2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Daejeon Convention Center
October 9-14, 2016, Daejeon, Korea

Inferring Human Body Posture Information from Reflective Patterns of Protective Work Garments

Rafael Mosberger, Erik Schaffernicht, Henrik Andreasson and Achim J. Lilienthal

Abstract— **We address the problem of extracting human body posture labels, upper body orientation and the spatial location of individual body parts from near-infrared (NIR) images depicting patterns of retro-reflective markers. The analyzed patterns originate from the observation of humans equipped with protective high-visibility garments that represent common safety equipment in the industrial sector. Exploiting the shape of the observed reflectors we adopt shape matching based on the chamfer distance and infer one of seven discrete body posture labels as well as the approximate upper body orientation with respect to the camera. We then proceed to analyze the NIR images on a pixel scale and estimate a figure-ground segmentation together with human body part labels using classification of densely extracted local image patches. Our results indicate a body posture classification accuracy of 80% and figure-ground segmentations with 87% accuracy.**

I. INTRODUCTION

Human detection is a fundamental problem in computer vision and plays an important role in the field of robotic perception. Especially in applications where robots need to operate safely in the vicinity of human workers and where human-robot interaction is required, robust human perception becomes a crucial prerequisite. However, it is in many situations not sufficient that a robot is aware of the presence and location of a person only. Additional attributes regarding the state of a person such as body posture, orientation, or the current activity are often required to execute planning tasks in human-robot interaction scenarios.

Our work builds on an active near-infrared (NIR) sensing approach proposed for the robust perception and localization of human workers wearing protective garments with retro-reflective markers [1]. The main idea behind the method is to base human detection entirely on the active sensing of the reflective markers and thereby remain independent from adverse lighting conditions which usually affect images acquired by conventional visible-light cameras. The method has been shown to reliably detect and track industrial workers in various indoor and outdoor environments.

Building on the robust sensing principle and focusing on applications that require close human-robot interaction, this article addresses the estimation of multiple entities of interest connected to the human body pose. We equip human workers with a two-piece set of conventional protective work garments comprising jacket and trousers with a total of 13 retro-reflective markers (cf. Fig. 1, left). We then present a method for estimating discrete human body posture labels (standing, sitting, lying, etc.) and the approximate upper body orientation, purely by analyzing the reflective patterns created by the safety garments in NIR images (cf. Fig. 1, middle). Furthermore, we describe a method for obtaining a pixel-level figure-ground segmentation with estimation of individual body part labels (cf. Fig. 1, right). Our approach uses established techniques such as shape matching based on the chamfer distance as well as sliding window classification.

Rafael Mosberger, Erik Schaffernicht, Henrik Andreasson and Achim J. Lilienthal are with the AASS Research Centre, School of Science and Technology, Örebro University, S-70182 Örebro, Sweden firstname.lastname@oru.se

Fig. 1: The figure illustrates the purpose of the proposed method. Human workers equipped with reflective work clothing (left) are captured by an active NIR camera. The resulting reflective patterns (middle) are analyzed to estimate body posture and orientation and to perform a pixel-wise image segmentation into to several human body parts (right). The RGB image is used for visualization purposes only.

Our article makes the following principal contributions: i) We study the problem of estimating human body posture labels and an approximate upper body orientation from patterns of reflective markers, for the case where the patterns cannot be designed but are given by conventional industrial work garments. ii) A method is proposed for estimating pixel-wise figure-ground and human body part labels by sliding window classification of local image patches. iii) We evaluate our approach on image sequences from a working scenario covering a broad range of different body poses.

II. RELATED WORK

The human sensing approach based on active NIR vision proposed in [1] stands in strong contrast to conventional vision-based methods relying on visible-light images. The combination of NIR camera, optical filter and active illumination allows the sensor device to acquire of images in which retro-reflective markers appear highly separated from the image background (cf. Fig. 3). Detection is then entirely based on the analysis and classification of reflective patterns.

The approach is limited to applications where it can be assumed that humans are equipped with retro-reflective garments. In return, the approach offers high robustness to

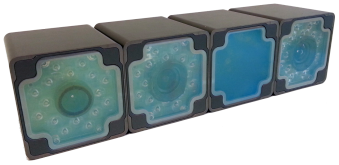


Fig. 2: Industrialized version of the active *RefleX* vision system proposed in [1], offering active stereo NIR and monocular RGB input. The camera module is designed for the robust perception and localization of retro-reflective markers under varying lighting conditions.

Fig. 3: Sensor data provided by the camera unit used in this work, comprising single RGB and stereoscopic near-infrared (NIR) images taken with active illumination. The NIR images show retro-reflective markers that appear highly separated from the image background.

strongly varying lighting conditions and copes with both low-light and back-light conditions where visible-light cameras struggle to produce images with good contrast. In addition, it copes with highly varying body positions. Applications where the sensing principle has been applied include detection and tracking of human workers at industrial work sites [1] and leader tracking for a walking logistics robot in off-road environments [2].

In [3], the approach has been further extended by processing the NIR images with a Hough forest detector [4]. Hough forests have been shown to be an efficient and flexible tool for vision-based human detection, localization and tracking [4], pose estimation [5], image segmentation [6], or action recognition [7]. With the introduction of the multi-band Hough forest detector in [3], the concept of Hough forests was adapted to the case where the same image scene is observed by multiple cameras from the same viewpoint but in different spectral frequency bands. The proposed extension simultaneously fuses NIR and RGB information and performs a generalized Hough transform that maps local observations of reflective patterns to the location of a defined reference point on the person appearing in the image.

While the work in [3] purely deals with the detection and tracking stage, our work addresses the question to what degree body pose information can be inferred from the reflective patterns. This problem is different from the one in human motion capture (mo-cap) with passive retro-reflective markers, in the sense that mo-cap systems most often observe the target from multiple viewpoints and involve strategic placement of distinguishable retro-reflective markers on specific human body landmarks (e.g. [8]). Here, we adopt a set of customary protective work clothes where neither the shape of the individual reflectors nor their spatial arrangement has been specifically designed.

The focus of this article is not on presenting a novel methodological approach. Instead, we concentrate on illustrating that posture classification and body orientation estimation for human workers can be performed even if only strongly reduced NIR sensor data is available. As opposed to more standard RGB input, the NIR input images used in this work depict solely the reflective markers of standard safety garments and the applied methods include established techniques such as sliding window classification and chamfer distance mathing [9].

III. METHOD

This section describes our method for estimating multiple entities of interest from NIR images depicting patterns of retro-reflective markers. The patterns originate from an active NIR camera setup that observes human workers wearing protective work garments (cf. Fig. 6c). The estimated entities of interest comprise:

1.) Body Posture: We estimate a discrete body posture label describing a person as either standing, stooping, squatting, crawling, sitting, kneeling or lying.

2.) Upper Body Orientation: We estimate the orientation of the human upper body in the horizontal plane and with respect to the camera, within a discretized angular space containing 8 different bins (cf. Fig. 5).

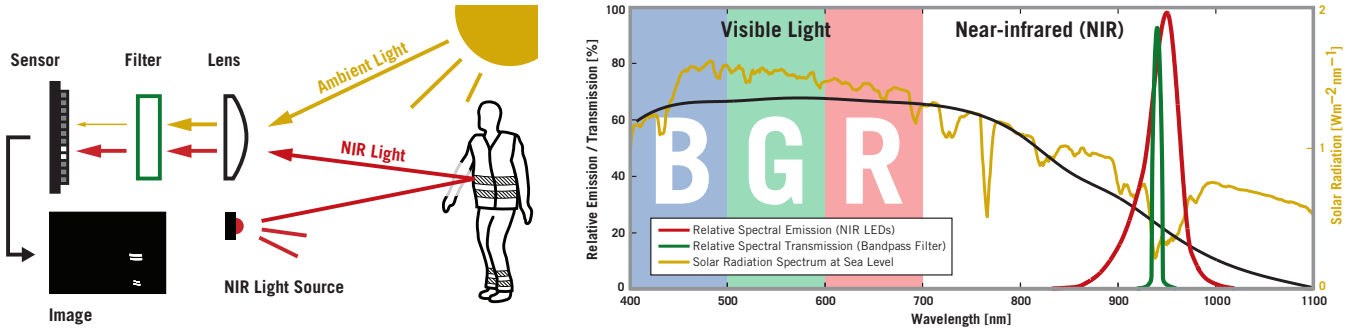
3.) Figure-ground Segmentation: We aim to obtain pixel-wise figure-ground labels with respect to the human silhouette.

4.) Body Part Labels: For all pixels classified as figure pixels, we further assign a label declaring them as either head, torso, left/right arm, or left/right leg.

Entities 1 and 2 describe attributes on a per-person level, and our approach for estimating them involves observing the entire reflective pattern of a person at once. In contrast, entities 3 and 4 describe information on a per-pixel level and have to be analyzed on a more local scale. The remainder of this section, summarizes the sensor and sensor data on which our method is based, and describes the method for estimating the defined entities of interest.

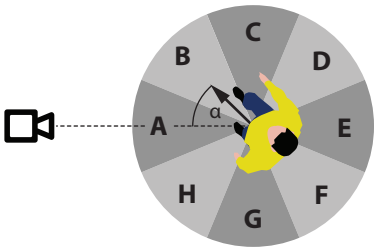
A. Hardware and Sensor Data

The sensor unit adopted in the underlying work is an industrialized version of the vision system described in [1]. The sensor, coined the *RefleX Vision System* (cf. Fig. 2), perceives and localizes retro-reflective markers by means of a stereo pair of near-infrared (NIR) cameras equipped with bandpass filter and active illumination. The module further comes with a single RGB camera placed in between the two NIR cameras. The functional principle and spectral specifications of the sensor are illustrated in Fig. 4 while Fig. 3 shows an example image triplet acquired with the camera. We emphasize at this point, that for the work presented in this article we make use of the RGB data solely for data annotation and visualization purposes.



(a) (b)

Fig. 4: Characteristics of the sensor module used in this work: (a) Operation principle of the NIR sensor designed to acquire images that discriminate objects with high reflectivity from the background. Sunlight and other ambient light (yellow) is largely filtered by the optical bandpass filter (green), resulting in a dark image background. The NIR light emitted by the active sensor unit (red) is backscattered by the retro-reflective markers of the high-visibility garments and transmitted by the bandpass filter, leading to brightly depicted reflectors in the image. (b) Relative spectral characteristics of the bandpass filter (green) and the active light source (red). The yellow curve represents the solar irradiation spectrum at sea level (Source: ASTM International). The operation wavelength of 940nm is chosen to exploit the negative peak in the sun spectrum and limit background illumination in outdoor applications.



Bin A $\alpha \in 0° \pm 22.5°$
Bin B $\alpha \in 45° \pm 22.5°$
Bin C $\alpha \in 90° \pm 22.5°$
Bin D $\alpha \in 135° \pm 22.5°$
Bin E $\alpha \in 180° \pm 22.5°$
Bin F $\alpha \in 225° \pm 22.5°$
Bin G $\alpha \in 270° \pm 22.5°$
Bin H $\alpha \in 315° \pm 22.5°$

Fig. 5: For the estimation of the upper body orientation, we divide the angular range into 8 bins and formulate the estimation of angle α as a classification problem.

B. Preprocessing and Data Annotation

Preprocessing involves extracting and localizing the reflective markers from each stereo pair of NIR images according to the approach described in [1]. Furthermore, we assume in our work that a Hough forest based detector as proposed in [3] has been applied to the NIR input image pair, resulting in individual human detections comprising the two-dimensional image coordinates of a defined object reference point and a characteristic scale. Here, we consider these entities as given and manually annotate the reference points in the input images while a scale is directly extracted from the depth of the reflective markers. We specifically adopt the term *reference point* (cf. Fig. 6b) instead of *object center* as used in [4], to emphasize that this point does not necessarily coincide with the centroid of the object silhouette nor with the center of the bounding box.

C. Global Template Matching

Here, we consider the problem of taking a square scale-normalized NIR images depicting the full reflective pattern of a worker and estimate the corresponding body posture and upper body orientation in the horizontal plane. We approach the problem by storing a set $\mathcal{T} = \{T_i\}$ of exemplary scale-normalized NIR image templates T_i (cf. Fig. 6) depicting the reflective patterns of human workers in a variety of different body positions and orientations with respect to the camera. The templates are extracted from a square region centered around the object reference points (cf. Fig. 6b) previously annotated during preprocessing and with a window size chosen such that the entire reflective pattern of a person is visible. Scale normalization is applied by exploiting the stereo depth measurements obtained during the preprocessing stage. The templates are further annotated with the ground-truth entities of interest for future use. In summary, each template $T_i = \{\mathcal{U}, \mathbf{r}, \mathcal{A}\}$ is characterized by the set of reflector edge points $\mathcal{U} = \{\mathbf{u}_j\}$ extracted during preprocessing (cf. Fig. 6d), the location of the object reference point $\mathbf{r}$ within the template, and the ensemble $\mathcal{A}$ of annotated ground-truth entities comprising a discrete body posture label and discrete upper body orientation according to Fig. 5.

We then consider a test image I depicting a human worker for which the body posture and upper body orientation is unknown and needs to be estimated from the set of templates $\mathcal{T}$. The test image $I = \{\mathcal{V}\}$ with reflector edges $\mathcal{V} = \{\mathbf{v}_k\}$ is assumed to be scale-normalized and centered around the object reference point in the same way as the templates, for instance by applying the Hough forest detector proposed in [3]. We then define the chamfer distance between the edge map $\mathcal{V}$ of the test image and edge map $\mathcal{U}$ of a given template as

$$d_{\text{chamfer}}(\mathcal{V},\mathcal{U}) = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{v}_k \in \mathcal{V}} \min_{\mathbf{u}_j \in \mathcal{U}} ||\mathbf{v}_k - \mathbf{u}_j|| \qquad (1)$$

which can be efficiently computed using the distance transform (DT),

$$d_{\text{chamfer}}(\mathcal{V},\mathcal{U}) = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{v}_k \in \mathcal{V}} DT_{\mathcal{U}}(\mathbf{v}_k) \qquad (2)$$

where $DT_{\mathcal{U}}(\mathbf{x}) = \min_{\mathbf{u}_j \in \mathcal{U}} ||\mathbf{x} - \mathbf{u}_j||_1$. We then identify

(a) (b) (c) (d) (e)

Fig. 6: Example template consisting of (a) the RGB image used for data annotation, (b) annotated entities comprising bounding box, reference point (black star), posture label, body orientation according to Fig. 5, as well as pixel-wise body part and figure-ground labels, (c) the input NIR image processed by our algorithm, (d) the extracted edge map describing reflector outlines, and (e) the corresponding distance transform used for computing chamfer distances.

the template T_{NN} with the most similar reflective pattern by performing a nearest neighbor search across the template set $\mathcal{T}$, while adopting the symmetric version of the chamfer distance for increased robustness:

$$T_{\mathrm{NN}} = \arg\min_{T_i} \left[d_{\mathrm{chamfer}}(\mathcal{U}, \mathcal{V}_i) + d_{\mathrm{chamfer}}(\mathcal{V}_i, \mathcal{U}) \right] \quad (3)$$

The final estimates for body posture and upper body orientation are directly inferred from the annotations $\mathcal{A}$ of the nearest-neighbor template T_{NN}. We emphasize at this point that no spatial alignment between the templates and a test image is performed. We assume that the reference point of a person has been previously identified and that both the templates and test images are centered around the reference point and can directly be compared without alignment.

D. Local Pixel-wise Image Segmentation

In a second step, we analyze the test image I on a pixel level and estimate a figure-ground map as well as a map of human body parts. To do so, we adopt the notion of local image patches connected to the object reference point as known from the Hough forest framework [4]. We start by enhancing each template of the previously established template set $\mathcal{T}$ with a segmentation map $S(\mathbf{x})$ that declares each pixel $\mathbf{x}$ as either background or member of a specific human body part (cf. Fig. 6b), namely torso, head, left/right arm or left/right leg. From the enhanced set of templates we extract a large set $\mathcal{P}$ of square local image patches, sampled from random locations in the templates that contain reflective markers. Each patch stores its local segmentation map, the subset of reflector edges that fall inside the patch borders, as well as an offset vector $\mathbf{d}$ that represents the position of the object reference point $\mathbf{r}$ in the template with respect to the upper-left corner of the patch.

Given a test image with known object reference point $\mathbf{r}$ but unknown pixel labels, we perform image segmentation by densely extracting and classifying local patches in sliding window fashion as illustrated in Fig. 7. Classification is carried out by a nearest neighbor search across the patch set $\mathcal{P}$ with the chamfer distance as distance measure. At each sliding window position $\mathbf{q}$, the matched patch in set $\mathcal{P}$ casts votes for pixel labels at the respective position of the sliding window. However, votes are restricted to matches which agree on the relative spatial position from the object reference point, by defining a weighting function for the vote casting, according to:

$$w(\mathbf{q}) = \left\{ \begin{array}{ll} 1 & \text{if } ||\mathbf{q} + \mathbf{d} - \mathbf{r}||_2 < t_c \\ 0 & \text{otherwise} \end{array} \right\} \quad (4)$$

Votes are accumulated in a 3-dimensional histogram where the first two dimensions represent the image space and the third dimension the labels to be estimated. The final segmentation is obtained in two steps. First, a pixel-wise search for the foreground label that accumulated the most votes is carried out. Second, the resulting image from step one is masked by setting all pixels to background for which the background channel has accumulated more votes than all foreground channels together.

IV. EXPERIMENTS

We report an experimental evaluation of the proposed method and assess to what extent the desired entities of interest can be estimated from the available NIR data. To the best of our knowledge, no datasets are publicly available that combine the type of sensor data and application scenarios addressed with our work. We therefore perform an experimental evaluation on a proprietary set of acquired video sequences containing an actual work scenario. All classification results are presented as confusion matrices where rows show the ground truth and columns the classification output.

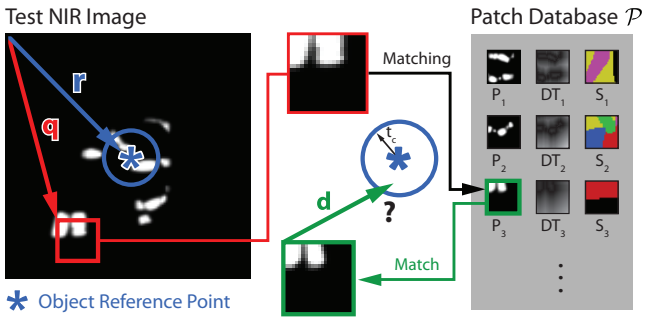


Fig. 7: The figure illustrates the sliding window based patch classification and offset vector verification (cf. Eq. 4).

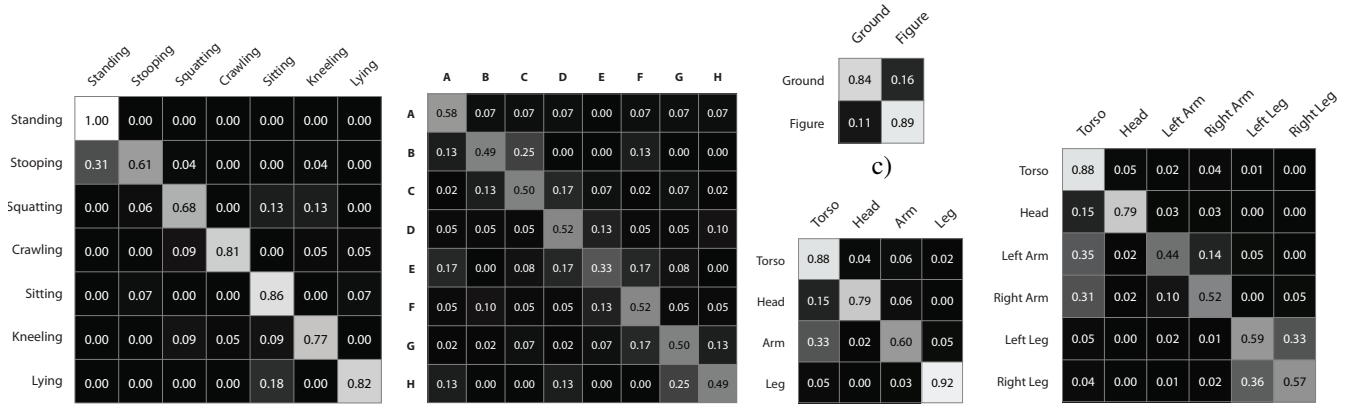


Fig. 8: Experimental results in terms of (a) human body posture classification, (b) upper body orientation classification (binning shown in Fig. 5), and (c–e) pixel-wise image segmentation. Segmentation is evaluated separately for (c) figure and ground pixels, (d) body parts without and (e) with discrimination of left and right limbs.

A. Data Collection and Preprocessing

Multiple video sequences are acquired with the *RefleX* sensor unit (cf. Sec. III-A) in an indoor working environment. The camera is mounted forward-facing at a height of 1.5 meters above ground on a mobile robot. The recorded sequences show a total of 3 persons engaged in a working scenario that involves moving, lifting and transporting objects and carrying out tasks in a range of body positions. All persons are equipped with the same set of unmodified off-the-shelf high-visibility garments consisting of jackets and trousers with a total of 13 retro-reflective markers.

Each time frame of a sequence consists of an image triplet with one RGB and two NIR pictures (cf. Fig. 3) at a resolution of 800×600 pixels. The RGB image is held back and used for data annotation purposes only. All sequences are preprocessing to extract the reflective markers and estimate their depth from each stereo pair of NIR images according to [1]. 75% of the acquired and preprocessed video sequences are used for training while 25% are held back as test data.

B. Body Posture Classification

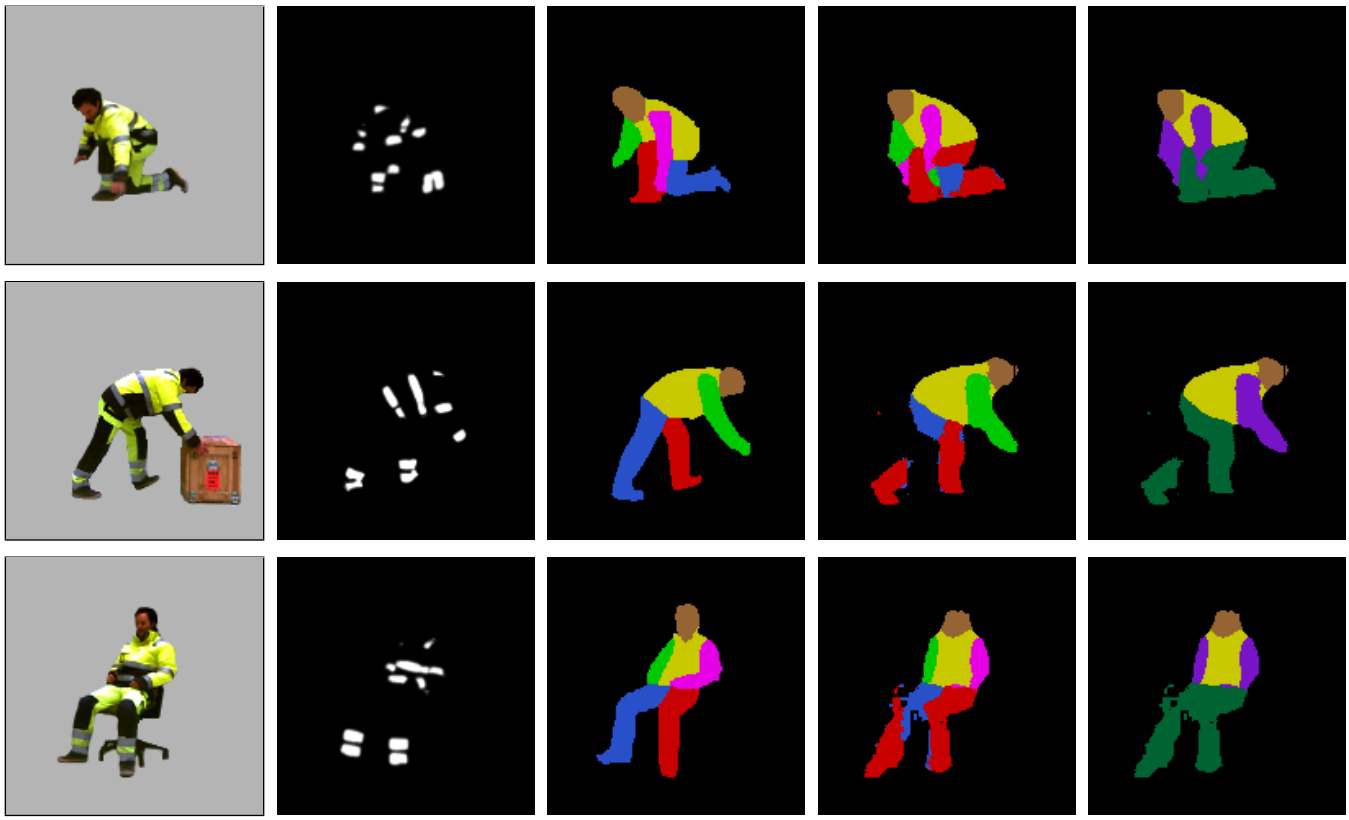
We assess the extent to which discrete human posture labels can be estimated from individual single-channel NIR frames by nearest neighbor matching of test images to a set of templates as described in Sec. III-C. A total of 750 templates of size 180×180 pixels are extracted at normalized scale and manually annotated with ground-truth body posture labels to form the template database $\mathcal{T}$. We then report classification performance on 250 test images in Fig. 8a. The results reveal a high classification performance with approximately 80% of correctly classified body postures. Most errors are made by classifying stooping persons as upright standing. A further uncertainty is observed in the estimation of the postures sitting, kneeling and squatting, which are postures with floating borders.

C. Upper Body Direction Classification

As a second entity, we estimate the upper body orientation in the horizontal plane with respect to the camera according to the discretized angular space illustrated in Fig. 5. The resulting classification performance is shown in Fig. 8b. The overall accuracy in estimating the correct orientation bin amounts to approximately 50% and lies significantly below the value achieved for the body posture labels. However, it is observed that 55% of the misclassified samples are assigned to an orientation bin next to the correct one. In average, the absolute angular estimation error amounts to 46°.

D. Figure-ground and Human Body Part Labels

We further evaluate the estimation of pixel-wise figure-ground and human body part labels by dense classification of local patches as described in Sec. III-D. As a much higher data annotation effort is required on a per-pixel level, we use a reduced data set of 300 templates and 100 test images. From the annotated templates, we extract a total of 10k local image patches of size 40×40 pixels from local regions depicting reflective markers in the NIR image. Segmentation is then performed by densely classifying feature patches from the test images as described in Sec. III-D using a center verification threshold t_c of 10 pixels (cf. Eq. 4). The segmentation accuracy with respect to figure-ground and body part labels is shown in Fig. 8c–e while Fig. 9 illustrates several qualitative examples of input and output images together with the ground-truth and estimated labels. The results illustrate that the segmentation allows for a better interpretation of the human body pose than the raw NIR image. Fig. 8d and Fig. 9e indicate that a discrimination of the four main groups of body parts considered here (head, torso, arms and legs) is possible to a large extent. However, Fig. 8e and Fig. 9d reveal a major difficulty in discriminating left and right limbs. This is hard to overcome as multiple body configurations can result in an identical reflective pattern. Further figure-ground segmentation deficiencies are observed at locations where no reflective marker is present.



(a) (b) (c) (d) (e)

Fig. 9: Qualitative results of the pixel-wise image segmentation based on pure NIR input: (a) RGB image used for data annotation, (b) NIR input image, (c) manually annotated ground-truth body part labels, (d) estimated labels, and (e) estimated labels if no discrimination between left and right limbs is made.

V. CONCLUSION AND FUTURE WORK

In this article we addressed the problem of recovering human body posture information and upper body orientation from sensor data that depicts exclusively patterns of reflective markers originating from the observation of humans wearing conventional protective work garments. Furthermore we estimate pixel-wise figure-ground labels and performed an image segmentation into multiple human body parts.

To the best of our knowledge, we are the first authors to propose a method for estimating body posture, orientation, figure-ground labels and body-part segmentation from the reflective patterns produced by standard work clothing. Our results illustrate that the estimation of these entities from the type of sensor data use in this work is principally possible. As a main limitation of the approach we identified the difficulty to resolve ambiguities between the left and right arms and legs during image segmentation. Nevertheless, the proposed method has a range of potential applications in industrial scenarios, owing to its high robustness to adverse ambient lighting conditions.

Future work involves investigating methods for matching templates and local image patches in more efficient ways by adopting tree-based search methods where large numbers of hypotheses can be excluded at an earlier stage.

REFERENCES

[1] R. Mosberger, H. Andreasson, and A. J. Lilienthal, "A customized vision system for tracking humans wearing reflective safety clothing from industrial vehicles and machinery," *Sensors*, vol. 14, no. 10, pp. 17 952–17 980, 2014.

[2] M. Perdoch, D. M. Bradley, J. K. Chang, H. Herman, P. Rander, and A. Stentz, "Leader tracking for a walking logistics robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 2994–3001.

[3] R. Mosberger, B. Leibe, H. Andreasson, and A. J. Lilienthal, "Multi-band hough forests for detecting humans with reflective safety clothing from mobile machinery," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 697–703.

[4] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 33, no. 11, pp. 2188–2202, 2011.

[5] I. Kostrikov and J. Gall, "Depth sweep regression forests for estimating 3d human pose from images," in *British Machine Vision Conference (BMVC)*, 2014.

[6] K. Rematas and B. Leibe, "Efficient object detection and segmentation with a cascaded hough forest ism," in *ICCV Workshop on Challenges and Opportunities in Robot Perception*, 2011.

[7] A. Eweiwi, M. S. Cheema, and C. Bauckhage, "Action recognition in still images by learning spatial interest regions from videos," *Pattern Recognition Letters*, vol. 51, pp. 8 – 15, 2015.

[8] C. Canton-Ferrer, J. R. Casas, and M. Pards, "Marker-based human motion capture in multiview sequences," *EURASIP Journal on Advances in Signal Processing*, 2010.

[9] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf, "Parametric correspondence and chamfer matching: Two new techniques for image matching," in *International Joint Conference on Artificial Intelligence*, vol. 2, 1977, pp. 659–663.