An Inexpensive Monocular Vision System for Tracking Humans in Industrial Environments

Rafael Mosberger and Henrik Andreasson
Centre for Applied Autonomous Sensor Systems (AASS), Örebro University, Sweden

***Abstract*—We report on a novel vision-based method for reliable human detection from vehicles operating in industrial environments in the vicinity of workers. By exploiting the fact that reflective vests represent a standard safety equipment on most industrial worksites, we use a single camera system and active IR illumination to detect humans by identifying the reflective vest markers. Adopting a sparse feature based approach, we classify vest markers against other reflective material and perform supervised learning of the object distance based on local image descriptors. The integration of the resulting per-feature 3D position estimates in a particle filter finally allows to perform human tracking in conditions ranging from broad daylight to complete darkness.**

I. Introduction

Recent years have seen an increasing interest in assistive technology and safety equipment for industrial vehicles that operate in the vicinity of human workforce in shared workspaces. Aiming at preventing work-related accidents, the industry demands reliable on-board human detection solutions that can then be integrated into driver assistance systems. At the same time, detecting humans is even a crucial prerequisite in the shift towards autonomous vehicles. While similar requirements in the field of road traffic safety have led to the development of advanced pedestrian protection systems, their adaption from the use in cars to the use in industrial machines has so far attracted less attention.

Our work therefore focuses on the task of people detection from construction and transportation machinery such as articulated trucks, wheel loaders, crawler tractors or forklift trucks that operate on industrial worksites like manufacturing areas, construction sites, warehouses, or storage yards. Due to the often bulky nature of these machines, the field of view from the driver's cabin is limited and worksite accidents often occur as a result of rear and blind-spot collisions. Equipping a vehicle with a human detection unit can help the driver navigate more safely around workers by being made aware of possibly dangerous driving situations. In the case of autonomous machines, a human detector delivers the necessary input for safe path-planning.

However, the requirements for reliable human detection from industrial machines are manifold and challenging. The machines often operate alternately indoors and outdoors, by day and night, and are thus exposed to a wide range of different weather and illumination settings. Workers to be detected appear at various distances and angles, in front of a potentially cluttered background, partly occluded and in a variety of body poses including standing upright, walking or sitting while possibly carrying objects of various size and shape. Especially on construction sites the vehicles are further confronted with rough terrain. The machines and the observed persons are typically in motion and the desired low reaction times entail considerable real-time constraints.

To facilitate the detection task, we exploit the fact that the use of a reflective safety vest (see Fig. 1a) by operators on industrial worksites is often a mandatory legal requirement. The retro-reflective vest markers redirect the light back along its incident direction and therefore increase the worker's visibility when illuminated by a light source close to the observer. Using this property, we introduced a single camera system (cf. Fig. 1b) equipped with an infrared (IR) flash that allows reliable detection of humans wearing a reflective vest [1]. By processing single image features and adopting a supervised learning based approach, a vest reflector is not only detected in the image but also approximately localized in 3D space [2].

In this paper we extend the work previously presented in [2]. We describe a complete monocular vision based tracking system with a wide field of view, able to maintain a 3D position estimate for a single observed person wearing a reflective vest and at distances up to 10 meters. Our work makes the following two major contributions. First, we show that by incorporating single per-feature position estimates (obtained according to [2]) in a particle filter with an appropriate measurement model, we are able to perform human tracking with an accuracy in the decimeter range using a single camera system. Second, we present a comprehensive experimental indoor and outdoor evaluation of the complete system and its individual processing steps in different environments and very diverging illumination settings. Thereby, we extend the range of evaluated feature descriptors to include also the recent rotation-invariant BRISK descriptor.

II. Related Work

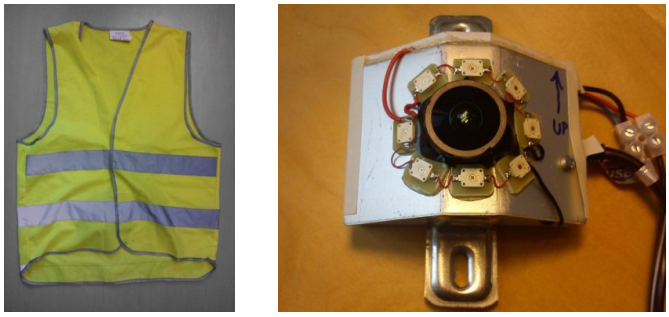
Our work is closely related to the field of pedestrian protection systems (PPSs) that represent a particular type of advanced on-board driver assistance systems developed for the automobile industry. The goal of a PPS is to detect the presence of pedestrians in a specific area around the vehicle and mitigate dangerous situations by providing the driver with an alert signal or by even taking counteractive measures. Most approaches described in literature are based on the input of cameras, working either in the visible, near infrared (NIR) or thermal infrared (TIR) spectrum. Alternative sensor modalities such as lidar and radar, as well as the fusion of different

sensors have also been investigated but have not gained the same popularity as vision-based systems. Recent surveys about achievements and adopted methods in pedestrian detection systems are given by [3], [4], [5] and [6].

For an analysis of the adopted methods in PPSs, Geronimo et al. [3] conceptually break down their architecture into several modules. The ones most related to our work are foreground segmentation, object classification and tracking. Foreground segmentation aims at extracting regions of interest from the raw input data. In vision-based approaches it is often achieved by considering criteria for color, intensity, texture or optical flow. Segmenting objects from an estimated ground plane is also common. Other algorithms use no explicit segmentation and perform exhaustive scanning over the entire image (e.g. the well-known HoG detector [7]). More recent approaches also successfully use stereo-based depth maps for segmentation. Object classification methods used in PPSs can be regrouped into template-based silhouette-matching approaches (e.g. Chamfer System [8]) and appearance based classification of feature descriptors (e.g. HoG [7]). Tracking modules either operate in the 2D image space or in 3D camera coordinates and most often use of Kalman and particle filtering.

Despite the similarities with pedestrian detection systems used in road traffic, the above methods are not directly applicable to industrial vehicles, due to a number of differences. First of all, the flat floor assumption is often not valid on many industrial sites as vehicles are faced with possibly very rough terrain. Therefore, detections cannot be restricted to certain image regions by exploiting ground plane constraints. Furthermore, the fact that machines often operate alternately in bright outdoor and poorly illuminated indoor areas leads to an even broader range of different illumination settings than in road traffic, hampering the adoption of standard vision based detectors that rely on good contrast or texture. The presence of various heat sources (engines, etc.) further makes the use of thermal vision approaches more challenging. When compared to road traffic, industrial vehicles are also bigger and typical motion scenarios comprise more acceleration and deceleration, sharper turns and reversing, resulting in dangerous front, rear and lateral zones. Consequently, the desired sensor coverage for human detection is considerably bigger than the relatively narrow cone observed in front of a regular car.

Based on the above observations and the fact that we in our work restrict detection to the case of people wearing a reflective vest, our system differs from the typical people detectors in a PPS, most notably on the sensory level and the methods employed for image segmentation. In fact, we enhance a standard monochrome imaging sensor with an NIR LED flash and an NIR bandpass filter. This camera setup allows the detection of the retro-reflective vest material through active illumination and subsequent high-intensity blob detection in the acquired image material. The technique is well-known from optical motion capture systems (e.g. [9], [10]) where it serves the detection of retro-reflective markers that are attached to the observed object. The approach works well in absence of any secondary IR light source but results in numerous false positives if the scene background is subject to external IR illumination, also caused by natural sunlight. To resolve this shortcoming, our algorithm processes an input image pair where only one of the two images is taken with an IR flash. This allows further refinement of the foreground segmentation by introducing selection criteria based on the image difference. The same approach that has been previously used in [11] to detect reflective ceiling markers.



(a) (b)

Fig. 1. **a)** Standard off-the-shelf reflective safety vest **b)** Single camera system equipped with a standard monochrome CMOS sensor, infrared bandpass filter (not visible in the image), fish-eye lens and IR-LED ring.

We further aim at locating reflective vest detections in 3D space. A popular vision-based approach is to use a stereo camera and triangulation while alternative monocular methods include depth from motion and depth from focus/defocus. Motivated by the desire to keep the sensor unit compact and inexpensive, we adopt a single camera approach in combination with an algorithm based on sparse local image feature descriptors and supervised learning. Similar approaches for depth estimation have been previously applied and provided good results (e.g. [12], [13]).

III. System Description

The reflective vest detection and tracking system presented in this paper consists of a single camera unit and an ensemble of processing steps that compare two distinct input images in order to detect a person wearing a reflective vest and track it over time in a 3D reference frame attached to the camera. Fig. 2 depicts a schematic overview of the complete algorithm. The core idea is based on the comparison of two input images, one of which is taken with IR flash while the other is captured without active illumination.

In a first stage, reflective material is detected by segmenting image regions that show a clear intensity difference between the two input images, exploiting the fact that reflective material appears significantly brighter in an image captured with IR flash following the strong back reflection. However, as a short time delay exists between the acquisition of the two images, simple image subtraction is inadmissible. Instead, we use a feature-based tracking approach to relate the two consecutive images to each other. Blob-like features are detected in the image taken with flash, then tracked in the corresponding non-flash image, and finally dismissed if their intensity difference

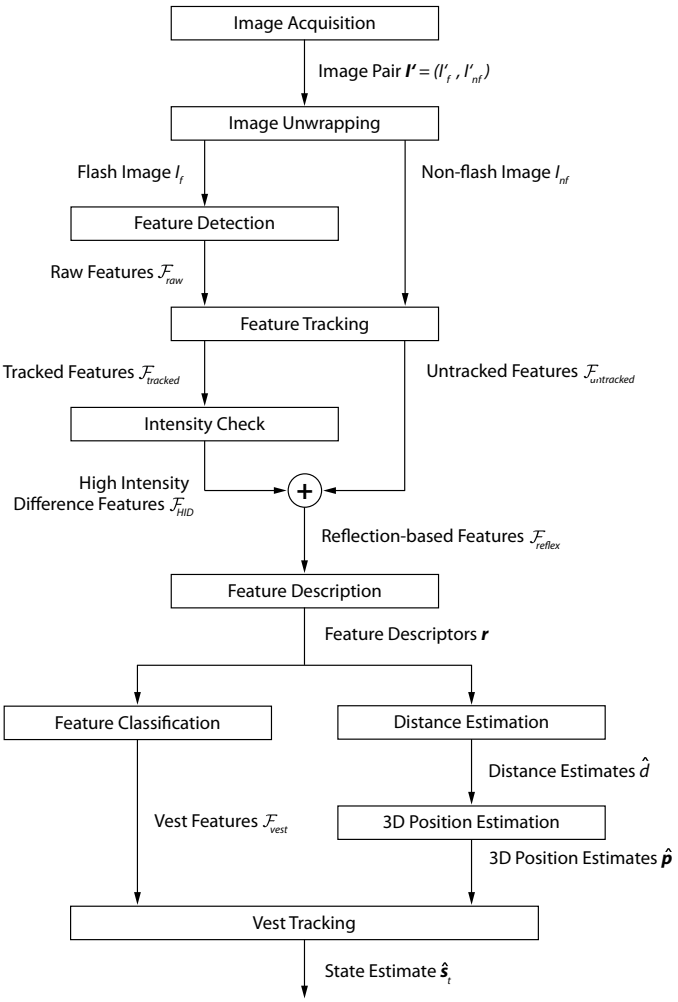


Fig. 2. Overview of the reflective vest detection and tracking system and data flow between the individual processing steps.

across the two images is low. The remaining set of features (i.e. those that could not be tracked) is declared as originating from reflective material and is further processed by extracting a local image feature descriptor from each feature's neighborhood.

The feature descriptors serve two purposes. Using a supervised learning approach, a Random Forest classifier is employed to classify the set of reflective features into features originating from a reflective vest and features originating from other reflective material possibly appearing in the images. Additionally, a Random Forest regressor is trained to estimate the distance between the camera and a reflective vest. Both the classifier and the regressor receive the feature descriptors as input. Using a feature's distance estimate and knowing its 2D location in the image then allows the projection in space and the estimation of a 3D position.

Finally, a particle filter continuously incorporates the single vest feature position estimates in order to maintain an overall estimate of the observed person's location.

A. Hardware and Camera Model

The camera unit (cf. Fig. 1b) consists of a standard monochrome CMOS sensor (IDS Imaging USB UI-1228LE) with a resolution of 752×480 pixels and a fish-eye lens with a field of view (FOV) of $180\,^\circ$. 8 IR LEDs with a wavelength of 850 nm are placed in a ring around the lens and a bandpass filter with a center wavelength of 852 nm and a full width at half maximum of 10 nm is mounted between the lens and the sensor.

We adopt the omnidirectional camera model by Scaramuzza et al. [14] that introduces the image projection function $\boldsymbol{g} : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ describing the relation between a 2D image coordinate pair $\boldsymbol{u'} = [u', v']^\top$, the metric coordinates $\boldsymbol{u''} = [u'', v'']^\top$ on the sensor plane, and a unit length 3D vector emanating from the camera's optical center O to the according scene point in 3D space,

$$\boldsymbol{g}(A\boldsymbol{u'} + \boldsymbol{t}) = \boldsymbol{g}(\boldsymbol{u''}) = (u'', v'', f(u'', v''))^\top \qquad (1)$$

with f a polynomial function, rotationally symmetric with respect to the sensor axis. The affine transformation $A\boldsymbol{u'} + \boldsymbol{t}$ accounts for the digitizing process as well as small axes misalignments.

B. Image Acquisition and Unwrapping

The image acquisition involves taking a pair of images, one with IR flash and one without. The time increment t_a between the capture of the two images is kept as short as possible in order to minimize the difference between the two images due to both changes in viewpoint due to camera motion and changes in the observed scene. The result of the image acquisition is a raw image pair $\boldsymbol{I'} = (I'_f, I'_{nf})$, consisting of the image I'_f taken with flash, and the image I'_{nf} taken without flash. The raw fish-eye images I'_f and I'_{nf} are then unwrapped to create a pair of undistorted panoramic images $\boldsymbol{I} = (I_f, I_{nf})$, containing the area of interest for people detection. The unwrapping is done using a panoramic projection function $\boldsymbol{h} : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ defining the relationship between a pair of panoramic image coordinates $\boldsymbol{u} = [u, v]^\top$ and a unit length vector pointing to the respective point in 3D space (see [15] for further details).

C. Feature Detection and Tracking

The back reflection of the emitted IR flash by the reflectors of a vest results in high intensity blob-like regions in the image I_f. Shape, size and blur of the blobs depend on various factors including the distance to the person, the body pose or occluding objects. Fig. 3 provides several examples of a reflective vest appearing in the image I_f and further illustrates that the background image intensity strongly depends on the presence of IR light sources other than the camera's flash, especially the sun during outdoor acquisitions.

In order to detect the appearance of a reflective vest in the input images, we first identify an initial set of potential interest points using a circular blob detector. We employ the computationally efficient STAR algorithm by Konolige et al. which is a speeded-up version of the Center Surround Extrema

a) b) c) d)

Fig. 3. Examples of image patterns resulting from the IR back reflection by the reflective vest markers during the acquisition of image I_f. Images taken in absence of any secondary IR light sources show bright reflectors in front of a dark background (a) and represent the ideal case. In contrast, images captured under exposure to sunlight are subject to higher background illumination (b–c), especially when the sun is directly facing the camera (d).

(CenSurE) feature detector [16]. The algorithm detects a set of raw high-intensity blob-like features $\mathcal{F}_{\text{raw}}$ in the image I_f taken with flash,

$$\mathcal{F}_{\text{raw}} = \left\{ f^{[i]} = \left\langle s^{[i]}, \boldsymbol{u}_f^{[i]} \right\rangle \mid i = 1, ..., N_f \right\} \quad (2)$$

with the feature scale s indicating the diameter of the circular blob and $\boldsymbol{u}_f$ the image coordinates of its center in image I_f.

The blob-like features in the set $\mathcal{F}_{\text{raw}}$ originate either from reflective material reflecting the IR flash or from another object that is illuminated by an external IR light source such as the sun (cf. Fig. 3b–d). In the latter case, the appearance of the features is very similar in the images I_f and I_{nf} as both were captured in short succession. In constrast, a clear intensity difference exists in the case of reflective material. We therefore aim at removing non-vest features by observing the intensity difference between I_f and I_{nf} in a neighborhood of the detected features.

To do so, we first need to identify the locations at which the features $f \in \mathcal{F}_{\text{raw}}$ appear in the image I_{nf}. That is, for a feature's location $\boldsymbol{u}_f$ in image I_f we seek its corresponding location $\boldsymbol{u}_{nf}$ in image I_{nf}. Especially under the influence of fast rotational camera motion the two locations can differ by several pixels. We employ a pyramidal implementation of the iterative Lucas-Kanade (LK) feature tracking method [17] to track $\boldsymbol{u}_{nf}$ for every feature $f \in \mathcal{F}_{\text{raw}}$. We submit all successfully tracked features to an intensity difference check $\epsilon(\boldsymbol{u}_f, \boldsymbol{u}_{nf})$ measuring the average intensity difference in a square neighborhood with side length corresponding to the feature scale s around the original and tracked feature locations.

Using the output of the feature tracker and the intensity difference check we split the initial set of features $\mathcal{F}_{\text{raw}}$ into two subsets $\mathcal{F}_{\text{reflex}}$ and $\mathcal{F}_{\text{non-reflex}}$ according to

$$\mathcal{F}_{\text{non-reflex}} = \{ f \in \mathcal{F}_{\text{raw}} \mid f \text{ tracked } \wedge\ \epsilon(\boldsymbol{u}_f, \boldsymbol{u}_{nf}) < \lambda \} \quad (3)$$

$$\mathcal{F}_{\text{reflex}} = \mathcal{F}_{\text{raw}} \setminus \mathcal{F}_{\text{non-reflex}} \quad (4)$$

where $\mathcal{F}_{\text{reflex}}$ is assumed to contain features originating from reflective material. There, we also explicitly include all features for which the tracker was unable to find any suitable match $\boldsymbol{u}_{nf}$, assuming that the failure to track a feature is due to very different intensity values that occur in the case of reflective material.

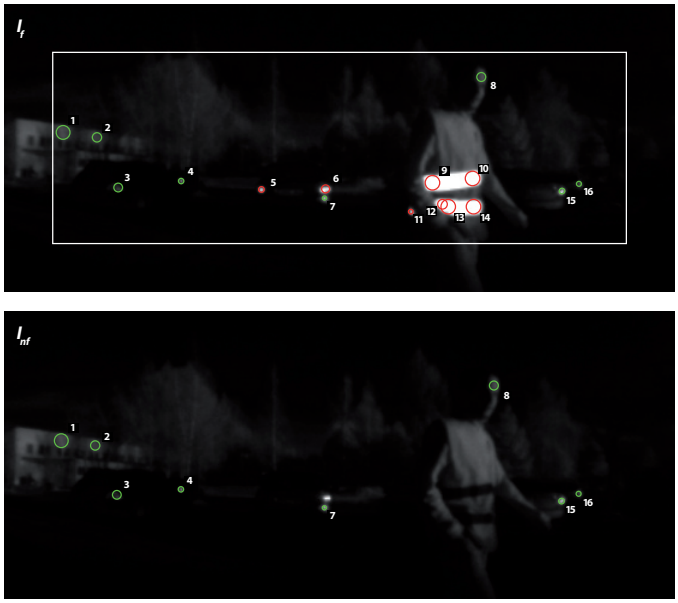


Fig. 4. Result of the feature detection, tracking and intensity difference check. Locations $\boldsymbol{u}_f$ where a high-intensity blob feature has been detected are indicated by a circle in image I_f taken with IR flash (above). The detection area in I_f is restricted to the white bounding box to assure that detected features are still in the camera's FOV when taking image I_{nf} (below), even under fast rotational movement. Features that are successfully tracked (1–4, 7-8, 15-16) are represented in green in image I_f and the corresponding tracked locations $\boldsymbol{u}_{nf}$ are indicated by a green circle in image I_{nf}. An intensity difference check reveals that all tracked features show very similar appearance in I_f and I_{nf}. Thus, they are collected in a set of non-reflective features $\mathcal{F}_{\text{non-reflex}}$. In contrast, features that failed to be tracked (5-6 and 9–14) are included in the set of reflection based features $\mathcal{F}_{\text{reflex}}$ (9–14 represent reflective vest markers and 5–6 reflective metallic surfaces on a car). This figure is best viewed in color.

Fig. 4 depicts the result of the feature detection, tracking and intensity difference check and illustrates that by constructing the feature set $\mathcal{F}_{\text{reflex}}$ using the described procedure described, the major part of raw features that do not correspond to reflective vest features are eliminated.

D. Feature Description and Classification

The set $\mathcal{F}_{\text{reflex}}$ contains features that originate from the reflection of the emitted IR flash on reflective material. To explicitely detect a reflective vest, it is important to distinguish between the reflective vest markers and other reflective objects such as metallic surfaces, windows or mirrors. Therefore, a supervised learning based binary classifier is included in the detection process, classifying all features $f \in \mathcal{F}_{\text{reflex}}$ into a set of vest features $\mathcal{F}_{\text{vest}}$ and a set of non-vest features $\mathcal{F}_{\text{non-vest}}$. The input to the classifier is a local image feature descriptor $\boldsymbol{r}$ computed from a square neighborhood of feature $f \in \mathcal{F}_{\text{reflex}}$ where the side length of the neighborhood corresponds to the feature scale s. Several state-of-the-art image feature descriptors including SURF [18], BRIEF [19] and the recent BRISK [20] descriptor have been evaluated in combination with the popular Support Vector Machine [21] and Random Forest [22] classifiers, where the latter have shown clearly and consistently superior performance.

E. Distance and Position Estimation

The same local feature descriptors $\boldsymbol{r}$ used to perform feature classification are further exploited to estimate the distance between the camera and the reflective vest using regression. The regressor model is again obtained using supervised learning where the learning algorithm is provided the ground-truth distance between the camera and the reflective vest.

Using the estimated distance $\hat{d}$ in combination with the feature's location $\boldsymbol{u}_f$ and the panoramic projection function $\boldsymbol{h}(\boldsymbol{u})$ we are able to obtain an estimate of the vest reflector's relative position in 3D space using:

$$\hat{\boldsymbol{p}} = \hat{d} \cdot \boldsymbol{h}(\boldsymbol{u}_f) \tag{5}$$

F. Vest Tracking

A reflective vest not only needs to be detected in the individual image pairs but has to be tracked over a sequence of input images. We therefore consider the scenario where image pairs $\boldsymbol{I}$ are repeatedly acquired and denote $\boldsymbol{I}_t$ the image pair acquired at time step $t \in \mathbb{Z}$. We further denote $\hat{\boldsymbol{p}}_t^{[i]}$ the estimated position corresponding to feature $f^{[i]} \in \mathcal{F}_{vest}$ at time t and introduce the set $\mathcal{P}_t$ of all position estimates obtained at the same time t, according to

$$\mathcal{P}_t = \left\{ \hat{\boldsymbol{p}}_t^{[i]} \mid i = 1, ..., N_{\mathcal{P}_t} \right\} \tag{6}$$

where $N_{\mathcal{P}_t}$ is the number of position estimates obtained at time t. $N_{\mathcal{P}_t}$ simply equals the size of the set $\mathcal{F}_{vest}$ at time t.

We aim at recursively estimating a state vector $\boldsymbol{s}_t$ comprising the position and speed relative to the camera by incorporating the single vest position estimates $\hat{\boldsymbol{p}}_t$. In addition to the position $\boldsymbol{p}_t = [x_t, y_t, z_t]^\top$ of a reflective vest at time t, the state also includes its velocity in the camera reference frame, denoted by the ensemble $\dot{\boldsymbol{p}}_t = [\dot{x}_t, \dot{y}_t, \dot{z}_t]^\top$:

$$\boldsymbol{s}_t = [\boldsymbol{p}_t \; \dot{\boldsymbol{p}}_t]^\top = [x_t, y_t, z_t, \dot{x}_t, \dot{y}_t, \dot{z}_t]^\top \tag{7}$$

The additional estimation of the velocity of an observed reflective vest allows to make a better prediction of the state transition from $\boldsymbol{s}_t$ to $\boldsymbol{s}_{t+1}$, as it is needed for the motion model.

To recursively estimate the latent state variable $\boldsymbol{s}_t$, we employ a particle filter in which the belief destribution over $\boldsymbol{s}_t$ is represented by a set of N_p particles,

$$\mathcal{S}_t = \left\{ \left\langle \boldsymbol{s}_t^{[k]}, w_t^{[k]} \right\rangle \mid k = 1, ..., N_p \right\} \tag{8}$$

with $\boldsymbol{s}_t^{[k]}$ denoting the k-th state hypothesis and $w_t^{[k]}$ the respective importance factor. Our implementation uses the standard sequential importance resampling algorithm [23] that sequentially incorporates the obtained vest position estimates $\hat{\boldsymbol{p}}$ by first producing a predictive particle set $\overline{\mathcal{S}}_t$ using the motion model $\overline{\boldsymbol{s}}_t^{[k]} = \boldsymbol{\psi}_{Motion}(\boldsymbol{s}_{t-1}^{[k]})$ before assigning each particle an importance factor according to the measurement model $w_t^{[k]} \sim p(\mathcal{P}_t | \overline{\boldsymbol{s}}_t^{[k]})$ and resampling according to the importance factors using a low variance resampler as proposed by [24]. An initial particle set $\mathcal{S}_0$ is generated by uniformly distributing the particles in the state space. Given the particle set $\mathcal{S}_t$ at time t, an estimate of the position of an observed person can be obtained using the weighted mean of the particle states.

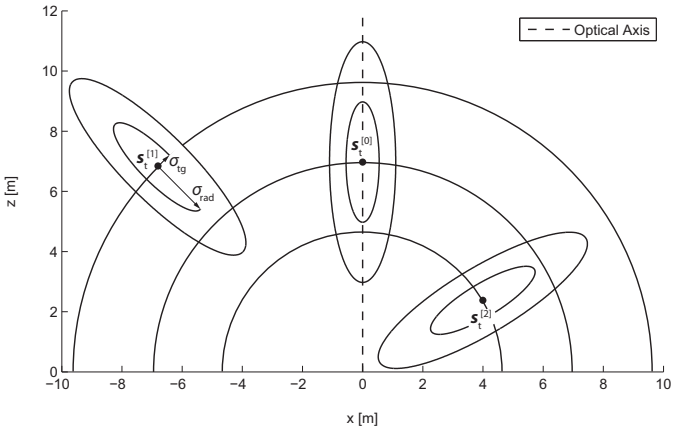


Fig. 5. 2D projection of the characteristic 3D measurement probability $p(\hat{\boldsymbol{p}}_t | \boldsymbol{s}_t)$ for three different example states $\boldsymbol{s}_t^{[0]}$, $\boldsymbol{s}_t^{[1]}$ and $\boldsymbol{s}_t^{[2]}$. The camera system is located at the origin. The measurement probability is modeled by a multivariate normal distribution with specific standard deviations σ_{rad} and σ_{tg} for the radial and tangential direction. The iso-lines show one and two standard-deviations.

1) Motion Model: The motion model $\boldsymbol{s}_t = \boldsymbol{\psi}_{Motion}(\boldsymbol{s}_{t-1})$ predicts the state transation from $\boldsymbol{s}_{t-1}$ to $\boldsymbol{s}_t$ using a description of the system dynamics. In our case, a change of the state $\boldsymbol{s}_t$ can be caused either by motion of the camera or by motion of the observed person. The possible movements of the observed person in an industrial environment are vast and include walking, running, turning on the spot, accelerating in any direction or performing abrupt turns. Furthermore, a person can move by means of a vehicle. Camera motion on the other hand can result from motion of the vehicle the camera is attached to. As everything is handled in the reference frame attached to the camera no decoupling is made into the two types of motion.

We employ a simple linear model assuming the target to move with constant speed between two time steps while modeling changes in speed and direction with the according process noise in the velocity vector $\dot{\boldsymbol{p}}_t = [\dot{x}_t, \dot{y}_t, \dot{z}_t]^\top$:

$$\boldsymbol{s}_{t+1} = \begin{bmatrix} \boldsymbol{I}_{\mathbf{3\times3}} & f_a^{-1} \cdot \boldsymbol{I}_{\mathbf{3\times3}} \\ \mathbf{0}_{\mathbf{3\times3}} & \boldsymbol{I}_{\mathbf{3\times3}} \end{bmatrix} \boldsymbol{s}_t + \begin{bmatrix} \mathbf{0}_{\mathbf{3\times1}} \\ \boldsymbol{w}_{\mathbf{3\times1}} \end{bmatrix} \tag{9}$$

with f_a the image pair acquisition rate and $\boldsymbol{w} = [w_x, w_y, w_z]^\top$ independent white noise processes of the form $\mathcal{N}(0, \sigma)$ for the different velocity components.

2) Measurement Model: The measurement model relates the set of observations $\mathcal{P}_t$ to the state vector $\boldsymbol{s}_t$ by the measurement probability $p(\hat{\boldsymbol{p}}_t | \boldsymbol{s}_t)$, describing the likelihood to make a single observation $\hat{\boldsymbol{p}}_t$ assuming that the state of the system is $\boldsymbol{s}_t$. Fig. 5 depicts the characteristic shape of the measurement probability $p(\hat{\boldsymbol{p}}_t | \boldsymbol{s}_t)$ in the x/z-plane for three different example states. Due to the processing scheme employed to obtain a position estimate $\hat{\boldsymbol{p}}_t$, the measurement uncertainty is different in radial and tangential direction and represented respectively by the standard deviations σ_{rad} and σ_{tg}. Uncertainty in radial direction mainly originates from

the estimation error committed by the distance regressor. In contrast, the variance in the detection of the tangential position arises from the fact that a reflective vest feature detected in the input images is not necessarily situated in the center of the reflective vest. Finally, measurement noise in the image material causes uncertainty in both directions as it influences the complete processing chain. Experimental results show that the values of σ_{rad} and σ_{tg} are relatively constant over the whole sensor range. The likelihood to make a single observation $\hat{\boldsymbol{p}}_t$, under the assumption of state $\boldsymbol{s}_t$, is then given by the multivariate Gaussian

$$p(\hat{\boldsymbol{p}}_t|\boldsymbol{s}_t) = \frac{1}{(2\pi)^{\frac{3}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\hat{\boldsymbol{p}}_t - \boldsymbol{p}_t)^\top \Sigma^{-1}(\hat{\boldsymbol{p}}_t - \boldsymbol{p}_t)\right) \tag{10}$$

where the covariance matrix Σ is obtained using

$$\Sigma = R_y(\theta)^\top R_x(\phi)^\top \Sigma_0 R_x(\phi) R_y(\theta) \tag{11}$$

with θ and ϕ the azimuth and altitude angles of position $\boldsymbol{p}_t$, R_x and R_y the rotation matrices around the x- and y axes respectively and Σ_0 the covariance matrix corresponding to states $\boldsymbol{s}_t$ situated on the camera's optical axis (cf. state $\boldsymbol{s}_t^{[0]}$ in Fig. 5), given by:

$$\Sigma_0 = \begin{bmatrix} \sigma_{tg}^2 & 0 & 0 \\ 0 & \sigma_{tg}^2 & 0 \\ 0 & 0 & \sigma_{rad}^2 \end{bmatrix} \tag{12}$$

Finally, the complete measurement probability defines the likelihood to make the full set of observations $\mathcal{P}_t$, given the state $\boldsymbol{s}_t$. Under the assumption that the noise in the individual measurements $\hat{\boldsymbol{p}}_t^{[i]}$ is independent, it is obtained by the product of the individual measurement likelihoods $p(\hat{\boldsymbol{p}}_t|\boldsymbol{s}_t)$:

$$p(\mathcal{P}_t|\boldsymbol{s}_t) = \prod_{i=1}^{N_{\mathcal{P}_t}} p(\hat{\boldsymbol{p}}_t^{[i]}|\boldsymbol{s}_t) \tag{13}$$

IV. Results

Our reflective vest detection and tracking system has been evaluated in four different test scenarios as listed in Table I. Evaluation has been performed on flat ground in order to facilitate the extraction of ground-truth data. The detection range with the current hardware setup is limited to roughly 10 meters, due to the illumination intensity and the camera resolution. A sensor unit consisting of the camera system and a 2D laser range scanner (SICK LMS-200) was used for the data acquisition. The sensor unit was mounted at a height of approximately 1.5 meter on a mobile platform with four hard rubber wheels.

Several training and validation data sets were acquired for each of the four scenarios by simultaneously recording the raw camera images and the 2D laser readings. Fig. 6a illustrates the characteristic appearance of the image material acquired in the different data sets. During the acquisition of all sets, a single person wearing a reflective vest according to Fig. 1a was constantly moving in the field of view of the camera in a distance range up to 10 meters. The mobile platform was kept in constant motion at a speed of approximately 0.5 m/s. Both a Random Forest classifier and regressor were trained on 50k extracted image descriptors with ground-truth distance labels obtained by the laser range scanner and manually assigned class labels.

TABLE I
Test scenarios

Scenario	Environment
1	Indoors, warehouse-like environment
2	Outdoors, car parking area, clear weather conditions
3	Outdoors, car parking area, direct sunshine into the camera
4	Outdoors, storage yard, light snowfall

The experimental results are summarized in Fig. 6b–e. The performance of the algorithm's major processing steps were assessed individually, namely the Random Forest classifier in Fig. 6b, the Random Forest distance regressor in Fig. 6c and the particle-filter based tracking in Fig. 6d–e.

V. Discussion

The degree of complexity of the different test scenarios in terms of people detection and tracking varies significantly as visualized in Fig. 6a. Scenario I represents the ideal case for successful vest tracking, as it is situated in an indoor environment with no IR light source present other than the camera's flash and no reflective objects other than the reflectors of the vest. Consequently, feature detections exclusively originate from the vest and the visibility is not limited by any disturbing factors. Scenario II is situated outdoors in clear weather conditions and the acquired images thus appear slightly brighter, due to the background illumination caused by the IR portion in the sunlight. The image material further contains many other reflective objects such as metallic surfaces and windows. In Scenario 3 and 4, the visibility is seriously restricted either by direct sunshine into the camera, which produces numerous lens artifacts (scenario III), or by snowfall (scenario IV) and the detection range is reduced to roughly 8 meters.

In terms of feature classification and distance regression, all the three evaluated feature descriptors yield fairly similar results with small differences in individual scenarios and at individual distance ranges (cf. Fig. 6b-c). The rotation invariance of the SURF and BRISK descriptors does not lead to a clear advantage over BRIEF. This can be understood by the fact that the observed patterns themselves show already a high degree of rotational symmetry. For scenarios I and II the accuracy of the distance estimation is relatively stable over the entire distance range considered in the evaluation and the accuracy with our single camera system is within a decimeter range. A slight tendency to overestimate the distance at short ranges and to underestimate it at higher ranges can be observed which is likely due to the fact that the distance has a lower bound of zero and no training data was provided

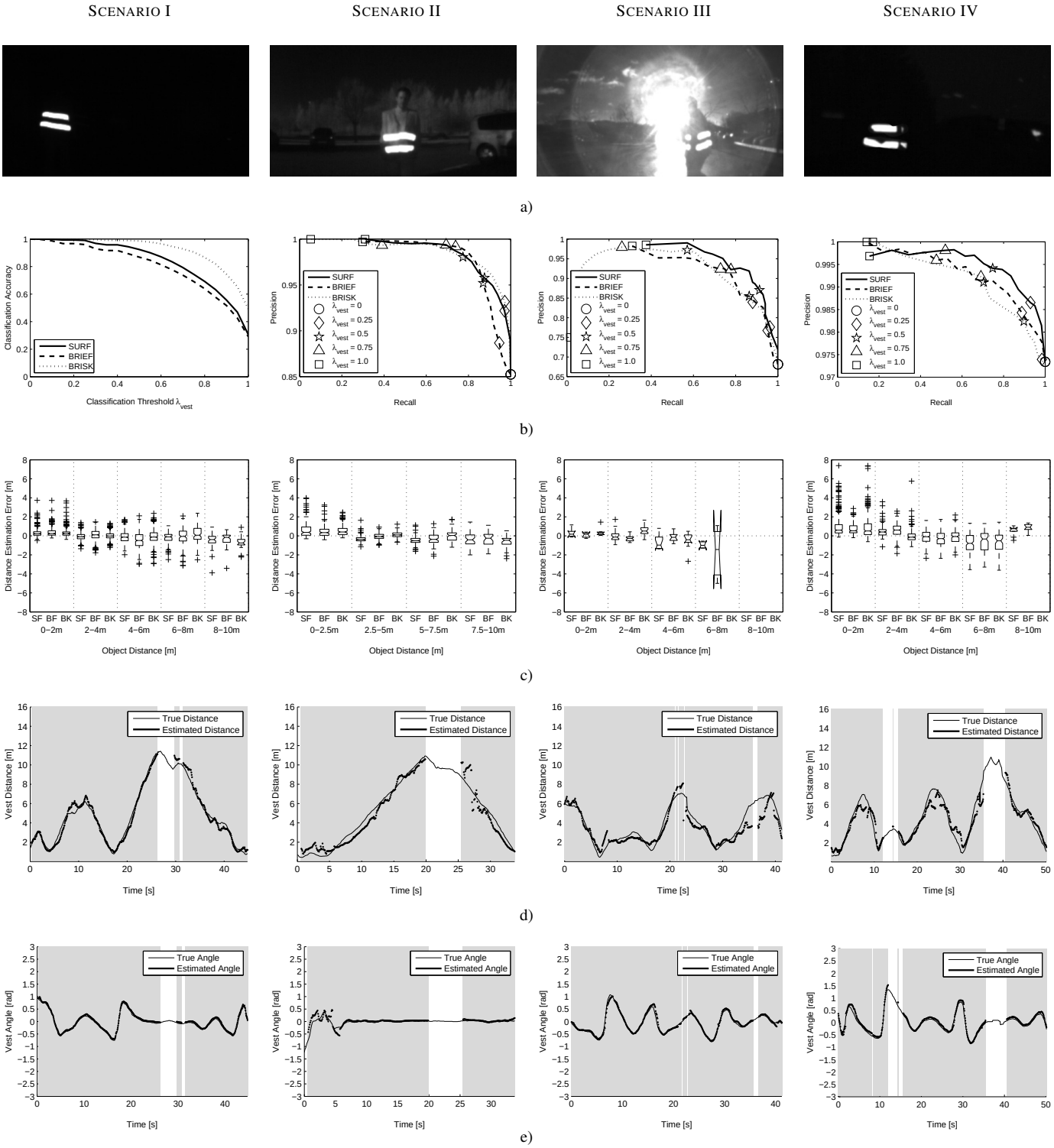


Fig. 6. Experimental results of the reflective vest detection algorithm for the test scenarios I–IV. **a)** The image illustrates the characteristic appearance of the images featured in the respective scenario, with scenario I offering the least and scenario III the most challenging conditions. **b)** Classification accuracy (scenario I) and precision-recall curves (scenario II–IV) describing the performance of the Random Forest classifier in classifying the feature set $\mathcal{F}_{reflex}$ into a set of vest features $\mathcal{F}_{vest}$ and a set of non-vest features $\mathcal{F}_{non-vest}$ based on the feature descriptors SURF, BRIEF and BRISK with varying classification threshold λ_{vest}. **c)** Boxplot of the regressor's per-feature distance estimation error at different distances ranges. The indications SF (SURF), BF (BRIEF) and BK (BRISK) specify the image descriptor on which the estimation is based. **d-e)** Temporal evolution of the ground-truth and estimated (filtered) distance and angle using classification and regression based on the SURF descriptor. Gray background indicates time periods during which the vest is considered as tracked.

with distances higher than 10 meters. The plots also report sporadic but large outliers indicating a distance estimation error of several meters. Further investigation revealed that most of the outliers originate from misclassification errors, namely cases where non-vest features are classified as vest features (false positives).

The tracking results presented in Fig. 6d–e show that the target is consistently tracked over large parts of scenarios I and II and over considerable parts of scenarios III and IV even though using per-feature position estimates with high error. The filtering effect becomes very clear, especially in the first two scenarios, where from position estimates with considerable outliers in the meter range, a position estimate is obtained whose error lies in the decimeter range for big parts of the image sequence.

VI. Conclusion

In this paper we presented a novel approach for detecting and tracking operators on industrial worksites. In contrast to existing people detectors we exploit the fact that industrial workers wear a reflective vest to enhance their visibility. Our approach uses a single-camera setup equipped with IR filter and flash to identify reflective vest markers in the input images using IR backscattering. We have shown that detecting humans through active illumination and detection of reflective vest markers has several key advantages over conventional vision-based methods. Using off-the-shelf hardware that costs a mere of €500, our camera system performs in broad daylight as well as in complete darkness and can be applied both indoors and outdoors in various weather conditions. The experiments have shown that the system performs well in detecting a single person up to 10 meters distance in an indoor warehouse-like environment as well as outdoors under direct exposure to the sun and in the presence of reflective objects other than the vest. Even though the performance is slightly decreased, the system still performs well in extreme conditions, namely when the sun is directly facing the camera.

VII. Future Work

Future work includes the extension of the described people tracking system towards robust multiple-person tracking. The underlying hardware will be extended to form a stereo camera unit that can obtain depth measurements not only using the learning based approach presented in this paper, but also through triangulation. For practical industrial applications it is further desirable to extend the detection range to 20 meters around the vehicle.

A comprehensive long-term evaluation of the system in different real-world industrial environments will be carried out. The evaluation will include an analysis of situations and conditions in which the presented tracking approach most clearly outperforms state-of-the art vision-based people trackers that work without the restriction of people wearing a reflective safety vest.

References

[1] H. Andreasson, A. Bouguerra, T. Stoyanov, M. Magnusson, and A. Lilienthal, "Vision-based people detection utilizing reflective vests for autonomous transportation applications." *IROS Workshop on Metrics and Methodologies for Autonomous Robot Teams in Logistics (MMART-LOG)*, 2011.

[2] R. Mosberger and H. Andreasson, "Estimating the 3d position of humans wearing a reflective vest using a single camera system," in *Proceedings of the International Conference on Field and Service Robotics (FSR)*, 2012.

[3] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems." *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 7, pp. 1239–58, 2010.

[4] B. S. P. Dollar, C. Wojek and P. Perona, "Pedestrian detection: A benchmark," in *CVPR*, June 2009.

[5] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 2179–2195, 2009.

[6] T. Gandhi and M. M. Trivedi, "Pedestrian protection systems: Issues, survey, and challenges," *Trans. Intell. Transport. Sys.*, vol. 8, no. 3, pp. 413–430, 2007.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *In CVPR*, 2005, pp. 886–893.

[8] D. M. Gavrila, J. Giebel, and S. Munder, "Vision-based pedestrian detection: The protector system," in *In IEEE Intelligent Vehicles Symposium*, 2004, pp. 13–18.

[9] J. Chung, N. Kim, G. J. Kim, and C.-M. Park, "Postrack: A low cost real-time motion tracking system for vr application," in *In International conference on Virtual Systems and MultilMedia*. IEEE, 2001, pp. 383–392.

[10] M. Ribo, A. Pinz, and A. Fuhrmann, "A new optical tracking system for virtual and augmented reality applications," in *IEEE Instrumentation and Measurement Technology Conference*, May 2001, pp. 1932–1936.

[11] Y. Nakazato, M. Kanbara, and N. Yokoya, "A localization system using invisible retro-reflective markers." in *MVA*, 2005, pp. 140–143.

[12] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *In NIPS 18*. MIT Press, 2005.

[13] J. Michels, A. Saxena, and A. Y. Ng, "High speed obstacle avoidance using monocular vision and reinforcement learning," in *In ICML*, 2005, pp. 593–600.

[14] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A flexible technique for accurate omnidirectional camera calibration and structure from motion," in *Proc. of The IEEE International Conference on Computer Vision Systems (ICVS)*, 2006.

[15] R. Mosberger, "Vision-based tracking of humans wearing a reflective vest using a single camera system," Master's thesis, École Polytechnique Fédérale de Lausanne EPFL, Lausanne, Switzerland, 2012.

[16] M. Agrawal, K. Konolige, and M. R. Blas, "Censure: Center surround extremas for realtime feature detection and matching," in *ECCV (4)*, ser. Lecture Notes in Computer Science, D. A. Forsyth, P. H. S. Torr, and A. Zisserman, Eds., vol. 5305. Springer, 2008, pp. 102–115.

[17] J.-Y. Bouguet, "Pyramidal implementation of the lucas kanade feature tracker description of the algorithm," 2000.

[18] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding (CVIU)*, vol. 110, pp. 346–359, 2008.

[19] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features." in *ECCV (4)*, ser. Lecture Notes in Computer Science, K. Daniilidis, P. Maragos, and N. Paragios, Eds., vol. 6314. Springer, 2010, pp. 778–792.

[20] S. Leutenegger, M. Chli, and R. Siegwart, "BRISK: Binary Robust Invariant Scalable Keypoints," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011.

[21] C. Cortes and V. Vapnik, "Support-vector networks," in *Machine Learning*, 1995, pp. 273–297.

[22] L. Breiman, "Random forests." *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[23] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation," *Radar and Signal Processing, IEEE Proceedings F*, vol. 140, no. 2, pp. 107–113, 1993.

[24] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005.