Reconstruction of 3D Information about Passing Vehicles

Petr Dobeš*
Supervised by: Adam Herout†

Department of Computer Graphics and Multimedia
Faculty of Information Technology, Brno University of Technology
Brno / Czech Republic

Abstract

This paper reports experiments performed during the work on my master's thesis which focuses on 3D reconstruction of vehicles passing in front of a traffic surveillance camera. Calibration process of surveillance camera is first introduced and the relation of automatic calibration with 3D information about observed traffic is described. Afterwards, a set of experiments with feature matching and Structure from Motion algorithm are presented and their results on images of passing vehicles are examined. Modifications to correspondence search stage of Structure from Motion pipeline are then proposed. Most importantly, instead of using SIFT features, DeepMatching algorithm (originally devised to find quasi-dense point matches in optical flow calculation) is used to obtain point correspondences for subsequent reconstruction phase. As a result of implemented modifications, the overall completeness of reconstructed point cloud model of passing vehicle has improved significantly.

Keywords: 3D Reconstruction, Structure from Motion, Traffic Surveillance, Traffic Analysis, Camera Calibration

1 Introduction

Deployment of high-resolution digital cameras in traffic surveillance has increased the need for computer vision algorithms that automatically extract data from captured video streams. When supplemented with computer vision methods, traffic surveillance cameras can serve a wide range of purposes, such as counting of passing vehicles, their classification, finding driving lanes, detecting traffic jams and discovering drivers in the opposite direction. Moreover, the primary aim of many traffic surveillance systems is to measure the speed of passing vehicles. Nevertheless, many of the tasks cannot be achieved without preceding camera calibration.

This paper addresses the problem of reconstruction of 3D information about vehicles passing in front of a surveillance camera. In existing algorithms developed for automatic traffic surveillance, the only obtained 3D data about a passing vehicle is its bounding box. This work therefore aims to devise a method that could acquire more precise 3D representation of a vehicle captured in a video stream. Such information is desired not only for visualisation purposes, but may also be utilized to infer the scale of the projected scene, and thus contribute to the camera calibration process.

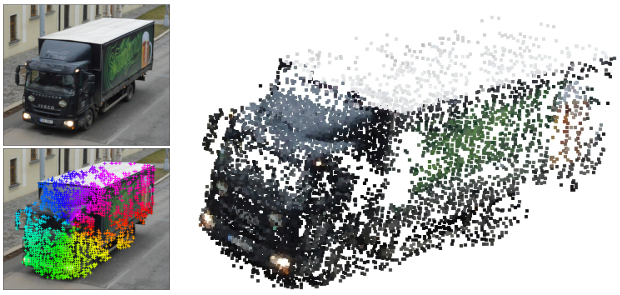

Figure 1: An example of points found by DeepMatching algorithm (left) and the corresponding result of 3D reconstruction of a passing truck (right).

Available tools for 3D reconstruction are first examined and tested to find out whether they could directly be used for the outlined task. Additionally, a set of experiments carried out with keypoint extraction is described. Lastly, a modification of the correspondence search stage of 3D reconstruction pipeline is proposed and implemented. An example of improved keypoint search and resultant 3D reconstruction is shown in Figure 1.

2 Calibration of Traffic Surveillance Camera

Monocular cameras can be utilized in numerous tasks of traffic analysis and surveillance, one of which is speed measurement of passing vehicles. Techniques for visual speed measurement have been developed by various authors [8, 4, 1, 3, 9]. Nevertheless, many of the traffic surveillance tasks, especially accurate speed measurements, require precise calibration of the particular roadside camera. This section therefore focuses on the approaches to calibration of monocular camera employed in

*dob.petr@seznam.cz
†herout@fit.vutbr.cz

traffic surveillance.

2.1 Camera Calibration Model

Traffic camera calibration can either be performed manually or fully automatically. As standard pattern-based approaches (such as the one developed by Zhang [13]) cannot be used, manual calibration requires user input of some information about the scene that is viewed by the camera. Such approach often relies on physical measurements in the scene or on placement of specific markers, and thus involves considerable amount of effort. This renders manual calibration impractical for large-scale deployment of roadside cameras, and it is therefore desirable for the calibration to be fully automatic [2, 8].

Standard camera calibration process involves finding its intrinsic parameters (matrix $\mathbf{K}$) and extrinsic parameters (matrix $[\mathbf{R}\,\mathbf{T}]$) that form the projection matrix $\mathbf{P}$:

$$\mathbf{P} = \mathbf{K}\,[\mathbf{R}\,\mathbf{T}] \tag{1}$$

However, for the purpose of speed measurement in visual traffic surveillance, it is more convenient to define the problem of camera calibration as finding the *intrinsic parameters*, determining the *road plane*, and finding the *scale of the road plane*. This approach is more suitable, as it enables direct speed measurement of vehicles driving on the road plane. This concept of camera calibration can be considered equivalent with the above mentioned standard camera model and methods exist to convert the obtained parameters from one model to the other [8].

When determining the intrinsic parameters, surveillance camera is assumed to exhibit zero pixel skew and to have principal point in the center of the image. The only remaining intrinsic parameter to determine is therefore its focal length. This parameter can be calculated using two vanishing points. Once the vanishing points are determined, the parameters of the road plane (without scale) can also be obtained. The scale of the road plane is thus the last necessary parameter to infer [8].

2.2 Automatic Calibration Using the Motion of Passing Vehicles

Whenever fully automatic calibration of surveillance camera is to be performed, it is suitable to extract the information necessary for obtaining the aforementioned calibration parameters from observed traffic flow.

Finding Vanishing Points and Road Plane

Methods, such as the one presented by Dubská et al. [2], first detect vanishing points using the observed motion of vehicles. Once positions of two vanishing points in image space are obtained, focal length of the camera can be calculated. Two vectors from the origin of the camera system can then be constructed from the coordinates of the vanishing points and the focal length. Cross product of these two vectors then yields the normal vector of the road plane. The only remaining parameter is thus the distance of the road plane from the camera which establishes the relation between the image and real-world units, i.e. the scale [8].

Determining the Scale of the Road Plane

If passing vehicles are to be used as the source of information to obtain the scale of the road plane, camera calibration inevitably becomes closely related to 3D structure of the vehicles. Two significant approaches to determine the scale of the road plane from observed traffic flow have been developed.

The first approach, presented by Dubská et al. [3], uses 3D bounding boxes of passing vehicles and statistical domain adaptation of their dimensions. The authors detect a vehicle blob and construct its 3D bounding box using lines that pass through vanishing points and that are tangent to the vehicle blob. Once image coordinates of corners of the bounding box are known, it is possible to project the base of the bounding box onto the road plane. As a result, coordinates of the bounding box base in 3D space are acquired. The distance between these coordinates, together with the information about the real-world dimensions of the vehicle, can be used to determine the scale of the road plane. In order to determine the scale factor, Dubská et al. [3] collected statistical data about sold cars and their dimensions, and subsequently formed a histogram of their bounding box dimensions. Scene scale was then determined by fitting statistics of known dimensions and the measured data from the observed traffic.

It is important to note that when extraction of 3D information about observed traffic is considered, bounding boxes have been so far the only 3D information obtained about passing vehicles. Moreover, the fact that bounding boxes are extracted using lines tangent to the vehicle blobs, whose edges tend to be bent, has negative influence on the overall accuracy.

The second approach to scale inference is proposed by Sochor et al. [8], who infer the scene scale by aligning rendered 3D models of frequently passing cars. They use fine-grained information abut vehicle type (i.e. make, model, variant, model year) and obtain 3D models for two vehicle types that are commonly observed. The method starts with classification of passing vehicles. When vehicle type with available 3D model is detected, the image of its 3D model is rendered in multiple different scales and its 2D bounding box is matched with 2D bounding box of the detected vehicle blob. Once rendered 3D model is aligned to the detected vehicle in the image, two points representing the front and the rear of the vehicle are projected to the road plane. Knowing the real-world distance of these points from available vehicle dimensions provides sufficient information for the scene scale to be calculated.

2.3 Prospective Contribution of This Work to Camera Calibration Process

As this work aims to reconstruct 3D information about vehicles passing in front of surveillance camera, the extracted 3D data can contribute to further improvement of camera calibration process. In particular, obtained 3D model could provide additional information for the calibration phase in which scale of the road plane is computed.

Fine-grained classification of detected vehicles could be used to distinguish between various vehicle models. Real-world dimensions would also be stored for each vehicle model. Once particular vehicle with known dimensions is recognized, its detailed 3D reconstruction could be created and utilized to infer the scene scale. Unlike the method where rendered 3D model alignment is used, this approach would only require the information about vehicles' dimensions to be saved in the traffic surveillance system, and no prior 3D model data would be necessary.

3 Utilized Computer Vision Methods

This section introduces computer vision algorithms that have been used throughout the work on this paper. First, Structure from Motion (SfM) algorithm is described. Secondly, the concepts of optical flow and DeepMatching are addressed. Lastly, a modification of SfM pipeline is proposed.

3.1 Structure from Motion

Structure from Motion (SfM) is an algorithm used for 3D reconstruction from image collections. Several implementations of this reconstruction strategy exist, such as COLMAP [6], Bundler [7] and VisualSFM [12]. This subsection introduces and describes individual phases of incremental Structure from Motion algorithm.

General pipeline of incremental Structure from Motion is shown at the top part of Figure 2. The input to SfM is a set of unordered images with projections of a scene that is to be reconstructed. The first stage of the SfM pipeline consists of correspondence search and is followed by the second stage that is represented by an iterative reconstruction component. The output of SfM is sparse 3D reconstruction in the form of point cloud.

Correspondence Search

The first stage of the Structure from Motion pipeline is correspondence search. This stage involves extraction of local feature points, identification of corresponding projections of the same points in overlapping images (matching), and subsequent geometric verification of the found matches.

Feature extraction encompasses detecting coordinates of feature points within every image and representing the points using descriptors. These points need to be distinctive in order to be uniquely recognized in multiple images, and thus SIFT [5] is a common choice in many implementations, including COLMAP. Next, sets of feature points are matched using similarity metric to find corresponding point pairs. Obtained point correspondences are then geometrically verified. Verification consists of estimating a transformation that maps a sufficient number of corresponding points between images, and the remaining point pairs are filtered out. Since corresponding point pairs are usually contaminated by outlier, estimation of the transformation requires techniques such as RANSAC. The result of this step is a geometrically verified set of image pairs and their associated inlier correspondences.

Incremental Reconstruction

The stage of incremental reconstruction receives the obtained set of image pairs with their point correspondences and performs iterative reconstruction of the scene. Initialization by two-view reconstruction is followed by a cycle in which additional images are registered to the already reconstructed model and new points are triangulated. An image is registered to the current model by solving the Perspective-n-Point problem using feature correspondences with already existing points in the model (2D-3D correspondences). Newly added image observes existing scene points in the model and can also increase the number of points in the model through triangulation. Once new scene point is observed from different angle by at least one more image, its coordinates can be triangulated and the point extends the current model.

Furthermore, bundle adjustment is employed to improve the precision of the model. This step is necessary to prevent reconstruction from drifting into non-recoverable state due to the accumulation of uncertainties in pose estimations and errors in point coordinates. In bundle adjustment, already reconstructed points are projected back into image space of their respective images. The aim of bundle adjustment is then to perform non-linear minimization of the reprojection error, and thus simultaneously refine the camera and point parameters.

3.2 Optical Flow and DeepMatching

Optical flow belongs to the set of algorithms used for motion estimation between two (or more) images. While other methods exist for simple movements, optical flow is the most general technique. The aim of optical flow is to compute an independent estimate of motion at each pixel. In other words, the task of optical flow is to find a vector for every pixel that defines the displacement of the pixel between two images [10].

In order to address the problem of large displacements contained within the two input images, some authors also incorporate descriptor matching component into the calculation process. The main idea is to guide optical flow estimation by providing correspondences from sparse descriptor matching. Weinzaepfel et al. [11] argue that even

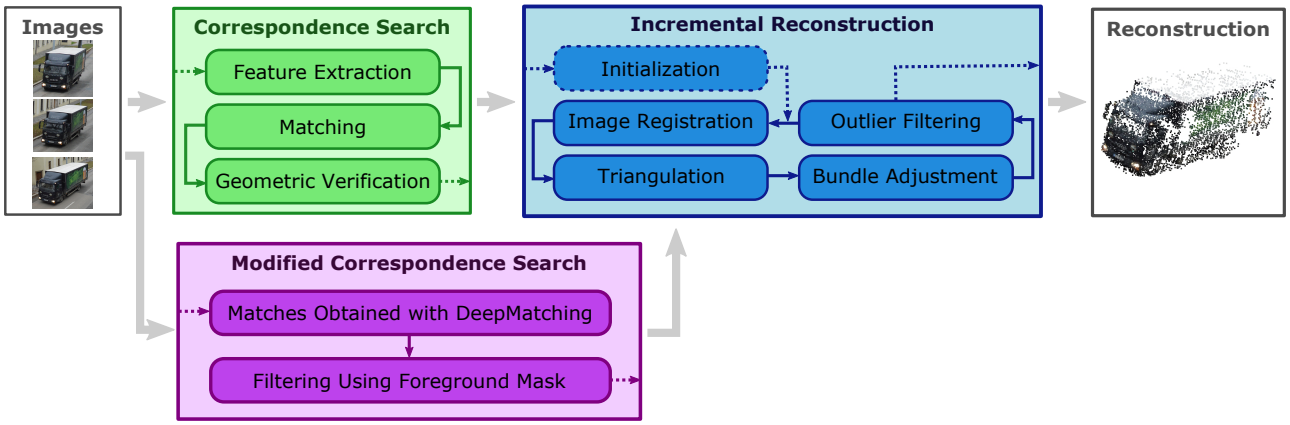

Figure 2: Standard pipeline of incremental Structure from Motion algorithm (top) and modified correspondence search stage (bottom), which employs DeepMatching to obtain quasi-dense point correspondences for the subsequent reconstruction phase.

though this modification significantly improves results of optical flow algorithm, standard methods for feature point extraction only produce points for salient image locations. Therefore, in their method for optical flow, named *DeepFlow*, Weinzaepfel et al. [11] enhance the variational approach with custom descriptor matching algorithm called *DeepMatching*.

The proposed DeepMatching algorithm aims to retrieve quasi-dense point correspondences for later optical flow calculation phase. DeepMatching is strongly inspired by non-rigid 2D warping and deep convolutional networks. SIFT descriptors based on histogram of oriented gradients with 4×4 cells are used. However, instead of keeping the fixed 4×4 grid, it is divided into 4 quadrants and each of the quadrants is allowed to move independently in order to yield non-rigid matching. This approach is then applied recursively together with max-pooling and convolution [11]. As a result, DeepMatching produces point correspondences with very high density.

3.3 Modification of Structure from Motion Pipeline using DeepMatching

As shown by experiments described in Section 4, the correspondence search stage of general Structure from Motion pipeline proved to be unsuitable for the outlined task of 3D reconstruction of passing vehicles. Therefore, a modification of this SfM stage was carried out in order to achieve improved reconstruction results.

Instead of using SIFT features in the correspondence search stage, DeepMatching is utilized to find high number of corresponding points within pairs of input images. Furthermore, filtering with foreground mask is performed in order to remove points that do not belong to the passing vehicle. Obtained matches are then fed to the incremental reconstruction phase, which remains unchanged. The modified SfM pipeline is presented at the bottom part of Figure 2. Implemented modifications are described in full details in Section 5.

4 Experiments with SIFT Features and Structure from Motion Tools

Throughout the first part of the work on this paper, a series of preliminary experiments was carried out in order to evaluate to what extent the current state-of-the-art Structure from Motion algorithms can be used when solving the problem of 3D reconstruction of passing vehicles. For this purpose, two Structure from Motion tools were selected, COLMAP [6] and VisualSFM [12]. However, before examining the performance of SfM tools, one more set of experiments was carried out. Since both of the selected SfM tools base their correspondence search stage on SIFT features [5], experiments were first performed to evaluate the behaviour of SIFT feature extraction and matching on images of vehicles.

In this section, data obtained for experimenting are first described. Next, experiments with SIFT features are discussed. Subsequently, the results of Structure from Motion reconstructions are presented.

4.1 Obtained Test Data

Several image sequences of passing vehicles were obtained for experiments presented in this section. In order to ensure sufficient quality and resolution, images were captured using stationary reflex camera (Nikon D3200 with Nikon AF-S DX 18-105mm f/3,5-5,6 G ED lens) used in burst mode. Therefore, images in each sequence represent frames that would be extracted from a video at different points in time. Each created sequence contains from 7 to 15 images. For the purpose of experiments, a sample containing sequences of 6 different cars, 2 vans, and 1 truck was selected. Additionally, all images in selected sequences were cropped to include the vehicle with only

small border containing the background. Examples from two image sequences are shown in Figure 3.

Considering the fact that Structure from Motion algorithms expect static scene and moving photographer, another set of image sequences was obtained using a stationary car with camera moving around. It is therefore possible to compare the results of inputs containing stationary and moving vehicles.

Figure 3: Examples of obtained sequences of images with a passing vehicle.

4.2 Experiments with SIFT Feature Extraction and Matching

Characteristics of extracted SIFT keypoints and correspondences were examined on obtained image sequences using SIFT implementation in OpenCV[1] library. First, positions of detected SIFT keypoints were inspected on single images. Secondly, found feature correspondences between pairs of images in each sequence were studied. In this case, various image pairs with different steps between images (i.e. different distance of the images within the sequence) were considered. All experiments were carried out on sequences of both stationary and passing vehicles, with equivalent results.

When SIFT keypoint detection algorithm is applied, the vast majority of obtained keypoints is located on the front part of the vehicle (mainly on grilles and license plate). The remaining parts of vehicle are covered very sparsely, as only low numbers of feature points are detected there. Furthermore, when feature point matching is performed, correct correspondences are often found only for small steps between the images in the particular sequence (i.e. small changes in vehicle orientation). Larger steps between images result into significant numbers of incorrectly calculated correspondences, especially for points which are not on the front part of the vehicle (grilles and license plate). An example of computed SIFT correspondences is shown in Figure 4.

The results of experiments with SIFT features indicate that algorithms for 3D reconstruction that rely on SIFT in their correspondence search stage are very likely to have only small numbers of feature points for subsequent reconstruction phase. Moreover, the number will probably be further reduced by incorrectly found correspondences.

[1] http://opencv.org/

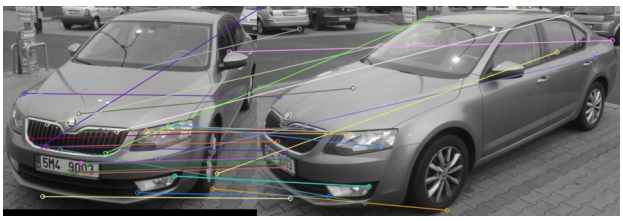

Figure 4: Example of found SIFT point correspondences on a static vehicle (30 best matches are shown). The vast majority of feature points is detected on the front part of the vehicle. Moreover, a significant number of incorrect matches can be observed.

4.3 Experiments with Structure from Motion

Experiments with 3D reconstruction were performed using COLMAP tool, which was released in 2016 and is currently the state-of-the-art Structure from Motion implementation [6]. Reconstruction process was tested for all created image sequences of both stationary and passing vehicles.

First, experiments with image sequences of stationary vehicle were performed. Out of 11 experiments, reconstruction was successfully completed in only six cases. In the remaining cases, SfM algorithm failed to produce any result at all, reporting that no good initial image pair was found. Only three of the successful reconstructions contained recognizable points that belong to the original vehicle. The best achieved result is presented in Figure 5. One of the remaining successful reconstructions shows an attempt of the algorithm to reconstruct the background scene instead of the vehicle, while other two successful reconstructions resulted in a point cloud with no meaningful structure.

Figure 5: The best obtained result using COLMAP Structure from Motion tool for a sequence of images containing a *stationary* vehicle. Point cloud model (on the right) includes partially recognizable front part of the vehicle (especially its license plate) and the front wheel. Remaining parts of the vehicle are not included at all, or reconstructed incorrectly.

Next, COLMAP was used on the image sequences of passing vehicles. Out of six image sequences of cars, only one reconstruction was successfully completed and point cloud model was produced, whereas all other reconstructions failed (again, the algorithm reported that no good initial image pair was found). As expected, only front part of the car is partially recognizable in the successfully created model. Reconstruction process also failed in case of image sequences of both vans. Nevertheless, a successful reconstruction was obtained for image sequence of passing truck, where significant portion of front part is recognizable. The resultant model is shown in Figure 6.

Figure 6: The best obtained result using COLMAP Structure from Motion tool for image sequence of *passing* vehicle. Resultant point cloud (on the right) contains recognizable front part of the truck.

Results of Structure from Motion algorithm confirm the conclusions drawn from the previous experiments with SIFT features. As expected, reconstructed models are often severely incomplete. In a vast majority of cases, the reconstruction process either failed entirely, or the resultant point cloud contained no meaningful structure. Apart from the presented tests using COLMAP, several experiments were also carried out with VisualSFM tool, producing comparable results.

5 Improvement of the 3D Reconstruction Process

Based on the experiments described in the previous section, two main aspects hindering the 3D reconstruction process can be identified. The first problem is insufficient number of feature correspondences, as standard SIFT features are not suitable input for reconstruction of passing vehicles. The second significant problem is represented by points and point correspondences located in the image background. In this section, changes to the reconstruction process are proposed and applied in order to improve the overall quality of the resultant 3D model.

5.1 Substitution of SIFT Features

In order to increase the number of point correspondences located on vehicle, it is necessary to substitute SIFT features with a different method for keypoint extraction and matching. In particular, a method producing matches with higher density is desirable. One option would be to use output of an algorithm for optical flow calculation, which would produce a vector that estimates movement of each pixel in an image pair. Nevertheless, in order to address large displacements, optical flow methods often utilize feature matching algorithms, too. It is therefore more beneficial to inspect the feature matching approaches used within optical flow, rather than entire methods for optical flow themselves.

As described in Subsection 3.2, optical flow algorithm DeepFlow employs a custom feature matching procedure, called DeepMatching, to calculate quasi-dense point correspondences before smoothing them using variational approach to obtain optical flow estimation. The power of DeepMatching algorithm, even though originally designed for optical flow, could be harnessed to provide a high number of point matches for subsequent 3D reconstruction of passing vehicles. An illustration of point matches found by DeepMatching algorithm is shown in Figure 8.

5.2 Filtering of Obtained Correspondences

The second necessary modification of the correspondence extraction procedure is removal of those point matches that belong to the scene background, as these points can be considered outliers, and thus negatively affect the reconstruction process. Obtained correspondences should therefore be filtered using a foreground mask of every individual image, so that only matches located on the vehicles in both images of particular image pair are taken as an input for reconstruction phase. An example of the original image and its respective foreground mask is shown in Figure 7, filtered correspondences are illustrated by Figure 8.

Figure 7: Original image of passing truck and its foreground mask.

Figure 8: Correspondences for two images of a passing truck calculated using the DeepMatching algorithm and filtered with foreground masks.

5.3 Application of the Proposed Modifications

Implementation of the proposed modifications requires a possibility of defining custom keypoint locations and point correspondences as an input for the following stage of incremental reconstruction. A suitable interface is offered by VisualSFM and application of presented modifications was therefore realized using the VisualSFM tool.

Correspondences were first calculated using DeepMatching algorithm for all possible pairs of images in an image sequence. Next, foreground masks were created and applied to perform filtering of point matches. A file with locations of matched points is then generated for every image. It should be noted that unlike standard SIFT keypoint detection, DeepMatching can obtain slightly different sets of points for one particular image when matching this image with several other images. Therefore, the obtained point sets are unified before the output file with keypoint coordinates is created. Furthermore, one file containing information about all found matches is generated. The described procedure replaces the first stage of the SfM pipeline, in which correspondence search is performed (as shown in Figure 2).

Information stored in the generated files was then used as the starting point for the 3D reconstruction stage of VisualSFM tool. An example of the resultant model can be seen in Figure 9. When compared to the reconstruction obtained with original SfM algorithm, the results of the proposed modifications significantly improve completeness of the resultant point cloud model.

Figure 9: Resultant 3D reconstruction of a passing truck obtained when proposed improvements to Structure from Motion pipeline are applied.

6 Conclusion

In this paper, a set of experiments with SIFT feature matching and Structure from Motion algorithm was carried out in order to examine their results on images of passing vehicles. SIFT features were found to be unsuitable for images of vehicles when 3D reconstruction is to be performed. This fact was also demonstrated by related experiments with Structure from Motion. Reconstructions using the Structure from Motion algorithm, which utilizes SIFT correspondences, often failed or produced point clouds with a minimal number of points belonging to the original vehicle.

Therefore, two modifications to the correspondence search stage of Structure from Motion pipeline were proposed. Firstly, SIFT features were substituted by DeepMatching. DeepMatching, which is originally intended for obtaining quasi-dense point matches for optical flow calculation, is utilized to obtain correspondences for the subsequent reconstruction phase. The second modification involves filtering of the computed correspondences using foreground masks in order to eliminate points that are not located on the vehicle. Implementation of both proposed modifications significantly improved the overall completeness of the reconstructed point cloud models of passing vehicles.

References

[1] F. W. Cathey and D. J. Dailey. A novel technique to dynamically measure vehicle speed using uncalibrated roadway cameras. In *Intelligent Vehicles Symposium*, 2005.

[2] M. Dubská, A. Herout, R. Juránek, and J. Sochor. Fully automatic roadside camera calibration for traffic surveillance. *IEEE Trans. on Intelligent Transportation Systems*, 16(3):1162–1171, 2015.

[3] M. Dubská, J. Sochor, and A. Herout. Automatic camera calibration for traffic understanding. In *Proceedings of BMVC 2014*, 2014.

[4] X. C. He and N. H. C. Yung. A novel algorithm for estimating vehicle speed from two consecutive images. In *WACV*, 2007.

[5] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[6] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016.

[7] N Snavely, S Seitz, and R Szeliski. Photo tourism: Exploring image collections in 3D. *ACM Transactions on Graphics*, 2006.

[8] Jakub Sochor, Roman Juránek, and Adam Herout. Traffic surveillance camera calibration by 3d model bounding box alignment for accurate vehicle speed measurement. Preprint submitted to Computer Vision and Image Understanding.

[9] Jakub Sochor, Roman Juránek, Jakub Špaňhel, Lukáš Maršík, Adam Široký, Adam Herout, and Pavel Zemčík. BrnoCompSpeed: Review of traffic camera calibration and a comprehensive dataset for monocular speed measurement. *IEEE Transactions on Intelligent Transportation Systems (under review)*.

[10] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[11] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. DeepFlow: Large displacement optical flow with deep matching. In *ICCV*, Sydney, Australia, 2013.

[12] Changchang Wu. Towards linear-time incremental structure from motion. In *International Conference on 3D Vision*. IEEE, 2013.

[13] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.