BIFOCAL STEREO FOR MULTIPATH PERSON RE-IDENTIFICATION

G. Blott[a,b]*, C. Heipke[b]

[a] Robert Bosch, Computer Vision Lab, Hildesheim, Germany, gregor.blott@de.bosch.com
[b] Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Germany

Commission II

KEY WORDS: Person Re-Identification, PRID, Bifocal Stereo, Multipath Person Re-Identification

ABSTRACT:

This work presents an approach for the task of person re-identification by exploiting bifocal stereo cameras. Present monocular person re-identification approaches show a decreasing working distance, when increasing the image resolution to obtain a higher re-identification performance. We propose a novel 3D multipath bifocal approach, containing a rectilinear lens with larger focal length for long range distances and a fish eye lens of a smaller focal length for the near range. The person re-identification performance is at least on par with 2D re-identification approaches but the working distance of the approach is increased and on average 10% more re-identification performance can be achieved in the overlapping field of view compared to a single camera. In addition, the 3D information is exploited from the overlapping field of view to solve potential 2D ambiguities.

1. INTRODUCTION

Person re-identification (re-id) is the challenge to re-identify persons in images taken from different perspectives. Solving this task is necessary e.g. to concatenate trajectories of persons in non-overlapping fields of view within a multi camera network, to re-associate person tracks for crowded scenes or to re-identify persons in the robotic domain such as robot owner re-id. In close range face recognition can achieve remarkable re-id results today, when persons look straight into the camera.[1] However, in real world situations such as railway stations, public spaces and airports such a set up can not be guaranteed. Hence, researchers try to solve this issue by exploiting full body appearance.

Activities in person re-id (PRID) are rapidly increasing. A survey of recent literature shows that about a single digit number of new papers are published every week. The work can be divided into five main groups:

1. Image data pre-processing, to make the image information more robust against illumination changes and ambiguous colors, e.g. salient color names and foreground segmentation (Yang et al., 2014) or a retinex transformation (Liao et al., 2014). As the results of person detection and tracking (Milan et al., 2016) are considered to be given in PRID, this step is also part of preprocessing.

2. Feature extraction, to find a discriminative descriptor that represents the appearance of the persons and further feature aggregation to fuse features over time for multi shot re-id. A detailed survey can be found in (Karanam et al., 2016).

3. Metric Learning, to increase the matching distance between images of two different persons while decreasing the one for images of the same person. A detailed survey can also be found in (Karanam et al., 2016).

4. Post re-ranking, to re-rank the n-first best matching results from a first trial to achieve a better ranking, e.g. with a different learned metric for only very similar looking persons. Surveys can be found in (Garca et al., 2015), (Leng et al., 2015) and (Zhong et al., 2017).

5. 3D information extraction from active sensor systems, e.g the Microsoft Kinect, where especially the distance between skeleton joints and the ground plane and gait analysis are frequently used for re-id. Detailed surveys can be found in (Imani and Soltanizadeh, 2016), (Liu et al., 2017) and (Wu et al., 2017).

In this work a new bifocal multipath approach for image based person re-identification is proposed. For this purpose the sensor is equiped with a fish eye and a rectilinear lens. Although general in nature, our work is applied in a security environment. Figure 1 gives an overview of the main input and improvements compared to current work. The highlights of this study are: (1) The first bifocal PRID approach for security camera application that combines near range 2D image, long range 2D image and in the mid-range additionally 3D information from the overlapping

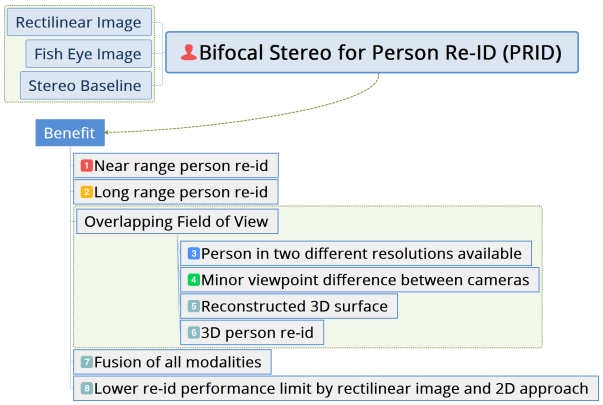

Figure 1. Input and benefit of bifocal PRID.

*Corresponding author
[1] face-rec.org/vendors

field of view. We call this approach *multipath PRID* to emphasize the different input data paths. (2) A validation of stereo precision for PRID. Besides, all 3D reconstruction experiments are carried out under different lighting conditions. (3) Practical implementation and validation of the approach with the best camera hardware with suitable stereo baseline for outdoor use cases according to the EMVA-1288.

The rest of the paper is organized as follows: In section 2 the problem statement is introduced. Further, in section 3 the approach to solve the problem statement is discussed including the architecture overview, and it is shown what can be expected from the approach in theory for single point measurements. In section 4 results of validation experiments are described and discussed. In this study no closed loop validation is done, instead the individual steps of the approach are validated independently. In section 5 a closed loop discussion follows. Finally, conclusions is given.

2. PROBLEM STATEMENT

2.1 Security Camera Mounting Location

As mentioned, we are interested in security applications. Security cameras located in airports, railway stations, religious places and public spaces are typically mounted four to five meters above the ground with a large field of view to observe as much area as possible. An example is shown in Figure 2. In this study the focus lies on a working distance of around ten meters to observe an area of around 100 square meters by using a field of view of around 90°. Additionally, cameras have a pitch angle towards the ground, that varies between a few degrees and 90° (bird's eye view).

2.2 Person Image Resolution vs. Working Distance

State-of-the-art PRID approaches, work as follows: A so called probe person image (queried person image) is compared to a set of known person images (gallery persons), typically of coarser resolution, to re-identify the probe person. The probe person image is then down-sampled to the size of the gallery images for a direct comparison. This means by the (coarser) resolution of the gallery image defines the working distance of the PRID approach and part of the higher information content of the probe image is lost during down-sampling. Obtaining a larger working distance or alternatively a higher re-id performance at a constant gallery resolution is addressed in this paper.

2.3 3D Re-Identification with Passive Stereo

In this study additionally 3D PRID based on dense image matching is addressed to support situations where persons can not be distinguished only by appearance. An often quoted challenge is to distinguish two persons of different height but wearing similar clothes, e.g. a black business dress, that can thus not be re-identified by only analysing color information.

Figure 2. Security camera image taken from 4m camera height, 30° pitch angle towards ground and 97° camera opening angle.

2.4 Difference to existing approaches

The approaches from the main working fields, introduced in Chapter 1, e.g. feature description and metric learning, form the foundation of this approach. Furthermore, on the one hand the bifocal stereo camera approach exploits proven 2D approaches and on the other hand additionally uses 3D information, automatically determined from the overlapping sensor field of view. In contrast to Kinect approaches a passive stereo system is used to acquire the 3D information and consequently a higher 3D reconstruction error is expected due to potential matching inconsistency. To the best of the authors' knowledge this article is the first PRID contribution that combines two sensors with different opening angles in a passive stereo system.

3. BIFOCAL PERSON RE-IDENTIFICATION

3.1 Architecture overview

Figure 3 introduces the flow chart of our PRID approach. It can be separated into the modules 2D near-, 2D long- and 3D re-id. For the 2D modules state-of-the-art approaches are employed to re-identify persons. Near to the camera, the fish eye image only is used since the persons are generally not visible in the rectilinear image, while for the long range the rectilinear image only is used because the resolution of the fish eye image is too poor. In the ranges in-between the fusion is performed. Investigation of the detection and tracking module is not part of in this article, see (Leal-Taixé et al., 2015) and (Milan et al., 2016) for reference. Further, rectification (de-warping) of the fish eye image to a rectilinear projection was done based on a unified projection model (Mei, 2007). The gallery, containing images of known persons in constant resolution, is shared by the long and the near range module.

For our study a hardware camera rig is used, consisting of two cameras with different lenses mounted with a fixed baseline (cf. Fig. 4). One camera is mainly responsible for the near range PRID and is equipped with a fish eye lens of small focal length with 180 degree opening angle. The second camera is mainly responsible for the long range PRID and is equipped with a rectilinear lens of a longer focal length. Images from both cameras are fused to generate 3D information in the overlapping field of

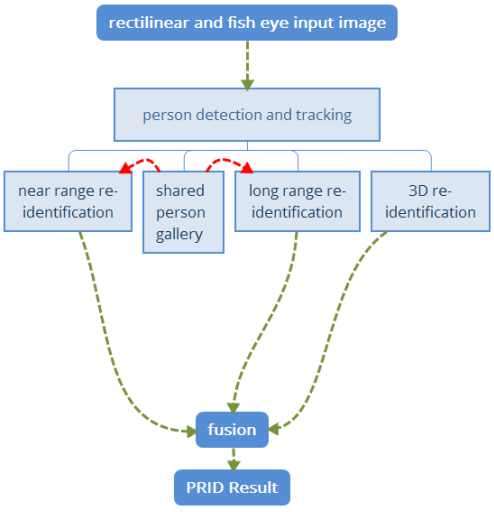

Figure 3. Bifocal PRID processing modules.

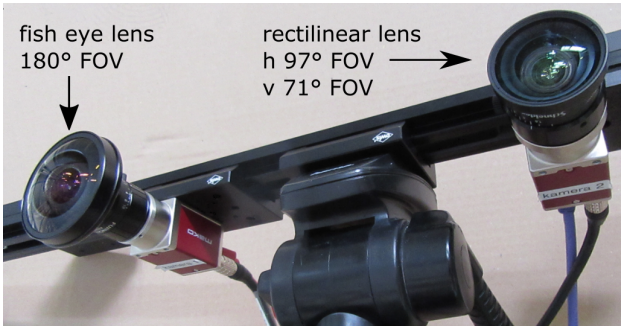

Figure 4. Bifocal camera rig with 21cm baseline.

focal length [mm]	4	6	8	10	12
hor. opening angle [°]	110	89	71	59	51
Z [m]	9	14	19	23	28

Table 1. Maximum distance (Z) to project a person of 1.75m into image space, having the same height as person images in the most cited PRID dataset VIPeR (128 pixels).

view. In addition, this set up enables new possibilities for PRID that are not part of this article, e.g. the use of new features based on two images of the same person in different resolutions.

The cameras selected are two AVT Mako G-234C (2.35 MP). They are equipped with infrared cut filters to suppress the spectrum that is not visible for humans. The effective foal length of the rectilinear lens is 5mm (97° horizontal and 71° vertical opening angle) and 2.5mm for the fish eye lens (180° opening angle). Further, the baseline of the stereo rig is set to 21cm as used in the Carnegie Robotics Multisense.[2] The camera itself was chosen because of the fact, that a Sony IMX249 imager is included. This imager complies with EMVA Standard 1288 (Standard for Charterization of Image Sensors and Cameras[3]) and is one of the score board leaders in the year 2016 for RGB quantum efficency, dynamic range, saturation capacity, absolute sensitivity threshold and temporal dark noise.[4]

3.2 Working Distance of Re-Id Approaches

We define the maximum working distance as the distance, were a person projected into an image has the resolution that is used in the person gallery. If the person is further away, up-sampling of the image is necessary, which however does not introduce new information. If the person is nearer to the camera, the captured image has to be down-sampled and information content is lost.

This maximum working distance Z can be computed, given the interior orientation of the camera and a mean person height, to $Z = f \cdot h_{pW} \cdot (h_{pI} \cdot p_{pitch})^{-1}$, where h_{pI} is the person height in pixels that is used in the gallery, h_{pW} is the person height in meters[5], f is the focal length and p_{pitch} is the pixel pitch, amounting to 5.86 μm for the used sensor. Table 1 illustrates the equation for the introduced AVT G-234C camera. As a result the working distance of the selected hardware is around 12m, since a 5mm rectilinear lens for the long range is used, and around 6m for the de-warped fish eye image.

[2] carnegierobotics.com/multisense-s21
[3] emva.org/wp-content/uploads/EMVA1288-3.0.pdf
[4] eu.ptgrey.com/support/downloads/10624
[5] (DIN EN 33402-2, 2005) gives a median of 1.75m for German men

3.3 Precision of the 3D reconstruction

For the re-identification of persons using 3D information more than one strategy can be used. (a) One strategy is to use a single 3D point, for example the highest point of the head or shoulder. In an interactive scenario, the respective positions are determined manually in one image, the conjugate points can then be found by image matching. For an automatic analysis a 2D algorithm able to detect the respective body parts is necessary. (b) Another strategy is to first reconstruct a person shape using dense stereo matching and then further process the resulting 3D point cloud. In this section the approximated 3D reconstruction precision is discussed for single point measurements (cf.(a)) since a closed theory exists to model the expected error. Further, in Chapter 4.2 experiments are done to validate dense matching (cf.(b)).

In the following section only the central image region of the fish eye image is considered. In this case the pixel content of a rectilinear projection and a fish eye projection is roughly the same. First, the central region is defined. The distance between the principle point and a point in space projected into a fish eye image can be approximated with equidistant projection to $R_{fe} = f_{fe} \cdot \alpha$ (Förstner and Wrobel, 2016), where α is the angle between the point in space and the optical axis as seen from the camera origin. The corresponding distance of the same point in rectilinear projection is $R_{rl} = f_{rl} \cdot \tan \alpha$. As long as points in the world are projected to roughly the same sensor area by both projection models, pixel information is roughly the same. The difference in sensor area can be approximated to $A_{difference} = \pi \cdot R_{fe}^2 - \pi \cdot R_{rl}^2$. For the selected hardware an image area corresponding to $\alpha < 10°$ results on a negligible difference. Consequently, the equations for rectilinear stereo projection are also valid for this central region of the fish eye image. Using the well known equations for the stereo precision for the normal case the standard deviation of the depth can be approximated to

$$\sigma_{Z_{cam}} = \frac{Z^2}{f \cdot B} \cdot \sigma_m, \tag{1}$$

where f is the focal length after the rectification, B is the stereo baseline, σ_m is the standard deviation of the image coordinate measurement (the matching error) and Z is the depth of the observed point in space. The standard deviations of the other two 3D coordinates can be determined to

$$\sigma_{X_{cam}} = \sqrt{(\frac{x - ppx}{f} \cdot \sigma_{Z_{cam}})^2 + (\frac{Z}{f} \cdot \sigma_x)^2}, \tag{2}$$

and

$$\sigma_{Y_{cam}} = \sqrt{(\frac{y - ppy}{f} \cdot \sigma_{Z_{cam}})^2 + (\frac{Z}{f} \cdot \sigma_y)^2}, \tag{3}$$

where $[x, y]'$ are the coordinates of the observed point in the reference camera (left camera), $[ppx, ppy]'$ is the principle point in x- and y-direction, $[\sigma_x, \sigma_y]'$ are the standard deviations of the matching error in x- and y-direction (they were set to one pixel for all following calculations).

In Figure 5 the standard deviations $\sigma_{X_{cam}}$, $\sigma_{Y_{cam}}$ and $\sigma_{Z_{cam}}$ at the image border are shown as functions of the depth to the camera, for three different matching errors. In a depth of 10m and for 0.5 pixel standard deviation of the matching, a depth precision of around 30cm as well as 38cm and 24cm for the other two coordinates are found. In a depth of 5m and for 0.5 pixel standard deviation of the matching the values are 7.6cm, 9.6cm and 6.0cm,

all at the image border. For the bifocal approach less precision is expected, especially towards the image border since the fish eye image is resampled.

Solving equation (1) for σ_m yields the necessary accuracy of image matching (or, equivalently, for the stereo parallax) for a given configuration. Table 2 shows which parallax accuracy in pixel units is necessary in our case to resolve a depth of 1cm, 2cm, 5cm and 10cm, respectively, in a distance Z of between seven and ten meters. In the addressed working distance of 10 meters

Z_{cam}	1cm	2cm	5cm	10cm
7m	0.033	0.066	0.165	0.327
8m	0.025	0.051	0.126	0.251
9m	0.020	0.040	0.100	0.198
10m	0.016	0.032	0.081	0.161

Table 2. Stereo parallax accuracy σ_m in pixel (col. 2-4) to resolve 1cm, 2cm, 5cm, 10cm in a distance Z_{cam} of 7-10m.

to the camera the parallax accuracy must be less than one sixth of a pixel to obtain a depth resolution of 10cm and one twelvth of a pixel for 5cm depth resolution. Reaching these values is hardly possible in real world applications (for a state-of-the-art matching accuracy evaluation see the Midddlebury[6] Stereo Evaluation and the KITTI[7] stereo evaluations).

3.4 Precision of the person height estimation

In this paragraph the person height estimation precision with respect to security camera orientations is discussed. Further, in (Barbosa et al., 2012) the person height had the most significant impact on the 3D re-id performance. The underlying assumptions of this paragraph are, firstly, a head reference point (the highest point of a person) in the images is given with high precision; secondly, the 3D reconstruction precision, discussed in 3.3, is correct, and thirdly, the ground plane is known very precisely. In this case the person height precision depends only on the head point precision.

In this subsection two error models are introduced to simplify two special cases (cf. Fig. 6): One model for a reference point next to the image

[6] vision.middlebury.edu/stereo/eval3/
[7] cvlibs.net/datasets/kitti/

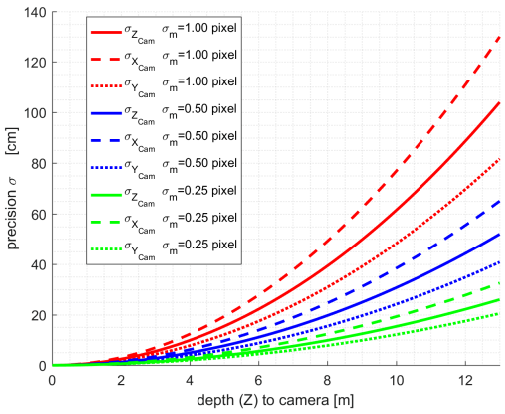

Figure 5. Standard deviations in 3D space at the image border in the camera coordinate system.

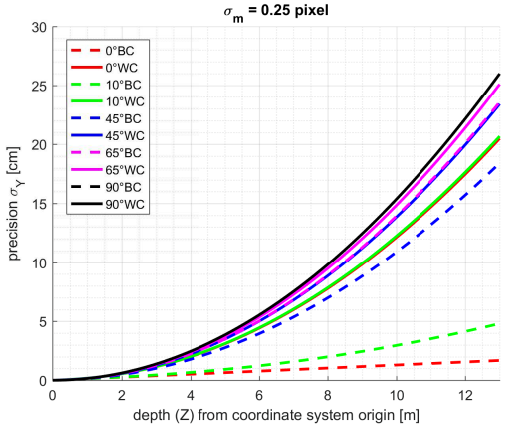

Figure 7. Precision with the BC and the WC error model for $0°$, $10°$, $45°$, $65°$ and $90°$ pitch angle and 0.25pixel matching error.

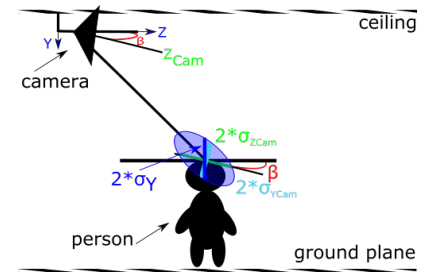

Figure 6. Illustration for the person height estimation error.

border, called Worst Case (WC) Model. The person is assumed to have arbitrary height and a camera mounted with a pitch angle β towards ground is introduced. The precision in Y-direction (cf. Fig. 6) as function of the camera pitch angle can be determined with

$$\sigma_Y = \sqrt{(\sin\beta \cdot \sigma_{Z_{cam}})^2 + (\cos\beta \cdot \sigma_{Y_{cam}})^2}. \quad (4)$$

In Figure 7 the resulting precision for five different camera pitch angles and 0.25 pixel matching error is shown. Depending on the pitch angle of the camera the precision of the person height changes. The highest precision can be reached for a pitch angle of zero degree, the lowest precision if the camera is mounted with 90 degrees (bird's eye view) to the ground. However, the best camera pitch angle for 2D PRID is around zero. In this case a person is completely visible in the image, whereas with increasing pitch angle less person parts can be seen. Furthermore, in the addressed working distance of 10m the Figure shows (a) for 0.25 pixel matching accuracy and zero degree camera pitch angle the person height can be measured with a precision of around 1cm along the optical axis (best case), but the precision decreases towards the image border to 12cm (worst case). (b) A more practical camera pitch angle is $45°$, here in the best case 11cm precision and in the worst case 14cm precision can be reached.

The outcome of this paragraph is, that both error models show, that the obtainable precision of the 3D person height is not very good. At the addressed working distance of 10 meter, even in the best case and a matching accuracy of 0.25 pixels, a person height precision of 20cm can be achieved. As Table 3 shows, 90% of the population of German men have a height between 1.65m (5% percentil) and 1.85m (95% percentil), also a range of 20cm. Consequently the person height obtained from a single measurement is not a distinctive feature for PRID. Furthermore,

	head	eye	shoulder	elbow
95% percentil	1.85	1.73	1.55	1.17
median	1.75	1.63	1.45	1.10
5% percentil	1.65	1.53	1.34	1.02

Table 3. Height to the ground plane for German men between 18-65 years. The Table is based on (DIN EN 33402-2, 2005).

as the table also shows the height of eye, shoulder and ellbow are also not beneficial for PRID.

4. EXPERIMENTS AND EVALUATION

In this section results of experiments conducted to validate the solution of the problem statement are discussed. Section 4.1 evaluates by how much percentage the PRID performance increases by using the bifocal approach. Section 4.2 focuses on the precision of dense bifocal 3D person reconstruction to obtain quantitative values which 3D precision is possible in the real world. Finally, section 4.3 tackles person candidate rejection by exploiting reconstructed 3D person height to reduce possbile person matches for a subsequent 2D PRID fusion approach.

4.1 Re-Id Performance on down-sampled images

The goal of this paragraph is to determine the PRID performance by using the bifocal approach. To validate the performance person images are down-sampled in image resolution to extrapolate how the bifocal approach will perform. For the experiment two recently published feature descriptor approaches, GOG (Matsukawa et al., 2016) and LOMO (Liao et al., 2014) are used. For the metric learning part XQDA (Liao et al., 2014) is exploited. For test and training data ten random splits (cf. (Gray et al., 2007)) of often cited datasets VIPeR (Gray et al., 2007), prid4502 (Roth et al., 2014), and i-lids (Wang et al., 2016) (Wang et al., 2014) were analysed. The same splits are used for every down-sampled trial. In Figure 8 the result is depicted. The abscissa presents the percentage of the original image size during down-sampling the ordinate shows the performance (rank#1 recognition rate), 100% is the original benchmark protocol resolution. For VIPeR the original resolution is not known, for the other datasets the original resolution of the person images varies between 76x30pixels and 267x137 pixels but was set to the benchmark protocol value here. By down-sampling from 100% to 50% the performance decreases on average by 10% for both PRID approaches (compare average performance difference between 100% and 50% for each PRID algorithm). Furthermore, by down-sampling from 100% to 25% the performance decreases by 33% for GOG and by 24% for LOMO. The best configuration was achieved with GOG and 100% image resolution. The outcome of this experiment is the confirmation, that a high resolution is mandatory for a higher performance. However, a high resolution also reduces the working distance of a monocular PRID system since sensor resolution is limited and only the focal length of a lens can vary. Consequently, the working distance can be doubled with our bifocal approach without PRID performance loss, and additionally an increased average performance of 10% in the overlapping field of view is feasible.

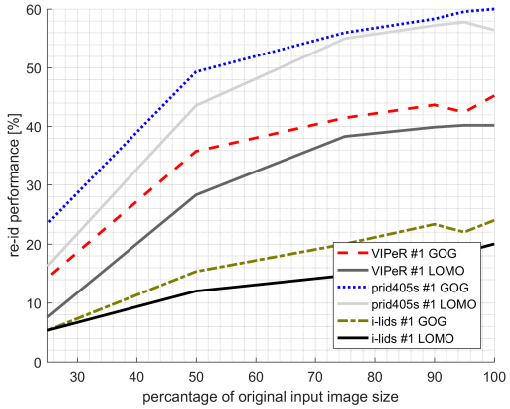

Figure 8. PRID rank#1 as function of the percentage image resolution. 100%, 95%, 90%, 75%, 50% and 25% are measured.

4.2 3D precision for different scene lighting

Section 3.3 addressed the expected 3D reconstruction precision for single measurements in a closed form. In this paragraph experiments for dense person shape reconstruction are described. The basic set up is shown in the left part of Figure 9, the right part shows images taken with the bifocal rig. Images of the depicted dummy were taken at different depths. Since a passive stereo system is used, the results depend on illumination. Here, all experiments were conducted with 300lux scene lighting, and with 1000lux. Corridor lighting is arround 100lux[8], office or room lighting is arround 400lux and TV studios work with 1000lux. The objective of the experiments is, firstly, to determine person point clouds to obtain a qualitative impression of the 3D shape reconstruction and, secondly, to identify reconstructed 3D positions taken from the point cloud and compare them against a reference. The interior orientation of the cameras and the base line were estimated based on (Strauß et al., 2014) in a bundle adjustment with 450 images for each lens. Semi Global Matching (Hirschmüller, 2008) was used as dense matcher.

As reference measurement system for the evaluation of the reconstructed stereo locations a Leica 3D DISTO laser was used, which has a precision of around 1mm for the working distance of 10m[9]. To register the laser data to the bifocal stereo camera coordinate system the transformation and rotation was estimated with 46 ground control points, including a RANSAC (Fischler and Bolles, 1981) procedure for outlier rejection. The root mean square error of the registration was 0.53cm.

The first experiment is discussed in the following. In Figure 10 the person point clouds are shown for 1000 lux office lighting, to give a first qualitatively impression of the results. The shape of the dummy remained unchanged during the experiments. As

[8] en.wikipedia.org/wiki/Lux
[9] lasers.leica-geosystems.com/eu/de/3d-disto/3d-disto

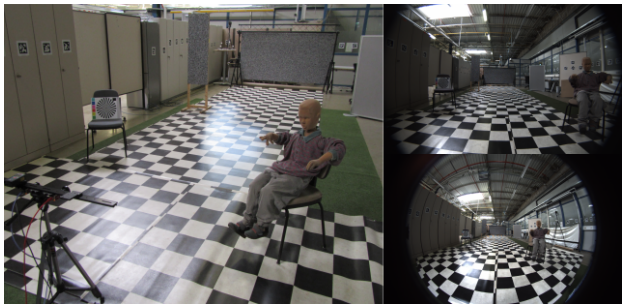

Figure 9. Overview of the experiment (left). Fish eye image (right, bottom) and rectilinear image (right, top).

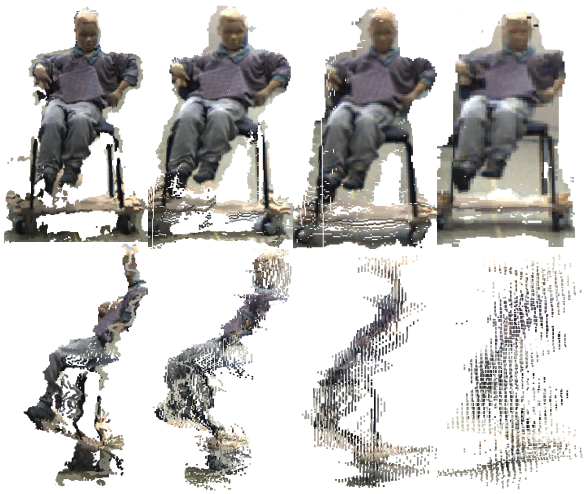

Figure 10. Point clouds for 4,6,8 and 10m distance and 1000lux lighting. The top row depicts four different distances to the camera in a front view. The second row shows the same point cloud from a side view.

can be seen it is difficult to differentiate e.g. between nose and forehead in a side view. Also, it is hardly possible to differentiate between arm and torso for distances larger than six meters. Moreover, the same person looks very different from two different distances. To distinguish between two different persons is harder as in 2D images, since the appearance enables more comparison possibilities than a point cloud in the obtained quality. In 10m working distance the reconstructed surface can consequently not be regarded as a discriminative feature for the re-id of persons.

In the following second experiment the reconstructed 3D positions are validated to obtain quantitative results for precision of one single point. The objective of this experiment is to determine the 3D surface reconstruction error for distances from 4m to 13m from a single stereo image pair and dense matching. Here, only one single point of the densely reconstructed surface is focused on. The person dummy was located at 20 different locations. For each location two stereo measurements were taken, one for 300lux and one for 1000lux scene lighting. The center point of the dummy forehead was the measurement target and could easily be re-identified in the point cloud and also with the laser reference system. The coordinates in the reference system coordinate system were transformed to the camera coordinate system and compared with the bifocal stereo coordinates, see Figure 11 for 300lux and Figure 12 for 1000lux. The bifocal stereo coordinates are shown as crosses (green) and the coordinates of the reference system as circles (red). The numbers represent the Euclidean distance between the two measurements and the error in Z-direction (depth). The error of the reference system and the interactive measurement is determined to less than 1cm each. In a working distance of around ten meters an error of better than 20cm (corresponding to around one third pixel matching accuracy) was achieved in the central image region, irrespective of illumination. At the border of the central region 48cm (300lux) and 18cm (1000lux) were obtained, and 32cm (300lux) / 28cm (1000lux) outside the central region.

The KITTI[10] Stereo Benchmark showed 0.6pixels *Average disparity / end-point error in non-occluded areas* (Avg-Noc) for Semi Global Matching. The pixel matching accuracy of one third

[10] cvlibs.net/datasets/kitti/eval_stereo_flow.php?benchmark=stereo&table=all&error=2&eval=est

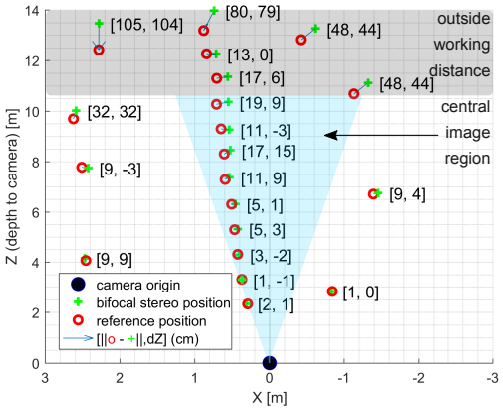

Figure 11. Visualization of the reconstructed 3D positions in a bird's eye view, for 300 lux lighting. Points within the bluish triangle are obtained from the central image region.

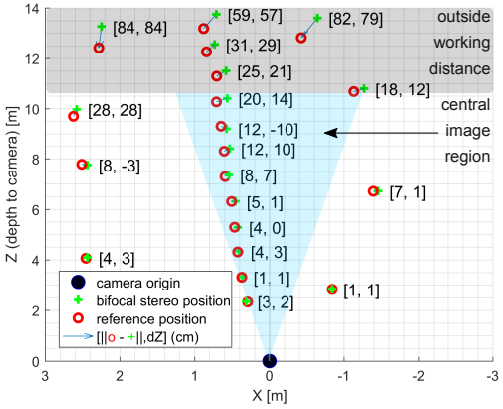

Figure 12. Visualization of the reconstructed 3D positions in a bird's eye view, for 1000 lux lighting. Points within the bluish triangle are obtained from the central image region.

is consequently plausible and confirms the magnitude of the resulting error.

Furthermore, the experiment shows: (a) for 12 out of 20 measurements (60%) the depth error for 1000lux lighting was smaller than for 300lux. The reconstructed 3D position was 13 times (65%) better with 1000lux. (b) The absolute error increased with higher distance to the central image region. (c) With increasing distance to the camera the depth respectivly 3D error increases. For a PRID exploration the achived values confirm, that 3D positions can normally not be used for discrimination. (d) In the addressed working distance of 10m the achieved accuracy was 19 cm in the best case.

4.3 Person Candidate Rejection by inacurate 3D Information

In this subsection the reconstruction accuracy necessary to actually use the bifocal approach for PRID fusion is discussed. One PRID example is addressed which can deal with inaccurate person height and supports a subsequent 2D PRID algorithm, namely person candidate rejection by using the person height reconstructed from 3D information. The fusion works as follows: Measured person heights are used in combination with an error range value to reject possible 2D person candidates and to reduce the number

of potential person candidates for a subsequent 2D PRID fusion approach.

In a first step consider a queried person p_P (probe person) in a first view and a gallery of persons $p_G = [p_1, p_2, ..., p_P, p_n]$, including the first person in a second view. The persons have random heights h_p. We assume that people with heights outside an interval of uncertainty, denoted as

$$h_{p_P} - 2 \cdot e \leq h_{p_P} \leq h_{p_P} + 2 \cdot e \tag{5}$$

where e is the precision of measurements, can be rejected as possible 2D PRID matching candidates. In a second step the remaining possible person candidates are in the focus. Consider the person gallery size s, the mean person height in Germany μ_P (DIN EN 33402-2, 2005) and the standard deviation of person heights in Germany σ_P (DIN EN 33402-2, 2005). The underlying assumption is the population is Gaussian distributed in person height. Furthermore, the percentage of persons that can not be differentiated in terms of person height is determined with the error function (integral over the Gaussian distribution, erf) to

$$p(h_p - 2 \cdot e \leq h_p < h_p + 2 \cdot e) = \frac{1}{2} \cdot erf(\frac{(h_P + e \cdot 2) - \mu_P}{\sqrt{2} \cdot \sigma_P}) - \frac{1}{2} \cdot erf(\frac{(h_P - e \cdot 2) - \mu_P}{\sqrt{2} \cdot \sigma_P}). \tag{6}$$

This equation shows the percentage of population which can not be distinguished. This means for PRID that the correct match is somewhere between rank#1 and the rank corresponding the maximum number of persons that can not be distinguished, called worst case rank. The worst case PRID rank from the person height distribution is denoted as $rank_{WC} = s \cdot p \cdot 0.01$. Since in the worst case, a probe person is on the last indistinguishable $rank_{WC}$ and can only be distinguished from other gallery persons that are not lying in the interval of uncertainty, the number of persons, that can not be distinguished, depends on the final person height. With increasing distance to the mean person height the absolute number of rejection candidates increases.

In Figure 13 the simulated Cumulative Matching Characteristic curve (CMC) is shown for a Gaussian distributed simulated person height dataset. The CMC curve is the commonly used metric to compare PRID approaches. The simulated distribution is again based on the German population, including males and females, between 18-65 years. On the ordinate the re-id performance is shown and on the abscissa the rank#. For the simulation for every person of the population the worst case rank was determined and the CMC curve was iteratively generated. In other words the diagram shows, how many re-id ranks have to be taken into account to be sure, that the correct match is among them. The visualized height estimation precision e is 1cm, 2cm, 5cm, 10cm and 20cm. The used gallery and probe size is 316 persons, since this is the gallery size for the most cited dataset, VIPeR. Absolute values of ranks bellow 100% re-id rate are not important in this context, so the person ranks below 99% were set to a random number between the perfect match and the worst case rank to illustrate which real re-id result can be expected. However, it is important to point out the rank, when the first time 100% performance is achieved because this information can finally be used to validate the person rejection only by height. From 316 persons in the gallery with a height error of 1 centimetre, it is possible to reduce the gallery by 78% to 68 remaining persons that can be re-identified with ordinary 2D PRID approaches. The removed 238 persons can also not contribute to wrong 2D person matches. With 2 centimetres height error 174 persons (55%) can be rejected, whereas an error of 5cm results in 46 persons (14%) being rejected. Further, Figure 13 depicts, that a height error of 10cm and 20cm is not beneficial for the person candidate rejection.

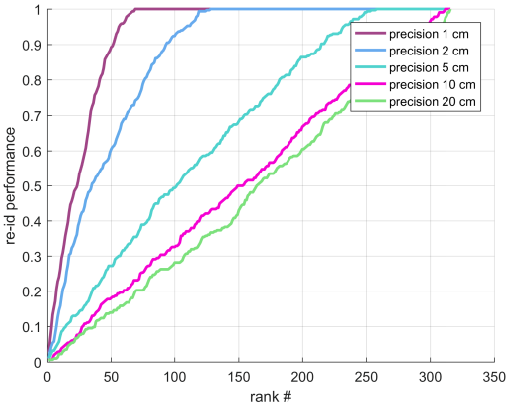

Figure 13. CMC curve for the simulation of German population.

The necessary precision of one, two and five centimetres can be reached with the bifocal approach only with multiple measurements (more than one stereo image) and filtering to get a higher precision or with a single measurement if the working distance of the approach is, by referencing to the 3D reconstruction experiment result, limited to around five meters (cf. Fig. 11,12).

5. CONCLUSION

In comparison to existing methods using mono cameras or the Kinect the new bifocal stereo approach proposed in this work increases both robustness and accuracy of PRID algorithms. The major extension is a new architecture with two cameras with different projections and viewpoints. Additionally, the system allows for the generation and exploitation of 3D data.

Section 4.1 showed that using this approach the working distance can be increased, since a person can be re-identified in near-range and long-range images. Moreover, the proposed hardware configuration increases the re-identification rate by 10% in the overlapping field of view and increases the working distance (c.f. section 3.1).

The previous sections showed in theory (cf. Sec. 3.3, 3.4) and experiments (cf. Sec. 4.2) that the reconstructed 3D surface does not have a high precision and it is hardly possible to use shape parameters derived from the reconstructed 3D point cloud to distinguish between persons. This result is not surprising since section 3.3 showed that the related necessary depth resolution corresponds to a parallax accuracy which cannot be reached in real world applications.

REFERENCES

Barbosa, I. B., Cristani, M., Del Bue, A., Bazzani, L. and Murino, V., 2012. Re-identification with rgb-d sensors. In: *Computer Vision - ECCV 2012. Workshops and Demonstrations*, Vol. 7583.

DIN EN 33402-2, 2005. *Ergonomie, Körpermaße des Menschen, Teil 2: Werte*.

Fischler, M. A. and Bolles, R. C., 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24(6), pp. 381–395.

Förstner, W. and Wrobel, B. P., 2016. *Photogrammetric Computer Vision – Statistics, Geometry, Orientation and Reconstruction*. Springer.

Garca, J., Martinel, N., Micheloni, C. and Gardel, A., 2015. Person re-identification ranking optimisation by discriminant context information analysis. In: *ICCV 2015*, pp. 1305–1313.

Gray, D., Brennan, S. and Tao, H., 2007. Evaluating appearance models for recognition, reacquisition, and tracking. In: *PETS*.

Hirschmüller, H., 2008. Stereo processing by semi-global matching and mutual information. *in IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(2), pp. 328–341.

Imani, Z. and Soltanizadeh, H., 2016. Person reidentification using local pattern descriptors and anthropometric measures from videos of kinect sensor. *IEEE Sensors Journal*.

Karanam, S., Gou, M., Wu, Z., Rates-Borras, A., Camps, O. I. and Radke, R. J., 2016. A comprehensive evaluation and benchmark for person re-identification: Features, metrics, and datasets. *CoRR*.

Leal-Taixé, L., Milan, A., Reid, I., Roth, S. and Schindler, K., 2015. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942 [cs]*. arXiv: 1504.01942.

Leng, Q., Hu, R., Liang, C., Wang, Y. and Chen, J., 2015. Person re-identification with content and context re-ranking. *Multimedia Tools and Applications* 74(17), pp. 6989–7014.

Liao, S., Hu, Y. and Li, S. Z., 2014. Joint dimension reduction and metric learning for person re-identification. *CoRR*.

Liu, H., Hu, L. and Ma, L., 2017. Online rgb-d person re-identification based on metric model update. {*CAAI*} *Transactions on Intelligence Technology* 2(1), pp. 48 – 55.

Matsukawa, T., Okabe, T., Suzuki, E. and Sato, Y., 2016. Hierarchical gaussian descriptor for person re-identification. In: *CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, IEEE Computer Society, pp. 1363–1372.

Mei, C., 2007. Couplage Vision Omnidirectionnelle et Télémétrie Laser pour la Navigation en Robotique/Laser-Augmented Omnidirectional Vision for 3D Localisation and Mapping. PhD thesis, INRIA Sophia Antipolis.

Milan, A., Leal-Taixé, L., Reid, I., Roth, S. and Schindler, K., 2016. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*. arXiv: 1603.00831.

Roth, P. M., Hirzer, M., Köstinger, M., Beleznai, C. and Bischof, H., 2014. Mahalanobis Distance Learning for Person Re-Identification. In: *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, Springer, pp. 247–267.

Strauß, T., Ziegler, J. and Beck, J., 2014. Calibrating multiple cameras with non-overlapping views using coded checkerboard targets. In: *ITSC 2014, Qingdao, China, October 8-11, 2014*, IEEE, pp. 2623–2628.

Wang, T., Gong, S., Zhu, X. and Wang, S., 2014. *Person Re-identification by Video Ranking*. Springer International Publishing, Cham, pp. 688–703.

Wang, T., Gong, S., Zhu, X. and Wang, S., 2016. Person re-identification by discriminative selection in video ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(12), pp. 2501–2514.

Wu, A., Zheng, W. S. and Lai, J. H., 2017. Robust depth-based person re-identification. *IEEE Transactions on Image Processing* 26(6), pp. 2588–2603.

Yang, Y., Yang, J., Yan, J., Liao, S., Yi, D. and Li, S. Z., 2014. *Salient Color Names for Person Re-identification*. Springer International Publishing, Cham, pp. 536–551.

Zhong, Z., Zheng, L., Cao, D. and Li, S., 2017. Re-ranking person re-identification with k-reciprocal encoding. *CoRR*.