Preliminary experiments in multi-view video stitching

Siniša Šegvić*, Marko Ševrović**, Goran Kos***, Vladimir Stanisavljević**** and Ivan Dadić**

* University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia
** University of Zagreb, Faculty of Transportation Sciences, Croatia
*** Institute of Traffic and Communications, Zagreb, Croatia
**** Tilda d.o.o, Zagreb, Croatia
E-mail: sinisa.segvic@fer.hr

***Abstract*—We address the problem of stitching together the three videos acquired by a special rig consisting of three high resolution cameras. The three cameras are placed in the horizontal plane on the top of the service vehicle in a way that the fields of view of the lateral cameras overlap with the field of view of the middle camera. In the presented approach, the transformations between the common parts of the corresponding video frames are approximated by planar projective mappings. The required mappings are estimated by aligning the common parts of the three views in corresponding video frames. The experiments have been performed on production EuroRAP videos provided by our industrial partner. The obtained results confirm that the presented approach would simplify the existing road inspection procedures relying on the recorded multi-view video.**

I. Introduction

Research on traffic accidents in many countries clearly shows that there is an intense need for increasing the safety in road traffic. Even in developed countries with well-designed and well-maintained traffic infrastructure, adequate traffic education, and strict law enforcement, the rates of serious road traffic injuries remain unacceptably high [1]. There are several campaigns over the world which promote conservative premises such as that on average 1 out of each 500 human reactions - is plainly wrong [2]. These campaigns advocate high standards in road infrastructure construction, which would provide enough protection in order to avoid fatal consequences. One of such campaigns is being carrried out through the international programme EuroRAP [3], [4].

In the scope of the EuroRAP programme, the road safety is assessed [5], [6] by analysing video footage acquired simultaneously by three high resolution cameras. The cameras are placed in the horizontal plane on the top of the service vehicle, in a way that the fields of view of the lateral cameras overlap with the field of view of the middle camera. A typical image triple acquired by such platform is shown in Figure 1. The acquired three videos are evaluated by certified experts in traffic safety, which estimate various risk factors for each section of the assessed road. The evaluation process results in a risk map for a given road network, which allows to quantify and compare the safety of the road sections along the considered route [3].

This research has been supported by the Faculty of Transportation Sciences, University of Zagreb.

Unfortunately, it has been found that the experts which perform the road traffic inspection find it qite difficult to follow three different videos at the same time. Thus, there is a necesity to find out a user-friendly arrangement of the three disjoint view onto the road scene. In this paper we take advantage of the fact that the focal points of the three cameras are quite close to each other when compared to the typical distances towards the imaged scene. Consequently, the transformation between the common parts of the corresponding video frames is approximated by a planar projective mapping, or homography [7]. The desired comprehensive view onto the scene is constructed by projecting pixels from the sidewise cameras onto the image plane of the middle camera (in computer vision literature this procedure is known as image stitching [8], [9]). The final result approximates an image which would be acquired by a middle camera equipped with a wide-angle lens.

The paper is organized as follows. Image stitching is briefly reviewed in Section II. Section III presents some details about the EuroRAP programme (including the geometry of the image acquisition rig). The employed lower level computer vision techniques are detailed in Section IV. The obtained experimental results are presented in Section V, while Section VI provides a conclusion and some directions for future work.

II. Image stitching

The purpose of image stitching or image compositing is to process multiple images of the same scene in order to create a high-resolution photo-mosaic in which the seams are as smooth as possible. Today, these techniques are routinely used to produce digital maps and satellite photos. They are also embeded in many digital cameras in order to enable shooting ultra wide-angle panoramas with a conventional inexpensive lens having a horizontal field of view of less than 45°. However, despite the maturity of the lower level building blocks, image stitching still can be a challenging task, which shall also be demonstrated in the rest of this paper.

Image stitching typically consists of the following two tasks: i) image alignment and ii) determining composite pixels. Image alignment tells us which pixels from original images map to a given pixel in the composite image. There are two main approaches to image alignment: feature-based and direct, or pixel-based [9]. Here we

Fig. 1. A typical image triple simultaneously acquired by the standard EuroRAP image acquisition platform.

only consider feature-based alignment since pixel-based approaches are not applicable for significantly displaced views such as as in our case (cf. Figure 1).

Feature-based alignment employs the point correspondences extracted by some wide-baseline matching approach [10], [11] (cf. IV-A) in order to recover parameters of the chosen alignment model. The choice of the alignment model depends on the spatial properties between the multiple viewpoints and the observed scene. It ranges from simple translation, 2D Euclidean transform, 3D rotation [8], projective transformation to the full structure and motion estimation [9].

Many previous works assume a small distance between the original viewpoints and describe the alignment by a homography induced by the plane at infinity. In normalized image coordinates [12] (which are available for calibrated cameras), this homography corresponds to the 3D rotation between the camera pose from which the original image was acquired and the camera pose of the composite image. Brown and Lowe [8] recover both the rotation parameters and the intrinsic camera parameters by a bundle adjustment of all detected features across all available views. Very nice results have been reported on images submitted by collaborating non-expert users. However, this approach can not fully compensate large inter-camera motions, leading to visually suboptimal results (cf. Section V).

Pixels of the composite image $q_C(\mathbf{x})$ are determined from the corresponding pixels $q_k(T_k(\mathbf{x}))$ in each of n original images $\mathbf{I}_k, k \in [0..n\rangle$. In the above expression T_k denotes the transformation between the composite image and the k-th original image. Often, the composite pixels are calculated as a weighted average of the original pixels:

$$q_C(\mathbf{x}) = \sum_k \frac{\alpha_k(\mathbf{x}) \cdot q_k(T_k(\mathbf{x}))}{\sum_l \alpha_l(\mathbf{x})}, \forall k, \mathbf{x} \quad . \qquad (1)$$

Pixels falling outside the original image do not contribute to the composite pixel:

$$\alpha_k(\mathbf{x}) = 0, \forall k, \mathbf{x} : T_k(\mathbf{x}) \notin \mathbf{I_k} \quad . \qquad (2)$$

If we set all remaining α_k to be equal, we obtain a technique called averaging:

$$\alpha_k(\mathbf{x}) = 1, \forall k, \mathbf{x} : T_k(\mathbf{x}) \in \mathbf{I_k} \quad . \qquad (3)$$

A more involved technique sets the contribution of original pixels to be proportional to its distance from the image border $d_{\mathrm{B}}(T_k(\mathbf{x}))$:

$$\alpha_k(\mathbf{x}) = d_{\mathrm{B}}(T_k(\mathbf{x})), \forall k, \mathbf{x} : T_k(\mathbf{x}) \in \mathbf{I_k} \quad . \qquad (4)$$

Recent works in image stitching focus on individual aspects of the problem which arise in specific applications. Uyttendaele et al. [13] considers the problem of extracting composite pixels with high dynamic range starting from original images acquired with different exposure parameters. Mills et al. [14] strive to detect moving objects in original images and avoid including them into the composite image. Chen et al. [15] construct a 360° panorama around a passenger car by stitching images obtained by four wide-angle fish-eye cameras mounted at the four sides of the car. Koo et al. [16] present an alignment model selection approach which finds the model which suits best the actual original images.

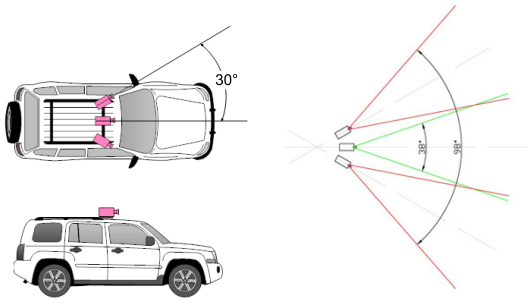
III. The EuroRAP programme

EuroRAP (European Road Assessment Programme)[1] is am international non-profit road safety organisation which aims to reduce traffic injuries on European roads [3], [4]. The programme has been founded by European automobilistic organizations and road authorities in order to improve the safety od the road traffic. Currently EuroRAP unites about 50 partners from 30 countries, including Croatia. The programme has also been supported by leading car manufacturers, as a sister programme to EuroNCAP, the European New Car Assessment Programme. EuroRAP designates different road sections with an easy-to-understand star rating and creates maps with risc factors derived from the accident history of the corresponding road section. EuroRAP also performs special road safety inspections [5] and proposes interventions which are expected to diminish the frequency and/or alleviate the consequences of traffic accidents. During 2008 and 2009 the first pilot EuroRAP projects have been conducted on the most important Croatian roads. The obtained star ratings were found to be in high correlation with the rate of traffic accidents with fatal consequences. This research has been recognized by the Croatian National programme for the road traffic safety, so that the range of operations is likely to be increased in the upcoming period until 2020.

The road safety inspections are performed according to the EuroRAP RPS (road protection score) protocols.

[1]EuroRAP web site is at http://www.eurorap.org.

These protocols define procedures and modalities of collecting and processing the road data. The protocol EuroRAP RPS 2.0 defines the data acquisition to be performed by a service vehicle equipped by three video cameras and a GPS receiver. The service vehicle acquires multi-view georeferenced video [6], which is consequently assessed by trained experts by estimating characteristic attributes of the road section. The three cameras are attached to a rig which ensures that the optical axes of the three cameras are placed in a horizontal plane above the vehicle. The rig also ensures that the middle camera is aligned with the longitudinal axis of the vehicle, while the other two cameras are rotated for 30° one towards the left and the other towards the right. The geometry of the multi-camera rig is shown in Figure 2.



(a) (b)

Fig. 2. The EuroRAP image acquisition platform is shown on the left (a). The angular configuration of the three cameras is shown on the right (b).

When combined with typical higher-consumer-grade cameras having a horizontal field of view of about 45°, the presented geometry of the multi-camera rig ensures a compound field of view of about 105°[2]. This configuration trades-off cost for simplicity: an equivalent industrial camera with a 105° wide-angle lens would probably cost less than three consumer-grade cameras, however the consumer cameras are much more easily procured and configured than custom imaging solutions.

IV. The Employed Computer Vision Techniques

In this section we briefly detail the required lower level computer vision techniques. We first address wide-baseline matching (cf. IV-A) which provides point correspondences required for feature-based alignment. Then we review camera calibration (cf. IV-B) which enables us to express the extracted correspondences in normalized image coordinates. This is useful for correcting radial distortion, and for being able to apply algorithms which assume the calibrated context. Finally we present the procedures for recovering parameters of the projective (cf. IV-C) and the rotational alignment model (cf. IV-D).

A. Wide-baseline matching

Wide-baseline matching is a technique for detecting point correspondences in arbitrary images of the common scene. The state of the art approaches are based on *invariant feature descriptors* [11] which are independently extracted in both images, at characteristic image locations called *keypoints* [17]. Usually the detected keypoints are locally distinctive with respect to position, scale and rotation, while some approaches even address affine invariance.

[2]Notice that this is somewhat smaller than the minimum driver's visual field which is usually prescribed at 120°

(a)

(b)

Fig. 3. The correspondences recovered between the left image and the middle image (a), and between the middle image and the right image (b).

In this paper, we employ a wide-baseline matching technique as described in [10], whereby the keypoints are detected as space-scale extrema of the DoG (difference of Gaussians) response. The obtained descriptors are exhaustively compared against the descriptors from the other image, typically with respect to L2 distance. The correspondences are established as distinctive pairs for which the best distance is less than 60% of the distance of the second-best match. Figure 3 shows the correspondences which are automatically extracted across the original images shown in Figure 1.

B. Camera calibration

Camera calibration allows to treat image pixels in normalized coordinates which provide an immediate connection with the corresponding optical rays. This is important since it i) corrects the radial distortion effects, and ii) allows one to recover metric measurements such as rotation and translation components of the underlying two-image geometry. We employ the usual model for transforming pixels into normalized coordinates comprising of a 5-DOF linear transformation and the fourth order radial distortion model [12]. We recovered calibration parameters for our cameras by employing our own implementation of the procedure with a planar calibration target described in [12]. The required relations are summarised in the following equations:

$$\mathbf{q} = \mathbf{K} \cdot f_R(\hat{\mathbf{q}}, k_1^{ud}, k_2^{ud}) \quad (5)$$

$$\hat{\mathbf{q}} = f_R(\mathbf{K}^{-1} \cdot \mathbf{q}, k_1^{du}, k_2^{du}). \quad (6)$$

The radial function f_R is defined as:

$$f_R(\hat{\mathbf{q}}, k_1, k_2) = \hat{\mathbf{q}} \cdot (1 + k_1 \cdot r_{\hat{\mathbf{q}}}^2 + k_2 \cdot r_{\hat{\mathbf{q}}}^4), \quad (7)$$

where $r^2_{\hat{\mathbf{q}}}$ stands for squared radius of the homogeneous image point $\hat{\mathbf{q}} = (\hat{q}_x, \hat{q}_y, 1)$ and is calculated as:

$$r^2_{\hat{\mathbf{q}}} = \hat{q}_x^2 + \hat{q}_y^2. \tag{8}$$

C. Homography estimation

It is well known that two perspective views of a planar scene can be related by a planar projective transform or homography [7]. If the translational displacement between the two viewpoints is small enough, the same procedure can be applied to arbitrary scenes. This approximation is used in most approaches to image stitching [9]. The standard procedure for estimating a homography from a given set of point correspondences is as follows [7]:

1) removing the outliers by random sampling [18] (four point correspondences are required for generating homography hypotheses);
2) reestimation by a standard linear algorithm, which also known as the direct linear transform [7];
3) iterative improvement by gradient optimization of the nine elements of the homography matrix.

D. Estimation of the 3D rotation

It is well known that the relation between corresponding points at infinity in two views can be described by a homography corresponding to the 3D rotation between the two views. This rotation can be easily recovered by solving the orthogonal Procrustes problem as shown in [19], [9]. The final procedure is analogous to the procedure for recovering the homographies sketched above:

1) removing the outliers by random sampling [18] (three point correspondences are required for generating rotation hypotheses);
2) reestimation based on SVD [19];
3) iterative improvement by gradient optimization of the quaternion representation of the rotation.

V. Experimental results

We first subjectively compare the composite images obtained by employing projective (cf. IV-C) and the rotational model (cf. IV-D), while the composite pixels are obtained by simple averaging. The results are shown in Figure 4. We notice that the seams are clearly visible and that the projective model results in better geometrical alignment of the compound image.

Subsequently, we repeat the previous experiment, but in the case where composite pixels are obtained by weighting contributions by border distance. The results are shown in Figure 5. We notice that the seams are much smoother, bu that the geometrical inadequacy of the rotational model is still visible.

In order to test the correctness of our implementation of the rotational model, we compare our results with the commercial software Autostitch [8]. The results are shown in Figure 6. We notice that the results are quite similar, although Autostitch does better job in photometric adjustment (photometric adjustment is out of the scope of this paper).

The projective model produced reasonably good results in most of the performed experiments. The results obtained in three more frames are shown in Figure 7.

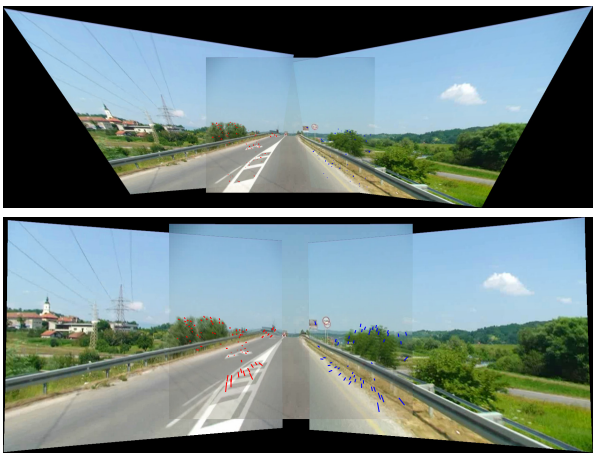


Fig. 4. The compound images obtained with projective (top) and rotational model (bottom). Composite pixels are obtained by simple averaging.

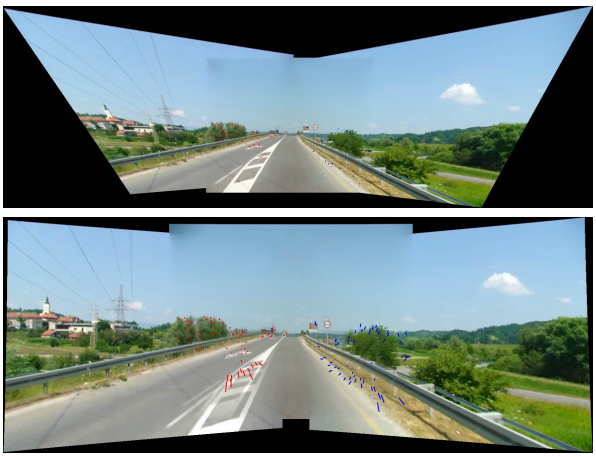


Fig. 5. The compound images obtained with projective (top) and rotational model (bottom). Composite pixels are obtained by weighting contributions by border distance.

Our initial ambition was to calibrate the two projective mappings beforehand, and to employ them to stitch video frames in near real time. Experiments have shown that this operation can be performed in about 150 ms on a modern quad-core processor. Unfortunately, we found that the employed rig hosting the three cameras does not prevent small interframe camera motions due to vehicle vibrations. Thus our best results have been obtained when the mappings are reestimated in each frame triplet, which typically takes about 10 s per triplet. A short video demonstration can be downloaded from:
http://www.zemris.fer.hr/~ssegvic/pubs/mipro11stitch.avi.

VI. Conclusion

We have presented a technique for improving the presentation of video footage acquired with the standard EuroRAP three-camera rig. The common parts of the corresponding video frames are first aligned by a suitable projective mapping, and consequently mapped onto the compound image. Despite the difficulties due to motion parallax, we have obtained encouraging results which would clearly be quite useful in assisting road safety inspection procedures.

It has been found out that enforcing a projectivity

Fig. 6. The compound images obtained with the rotational model (top) and by the application AutoStitch (bottom).

Fig. 7. Three more results with projective model and border-weighted contributions.

induced by the plane at infinity results in compound images with inferior aesthetical quality. This is probably caused by significant distance between the projection centers of the three cameras, which effectively invalidates the assumptions which are required for this model. General projectivities achieve better alignment since in most images the 3D configuration of the scene is better approximated by some other 3D plane.

Further improvements in the presentation of the multi-view EuroRAP video footage would have to take into account the 3D structure of the scene. In the simpler approach, the scene would be approximated by a piece-wise planar model. Image regions projected from different planes could be identified either dynamically (by robust statistical analysis of the point correspondences), or statically by a machine learning approach. Finally, best results would be obtained by recovering depths of all point correspondences, and constructing the compound image by employing the extrapolated depth information in each source pixel.

REFERENCES

[1] L. Fletcher, N. Apostoloff, L. Petersson, and A. Zelinsky, "Vision in and out of vehicles," *IEEE Intelligent Systems*, vol. 18, no. 3, pp. 12–17, 2003.

[2] R. B. Whittingham, *The blame machine: why human error causes accidents*. Elsevier Butterworth-Heinemann, Oxford, 2004.

[3] D. Lynam, J. Castle, J. Martin, D. Lawson, J. Hill, and S. Charman, "Eurorap 2005-06 technical update," *Traffic Engineering and Control*, vol. 48, no. 11, pp. 477–484, 2007.

[4] H. Stigson and J. Hill, "Use of car crashes resulting in fatal and serious injuries to analyze a safe road transport system model and to identify system weaknesses," *Traffic injury prevention*, vol. 10, no. 5, pp. 441–450, 2009.

[5] J. L. Cardoso, C. Stefan, R. Elvik, and M. Sřrensen, "Road safety inspection - best practice guidelines and implementation steps," tech. rep., Deliverable D5 of the EU FP6 project RIPCORD - ISEREST, 2007.

[6] S. Šegvić, K. Brkić, Z. Kalafatić, V. Stanisavljević, M. Ševrović, D. Budimir, and I. Dadić, "A computer vision assisted geoinformation inventory for traffic infrastructure," in *Proceedings of the IEEE Intelligent Transportation Systems Conference*, (Madeira, Portugal), Sept. 2010.

[7] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[8] M. Brown and D. Lowe, "Automatic panoramic image stitching using invariant features," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 59–73, 2007.

[9] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer, 2011.

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

[11] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern recognition and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[12] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern recognition and Machine Intelligence*, vol. 22, pp. 1330–1334, Nov. 2000.

[13] A. Eden, M. Uyttendaele, and R. Szeliski, "Seamless image stitching of scenes with large motions and exposure differences," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 2498–2505, 2006.

[14] A. Mills and G. Dudek, "Image stitching with dynamic elements," *Image and Vision Computing*, vol. 27, no. 10, pp. 1593–1602, 2009.

[15] Y.-Y. Chen, Y.-Y. Tu, C.-H. Chiu, and Y.-S. Chen, "An embedded system for vehicle surrounding monitoring," in *International Conference on Power Electronics and Intelligent Transportation Systems*, vol. 2, pp. 92–95, Dec. 2009.

[16] H. I. Koo, B. S. Kim, and N. I. Cho, "A new method to find an optimal warping function in image stitching," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1289 –1292, Apr. 2009.

[17] T. Tuytelaars and K. Mikolajczyk, *Local Invariant Feature Detectors: A Survey*. Hanover, MA, USA: Now Publishers Inc., 2008.

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

[19] G. H. Golub and C. F. V. Loan, *Matrix Computations*. The Johns Hopkins University Press, 1996.