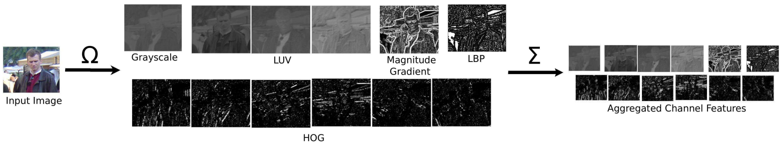


Figure 4: Aggregated Channel Features computation.

every scale s for the corresponding rescaled image I_s. The scales are organized per octave of 8 scales evenly sampled in log-space between one given scale and the next scale with half its value. Instead of generating each scale computing $C_s = \Omega(I_s)$, a feature channel scaling approximation is used :

$$C_s = R(C,s) \cdot s^{\lambda_\Omega} \tag{1}$$

where $R(C,s)$ denotes C re-sampled by s and λ_Ω, a lambda coefficient specific to the channel transformation Ω in order to scale the feature channel C using a power law. An iterative method can be deduced to efficiently generate a feature pyramid approximating channel features at intermediate scales s using the computed channel features at the closest scale s'as $C_s = R(C,s/s') \cdot (s/s')^{\lambda_\Omega}$. The best compromise consists to compute one scale per octave and approximate the 7 others.

This approach, pioneered by Dollar et al. (2014), combined with the presented ACF features and soft cascade classifier enables efficient sliding window based detection implementation. This leads to fast and highly accurate detector as will be shown in section 5.

4 UPPER BODY POSE CLASSIFICATION

The objective of this modality is to estimate the orientation of a person's upper body with respect to the camera. The adopted body pose classification approach has initially been applied for pedestrian direction estimation (Chen et al., 2011).

4.1 Upper Body Pose Representation

Tackling the upper body pose determination as a classification problem–rather than a regression problem due to limited training dataset–and in accordance with the literature, we consider eight evenly spaced directions in steps of 45 degrees (N, NE, E, SE, S, SW, W, NW) as the possible orientations, figure 5.

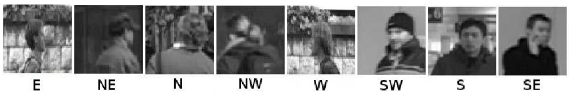


Figure 5: The eight upper body pose classes.

In the state-of-the-art work of (Chen et al., 2011), a multi-level HOG feature is generated to extract information from the human bounding box. Three different levels are generated using different cell sizes (respectively 8×8, 16×16 and 32×32) and the gradient orientation is quantized in 9 bins. In this work, we propose to use the ACF features representation in a sparse representation for pose classification. We also take advantage of the fast feature pyramid framework introduced in section 3.3 to generate multiscale ACF at three different levels similar to those used on the multi-level HOG feature representation. So we generate aggregated channels at three scales without approximating intermediate scales in the octave.

4.2 Classification by Sparse Representation

The upper body pose classification method adopted is the same as in (Chen et al., 2011) inspired by the work in face recognition in (Wright et al., 2009). We generate $\mathbf{F} = [F_1, F_2, ..., F_k]$, the training set matrix with $F_i (1 < i < k) \in \mathbb{R}^{m \times n_i}$, the matrix associated to the pose label i composed of n_i feature vectors of size m. A new feature vector $\mathbf{y}$ can be expressed as a linear combination of the training features:

$$\mathbf{y} = a_1 F_1 + a_2 F_2 + ... + a_k F_k = \mathbf{Fa}_0 \tag{2}$$

where $\mathbf{F} = [F_1, F_2, ..., F_k]$ and $\mathbf{a}_0 = [a_1, a_2, ..., a_k]^T$ are the concatenation of the coefficient vectors associated to each class. The coefficients of $\mathbf{a}_0$ obtained by solving the equation $\mathbf{Fa} = \mathbf{y}$ are non-zero coefficients if they are related to the actual pose of the new feature vector demontrasting the sparsity of this decomposition. An easy way to solve this problem is to use the l_1-minimization approach:

$$\mathbf{a*} = \arg\min \|a\|_1 \text{ subject to } \mathbf{Fa} = \mathbf{y}, \tag{3}$$

which can be solved using the pseudoinverse of $\mathbf{F}$. Giving this decomposition, for each class we can calculate its pose probability $\rho_k(\mathbf{y})$ as :

$$\rho_k(\mathbf{y}) = \sum a_i^* / \|a*\|_1, \tag{4}$$

Then classifying $\mathbf{y}$ consists of finding the maximal pose probability defined by:

$$class(\mathbf{y}) = \max \rho_k(\mathbf{y}) \tag{5}$$

5 EXPERIMENTS AND RESULTS

In this section, the different experimental considerations and carried out evaluations along with obtained results are presented in detail.

5.1 Features Set Considerations

The ACF features are primarily used in both the detection and body pose classification modules. Utilizing the same family of features is advantageous as it avoids further computational overhead in the global process. To see the effects of different channels on the detector and pose classifier modules, we evaluate different constituent channel feature combinations listed below:

- Magnitude gradient and six channels of histogram of oriented gradients (GM+HOG or **a**),
- Color channels, magnitude gradient, and histogram of oriented gradients (Clr+GM+HOG or **b**),
- Color channels, magnitude gradient, histogram of oriented gradients, and local binary pattern (Clr+GM+HOG+LBP or **c**),
- Magnitude gradient, histogram of oriented gradients, and local binary pattern (GM+HOG+LBP or **d**),

In the body pose classifier, the gray-scale channel is used instead of the 3 LUV color channels because current publicly available training sets for pose classification are mainly composed of gray-scale images.

5.2 Dataset and Implementation Specifications

In both modules, the upper body windows considered are cropped square windows from standard pedestrian windows–the same width and a third of their height from top (Fig. 5). Hence, a 64×64 base window dimension is adopted. It has been observed than using a square window with a third of the pedestrian window rather than half leads to marginal loss in overall detection performance on public datasets, but it is able to detect head and shoulders in situations where the person is close to the robot.

5.2.1 Upper Body Detector

Dataset. To train the upper body detector, we use the INRIA Dataset (Dalal and Triggs, 2005). The dataset contains 614 positive images containing 1237 annotated pedestrians in various situations including crowds. The negative samples are randomly selected from the 1218 people-free images. As in (Ferrari et al., 2008), the training set is augmented by perturbing the positive samples applying slight rotations (3 rotations in steps of 3^o) and mirror reflections. The positive training set is hence augmented 6 times consisting in total around 7000 samples. Introducing these variations allows better generalization of the classifier.

To test the trained detector, we use the InriaLite dataset, a subset of the INRIA person dataset containing 145 outdoor images with 219 persons in total, most of them entirely visible and viewed approximately from the front or from the back.

Detector Training. The training of the soft cascade used 2048 depth-two trees in four bootstrapping rounds where the misclassified negative samples are reused in the training process.

5.2.2 Upper Body Pose Classifier

Dataset. To train and test the body pose classifier, we use the TUD Multiview Pedestrians dataset (Andriluka et al., 2010). The dataset contains 400 to 749 annotated pedestrians in each class. It also contains a total of 248 validation and 248 test annotated images which, in this work, are used for test purposes resulting in combined 496 test instances.

Pose Classification Training. As presented in section 4.2, the training matrix **F** depends of the number of training samples per class n_i and the dimensionality m of the feature vector which is varied through our feature evaluation process. The parameters affecting m are specific to the features used. For the ACF, it depends of the number of channels which varies between 7 and 9 in the classification module and of the block size which divides the channels in blocks during the summation step. Here, we use a fixed block size of 4×4 pixels. To make sure that the body pose classifier does not over-fit, we test a classifier trained with differing samples, n_i, of equal sizes per each class varied between 200 and 400. Finally, the classifier is trained using the whole training set (variable number of samples per class).

5.3 Evaluation Metrics

Detection Evaluation. The detection evaluation protocol used is the same as in the PASCAL Visual Object Classes Challenge (Everingham et al., 2010). The detection system returns a series of detected windows after analyzing the image. These windows are obtained after multiscale detection and non-maximum suppression which avoid the accumulation of nearby windows around a person. Given a detected bounding box (BB_{dt}) and a ground truth bounding box (BB_{gt}), we consider the detection as correct if the overlap between the two windows exceed 50%:

$$overlap = \frac{BB_{dt} \cap BB_{gt}}{BB_{dt} \cup BB_{gt}} > 0.5 \qquad (6)$$

The non assigned BB_{dt} and BB_{gt} are counted respectively as false positives and false negatives. The comparison of the detectors is realized as in (Dollár et al., 2012), plotting the log-average miss rate (MR) against the false positives per image (FPPI).

Pose Classification Evaluation. The classification is evaluated using 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 preferred 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 ($acc.1$). We consider a second criteria, accuracy 2 ($acc.2$), where the predictions to one of the two adjacent classes are considered as correct as in (Andriluka et al., 2010).

5.4 Results

Upper Body Detection. To compare the performance of the proposed upper body detector variants (using different constituent channel features) with notable approaches in the literature, we evaluated three others upper body detectors including the OpenCV implementation of the Viola and Jones detector (Viola and Jones, 2001), our implementation of the Dalal and Triggs (2005) HOG-SVM detector, and the Deformable Parts Model (DPM) based detector proposed by the Calvin group (Eichner et al., 2012). All detectors are trained with identical training set.

As can be seen from the results depicted in figure 6, all the ACF based detectors outperform the HOG-SVM and Viola and Jones based upper body detectors significantly.

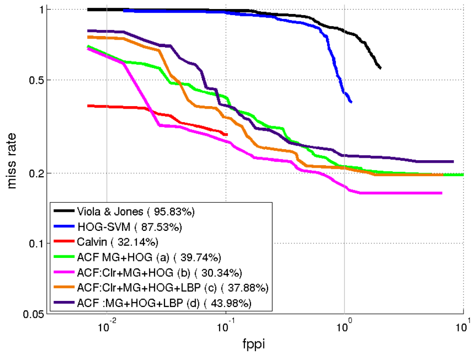
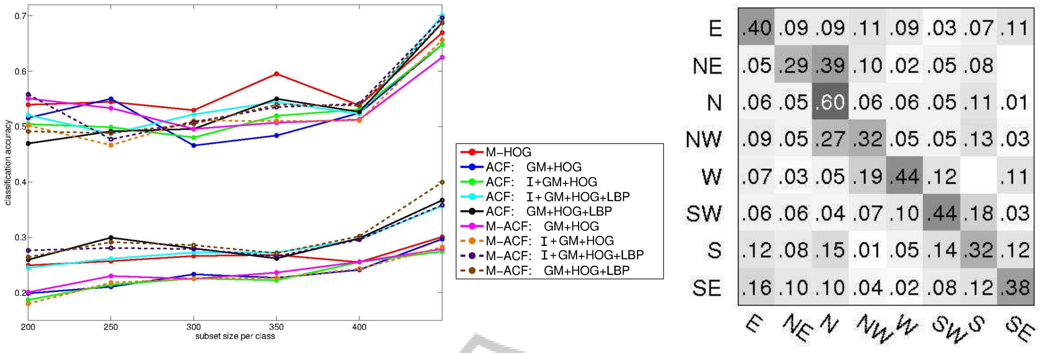


Figure 6: Log-average miss rate on InriaLite dataset.

The best results are obtained by the original combination of ACF combining the color, the magnitude gradient, and the histogram of oriented gradient channels which also outperforms the Calvin detector–the best upper body detector in the literature–achieving a 30.34% log-averge miss rate. This detector records a 0.18 Miss Rate at an average of 1 FPPI which can be further ameliorated by using a filtering mechanism. The combination of magnitude gradient and histogram of oriented gradients is the simplest channel combination used but also illustrates how informative the gradient orientation features are. Adding only the LBP channel decrease the performances of the detector whereas using all the channel features presented improves it relatively but overall it tends to over-fit the detector due to the increased number of feature sets. Computationally, with unoptimized Matlab based code, all the ACF based detectors run at approximately 12.5 fps on 640 × 480 images on an Intel core i7 machine using a single thread. This is sufficient for most real time robotic application requirements and is much higher than the 0.63 fps achieved by the Calvin detector.

Upper Body Pose Classification. The upper body pose classifier is evaluated using the combined 496 annotated samples from the TUD multiview pedestrian dataset. We generate classification matrices for each type of features (11 in total) and consider the two accuracies: $acc.1$ and $acc.2$. We evaluate the performance of the proposed approach based on ACF along with their multiscale variants and compare it with the approach using multi-level HOG features (Chen et al., 2011). Corresponding results are shown in figures 7(a) and 7(b). Figure 7(a) shows the variations in the performances of the classifiers on the test set as the data used for training is varied from 200 to 400 samples per class. The results confirm that as the training dataset is increased, there is no observed over-fitting. And indeed, the best results are obtained when using all the data in the training set (which corresponds to extreme data points in the plot). The confusion matrix depicted in figure 7(b) corresponds to the best classifier which is the multiscale ACF: GM+HOG+LBP.

Table 1 shows the dimensionality and the accuracies obtained for the upper body pose classification with the different combination of ACF using multiscale computation or not, in addition of the multilevel HOG features. The combination of magnitude gradient and histogram of oriented gradients results are close to those using the multi-level HOG features as we could expect because they use similar infor-



(a)

	E	NE	N	NW	W	SW	S	SE
E	.40	.09	.09	.11	.09	.03	.07	.11
NE	.05	.29	.39	.10	.02	.05	.08	
N	.06	.05	.60	.06	.06	.05	.11	.01
NW	.09	.05	.27	.32	.05	.05	.13	.03
W	.07	.03	.05	.19	.44	.12		.11
SW	.06	.06	.04	.07	.10	.44	.18	.03
S	.12	.08	.15	.01	.05	.14	.32	.12
SE	.16	.10	.10	.04	.02	.08	.12	.38

(b)

Figure 7: (a) Classification accuracy for upper bodies, *acc*1 (lower curves), *acc*2 (upper curves), plotted as a function of the amount of samples per each class; the extreme data points correspond to using the whole training set. (b) Confusion matrix for upper body pose estimation using multiscale ACF: GM+HOG+LBP.

Table 1: Upper body pose classification.

		upper body (64x64)		
Approach		dim.	acc. 1	acc. 2
Multi-level HOG		756	0.3	0.67
Aggregated Channel Features	(a)	1792	0.3	0.65
	(b)	2048	0.27	0.65
	(c)	2304	0.36	0.7
	(d)	2048	0.37	0.69
Multiscale Aggregated Channel Features	(a)	2352	0.28	0.63
	(b)	2688	0.28	0.66
	(c)	3024	0.36	0.7
	(d)	2688	0.4	0.7

mations. But, contrary to the detection module, the improvement of the results is given by the texture information from the LBP channel whereas the addition of the grayscale color channel decrease the results. Then the utilization of the multiscale information keeps improving the accuracies. The best exact precision is also obtained by the multiscale ACF: GM+HOG+LBP.

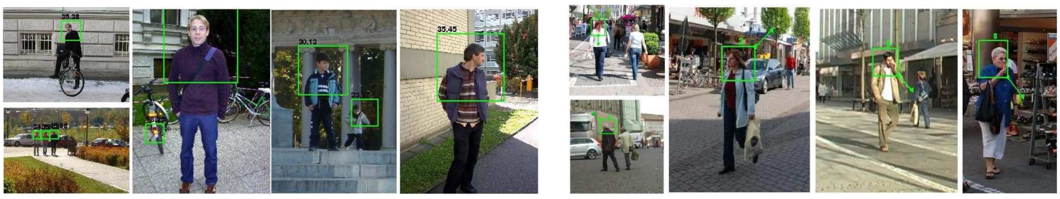
The classification matrix is quite full with a concentration of the scores on the diagonal. Some errors come from misclassified estimations for adjacent classes. The North-East and North-West orientations are often confused with North orientation for example. These errors are considered in the second accuracy criteria. This criteria has the same order of magnitude around 66 per cent of the estimates whatever the kind of features used. We can also quote the fact that the best pose estimation is the North orientation when the face is not visible whereas the score are lower for the other orientations when the face is visible. Even the symmetric confusions are less visible because of all the weak estimations along one orientation. These estimations allows to have a first idea of the context in the image but would need to be improved. Sample correct/incorrect classification and pose estimates are shown in figure 8.

6 CONCLUSIONS

In this work, we presented two important perceptual components using 2D images–upper body detection and upper body pose classification–that have pertinent applications in machine perception of humans. The presented upper body detector based on ACF features results in state-of-the-art detection result improving the previous best detector by a 2% average miss rate while improving computation speed 20x. We also presented upper body pose classification based on sparse representation using single scale and multiscale ACF features. Generally, the pose classification results showed comparable accuracy compared to the best approach in the literature. The results were further improved by the addition of the LBP channel features in the ACF framework leading to a 70% accuracy (*acc*.2), outperforming the best approach in the literature. Hence, the implemented perceptual functionalities based on ACF lead to state-of-the-art performance taking both accuracy and speed into account. In future works, the presented functionalities will be coupled with stochastic filtering approach to further improve the results and will be ported on our mobile robot platforms for human intention detection.

ACKNOWLEDGEMENTS

This work has been funded by ROMEO2 (http://www.projetromeo.com/) and RIDDLE projects – BPIFrance in the framework of the Structuring Projects of Competitiveness Clusters (PSPC), and a



(a) (b)

Figure 8: Some illustrations of upper body detections from the Inrialite dataset (a), and pose estimations from the TUD Multiview Pedestrians Set (b).

grant from the French National Research Agency under grant number ANR-12-CORD-0003 respectively.

REFERENCES

Andriluka, M., Roth, S., and Schiele, B. (2010). Monocular 3d pose estimation and tracking by detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 623–630.

Baltieri, D., Vezzani, R., and Cucchiara, R. (2012). People orientation recognition by mixtures of wrapped distributions on random trees. In *European Conference in Computer Vision (ECCV)*, pages 270–283.

Bazzani, L., Cristani, M., Tosato, D., Farenzena, M., Paggetti, G., Menegaz, G., and Murino, V. (2013). Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems*, 30(2):115–127.

Bourdev, L. and Brandt, J. (2005). Robust object detection via soft cascade. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 236–243.

Chen, C., Heili, A., and Odobez, J.-M. (2011). Combined estimation of location and body pose in surveillance video. In *IEEE Advanced Video and Signal-Based Surveillance (AVSS)*, pages 5–10.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE Conference onComputer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893.

Dollár, P., Appel, R., Belongie, S., and Perona, P. (2014). Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1532–1545.

Dollár, P., Wojek, C., Schiele, B., and Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761.

Eichner, M., Marin-Jimenez, M., Zisserman, A., and Ferrari, V. (2012). 2D articulated human pose estimation and retrieval in (almost) unconstrained still images. *International Journal of Computer Vision*, 99(2):190–214.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338.

Ferrari, V., Marin-Jimenez, M., and Zisserman, A. (2008). Progressive search space reduction for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

Hu, R., Wang, R., Shan, S., and Chen, X. (2014). Robust head-shoulder detection using a two-stage cascade framework. In *International Conference on Pattern Recognition (ICPR)*.

Jafari, O. H., Mitzel, D., and Leibek, B. (2014). Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras. In *International Conference on Robotics and Automation (ICRA'14)*.

Katzenmaier, M., Stiefelhagen, R., and Schultz, T. (2004). Identifying the addressee in human-human-robot interactions based on head pose and speech. In *International Conference on Multimodal Interfaces (ICMI)*, pages 144–151.

Li, M., Zhang, Z., Huang, K., and Tan, T. (2009). Rapid and robust human detection and tracking based on omega-shape features. In *International Conference on Image Processing (ICIP)*, pages 2545–2548.

Mekonnen, A. A., Lerasle, F., and Zuriarrain, I. (2011). Multi-modal person detection and tracking from a mobile robot in a crowded environment. In *International Conference on Computer Vision Theory and Applications (VISAPP'11)*, pages 511–520.

Sheikhi, S. and Odobez, J.-M. (2012). Recognizing the visual focus of attention for human robot interaction. In *Human Behavior Understanding*, pages 99–112. Springer.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–511.

Weinrich, C., Vollmer, C., and Gross, H.-M. (2012). Estimation of human upper body orientation for mobile robotics using an svm decision tree on monocular images. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2147–2152.

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., and Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227.

Zeng, C. and Ma, H. (2010). Robust head-shoulder detection by pca-based multilevel HOG-LBP detector for people counting. In *International Conference on Pattern Recognition (ICPR)*, pages 2069–2072.