Estimating Target Orientation with a Single Camera for Use in a Human-Following Robot

Michael Burke
Mobile Intelligent Autonomous Systems
Council for Scientific and Industrial Research
Pretoria, South Africa
Email: michaelburke@ieee.org

Willie Brink
Applied Mathematics
Department of Mathematical Sciences
University of Stellenbosch, South Africa
Email: wbrink@sun.ac.za

***Abstract*—This paper presents a monocular vision based technique for extracting orientation information from a human torso for use in a robotic human-follower. Typical approaches to human following use an estimate of only human position for navigation, but we argue that a better navigation scheme should include directional information. We propose that the pose of a walking person's upper body typically indicates their intended travelling direction, and show that a simple planar fit to the back of a human torso contains sufficient information for the purpose of inferring orientation. We obtain this planar fit using only 2D image points. Results showing the efficacy of this approach are presented, together with those of a simple human-following controller incorporating the pose estimate.**

I. Introduction

The ability of a mobile robot to track and follow a human is required in a wide variety of applications, particularly in service robotics. Human-following robots not only need to detect, recognise and track their targets in real time but also navigate towards them in an intelligent manner.

These robots are typically equipped with a diverse and varying combination of sensors for locating and recognising targets. Light detection and ranging (LiDAR) [1], for example, provides accurate distance measurements but may lead to potential ambiguity in target recognition. Electronic tethering techniques that use radio frequency identification (RFID) [2] are effective but require that the human followed wear a tracking device and often need a secondary sensor for greater measurement accuracy. As a result many systems employ vision, selected for its ability to provide abundant information about the robot's environment, in a passive manner, at relatively high speeds and low cost.

Early examples of vision based human-following robots made use of simple template matching schemes [3] or colour based blobs with contour models [4] for target detection and recognition. The latter approach uses stereo-vision in order to obtain an absolute scale representation of the human's position. In the single camera case, however, absolute scale is typically not available and alternative distance measurements, such as the size of detected blobs [5], are incorporated for navigation. These approaches and other early ones suffer potential target ambiguity in the presence of multiple persons or cluttered environments, mainly due to detection and recognition schemes that are not particularly robust. More recently, feature based approaches have been applied to the problem with impressive results [6].

The above-mentioned approaches, and many others, normally use merely some form of position measurement for navigation. More intelligent navigation schemes could be implemented, however, if some knowledge of the intended motion of the human target was incorporated. In fact, results of a preliminary study on the social acceptance of human-following approaches [1] indicate that the following of direction is more acceptable to people than point-to-point path following.

A human-following system that includes orientation information requires that some measure of human body pose be made. Unsurprisingly, human pose estimation is a popular topic within the computer vision research community and a large body of work on the subject is available. Common approaches fit complex articulated body models to image scenes [7]. While these and similar techniques are effective and produce commendable results, our focus is on simplicity and speed. Moreover, these techniques typically produce a large amount of information (e.g. individual limb positioning) where a single orientation angle would be sufficient for our purpose of controlling a wheeled robot.

In this paper we present a means of extracting human orientation for use in a human-following robot. A feature based matching scheme is chosen to detect and recognise the human target, in an effort to minimise the likelihood of tracking ambiguity, and an approach to extracting human pose information from a single image is explained. A measure of certainty in the pose estimates is introduced and we present results showing the efficacy of our approach. Results of human-following using a simple controller that incorporates our estimate are presented and discussed.

II. Our Method

Our system operates under the assumption that the pose of a walking person's upper body typically indicates travelling direction. Although humans are capable of walking in directions opposed to that indicated by their torsos, this is certainly not the norm and, intuitively, the assumption seems valid. It is justified further by the work of [8].

The authors of [8] attempted to simulate human walking on level ground using a three-dimensional, neuromusculoskeletal

model of the body together with dynamic optimisation theory. They compared their model with data captured from a variety of sources in human walking trials. This data showed that the deviation in back angle of a walking person typically remains within 10°. As we are interested only in the approximate facing direction of the human, a complex model of body shape and limb position is not required. A simple planar fit to the back of the torso should contain sufficient information for us to infer travelling direction.

The proposed method allows a planar fit to a torso to be obtained from a single image obtained by a perspective camera mounted on a robot. Note that the system requires that relatively salient clothing be worn by the human because the detection is feature based. Initially, matches relating the robot's current view of the human to some desired view are obtained. Here the desired view would typically be a fronto-parallel image of the back of the human's torso. A planar homography mapping the features in the current view to the desired view is then estimated, from which pose measurements are extracted.

Although features detected on the back of a human torso are usually not strictly coplanar, a sufficiently robust method of homography estimation is able to discard errors induced by this assumption. In addition, a robust measure of homography is also required to reduce errors caused by the deformable nature of clothing, which may ripple and warp during motion.

A. Detection

Point correspondences between the current and desired frame are obtained from the speeded-up robust features (SURF) of Bay *et al.* [9]. The SURF descriptor performs similarly to the widely used scale invariant feature transform (SIFT) [10], but uses first-order Haar wavelet responses instead of gradients. SURF also makes use of integral images for filter convolutions, thereby greatly improving processing speed. We selected this approach because of its high speed as well as the good detector repeatability over varying blur, scale and viewpoint angle. It is possible to further reduce the computation time by limiting the scale range over which interest points are detected.

Matching features between the current frame and desired view is accomplished by conducting a nearest neighbour search on the SURF descriptors calculated at the interest points. With this method we obtain good matching results for a wide variety of torso motions, despite the fact that SURF was not specifically designed for affine invariance.

B. Homography Estimation

Our goal is to find a means of fitting a plane through keypoints detected on the back of a human's torso. This is a relatively simple task if the 3D locations of features on the torso are known, but the only information available when a single perspective camera is used is the projected 2D locations of detected features on the image plane.

Suppose $\mathbf{x}_1$ and $\mathbf{x}_2$ represent the projections in two images of some point on a plane, in homogeneous coordinates. These projections can be related by means of a 3×3 homography matrix $\mathbf{H}$, as follows:

$$\mathbf{x}_1 = \mathbf{H}\,\mathbf{x}_2. \tag{1}$$

Note that this relationship assumes an ideal pinhole camera model and therefore requires images to first be dewarped with respect to lens distortion.

If we were to estimate the homography between two views of a human torso, we would effectively be measuring the rotation and translation between two planar approximations of the torso.

The normalised direct linear transform (DLT) [11] can be used to find the homography from at least four available point correspondences because every correspondence provides two equations and, since $\mathbf{x}_1$ and $\mathbf{x}_2$ are specified in homogeneous coordinates, $\mathbf{H}$ is specifiable up to scale.

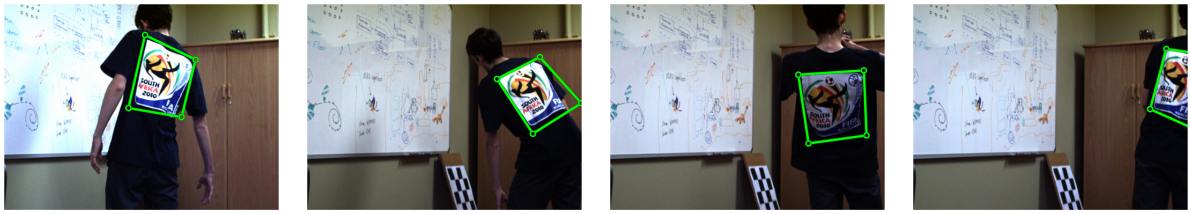
The problem is likely to be over-specified as typically more than four correspondences are found by the SURF matching scheme. Many correct matches would be useful in solving for the homography in a least-squares sense but, unfortunately, incorrect matches (outliers) can have a drastic negative effect on such a solution. We therefore opt for an iterative RANSAC based approach [12], in an effort to find a homography that minimises a re-projection error.

In our context RANSAC, short for random sample consensus, operates as follows. From the set of all available point correspondences a random subset of four is drawn and a model homography is determined using DLT. This homography is used with a re-projection error to determine which of the remaining points agree with the model, thereby forming a consensus set. In our case the re-projection error measures the error between the original coordinates of matched points and those projected in both directions under the model homography. If the consensus set is large enough the final homography is calculated from it as a least-squares solution. If not, a new subset is chosen and the process is repeated until a large enough consensus set is obtained or a specified number of iterations is reached, in which case the final homography is calculated from the largest consensus set found.

This robust RANSAC based homography estimation is extremely effective at obtaining homographies in the presence of a large number of outliers. This property is especially desirable as many outliers could be present in our system due to the deformable nature of clothing, the occasional mismatched feature and the slight curvature (or deviation from planarity) of a human torso.

Also, and importantly, the use of RANSAC based homography estimation allows for a measure of certainty to be obtained. We define this certainty measure as the ratio of inliers used for homography estimation to the total number of detectable features on a target (the number of features marked by the SURF algorithm on the template image). This ratio implies that as the size of the consensus set increases so does the trust in the estimated homography.

(a) reference template (b) approx. 30° yaw (c) rippled shirt (d) approx. −45° yaw



(e) yaw and roll (f) forward tilt (g) backward tilt (h) roll, partial occlusion

(i) near, partial occlusion (j) distant, blurred (k) outdoors (near) (l) outdoors (far)

Fig. 1: Results of the single-view homography based pose measurement system on a range of test cases. The template image is shown in (a). The superimposed green quadrilaterals in (b)–(l) show the estimated planar approximations from which position and orientation, relative to the template, are extracted.

C. Homography Decomposition

Once the homography has been determined the various pose parameters, mapping the current camera coordinate system to the desired (template) camera coordinate system, can be retrieved from the decomposition

$$\mathbf{H} = \mathbf{K}\left(\mathbf{R} + \mathbf{t}\,\mathbf{n}^T\right)\mathbf{K}^{-1} \tag{2}$$

[13], where $\mathbf{K}$ is the intrinsic camera calibration matrix, $\mathbf{R}$ a rotation matrix, $\mathbf{t}$ the translation of the camera and $\mathbf{n}$ a vector normal to the target surface. There are eight degrees of freedom: three in the rotation and five in the surface normal and camera translation (which is extractable up to scale).

We use the algorithm of Faugeras and Lustman [13] to calculate the pose parameters in (2) from a given homography. Camera effects are removed from $\mathbf{H}$ and the singular value decomposition (SVD) of the result is obtained, as

$$\widehat{\mathbf{H}} = \mathbf{K}^{-1}\,\mathbf{H}\,\mathbf{K} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^T. \tag{3}$$

Here $\mathbf{K}$ indicates the camera calibration matrix, obtainable in an offline calibration procedure.

The diagonal matrix $\mathbf{\Sigma}$, containing singular values of $\widehat{\mathbf{H}}$, can be decomposed into the various pose parameters with relative ease, such that

$$\mathbf{\Sigma} = \widetilde{\mathbf{R}} + \widetilde{\mathbf{t}}\,\widetilde{\mathbf{n}}^T. \tag{4}$$

This decomposition is rather lengthy, however, and the reader is referred to [13] for details. The algorithm can provide up to eight different solutions but, fortunately, not all are physically possible. The solution set is immediately reduced to four by including the constraint that both image frames must be located on the same side of the target object or, in other words, that the object viewed cannot be transparent. A second constraint, enforcing that visible points must be in front of both cameras, reduces the set to two solutions. Finally a single solution is obtained by incorporating assumed knowledge of the surface normal in the desired view.

The final decomposition elements of $\mathbf{H}$ are then calculated according to

$$\mathbf{R} = \mathbf{U}\,\widetilde{\mathbf{R}}\,\mathbf{V}^T, \quad \mathbf{t} = \mathbf{U}\,\widetilde{\mathbf{t}}, \quad \mathbf{n} = \mathbf{V}\,\widetilde{\mathbf{n}}. \tag{5}$$

The translation vector is returned up to scale, because a single camera is used. However, for the purposes of control, this ambiguity is not a problem as long as the translation components remain monotonic. The controller will minimise error in translation by generating proportional motion commands so, in a sense, the unknown scale is incorporated in the controller gains.

Only three parameters in $\mathbf{R}$, $\mathbf{t}$ and $\mathbf{n}$ are of particular interest for wheeled platform control: the target yaw and the 2D translation to target. The ability to extract the three parameters of interest independently of the unnecessary degrees of freedom

is important though, because it implies that a certain amount of invariance to uneven terrain is present.

Target yaw or orientation is extracted from the rotation matrix and represents a rotation about a vertical axis in the camera coordinate frame. The translations of interest are the shift of the horizontal camera frame axis, t_x and the optical axis shift, t_z. Note that these parameters are measured relative to some template image or reference frame. For our applications we assume that the reference frame is an approximately fronto-parallel image of the target human's back.

III. Results

This section aims to show that the homography based pose estimate provides translations and a measure of orientation that is useful for the purposes of wheeled robot control. While it is difficult to provide insight into the accuracy of the measurement, as no ground truth is available, we aim to show that the homography based plane fit provides a believable estimate of a human torso's facing direction.

A. Pose Estimation

Fig. 1 confirms that the pose estimate is conceptually correct, through examples of planes fit through a human torso using the homography pose estimate. These examples show that a plane fit to the torso appears to capture the facing direction. The reference image of the shirt worn during experiments is shown in Fig. 1(a). All pose estimates obtained are measured relative to this view and the goal of a human-following task would be to generate platform control signals that recreate this view. The superimposed green quadrilateral in (b)–(l) shows the estimated planar approximation to the back of the torso.

As the figure shows, the system is robust when subjected to some extreme human motions and deforming clothing. Valid pose measurements are also obtained when the torso undergoes partial occlusions and over large scale changes. The images obtained outdoors are of poor quality and affected by glare, but illustrate that the system still functions effectively in challenging environmental conditions.

These images show that the pose estimate contains information regarding a person's position and orientation, but do not provide any information as to the accuracy of an estimate. As discussed earlier, only three pose parameters are of interest for the motion control of a wheeled platform: target yaw, the shift of the horizontal camera frame axis, t_x and the optical axis shift, t_z. Results of experiments conducted to test the reliability of these measurements are now presented.

Fig. 2 shows the relationship between actual variations of horizontal target motions and those obtained by the homography based pose estimate. Image sequences of a stationary human target in a fronto-parallel configuration were captured at three positions approximately 2 m from a camera. Note that the homography based pose estimate does not continually provide the same estimate when viewing a target in a static scene, as noise in images causes changes in the features used for pose estimation. This variation in measurement is

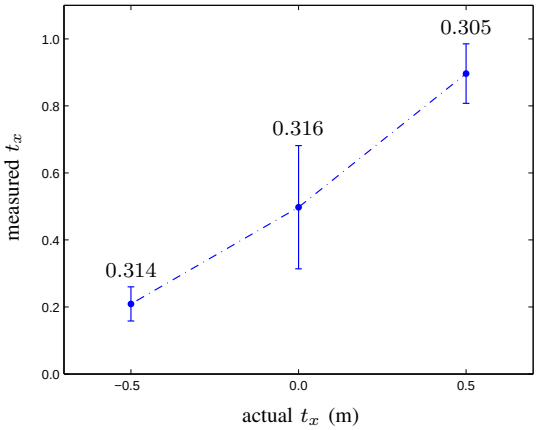


Fig. 2: Mean (dots) and standard deviation (bars) of the measured horizontal translations by our homography based pose estimation, plotted against ground truth. Average certainty measures are also shown (the annotations above the error bars).

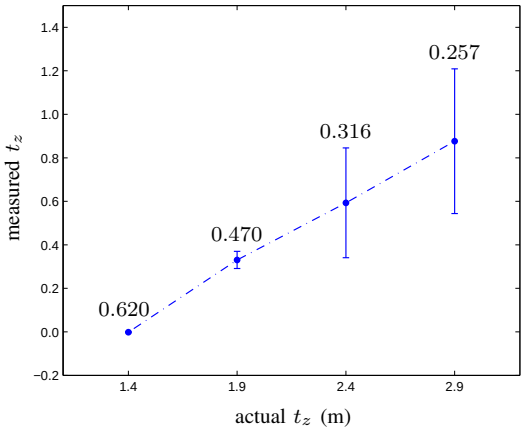


Fig. 3: Mean and standard deviation, and average certainties, of measured optical axis translations against ground truth.

quantified by the standard deviation error bars displayed in the figure, with a line fit through the mean of estimates. The average certainty measures for each position are also noted in the figure.

Fig. 3 shows the relationship between actual variations of target motions along the camera optical axis and those obtained by the homography based pose estimate. As before, image sequences of a stationary human target in a fronto-parallel configuration were captured at incrementing 0.5 m intervals. The variation in measurement, quantified by the standard deviation error bars displayed in the figure, shows that the estimate becomes less reliable as the target moves away from the camera. It also confirms the usefulness of our certainty measure for each position.

Practical experimentation shows that the certainty measure rarely exceeds 0.6, with a measure greater than 0.1 corresponding to a reliable parameter estimate. Note that the

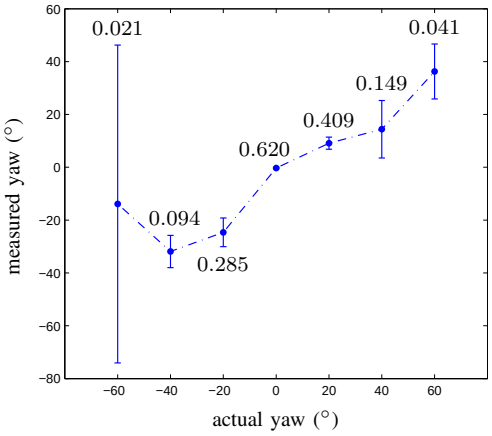


Fig. 4: Mean and standard deviation, and average certainties, of measured target yaw against ground truth.

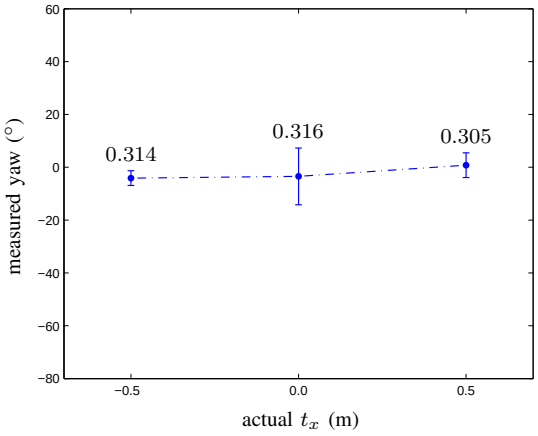


Fig. 5: Effect of actual horizontal target translations on target yaw estimates. Average certainty measures are also shown.

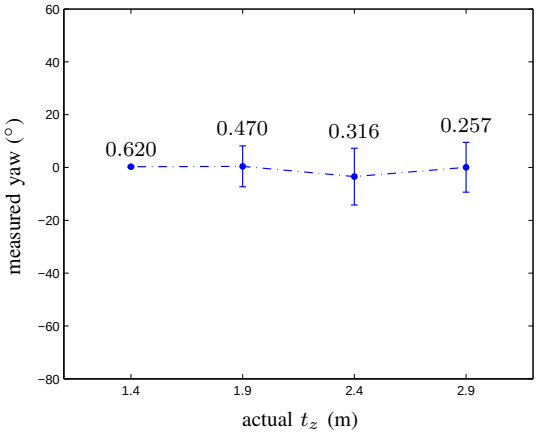


Fig. 6: Effect of actual optical axis target translations on target yaw estimates.

estimated translations are not actually equivalent to the actual translations used for test purposes, but are linearly related by means of a scale and shift. This is unimportant for the purposes of controlling a platform using these parameters, where bias in controller set-point is added to account for any shift and, as mentioned before, scales are incorporated into controller gains. In fact, suitable platform control can be obtained as long as the estimate is monotonic. This is clearly the case for the given translations, indicating that the estimates can be used for the purposes of control.

Fig. 4 shows the relationship between actual variations of target yaw and those obtained by the homography based pose estimate. These estimates were obtained by performing pose estimation on image sequences of a human target with varied orientation. Orientation was controlled by marking $20°$ intervals directly in front of a camera, and capturing image sequences of a human target facing in each these directions. The variation in measurement, quantified by the standard deviation error bars displayed in the figure, shows that the estimate becomes less reliable as the target rotates away from the camera. At $\pm 60°$, near the measured limits of the feature based recognition, the estimate becomes untrustworthy. Again, the average certainty measures confirm this.

Once more, it is important to note that while the yaw estimates are not exactly that of the input system, this is unimportant for the purposes of control. As long as the measurements are monotonic, a suitable controller will still result in corrective motions that cause the magnitude of target yaw to decrease. This results in target pose estimates that are more trustworthy, as indicated by the certainty measures in Fig. 4, which in turn allows for more accurate orientation control.

Figs. 5 and 6 show the variation in target yaw, given pure translations. Ideally, the estimate should be independent, but in practice this is not achieved. Fortunately, this variation is not significant and does not affect the estimate's use in a control system. Recall that the use of this estimate of human pose is still based on the assumption that a human's upper body represents a good measure of their travelling direction. The work of [8] reiterates this, with the finding that a human torso deviates within approximately $10°$ during a walking task. This finding implies that the non-ideal variation of target orientation given translations typically falls within the estimate noise floor, and is thus not overly significant.

B. Human Following

We now show that the pose information extracted is of use, by briefly presenting results of a simple controller used for human-following. The controller implemented aims to minimise errors in the relative target orientation and translations to the reference frame. This controller behaviour essentially results in platform motion commands that attempt to move the camera in such a way as to recreate the reference or template image view, against which all measurements are made.

The results presented here were captured using the odometry measurements of a Pioneer P3-AT test platform. Although

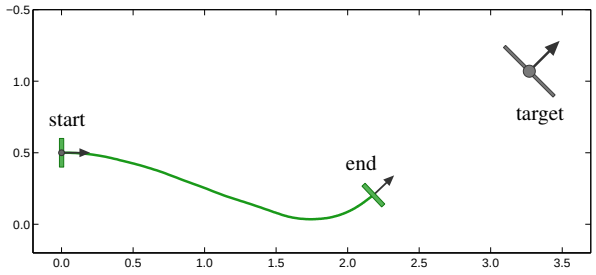


Fig. 7: Step response (in m) of the robot to a target offset in both position and orientation.

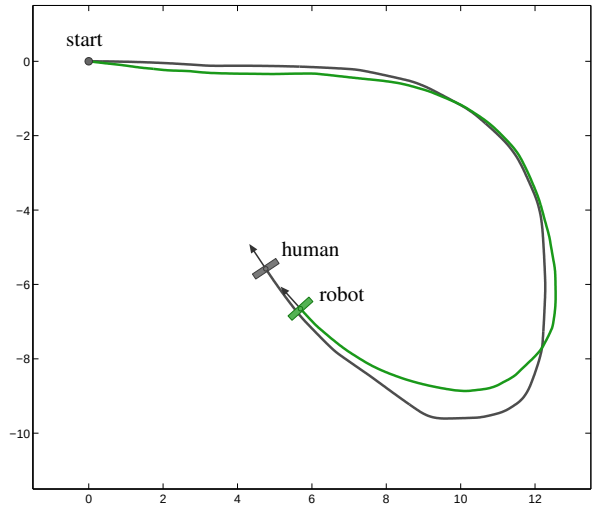


Fig. 8: Path of a robot following a human. As no ground truth for human motion is available, an attempt is made to obtain this using the odometry of a robot driven along the same path prior to the human-following task.

odometry is subject to drift, it is still reasonably accurate over short distances and provides a good idea of the platform's motion.

Fig. 7 shows the controller behaviour when responding to an offset straight line, with differing orientation to that of the platform. It shows that the controller causes the platform to move in such a way as to always be behind the human target, allowing for improved human-following trajectories.

Fig. 8 shows results of a human-following task incorporating the orientation and translation estimates presented here. The results show that the information extracted through the homography based pose estimate is indeed useful, as the robot is able to track and follow a human target using a controller incorporating this information.

IV. Conclusions

We have presented an approach to robotic human-following that makes use of feature matching for detecting and recognising a human target from a single-camera view, and then extracts pose information. Our approach approximated a planar fit to the target using a robust measure of a homography mapping between 2D image coordinates. The results of this technique were presented and shown to incorporate sufficient information for the purposes of platform motion control. Moreover, our approach allowed for a certainty measure to be defined for an estimated pose which may be extremely useful in a control situation.

A simple controller that illustrates the benefit of a human pose measurement that includes basic orientation information was implemented to confirm the validity of the estimate. Though this controller may not be optimal for human-following, it showed that the information extracted through the homography based pose measurement was useful, and could be applied to a more complete human-following system.

Future work may involve the design of navigation schemes that optimally follow a human target, given the constraints of the feature based pose estimate, and the incorporation of additional target recognition strategies for added redundancy.

Acknowledgment

This work was funded by the Council for Scientific and Industrial Research, South Africa.

References

[1] R. Gockley, J. Forlizzi, and R. Simmons, "Natural person-following behavior for social robots," in *HRI '07: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2007, pp. 17–24.

[2] T. Germa, F. Lerasle, N. Ouadah, V. Cadenat, and M. Devy, "Vision and RFID-based person tracking in crowds from a mobile robot," in *IEEE/RSJ International Conference on Intelligent Robotics and Systems*, 2009, pp. 5591–5596.

[3] N. Hirai and H. Mizoguchi, "Visual tracking of human back and shoulder for person following robot," in *IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, 2003, pp. 527–532.

[4] C. Schlegel, J. Illmann, H. Jaberg, M. Schuster, and R. Worz, "Vision based person tracking with a mobile robot," in *British Machine Vision Conference*, 1998, pp. 418–427.

[5] H. Latif, N. Sherkat, and A. Lotfi, "Fusion of automation and teleoperation for person-following with mobile robots," in *IEEE International Conference on Information and Automation*, 2009, pp. 1240–1245.

[6] Z. Chen and S. Birchfield, "Person following with a mobile robot using binocular feature-based tracking," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2007, pp. 815–820.

[7] Y.-R. Chen, C.-M. Huang, and L.-C. Fu, "Upper body tracking for human-machine interaction with a moving camera," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 1917–1922.

[8] F. C. Anderson and M. G. Pandy, "Dynamic optimization of human walking." *Journal of Biomechanical Engineering*, vol. 123, no. 5, pp. 381–390, October 2001.

[9] H. Bay, A. Ess, T. Tuytelaars, and L. van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[11] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.

[12] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[13] O. Faugeras and F. Lustman, "Motion and structure from motion in a piecewise planar environment," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 2, pp. 485–508, 1988.