Abstract

Our work addresses the problem of accurately 3D localising specific types of road fixtures, such as manhole covers. The surveying task for manholes has to be done for millions of kilometres of road. We propose an efficient pipeline that starts from images captured by vans, which are then used to detect, recognise and localise the manholes. Challenges come from frequent occlusions, regular changes in illumination conditions, substantial viewpoint variance, and the fact that the actual manhole covers are far less similar than one might expect. We combine 2D and 3D techniques and demonstrate high performance for large amounts of image data. To our knowledge this is the first published work for manhole mapping based solely on computer vision techniques and GPS.

1. Introduction

Mobile mapping as a field deals with tasks such as the creation of 3D city models for navigation, or digital surveying campaigns for public authorities to digitise their existing maps. A particular such goal is to automatically detect and localise road fixtures such as traffic signs [16], road markings (crosswalks, dashed lines) [13], manholes [4], street lights [7], cracks and anomalies in the pavement [18], or sidewalk edge deteriorations [17]. This paper describes a pipeline for the efficient detection and recognition of manholes, which is achieved by detecting their covers.

This may seem a simpler problem than it actually is. Challenges come from frequent occlusions (e.g., partially covered by a parked car), regular changes in illumination conditions (or even shadows cutting across the manhole), substantial viewpoint variance given different positions on the road and rotations about the vertical axis of the cover, and – last but not least – the fact that actual manhole covers are far less similar than one might expect (they vary by country, city, and the time of manufacture as seen in Figure 1; also the reason why some people seem to even collect them). Moreover, as with many detection systems, the errors in terms of false detections and missed manholes have to be very low for automated methods to be useful at all. This is the main reason why human operators are currently still carrying out most of this work.

Mobile mapping technology for road data acquisition has a long history [15]. Most such platforms integrate global positioning systems (GPS) and inertial navigation systems (INS), combined with lasers. Such solutions are quite expensive. In contradistinction, we propose a cheaper combination of cameras with GPS. The latter is mainly used to geo-reference the data, while the camera images are used to create a 3D point cloud by the use of Structure-from-Motion. As a by-product the latter also provides the positions and orientations of the cameras along the van’s trajectory. Internal calibration is performed before starting the mobile mapping campaign. The camera calibrations and the images were provided to us by GeoAutomation 1.

The current state-of-the-art in manhole localisation, like most of the work in this field, uses the electromagnetic properties of the material from which manhole covers are made (usually a metal). The shift in the phase and amplitude response are captured by a loop coil-type antenna ([19], [1], [4]). Compared to a vision-based solution, it is more difficult to have human operators check and correct such results. Moreover, other structures may also mislead such system, like drainage covers. We will come back to this option later though. Radio-frequency identification (RFID) tagging [3] is another way of assuring easy detection and mapping. Yet, this approach is inconvenient, as it requires placing tags in all existing manholes. Our work aims to map manholes by solely using computer vision techniques.

Vision has already been applied for manhole detection from aerial images [8], mainly with success for circular manhole covers. In [14], a dual-light system was proposed for pavement inspection, where a camera and two lights with different settings were used to obtain depth information, allowing the cracks in pavement to be detected. This can be used for manhole detection as a by-product of crack classification.

1http://www.geoautomation.be
Our system uses calibrated cameras for a 2D detection on the ground surface, in combination with a multi-view improvement of the detection scores and 3D mapping. We built a detector which first gradually prunes the candidate locations, based on the peculiarities of manholes, to then make final 2D decisions on the basis of a generic part-based detector [10], which strongly depends on the shape and edges of the manholes. We experimentally demonstrate the importance of capturing the manholes in several views (we use view and image interchangeably in this paper). Performance seems to saturate when at least 4 views are available for each physically distinct manhole cover.

The remainder of the paper is structured as follows. Section 2 provides an overview of our proposed system. Section 3 introduces the single-view processing steps, while Section 4 describes the multi-view related steps. The conducted experiments and achieved results are discussed in Section 5 and the conclusions are drawn in Section 6.

2. System overview

We first describe the data capturing equipment we used, before describing each of the pipeline steps in the subsequent sections. Our images are taken by a GeoAutomation van carrying a GPS sensor and 8 downward looking cameras. The latter are mounted as 4 stereo pairs on the roof, one pair looking out from each roof edge and thus together covering the street surface around the van. The van was driven through the streets and for every metre, each of the cameras simultaneously took a 1628 \times 1236 pixel image. The average speed of the van was about 35 kilometres per hour. The internal calibration of the cameras and their relative positions are known. The egomotion of the van is obtained by combining Structure-from-Motion with GPS readings.

We propose an off-line system that per image capture time looks for manholes within a limited region on the ground plane, in front of each camera. The system first processes single images independently, keeping the detection rate high and the number of false positives (FP) reasonably low. Single-view manhole detection in conjunction with the use of scene geometry subsequently allows for a global optimisation, which performs 3D localisation and refinement simultaneously. Since we deal with hundreds of thousands of high-resolution images we quickly throw out most of the background, and then invest increasing amounts of time in whatever patterns survive the previous steps.

Figure 2 gives an overview of the different steps. The processing is separated in single-view and multi-view processing. In the single-view stage, we start by estimating the position and inclination of the ground plane relative to the
3. Single-view processing

3.1. Ground plane projection

Under the assumption that the manholes are lying on the ground, and the ground plane surface is not curved or can be locally approximated by a plane, we fit a plane beneath the wheels of the van and project the camera view on it. We thus generate a bird’s-eye view (see Figure 3) which restores the true geometry of the objects on the ground. This is a normalization step: after projection the manholes are comparable in size and shape, reducing the variations in scale and viewing angle to planar similarity transformations.

We consider an area of 7 × 7 metres on the ground in front of the camera, which corresponds to 1024 × 1024 pixels in the projected image, where the original image is 1628 × 1236 pixels. The image quality of the projection degrades towards the top of the image, where the ground lies farthest from the van. This effect results in lower quality texture descriptors extracted from the top region.

If the ground is not planar then small distortions are introduced in the projected image. This may lower the quality of the texture descriptors. Yet, in our detectors we do not assume perfect shapes. As a result, violations of the ground plane assumption have a reduced impact on the overall performance.

3.2. Mean-shift colour segmentation

To reduce the number of candidates we employ mean-shift colour segmentation [5], due to its robustness and performance at an acceptable computational time. It is assumed that there is at least one colour-segmented region inside each manhole that is distinguishable from the surrounding area. Thus the set of potential candidates becomes the set of colour segmented regions. The mean-shift procedure detects modes in the density function given by similarly coloured image pixels.

For speed purposes, the 1024 × 1024 images are down-scaled to 102 × 102 and the EDISON mean-shift segmentation implementation is used in the L*u*v* colour space. As EDISON parameters we use ‘synergistic segmentation’, filtering and region fusion, with a minimum segment area of 32 pixels (thus imposing a minimum area of 3200 pixels at the original scale), a segmentation spatial radius of 7, a segmentation feature space radius of 6.5, the SpeedUp flag set to medium 2, a gradient window radius of 4, the mixture parameter set to 0.3, and the edge strength threshold to 0.3.

The resulting segments are now subjected to a cascade of tests. Only segments surviving all prior tests go to the next test. All other parameters are the default ones.

3.3. Area

The results are then scaled up to the original resolution. The segments having too few (< 3200) or too many pixels (> 32000) are dropped. Please note that the mean-shift segmentation already imposed the minimum area of 3200 pixels in fact. The area size constraints correspond roughly to manhole covers with an area between 0.15 and 1.5 square metres. The number of segmented candidates is maximally 325, but usually in the range of tens. This is a huge improvement over a sliding window approach that would require checking 10^6 up to 10^9 candidates.

3.4. Aspect ratio

Segments with an appropriate area are next tested for their aspect ratio. Extreme aspect ratios (< 0.5, > 2) for the surrounding bounding boxes eliminate a further set of segments.

3.5. Intensity variance

As a next filter, we demand that the variance in pixel intensities within a candidate/segmented region containing a manhole should exceed a certain threshold (here taken 90).

3.6. Radial symmetry

The large majority of manholes have a pronounced symmetry in their shape (circle, square) and texture. We compute a radial symmetry score [11] for each pixel in the image. We then prune the candidates/segmented regions with a low maximum radial symmetry pixel score. In our experiments, we run FRST (Fast Radial Symmetry Transform) on greyscale 1024 × 1024 projected images within a radius of 44 pixels (corresponding to 30cm on the ground plane) and a radial symmetry pixel score of at least 10.

---

2http://coewww.rutgers.edu/riul/research/code/EDISON/index.html
3http://www.nada.kth.se/~gareth/homepage/local_site/code.htm
3.7. Detection with LBP-based models

The next filter tests on texture. Local Binary Pattern (LBP) [12] feature vectors are computed by coding a neighbourhood in a specific order. For example, for a given neighbourhood, the centre’s pixel value is compared with each neighbouring pixel considered and the code in the vector is “1” for a greater value and “0” otherwise. In our case, uniform rotation-invariant LBP [12] yields our basic texture descriptor. We apply such LBP to $(16, 2)$ neighbourhoods in the intensity version of the $1024 \times 1024$ ground projected images. This neighbourhood selection yields a good trade-off between performance and complexity, as shown by experiments in [12] and our own empirical validation. At the segment level we simply take the average over the LBP histograms computed for all segmented pixels, as the texture descriptor.

As a further test we check whether the LBP histogram of a segment sufficiently resembles that of positive training samples (found as those segments with enough overlap with the manual annotations). Other, ‘negative’, segments correspond to false positives.

This resemblance is checked between the segment histogram and both a model histogram for the manholes and a model histogram for the background. These model histograms are formed by clustering the histograms for positive and negative segments, resp. As distance measure for the clustering we use symmetric cross entropy

$$S(u, v) = S(v, u) = H(u, v) + H(v, u)$$  \hspace{1cm} (1)

where $H$ stands for the cross entropy

$$H(u, v) = -\sum_i u(i) \log v(i)$$  \hspace{1cm} (2)

and $u$ and $v$ are the normalised LBP histograms (i.e. summing to 1). This results in a small set of reference histograms for manholes and background each (40 and 120, respectively).

In testing, we compute the best cross entropy for each candidate with the manhole representatives, the best with the background representatives, and the ratio between them. Based on thresholding this ratio the classification decision is taken. Detected manholes can be assigned to the trained representatives with which they have the best cross entropy.

3.8. Detection with part-based models

Another option is to replace the segmentation/LBP-based detection part by the generic, discriminatively trained part-based model for object class detection from [10] as provided by the authors. This detector is the top performer of

---

4http://www.cse.oulu.fi/MVG/Downloads/LBPMatlab

5http://people.cs.uchicago.edu/~pff/latent/
the PASCAL VOC Challenge 2009 [9]. The basic features employed by this method are the histograms of oriented gradients (HOG) [6].

We use the same split for data (as for the segmentation-LBP detection) and we train a model with one component corresponding to circular and rectangular manhole covers. We chose one component for speed reasons and because many of the circular manhole covers are surrounded by a rectangular rim. Furthermore, we restrict the detector to operate only for a reduced range of sizes/scales on the ground projected images. The root, the sub-parts and the deformation for each sub-part of our trained model are depicted in Figure 3. The LBP texture models can be used as postfiltering/classification step.

4. Multi-view processing

4.1. Grouping

For each detection we compute the average pixel position and then the corresponding 3D position based on the camera parameters and the fitted ground plane. The detections are grouped based on proximity (less than 1 metre distance in 3D space) and texture likelihood. This can be solved by a Minimum Description Length formulation as in [16].

In our case, a simple greedy procedure was sufficient. We pick the detection with the highest confidence, then assign all other detections that agree in texture within 1 metre radius to the same cluster, and repeat the procedure with the remaining detections until all the detections are grouped, further averaging the 3D positions.

4.2. GraphCut accurate segmentation

In order to get accurate segmentations of the potential manhole detections against the surrounding background in the ground projected images, as needed in the next section, we use a GraphCut-based method [2], (rather than the less precise mean-shift, which would have to be redone here on the texture features used here).

A window of 240 × 240 pixels (corresponding roughly to 1.71m × 1.71m in the real world) centred on the candidate is considered. The central 40 × 40 patch (corresponding to 27cm × 27cm) is assumed to belong to the manhole, while the central 200 × 200 (1.37m × 1.37m) area is supposed to include the entire manhole, but probably also undecided pixels (not clear a priori whether manhole or background). The pixels outside this area are taken as background.

From the background area we extract statistics for an 8 × 8 × 8 binned RGB colour space. By dividing by the number of pixels we get a probabilistic estimate for background colours. The background LBP histogram is computed as an average over the already computed LBP descriptors of the pixels considered background and is $l_1$ normalized. We do the same with the patch assumed to belong to the object.

For each undecided pixel, $p$, we compute an LBP histogram, $L(p)$, as the average over the already computed LBP descriptors in a surrounding patch of 20 × 20 pixels and then $l_1$-normalize it. Also, we compute the symmetric cross entropy with the object and the background LBP histograms, $S(L(p), L('bkg'))$ and $S(L(p), L('obj'))$. Moreover, from the binned RGB spaces we extract the colour probabilities for each undecided pixel $p$ to belong to the object, $Pr(p|'obj')$, vs. the background, $Pr(p|'bkg'$).

We formulate the segmentation as a standard energy minimization

$$E(f) = \sum_{p \in P} D_p(f_p) + \lambda \sum_{(p,q) \in N} V_{(p,q)}(f_p, f_q)$$  \hspace{1cm} (3)

where $f_p$ is the labelling function taking 2 values, “obj” or “bk”, $N$ are the edges in the four connected image grid and $\lambda$ is a weighting parameter.

The unary costs are computed as

$$D_p(f_p) = -\log Pr(p|f_p) + \kappa(S(L(p), L(f_p)) - S_{min})$$  \hspace{1cm} (4)

where $S_{min}$ is the minimum symmetric cross entropy computed for the undecided pixels and $\kappa$ is a weighting parameter.

The pairwise costs are obtained for each edge $\{p,q\}$

$$V_{(p,q)}(f_p, f_q) = \delta(f_p, f_q) \exp(-\frac{(I_p - I_q)^2}{2\sigma^2})$$  \hspace{1cm} (5)

where $\delta(f_p, f_q)$ is 1 when $f_p = f_q$ and 0 otherwise, $I_p$ is the intensity value of $p$, and $\sigma$ stands for the noise. In our experiments we obtained improved segmentations when the channel $I$ in equation (5) is replaced with images convolved with Gaussian filters or a combination of Gaussian and Sobel filters.

We are using GraphCut with $\alpha$-expansion $^6$ for solving (3) and, thus, binding the undecided pixels to object or background (see Figure 4).

4.3. Accurate 3D localisation

Accurate 3D localisation implies accurate pixel localisation of manholes in single views.

GraphCut segmentation (see Figure 4 and Section 4.2) is applied with a single iteration only for the most favourable candidates (in terms of closeness to the centre of the bottom part of the ground projected image) from each cluster of candidates until 5 are found with plausible segmentations or all the available views were used. A plausible segmentation is one with a width/height ratio close to the expected ratios in the training material (here close to 1), and with an area of the manhole in the expected value range. Failure to obtain plausible segmentations is an indicator for a potentially false detection and is used to prune such cases.

$^6$http://www.csd.uwo.ca/faculty/olga/OldCode.html
The GraphCut segmentations are used for the estimation of the 3D position of the centre, by getting first the centre of mass of the segmented pixels in the ground projection and then moving it into the 3D world space. Here we just take the average of these 3D reconstructed centres as the 3D localisation of the manhole cover.

5. Experiments

5.1. Ground truth data

The data set consists of a training set of 231 manhole annotations and a testing set of 1046 annotations, the latter for 359 distinct manholes. The data is recorded in Flemish towns. In our material the manhole covers exhibit different shapes, textures, and sizes, and appear under variable viewpoints. Also, the manholes suffer from occlusions, differences in light exposure, rust, and dirt. For the evaluation of 3D mapping we have sequences simultaneously captured by 8 downward-looking cameras on the van.

5.2. Single-view evaluation

For single-view detection the results are poor (see Figure 5, “1 view”). Only 40% are detected by the LBP-based model, while the part-based model reaches 70% at a very low FP per image (FPPI) rate. Most of the missed manholes are due to their blurry appearance in the ground projected image and because they largely fall outside the image. Often we do not get the entire manhole by means of mean-shift segmentation, which was to be expected because of dirt, shadows, and confusion with the surroundings. Figures 6c) and d) show the complementary behaviour of the two proposed detectors at an FPPI of 0.05, the LBP-based detector copes better with loose boundaries while the part-based detector better handles situations when the texture is corrupted or unreliable. By applying the GraphCut-based manhole segmentation the accuracy of the detection is increased. We usually obtain pixel level accuracy for the centre of mass of the segmented manhole in the image.

5.3. Multi-view evaluation

For each detection in the projected image, we can trace it back to the 3D space and thus obtain the 3D mappings of our detections. Grouping the detections and averaging the 3D positions of the best detections (after applying the refinement step involving GraphCut segmentation) we get very good 3D mappings of the physical manhole corresponding to the grouped detections. In our testing data, the localisation precision is on average better than 10 centimetres according to our hand labelled ground truth.

In Figure 5, one finds the experimental validation of the usefulness of having manholes captured in more camera views. A manhole is considered detected if any of its occurrences (in different views) is successfully detected. If we have four views for a manhole, our system has a detection performance of more than 93% by using either LBP or the part-based model (i.e. choosing either single-view path in Figure 2), while having a very low FP rate (see Figure 5). The part-based detector performs better than the LBP-based detector, and has the advantage of being more general because it works on edges/shapes. On the other hand, the LBP-based detector (non-optimised code) is 2 to 3 times faster than the part-based counterpart and has a similar performance as soon as at least 4 views are available per manhole, which is usually the case. On the first three rows of Figure 6a) and b) we see that for the same manhole covers there are favorable views (a) that allow detection and views where detection fails (b) due to various causes such as occlusions, dirt, or light conditions. Note that the reported performance in Figure 5 is without applying GraphCut segmentation (see Section 4.2) or accurate 3D localisation (see Section 4.3) and, thus, without enjoying their FP pruning benefits.

Since we need the camera calibrations, the multi-view 3D mapping is an offline process. We ran our pipeline on long sequences, and found FPPI rates lower than $10^{-5}$ for mapping nearly all the manholes. A demo running the LBP-based detector without using accurate 3D localisation on a 1000 $\times$ 8 frames sequence where we have 0 falses and 1 missed manhole, is available at: http://homes.esat.kuleuven.be/~rtimofte

5.4. Discussion

This work investigated the problem of 3D manhole cover mapping based on computer vision. We succeeded in pro-
providing a high performance vision system and in showing convergence in the performance achieved with different detectors under the multi-view, 3D formulation.

The RFID tagging based methods [3] are intrusive, requiring prior tagging of the manholes. Moreover, just like loop-coil methods, they do not allow for an easy visual check of the results.

We were not able to provide a direct comparison with a loop-coil-type antenna based system [19], which uses the electromagnetic properties of the manhole cover material. Here we see more potential for a fused material-vision approach. Both approaches have complementary advantages and drawbacks. The vision system can have problems with manholes visually undistinguishable from the surroundings or partly buried. The material-based system can spot these but along with many FP, or can be unreliable when the surroundings contain objects from the same or a similar material (e.g. iron).

6. Conclusions

Manhole cover localisation is a challenging problem. We have proposed a multi-view scheme, which combines 2D and 3D analysis. Following a principle of spending little time on the bulk of the data, and reserving a more refined analysis for the promising parts of the images, the proposed system combines efficiency with good performance. The main contribution of the paper is that for the first time a system combines efficiency with good performance. The RFID tagging based methods [3] are intrusive, requiring prior tagging of the manholes. Moreover, just like loop-coil methods, they do not allow for an easy visual check of the results.

We were not able to provide a direct comparison with a loop-coil-type antenna based system [19], which uses the electromagnetic properties of the manhole cover material. Here we see more potential for a fused material-vision approach. Both approaches have complementary advantages and drawbacks. The vision system can have problems with manholes visually undistinguishable from the surroundings or partly buried. The material-based system can spot these but along with many FP, or can be unreliable when the surroundings contain objects from the same or a similar material (e.g. iron).

Acknowledgements. This work has been partly supported by the European Commission FP7-231888-EUROPA project. We thank GeoAutomation for providing the images and for their support.

References


Figure 6. The LBP-based and the part-based detectors are complementary.


