Distraction Suppression for Vision-Based Pose Estimation at City Scales

Colin McManus, Winston Churchill, Ashley Napier, Ben Davis, and Paul Newman

***Abstract*— This paper is concerned with the problem of egomotion estimation in highly dynamic, heavily cluttered urban environments over long periods of time. This is a challenging problem for vision-based systems because extreme scene movement caused by dynamic objects (e.g., enormous buses) can result in erroneous motion estimates. We describe two methods that combine 3D scene priors with vision sensors to generate *background-likelihood images*, which act as probability masks for objects that are not part of the scene prior. This results in a system that is able to cope with extreme scene motion, even when most of the image is obscured. We present results on real data collected in central London during rush hour and demonstrate the benefits of our techniques on a core navigation system — visual odometry.**

I. INTRODUCTION

For vision-based navigation systems, operating in highly dynamic environments is a challenging problem as extreme scene motion can degrade standard outlier rejection schemes and result in erroneous motion estimates. In this paper, we approach the problem of pose estimation in heavily populated urban environments by leveraging knowledge of prior 3D structure for distraction suppression in images. In other words, given prior knowledge of how the world "should look", our system is able to focus its attention on just the static parts of the scene for motion estimation, even in situations where most of the image is completely obscured by dynamic objects (see Figure 1 for an example).

Although one may approach this problem with a trained detector/tracking system (e.g., [1], [2], [3]), these techniques require a great deal of time to train, are challenging to implement, and require knowledge of all of the various distraction classes. In contrast, we present two straightforward and effective vision-based methods that exploit prior 3D structure to generate *background-likelihood images*, which effectively mask ephemeral objects of any type.

Thus, we are not specifically interested in object detection per se, but rather, *scene relevance* — what should we be focusing on in the scene, given that we have prior knowledge of its structure and simply wish to localise and perform egomotion estimation. We present results on kilometres of data collected in busy urban environments, demonstrating how these techniques can improve the robustness of visual odometry (VO).

II. RELATED WORK

In the area of road-vehicle navigation, leveraging prior surveys to improve motion estimation is a common approach.

Mobile Robotics Group, University of Oxford, Oxford, England; {colin,winston,ashley,ben,pnewman}@robots.ox.ac.uk

(a) Image taken in central London during the Olympics. Large parts of the scene are occupied by dynamic objects, which can distract and impede egomotion estimation. We present two techniques that leverage knowledge of prior structure to enable robust pose estimation, even in cases where most of the scene is moving.

(b) Using knowledge of prior 3D structure, we can generate probability masks that indicate which regions in the image are likely to belong to the static background (white). These masks are used in our front-end visual odometry pipeline to improve pose estimation in the presence of significant scene motion.

Fig. 1. This paper presents two methods that exploit knowledge of prior 3D structure to enable accurate pose estimation in heavily cluttered, highly dynamic urban environments. Having driven a route at least once, we are able to leverage prior information to suppress distracting objects in the image and focus on just the static parts of the scene, which is represented as a *background-likelihood image* (see 1(b)). This background-likelihood image is used to mask ephemeral objects, thereby enabling accurate feature matching, even in situations where most of the scene is moving.

Numerous techniques exist for both vision and laser, and include: (i) combining vision with aerial images [4], [5], synthetic overhead images [6], or prior visual experiences [7], (ii) combining 2D laser rangefinders with 2D priors [8] (iii) combining 2D laser laser rangefinders with 3D priors [9], (iv) combining 3D laser rangefinders with 3D priors [10], and (v) combining vision with 3D priors [11]. In our work, we consider the latter case of using vision sensors in conjunction with a prior 3D survey generated from a laser scanner. Our goal is to identify areas in an image that have a high likelihood of belonging to the static background, even if 90% of the image is obscured by dynamic objects. These background-likelihood scores are used in our front-end VO system to mask features detected on ephemeral objects and thus, improve outlier rejection in our VO system.

The methods described in this paper rely on the idea of background subtraction, which have traditionally been applied to static camera systems for surveillance operations [12], [13]. The typical approach is to learn a statistical model of the background (e.g., Mixture of Gaussians (MoG) for each pixel in the image) and compare current views with the background model to identify large discrepancies (see Piccardi et al. [14] for a review of the various statistical models that have been used).

Various techniques for moving systems have been proposed, such as estimating a planar homography and applying

Fig. 2. A laser-generated 3D point cloud in central London. This data was collected with a mobile sensing suite mounted on a commercial vehicle, quipped with a stereo camera, planar laser rangefinder, and GPS. At runtime our system uses the stereo input to compare the observed structure of the world with the prior structure to identify ephemeral objects.

standard statistical techniques for foreground/background detection [15], [16]; however, these methods are only valid under rotational motion. Plane-parallax constrains were introduced to compensate for rotational and translational motions [17], [18], but assume that a dominant 3D plane is present. Sheikh et al. [19] offer a different solution that estimates a background trajectory based on a rank constraint for a sequence of tracked point trajectories. However, these methods have only demonstrated results under modest displacements and not in outdoor settings with fast-moving vehicles in cluttered environments. Additionally, these methods do not combine two different sensing modalities.

Taneja et al. [20] presented an offline monocular-based technique for detecting and updating changes in a prior 3D model. Their method uses prior structure to reproject pixels from the current camera frame into a collection of neighbouring frames to identify geometric inconsistencies. The problem is then formulated as the minimisation of a Gibbs energy function to find the optimal labelling of their voxelised prior (i.e., changed or unchanged). Similar to this work, we use prior 3D structure to identify regions of change in camera images, but take two very different approaches that attempt to account for uncertainties resulting from localisation errors. Furthermore, we demonstrate how to generate background-likelihood images and integrate them into a visual odometry pipeline.

III. SYSTEM OVERVIEW

Our system operates under the requirement that the environment in which we wish to operate has been pre-mapped by a survey vehicle equipped with high-quality 3D laser sensors, cameras, and a INS. More specifically, we assume that dense 3D point-clouds and stereo imagery of the environment will be available (see Figure 2). Furthermore, we assume that these point clouds are free of most/all ephemeral objects.

At a high-level, our system works as follows. At runtime, we match live stereo images against prior visual experiences (i.e., visually distinct image sequences) using an Experience-Based Navigation (EBN) system [7]. Since each prior visual experience has an associated 3D point-cloud, we are able to synthesise depth images from estimated camera poses in this 3D prior. These synthetic depth images are then used to compare the current structure of the scene (given by our live imagery) with the static structure of scene (given by the prior) to identify large discrepancies. This provides us with a clean segmentation of the image into foreground and background elements, without the need for an object detection system.

We present two vision-based techniques for a moving platform that exploit prior 3D structure from a laser-generated point cloud in order to detect ephemeral objects for distraction suppression. We show how to produce background-likelihood images that provide pixel-wise likelihood scores for belonging to the background. These likelihood images are used in the front-end of our VO pipeline to reject candidate features on ephemeral objects and improve pose estimation, which is of great importance for autonomous vehicles operating in urban environments.

The proceeding sections will describe in more detail how our system generates synthetic camera views in the 3D scene prior, how this information is used to generate background-likelihood images, and how these likelihood scores are incorporated in our VO pipeline.

A. Generating Synthetic Camera Views

At runtime, our EBN system provides an estimate of the pose of the vehicle, denoted by a 6×1 column vector $\mathbf{x}$, within the 3D scene prior. Using this estimated pose, we reproject all of the points from the 3D scene prior into the camera frame, producing a synthetic depth image (see Figure 3). For reasons of efficiency, we restrict the size of the 3D scene prior by using a sliding window about the estimated camera position (we have used a window of 40 m in our experiments). Thus, for every pixel, i, in the image, we compute the estimated depth, z_i, in the local map according to the localisation estimate,

$$z_i = z_i(\mathbf{x} + \delta\mathbf{x}), \quad \delta\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{P}_x), \tag{1}$$

where $\delta\mathbf{x}$ is normally distributed noise given by covariance $\mathbf{P}_x$, which represents our localisation and calibration uncertainty.

Due to the sparsity and sub-pixel values of the reprojections in the image, we perform bilinear interpolation and then apply a median filter for smoothing. We only perform interpolations on pixels that are within a specified threshold of their reprojected neighbours. Note that as we have preprocessed the prior to remove ephemeral objects, this depth image contains only the static/invariant components of the scene.

B. Disparity-Based Distraction Suppression

As the vehicle to be localised has a stereo camera, we can, online, perform dense stereo to generate a live disparity image [21]. Using the background depth image from III-A, we can also generate a synthetic disparity image containing only the background. Thus, assuming that the estimate of the camera pose used to generate the synthetic prior is reasonably accurate, any discrepancies between the real and

(a) Camera image of the scene for reference.

(b) 3D scene prior coloured with corresponding laser intensity values.

(c) Reprojected laser-intensity image at the estimated camera pose in the prior.

(d) Reprojected depth image, where lighter colours represent larger depths.

Fig. 3. Illustration of generating a synthetic depth image. Using the estimated camera pose in the point cloud, all points within a local window (e.g., 40 m window) are reprojected into the image plane. As these reprojections fall within sub-pixel values, bilinear interpolation is performed with neighbouring points, provided they are within a closeness threshold.

synthetic disparity images represent ephemeral objects in the live stream (see Figures 4(a), 4(b), and 4(c)).

Although it is tempting to simply take the difference between the disparity images, there are two problems with this approach. Firstly, we note that calibration and localisation errors can lead to large disagreements in the foreground because of the inverse relationship between depth and disparity (i.e., noise on smaller depth values will produce large noise in disparity; see Figure 4(d)). Secondly, disparity differences for distant objects will naturally be smaller, meaning that we need some way of amplifying these weaker signals. By accounting for the uncertainties in generating the synthetic depth images, it turns out, that we are able to address both of these issues.

It is important to note that because we only have a single 3D prior as our background, we are unable to learn a statistical model for the background as is done in most background subtraction methods (i.e., we only have a single sample of the prior). Additionally, since we are using two completely different sensor modalities with different noise models, book-keeping of these uncertainties is important.

We therefore take a probabilistic approach and weight the disparity differences by their associated measurement uncertainties. For every pixel, i, in the image, we define a disparity measurement from the dense-stereo algorithm, d^c, and synthetic depth image, d^s, as follows,

$$d_i^c := \bar{d}_i^c + \delta d_i^c, \qquad \delta d_i^c \sim \mathcal{N}(0, \sigma_{d_i^c}^2), \tag{2}$$

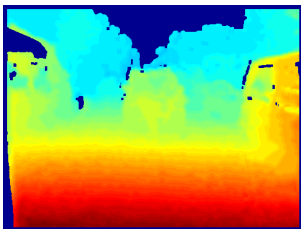
$$d_i^s := \frac{fb}{z_i^s(\mathbf{x} + \delta\mathbf{x})}, \qquad \delta\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{P}_x), \tag{3}$$

where δd_i^c is normally distributed pixel noise with standard deviation $\sigma_{d_i^c}^2$, $\{f, b\}$ are the intrinsic focal length and baseline, $z_i^s(\cdot)$ is the synthetic depth produced by a map-localisation estimate, $\mathbf{x}$, with normally distributed noise given by the covariance matrix $\mathbf{P}_x$. Dropping the pixel subscript for convenience, we now define a disparity difference measurement as,

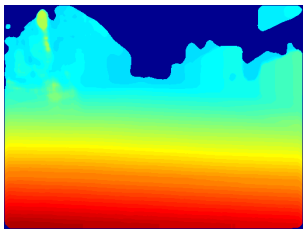
$$e_d := d^c - d^s \approx \underbrace{\bar{d}^c - \bar{d}^s}_{=:\bar{e}_d} + \underbrace{\delta d^c + \frac{fb}{(\bar{z}^s)^2}\left(\frac{\partial z^s}{\partial \mathbf{x}}\right)\delta\mathbf{x}}_{=:\delta e_d}, \tag{4}$$

where $\bar{z}^s := z^s(\bar{\mathbf{x}})$, $\bar{d}^s := fb/\bar{z}^s$, and we have performed a first-order taylor series expansion on the inverse depth term.

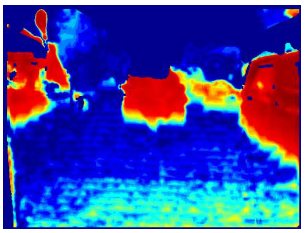
(a) Camera image for reference.



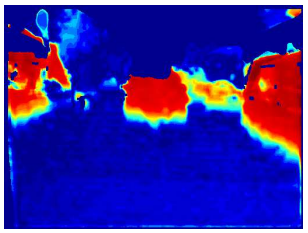
(b) Stereo disparity image using the method of Geiger et al. [21].



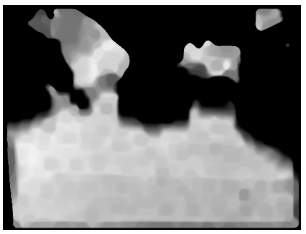
(c) Synthetic disparity image generated from the 3D scene prior.



(d) Disparity difference image (i.e., the absolute value of the difference between 4(b) and 4(c)).



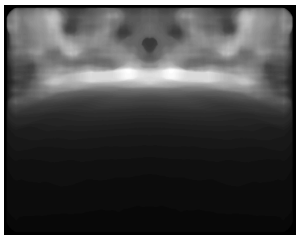
(e) Uncertainty-weighted disparity difference image after applying a max filter to amplify the signal and a low-pass filter for smoothing. Note that the signals in the near field that are present in Figure 4(d) are significantly down weighted when taking the uncertainty into account.



(f) Background-likelihood image, where black represents a likelihood of 0 and white represents a likelihood of 1. This image is used to weight the feature detection scores in the front-end of our VO pipeline.

Fig. 4. Generating a disparity-based background-likelihood image. Beginning at the top, 4(b) shows the true disparity image captured from a live video stream, while 4(c) shows the synthetic disparity image generated by using the 3D scene prior (see Figure 3). Since this scene prior is absent of dynamic objects, there is a clear visual dissimilarity between the true and synthetic disparity images. Taking the difference of these images and weighting by the uncertainties resulting from localisation errors, we obtain a clean segmentation of the foreground and background elements 4(e), allowing us to create a background-likelihood image 4(f) to be used in our VO pipeline.

(a) Representative Jacobian image given by Equation (6) (i.e., evaluating $\partial z^s/\partial \mathbf{x}$ for each pixel). Camera image provided for reference.



(b) An average depth-Jacobian image produced by averaging over 500 depth-Jacobian images.

Fig. 5. An illustration of a representative depth-Jacobian image 5(a) and the average depth-Jacobian image 5(b). Light colours represent larger sensitivities to pose changes.

The associated measurement noise is given by the following,

$$\begin{aligned}\sigma^2_{e_d} &:= \mathrm{E}(\delta e_d \delta e_d^T) \\ &= \sigma^2_{d^c} + \frac{(fb)^2}{(\bar{z}^s)^4}\left(\frac{\partial z^s}{\partial \mathbf{x}}\right)\mathbf{P}_x\left(\frac{\partial z^s}{\partial \mathbf{x}}\right)^T. \end{aligned} \quad (5)$$

Note that the Jacobian, $\partial z^s/\partial \mathbf{x}$, represents the change in depth that occurs given small perturbations of the vehicle's pose. At present, we have no efficient means of computing this quantity. Ideally, we wish to move towards creating an implicit surface model of our environment offline, which could allow us to compute these Jacobians analytically. Unfortunately, this option is not available and numerical techniques would be too slow. As such, we use the following approximation. To begin, let us define,

$$Z_x := \sqrt{\left(\frac{\partial z^s}{\partial \mathbf{x}}\right)\mathbf{P}_x\left(\frac{\partial z^s}{\partial \mathbf{x}}\right)^T}, \quad (6)$$

which provides an estimate of the depth change at a particular pixel location, given the localisation uncertainty. Figure 5(a) shows an example image where Z_x has been numerically computed for each pixel location. Examining this image, it becomes clear that the regions with the most uncertainty occur at large depths (due to the oblique angle between the plane and the optical axis), as well as non-smooth surfaces (e.g., trees). To approximate this Jacobian, we precomputed an *average depth-Jacobian image* by averaging over 500 keyframes from a separate dataset. This depth-Jacboian image is shown in Figure 5(b). It should be noted that this approximation works well because we are operating in urban environments, where the structure of the scene remains relatively constant. Denoting this approximation as $\hat{Z}_x$, we have

$$\sigma^2_{e_d} \approx \sigma^2_{d^c} + \frac{(fb)^2}{(\bar{z}^s)^4}\hat{Z}^2_x, \quad (7)$$

allowing us to define our Mahalanobis disparity difference measurement as,

$$\tilde{e}_d := \sqrt{e_d^2/2\sigma^2_{e_d}}. \quad (8)$$

Figure 4(e) shows the result of applying our measurement uncertainty to get the uncertainty-weighted disparity difference. The effect is that errors in the near field are down-weighted, which naturally brings out differences with objects that are farther away (i.e., the weaker signals for distant objects appear stronger since the foreground noise is reduced). The background-likelihood image is then obtained by thresholding the uncertainty-weighted disparity (i.e., set $\tilde{e}_d > \tau_d = \tau_d$ for all pixels), using a max-filter to amplify the disparity disagreements, scaling the image between $[0, 1]$, and taking the complement (see Figure 4(f)).

(a) Camera image of the scene for reference. Note that all vehicles in this scene are in motion.

(b) Synthetic camera image generated by reprojecting the coloured point cloud into the image plane. Large residuals with the true camera image (see left) are highlighted in yellow.

Fig. 6. Illustrating the generation of a synthetic camera image based on prior 3D structure. The motion estimate between time t_{k-1} and t_k is applied and the coloured points are reprojected into the current camera frame. In this example, the camera hardly moved between frames, meaning that most points reprojected in roughly the same place in the image. However, as the vehicle on the right was actually in motion, there is a large discrepancy between the synthetic camera image and the true image.

C. Flow-Based Distraction Suppression

This section presents an alternative method for generating a background-likelihood image, which relies on optical flow instead of dense stereo, making it applicable to monocular-based systems. To create a synthetic optical flow image at time t_k, the synthetic depth image and camera image at t_{k-1} are used to create a coloured point cloud. The motion estimate between times t_{k-1} and t_k, denoted by the 4×4 SE(3) transformation $\mathbf{T}_{k,k-1}$, is applied and the coloured point cloud is reprojected into the estimated camera pose at time t_k to create a synthetic camera image (see Figure 6). Regions without any data (i.e., pixel locations where the nearest reprojected point is beyond a certain distance) are filled in with the intensity values from the true camera image. This is necessary in order to ensure that we can create a full image without missing data, otherwise the optical flow algorithm will produce an extremely noisy result. After reprojecting the coloured point cloud and filling in missing regions, we apply bilinear interpolation, followed by a Gaussian low-pass filter to smooth the image.

Once we have generated a synthetic intensity image at time t_k, we use the method of Liu [22] to compute the expected optical flow (i.e., between the true image at t_{k-1} and the synthetic image at t_k) and the true optical flow (i.e., between the true image at t_{k-1} and the true image at t_k); see Figure 7(b) and 7(c) for an example. We define the true optical flow measurement, f^c, and synthetic optical flow measurement,

f^s, for pixel i as,

$$f_i^c := \bar{f}_i^c + \delta f_i^c, \qquad \delta f^c \sim \mathcal{N}(0, \sigma^2_{f_i^c}), \quad (9)$$

$$f_i^s := f_i^s\left(z^s(\mathbf{x}+\delta\mathbf{x})\right), \qquad \delta\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{P}_x). \quad (10)$$

In a similar fashion as before, and dropping the subscript, we define a difference measurement and its associated uncertainty as,

$$e_f := f^c - f^s \approx \underbrace{\bar{f}^c - \bar{f}^l}_{=:\bar{e}_f} + \underbrace{\delta f^c - \frac{\partial f^s}{\partial z^s}\left(\frac{\partial z^s}{\partial \mathbf{x}}\right)\delta\mathbf{x}}_{=:\delta e_f}, \quad (11)$$

$$\sigma^2_{e_f} := \sigma^2_{f^c} + \left(\frac{\partial f^s}{\partial z^s}\right)^2 \left(\frac{\partial z^s}{\partial \mathbf{x}}\right)\mathbf{P}_x\left(\frac{\partial z^s}{\partial \mathbf{x}}\right)^T. \quad (12)$$

Unfortunately, this derivation introduces another Jacobian term, $\partial f^s/\partial z^s$, which represents changes in optical flow due to changes in depth. This Jacobian term is far from smooth and we have no clear means of computing it at present; it involves reprojecting coloured points, interpolating a grayscale image, and running it through an optical flow algorithm that computes local spatial and temporal derivatives. We therefore adopt an alternative solution based on the intuition that scaling 2D flow fields by their associated depth approximates the 3D velocity [23]. In our case, we scale the difference between the expected and observed flow by the expected depth to amplify large differences:

$$\tilde{e}_f := e_f z^s. \quad (13)$$

Although this approach is slightly more crude in that we are not explicitly accounting for uncertainties in the flow difference, we found this to work well in practice. Figures 7(e) shows the depth-weighted flow difference and Figure 7(f) shows the resulting background-likelihood image, which is formed in the same manner as described earlier. The next subsection will discuss how we use these background-likelihood images in our front-end VO pipeline.

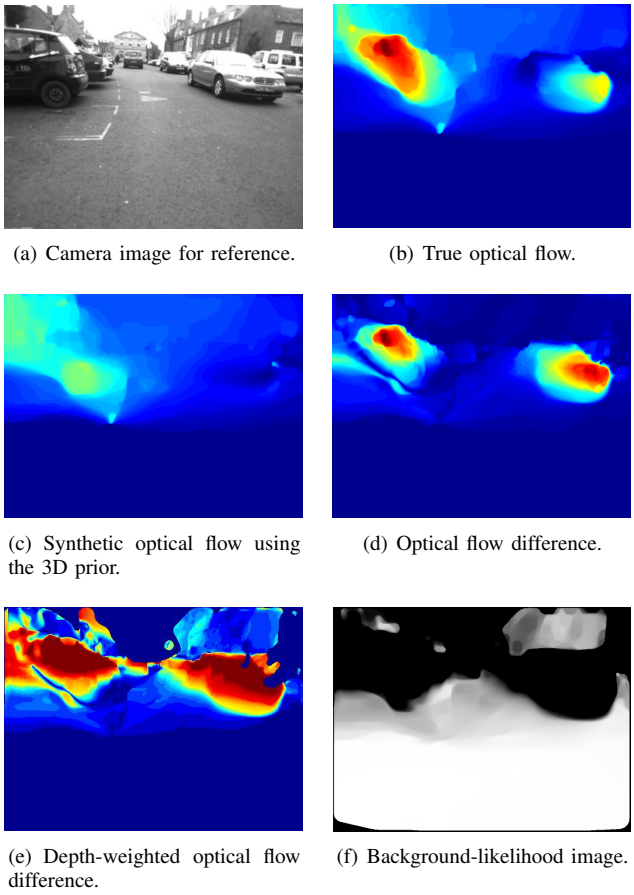
D. Feature Score Reweighting

For feature extraction in our VO front-end, we use the FAST corner detector [24] with a low threshold to obtain thousands of candidate features. As it would be intractable to perform feature matching on all of these candidates, our system takes the top N features, ranked by their corner score, s_i. In order to ensure that the features are well distributed spatially, the image is partitioned into a number of quadrants and the desired number of features N, is divided equally among each quadrant.

The background-likelihood images are then used to re-weight each corner score by looking up the closest likelihood weight, b_i, and re-weighting according to the following

$$\tilde{s}_i = \begin{cases} 0 & \text{if } b_i < \tau_b \\ b_i s_i & \text{otherwise} \end{cases},$$

where τ_b is a threshold for the minimum required likelihood. This threshold is needed because our system will always seek to find a minimum number of features in each quadrant, provided that the corner scores are above zero. This means that there could be a quadrant with very low likelihood scores (close to zero, but not exactly zero), yet, the target number of features will still be taken since all scores have decreased by a proportional amount.



(a) Camera image for reference. (b) True optical flow. (c) Synthetic optical flow using the 3D prior. (d) Optical flow difference. (e) Depth-weighted optical flow difference. (f) Background-likelihood image.

Fig. 7. Generating a flow-based background-likelihood image. Beginning at the top, 7(b) shows the true optical flow from a live video stream, while 7(c) shows the synthetic optical flow generated by using the 3D scene prior (see Figure 6). Since this scene prior is absent of dynamic objects, there is a clear visual dissimilarity between the true and synthetic flow fields. Taking the difference of these images and weighting by the synthetic depth, we obtain a clean segmentation of the foreground and background elements 7(e), allowing us to create a background-likelihood image 7(f) to be used in our VO pipeline.

IV. EXPERIMENTS

A. Hardware & Setup

We present experimental results from two different urban areas in the UK: Woodstock and London. The Woodstock datasets were collected with our Bowler Wildcat mobile platform, equipped with a Bumblebee 2 stereo camera, a SICK LMS-151, and an Oxford Technical Solutions (OxTS) RT-3042 Inertial Navigation System (INS) for groundtruth. Our London dataset was gathered with a self-contained, vehicle-mounted mobile sensing suite, equipped with a Bumblebee 2 stereo camera, a SICK LMS-151, and a Trimble R8 GPS for groundtruth. The 3D priors were generated from the SICK lasers with VO for pose estimation; however, we wish to stress that the 3D prior could have been generated with a

TABLE I

SYSTEM PARAMETERS

Parameter	Description	Value
$\sigma_{d^c}^2$	Stereo disparity noise covariance [pixels^2]	0.05^2
τ_d	Mahalanobis distance threshold for the disparity-based method	1
τ_f	Depth-adjusted error threshold for the optical flow-based method	20

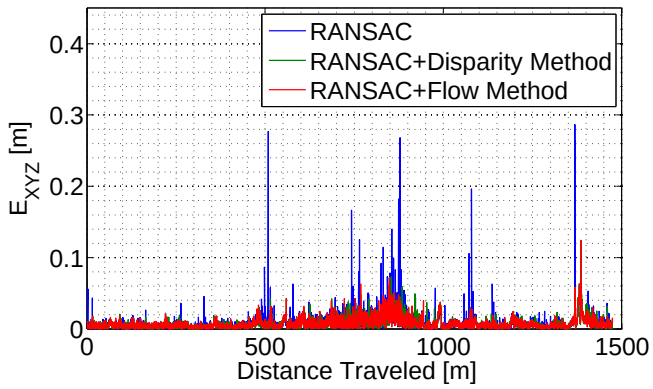


Fig. 8. Frame-to-frame translational errors measured by GPS groundtruth (Woodstock results). Note that we are, on average, always outperforming our baseline system and that both distraction-suppression techniques perform comparably. The blue spikes represent cases where the baseline system failed. Representative cases are shown in Figure 9.

more sophisticated lidar sensor, such as the Velodyne. In addition, we use the 3D object detection and classification method of Wang et al. [25] to postprocess our prior maps and remove most of the dynamic objects in the scene, which include pedestrians, vehicles, and cyclists.

To localise against the scene prior, we used the EBN system of Churchill et al. [7], which provides a transformation estimate against a prior visual experience (recall that one of the requirements for the 3D prior is that it have an associated image sequence). It should be noted that at present, this processing is done offline; however, implementing the online version is the focus of ongoing work. Lastly, Table I provides the system parameters used in our experiments, corresponding to the notation introduced in the previous section.

B. Visual Odometry

The goal of this section is to illustrate the improvements to our VO system by incorporating our background-likelihood images, as described in Section III-D. We present results from Woodstock and central London during the Olympics.

1) Woodstock: Two separate datasets were collected from the Begbroke Science Park research lab to the town of Woodstock. A subset of the busiest sections of these datasets were processed, totalling approximately 2 km. As our EBN system was only able to localise against one prior visual experience, there are regions where we were unable to localise — as such, we only present results on sections with a successful localisation, which is approximately 1.5km. In the future, we plan to collect enough datasets of these trajectories to saturate our EBN system so that we are always well localised regardless of the appearance change of the scene.

To compute localisation error, we measure the difference between the estimated frame-to-frame pose changes and the INS measured pose change. This is a more appropriate measure than looking at cumulative errors since a orientation error in one frame can skew the results for the rest of the trajectory. Denoting the true frame-to-frame translation as ρ_t and the estimated as ρ_e, we define a frame-to-frame error error measure as,

$$\mathrm{E_{XYZ}} := \left| \; ||\rho_e||_2 - ||\rho_t||_2 \; \right|. \tag{14}$$

We computed this error measure for three implementations: (i) our baseline VO system using RANSAC[1], (ii) our disparity-based method with RANSAC, and (iii) our flow-based method with RANSAC. Figure 8 shows the GPS groundtruth errors for each implementation versus distance traveled. Note the blue spikes in the plot, which represent frame-to-frame failures for the baseline system. To reiterate, the method presented in this paper provides an extra step of outlier rejection before proceeding with RANSAC, which is why we still require RANSAC in our pipeline. The goal is to illustrate the improvements in VO by incorporating these likelihood images for feature reweighing. Figure 8 shows the error percentages for our disparity-based and flow-based distraction suppression techniques against our standard VO system, where we see a noticeable improvement in accuracy.

A number of representative cases where our methods outperform the baseline are shown in Figure 9 and occur when there are many strong candidate feature matches on moving vehicles. Although one may argue that motion segmentations systems could potentially resolve some of these issues, we note that there are several cases where most of the scene was initially static but began moving (e.g., pulling up to traffic stopped at a red light). The strength of our technique is that regardless of how much of the image is obscured, we are able to focus our attention on just the portions of the image that belong to the static background.

2) London: For our London datasets, we collected three 10 km loops around several landmarks sites, such as the Houses of Parliament, Trafalgar Square, and St. Paul's Cathedral. For these experiments, signal-strength issues resulted in poor GPS measurements, which are not accurate enough to groundruth our motion estimates. We note that this is in fact a common problem in urban environments, strengthening the case that improving the robustness of relative motion estimation is a vital pursuit. Owing to this lack of groundtruth, we present qualitative evidence of our algorithms working in situations with extreme scene motion[2] (see Figure 10).

V. DISCUSSION AND FUTURE WORK

We have presented an alternative approach to improving the robustness of our VO system that does not rely on a trained object detection system, but rather, two straightforward methods that exploit prior 3D structure to identify

[1] 3-point RANSAC with Horn's method [26] for hypothesis generation.

[2] Video results of these techniques can be viewed at `http://www.youtube.com/watch?v=7ie9fNvcDC4&feature=youtu.be`

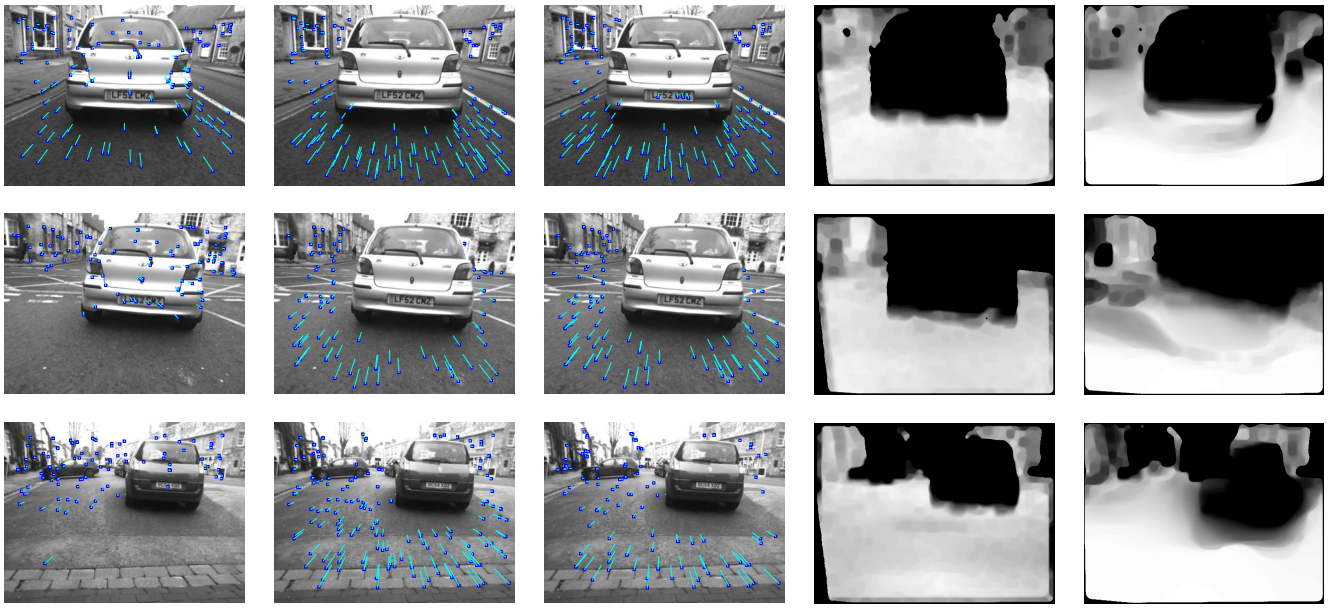


Fig. 9. Results from our Woodstock dataset. The top two rows showcase examples where we drove behind a vehicle that was initially at rest, but then began to move. As the vehicle makes up a large portion of the image and has distinctive features, our baseline system matched features on the vehicle across subsequent frames, leading to an erroneous motion estimate. In contrast, our distraction suppression systems ignored this vehicle and produce an accurate estimate. The last row shows a situation where RANSAC yielded a poor initial guess and the baseline system converged to an inaccurate estimate. Once again, this was not an issue with our distraction suppression methods, which can easily distinguish the foreground and background objects.

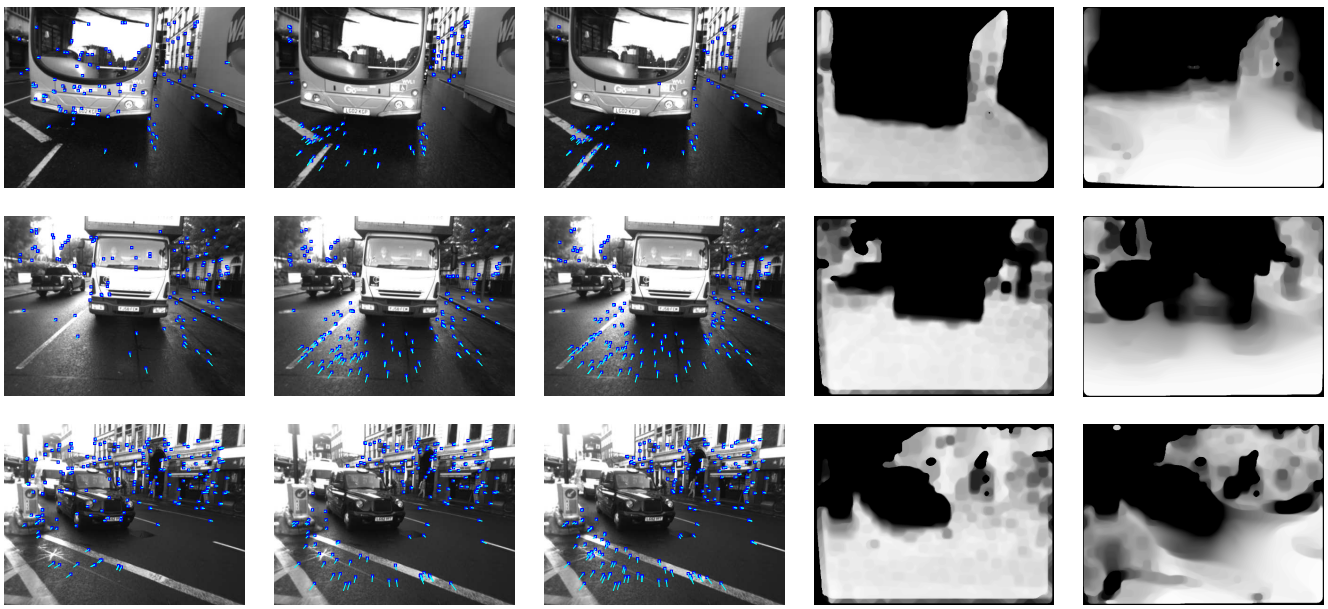


Fig. 10. Results from our London dataset. The top row illustrates an example of a large bus obscuring the image and very slowly approaching as our vehicle began to move. Our baseline system tracked features on the bus instead of the road surface, leading to an incorrect motion estimate. We wish to stress that even though most of this image is obscured by foreground objects, our distraction suppression techniques are able to focus on the static parts of the scene, resulting in more robust estimates. The bottom two rows illustrate other examples of our baseline system (i.e., without distraction suppression) incorrectly tracking features on moving vehicles and producing erroneous estimates.

geometric inconsistencies with the static background. Our results showed that both the disparity-based and flow-based methods outperformed our standard VO system with comparable results (see Figure 8). The only observable difference between the two approaches regarded the detection of stationary objects. For the flow-based method, stationary objects where only identified if the camera was in motion, otherwise, the objects would reproject to the exact same location, which would not produce a flow difference. In contrast, the disparity-based method was able to detect stationary objects

regardless of whether or not the camera was in motion. However, since tracking features on stationary objects does not directly impact the performance of egomotion, these two techniques ended up performing comparably.

It is worth noting that although we employed 3-point RANSAC in our VO pipeline for outlier rejection, there exist other, more efficient techniques, such as the 1-point RANSAC method by Scaramuzza [27]. However, regardless of what RANSAC technique is used, if most of the image is obscured by a moving vehicle, all RANSAC-based approaches will suffer since the majority of coherent features will be outliers.

The benefits of our proposed methods are not restricted solely to improving VO performance, but are in fact wide ranging and there are a number of other exciting avenues that we wish to explore. For instance, we plan on incorporating the optical flow distraction-suppression method into our localisation system called LAPS [11], to obtain a monocular-only system that is capable of robustly localising against a 3D scene prior. We also wish to apply these distraction suppression methods to our EBN long-term navigation framework, which would improve the quality of each visual experience. Lastly, we also aim to upgrade our current techniques to run online and test them in a closed-loop autonomous system.

VI. CONCLUSION

This paper has presented two novel techniques for distraction suppression in image data by leveraging knowledge of prior 3D structure. As our method does not rely on trained detectors, we are able to cope with arbitrary object types, even if they are obscuring a majority of the image. We have detailed how to produce background-likelihood images using just camera imagery and a 3D scene prior as well as how to incorporate these likelihood images into a visual odometry pipeline for distraction suppression. Lastly, we validated our approach in busy, cluttered, and distracting urban environments.

VII. ACKNOWLEDGMENTS

The authors wish to acknowledge the financial support provided by the NISSAN Motor Company and the EPSERC Leadership Fellowship Grant. Additionally, we wish to acknowledge BAE Systems for providing the Bowler Wildcat research platform and for their ongoing support.

REFERENCES

[1] E. Horbert, K. Rematas, and B. Leibe, "Level-set person segmentation and tracking with multi-region appearance models and top-down shape information," in *Proceedings of the International Conference on Computer Vision*, 2011.

[2] A. Ess, K. Schindler, B. Leibe, and L. van Gool, "Object detection and tracking for autonomous navigation in dynamic environments." *International Journal of Robotics Research*, vol. 29, 2010.

[3] B. Leibe, K. Schindler, N. Cornelis, and L. van Gool, "Coupled detection and tracking from static cameras and moving vehicles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1683–1698, 2008.

[4] A. Napier, G. Sibley, and P. Newman, "Real-time bounded-error pose estimation for road vehicles using vision," in *Intelligent Transportation Systems Conference (ITSC)*, 2010.

[5] O. Pink and C. Stiller, "Automated map generation form aerial images for precise vehicle localization," in *Intelligent Transportation Systems Conference (ITSC)*, 2010.

[6] A. Napier and P. Newman, "Generation and exploitation of synthetic overhead images for road vehicle localisation," in *Proceedings of the International Conference on Robotics and Automation*, RiverCenter, Saint Paul, Minnesota, USA, 14-18 May 2012.

[7] W. Churchill and P. Newman, "Practice makes perfect? managing and leveraging visual experiences for lifelong navigation," in *Proceedings of the International Conference on Robotics and Automation*, Saint Paul, Minnesota, USA, 14-18 May 2012.

[8] M. Bosse and R. Zlot, "Map matching and data association for large-scale two-dimensional laser-based slam," *International Journal of Robotics Research*, vol. 27, no. 6, 2008.

[9] I. Baldwin and P. Newman, "Road vehicle localization with 2d push-broom lidar and 3d priors," in *Proceedings of the IEEE International Conference on Robotics and Automation*, Saint Paul, Minnesota, USA, May 14-18 2012.

[10] J. Levinson, M. Montemerlo, and S. Thrun, "Map-based precision vehicle localization in urban environments," in *Proceedings of Robotics Science and Systems*, 2010.

[11] A. Stewart and P. Newman, "Laps - localisation using appearance of prior structure: 6-dof monocular camera localisation using prior pointclouds," in *Proceedings of the International Conference on Robotics and Automation*, Saint Paul, Minnesota, USA, 14-18 May 2012.

[12] I. Haritaoglu, D. Harwood, and L. Davis, "W4: Real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809–830, 2000.

[13] C. Wren, A. Azarbayejani, T. Darell, and A. Pentland, "Pfinder: Real-time tracking of human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997.

[14] M. Piccardi, "Background subtraction techniques: a review," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, 2004.

[15] Y. Ren, C.-S. Chua, and Y.-K. Ho, "Statistical background modeling for non-stationary camera," *Pattern Recognition Letters*, vol. 24, 2003.

[16] E. Hayman and J. olof Eklundh, "Statistical background subtraction for a mobile observer," in *IEEE International Conference on Computer Vision*, 2003.

[17] C. Yuan, G. Medioni, J. Kang, and I. Cohen, "Detecting motion region in the presence of strong parallax from a moving camera by multiview geometric constraints," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

[18] M. Irani and P. Anandan, "A unified approach to moving object detection in 2d and 3d scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998.

[19] Y. Sheikh, O. Javed, and T. Kanade, "Background subtraction for freely moving cameras," in *IEEE 12th International Conference on Computer Vision*, 2009.

[20] A. Taneja, L. Ballan, and M. Pollefeys, "Image based detection of geometric changes in urban environments," in *Proceedings of the International Conference on Computer Vision*, 2011.

[21] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Asian Conference on Computer Vision (ACCV)*, Queenstown, New Zealand, November 2010.

[22] C. Liu, "Beyond pixels: Exploring new representations and applications for motion analysis," Ph.D. dissertation, MIT, 2009.

[23] A. Taludker, S. Goldberg, L. Matthies, and A. Ansar, "Real-time detection of moving objects in a dynamic scene from moving robotic vehicles," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robotics and Systems*, 2003.

[24] E. Rosten, G. Reitmayr, and T. Drummond, "Real-time video annotations for augmented reality," in *Advances in Visual Computing*, 2005.

[25] D. Wang, I. Posner, and P. Newman, "What could move? finding cars, pedestrians and bicyclists in 3d laser data," in *Proceedings of the IEEE International Conference on Robotics and Automation*, Saint Paul, Minnesota, USA, May 14-18 2012.

[26] B. Horn, "Closed-form solution of absolute orientation using unit quaternions," *Journal of the Optical Society of America*, vol. 4, no. 4, pp. 629–642, 1987.

[27] D. Scaramuzza, "Performance evaluation of 1-point-ransac visual odometry," *Journal of Field Robotics*, vol. 28, no. 5, pp. 792–811, 2011.