Constraining Relative Camera Pose Estimation with Pedestrian Detector-Based Correspondence Filters

Emanuel Aldea
SATIE – CNRS UMR 8029
Paris-Sud University
Paris-Saclay University
France
emanuel.aldea@u-psud.fr

Thomas Pollok Chengchao Qu
Fraunhofer IOSB
Fraunhoferstr. 1
76131 Karlsruhe
Germany
firstname.lastname@iosb.fraunhofer.de

Abstract

A prerequisite for using smart camera networks effectively is a precise extrinsic calibration of the camera sensors, either in a fixed coordinate system, or relatively to each other. For cameras with partly overlapping fields of view, the relative pose estimation may be directly performed on or assisted by the video content obtained during scene analysis. In typical conditions however (wide baseline, repetitive patterns, homogeneous appearance of pedestrians), the pose estimation is imprecise and very often is affected by large errors in weakly constrained areas of the field of view. In this work, we propose to rely on progressively stricter constraints on the feature association between the camera views, guided by a pedestrian detector and a re-identification algorithm respectively. The results show that the two strategies are effective in alleviating the ambiguity which is due to the similar appearance of pedestrians in such scenes, and in improving the relative pose estimation.

1. Introduction

Nowadays, there is a growing interest for the setup of surveillance camera systems aimed to improve the safety of large public events. By definition, the ephemeral character of such events implies that the frequent deployment and calibration of such systems should require as little expert supervision as possible. This covers the use of specific calibration objects, or other techniques implying human involvement for guiding the estimation of camera poses towards high quality solutions. For synchronized camera systems, the direct use of the urban scene being viewed is possible, irrespective of the static or dynamic nature of the objects being present. The pedestrian projections may thus contribute directly to solving the relative pose estimation between pairs of cameras with overlapping fields of view, instead of being solely the object of the video analysis.

Many algorithms for estimating the relative pose from image pairs exist ([8, 19]), but most of them are challenged by urban scenes observed with wide baselines, since most of the potential correspondences are unreliable. In such contexts, it is difficult to get a robust estimation from a single image pair, thus a natural extension is to integrate observations temporally from video streams. Some techniques have already been proposed for exploiting the pedestrian trajectories and the generated visual cues [2, 13], but they are mainly limited by two aspects. First of all, dominant planar trajectories introduce degenerate configurations for the epipolar geometry estimation, thus the use of matches from the entire scene is essential. Secondly, the pedestrians tend to have a homogeneous appearance and are prone to generating outlier observations in the form of wrong correspondences. The solution provided by standard guided matching techniques [18] tends to drift progressively and include outliers, but recently [11, 12] proposed a guided matching algorithm aiming to enforce a uniform selection of matches in the common field of view of the cameras. This in turn provides a high-quality estimation locally as well. The main limitation of [11] is that it relies significantly on pedestrian generated matches which are unreliable, since the moving pedestrians are the only objects generating new interest points. They do not perform however a higher level analysis in order to locate pedestrians and potentially match them at object level by re-identification in order to validate the low level interest point associations.

In terms of a higher-level interpretation of the scene content, the topic of semantic segmentation [10, 1] emerged as well recently and its output may also be used in order to guide a low level matching algorithm away from spurious correspondences. There are two main issues related to this approach, the first being the fact that matches are often

located close to the object boundaries, where the segmentation algorithms are the most unreliable. Secondly, when estimating the pose from a video sequence, the pedestrians remain the most useful high-level objects providing new visual cues, the rest of the image content being redundant in the current and in the initial frame. Our work focuses thus on the exploitation of high-level cues related specifically to pedestrians.

The objective of our current work is to integrate additional filtering steps into the general guided matching framework in order to benefit from the latest progress in pedestrian detection and re-identification. As our main contribution, we show that a relatively straightforward procedure for adding the object level filtering into a camera pose estimation framework further improves the accuracy of the solution.

The next section describes the general framework and the filtering we propose. Then, we detail the the utilized evaluation protocol, and continue with the presentation of the experiments. Finally we conclude with a discussion of the obtained results and with some further perspectives.

2. Method

2.1. Guided matching from video

The core idea of the guided matching from video introduced in [11] is that in image areas not covered by matches, the search for new matches should be performed within an epipolar band with a width reflecting the absence of matches constraining the search reliably. This assertion may be expressed as:

$$\Sigma_l = J_F \Sigma_F J_F^T + \sigma^2 J_p J_p^T. \quad (1)$$

where J_F and Σ_F are the Jacobian and the covariance of the fundamental matrix F, J_p is the Jacobian of the interest point p location, and σ is a location uncertainty which is related to the local interest point density. Depending on a local density estimation based on the DBSCAN algorithm [3], the uncertainty is set to a high value σ_H which will promote the selection of matches from new image pairs in that area, or conversely it is set to a low value σ_L which will prevent the selection of too many points into an already well covered area. This last scenario is detrimental and needs to be avoided as it constrains the fundamental matrix to locally optimal configurations.

Each new image pair in the video sequence provides new matches, mainly in the low density areas, which are integrated into the existing match set and validated using a robust estimation algorithm such as RANSAC [4] or ORSA [9]. However, due to pedestrian homogeneity or to their unfavourable instantaneous layout, outlier observations may still be progressively added to the solution, degrading it. In the following, we assume that we dispose of either (a) a pedestrian **detection** algorithm providing a rough border $\mathcal{B}$ for each pedestrian (*e.g.*, a bounding box), or (b) a pedestrian **re-identification** algorithm providing, in addition to $\mathcal{B}$, a unique identifier `id` across all the camera views. We will summarize in the next paragraphs the straightforward use of these additional algorithms.

Box filtering The most common scenario is the one using a pedestrian detector, since this type of algorithm may run on lower resolution data. Assuming that we intend to validate a new match defined by two interest points in the image pair $m = (p, p')$, the additional constraint that must be checked by m before being included into the current inlier set is:

$$\begin{aligned}(\exists \mathcal{B}, p \in \mathcal{B} \Rightarrow \exists \mathcal{B}', p' \in \mathcal{B}') \vee \\ (\nexists \mathcal{B}, p \in \mathcal{B} \Rightarrow \nexists \mathcal{B}', p' \in \mathcal{B}')\end{aligned} \quad (2)$$

Re-id filtering In addition to the previous condition, one can add a stricter constraint for corners being located on pedestrians whenever `id` is available:

$$\begin{aligned}(\exists \mathcal{B}, p \in \mathcal{B} \Rightarrow \exists \mathcal{B}', p' \in \mathcal{B}' \wedge id(\mathcal{B}) == id(\mathcal{B}')) \vee \\ (\nexists \mathcal{B}, p \in \mathcal{B} \Rightarrow \nexists \mathcal{B}', p' \in \mathcal{B}')\end{aligned} \quad (3)$$

In the following part of this section, we will discuss the specific detection and re-identification algorithms considered in this work.

2.2. Association refinement

Pedestrian detectors Generic object detection has gained a huge leap thanks to the recent advances of convolutional neural networks (CNNs) in the past few years. Following the traditional workflow for object detection with region selection, feature extraction and classification, the seminal works of R-CNN families [6, 5, 17] achieve accurate localization for diverse object classes using a two-stage approach consisting of region proposals and fine localization and classification. You Only Look Once (YOLO) [14], in comparison, is famous for its real-time capability by virtue of the single network architecture, with the compromise of lower detection rate especially for small targets. However, later versions YOLOv2 [15] and YOLOv3 [16] successfully address this problem using a number of improvements, such as more efficient network backbones, multi-scale training and feature pyramids, where the performance is close to state-of-the-art two-stage methods with large advantage in efficiency.

Hence, we apply YOLOv3 for detecting humans in our camera images. A noticeable challenge in some surveillance data, such as one of the datasets considered in our experiments, is the fact that the video information is in single-channel grayscale format, while the YOLOv3 models are

obtained based on RGB images. Moreover, the high contrast with overexposed ground and cast shadow increase the difficulty of the detection task for the pretrained network.

Re-identification algorithms Person re-identification, or re-id in short, is a high-level task for finding occurrences of the same person given a query image based on the full-body appearance. Often as a follow-up task after person detection in the analysis pipeline, re-id provides the same ID to a person that has been seen before, offering more fine-grained association information than with pure detections alone. This is of special interest for the work in this paper thanks to the exact correspondences of persons seen across the cameras, which is able to introduce an additional restriction as detailed in Equation 3.

One of the most relevant features for re-id is the color information, often complemented with low-level information such as contour or texture, and additional semantic cues, *e.g.*, pose and soft-biometric attributes. Recently, CNNs are gaining popularity as the most successful feature extractor for re-id. The objective of training the network is to ensure high discriminative power for the output feature vector given a full-body image crop as input, such that the distance of the embeddings in the feature space for the same person is close, and large for different persons, even with the existence of several disturbing factors such as partial occlusion, different viewpoints and body poses.

This work leverages the `person-reidentification-retail-0076` model from the OpenCV model zoo[1], which offers state-of-the-art accuracy on benchmark datasets. Given a crop of a detection bounding box resized to 128×384 pixels as the input, the re-id head outputs a 256D embedding vector after going through the RMNet backbone. Afterwards, distances of the embedding vectors for all detected person pairs across the camera views are computed using the Cosine metric. Image pairs with a distance below a predefined threshold are considered to belong to the same person.

To conclude this section, we underline that for our filtering approach we considered two mainstream algorithms which have the merit of being validated and widely used by the community. While novel solutions for detection and re-identification are regularly proposed, it is expected that their improved performance would also have an immediate positive impact on the filtering as well. A broader study encompassing various families of detectors and re-id algorithms is out of the scope of our current paper, and will be done in future work.

[1] https://github.com/opencv/open_model_zoo

3. Evaluation metric

When estimating a relative pose, it is crucial to have a reliable ground truth but in the case of large scale urban scenes, there is no simple strategy to acquire such ground truth. Thus, we rely on a manual annotation procedure which allows to evaluate to the best of our capabilities the accuracy of the estimated poses.

In a first step, a large number of manual matches are performed on convenient image pairs from the video stream, guided by a grid overlaid on the image space in order to enforce a uniform distribution. Let us denote this initial set of manual matches as $\mathcal{S}_0$.

Despite the annotator's carefulness, occasional gross errors may be present in $\mathcal{S}_0$. Therefore, an iterative process is performed, in which a RANSAC algorithm with a very small tolerance threshold (0.5 px) is applied in order to identify the largest errors, and the corresponding matches are either corrected or removed. In the end, we get a set of high quality matches denoted as $\mathcal{S}$.

Assuming that a relative pose estimation algorithm provides a fundamental matrix F encoding the relationship between the two analyzed cameras, we compute, based on $\mathcal{S}$, the RMSE of this set with respect to the estimated F [7]:

$$\begin{aligned} RMSE_F(\mathcal{S}) &= \sum_{(p,p')\in\mathcal{S}} (d^2(p,l) + d^2(p',l')) \\ &= \sum_{(p,p')\in\mathcal{S}} \left(\frac{1}{l_1^2+l_2^2} + \frac{1}{l'^2_1+l'^2_2}\right) (p'^T F p)^2 \end{aligned} \tag{4}$$

where l and l' represent the epipolar lines corresponding to p and p' respectively.

Additionally, we also compute the Maximum geometric error in the form of:

$$ME_F(\mathcal{S}) = \max_{(p,p')} \max \left(\frac{p'^T F p}{\sqrt{l_1^2+l_2^2}}, \frac{p'^T F p}{\sqrt{l'^2_1+l'^2_2}} \right) \tag{5}$$

While the use of the RMSE is rather classical, the use of the Max Error is the strictest possible metric that one can use for validating F against $\mathcal{S}$, and is essential for highlighting the local quality of a pose estimation. On the contrary, the RMSE illustrates the overall fitness of F across the field of view, but it may hide localized errors.

4. Experimental results

4.1. Datasets and implementation

For validating the proposed algorithms, we use two datasets. The first dataset, denoted as Dataset1 or D1 in the rest of the section, is a real world sequence of a medium density courtyard of a mosque, around prayer time,

(a) (b) (c) (d)

Figure 1: (a) and (b) Synchronized image pair from Dataset1 (c) and (d) Synchronized image pair from Dataset2

Figure 2: A typical example of sample boxes provided by the pedestrian detector.

also used by the baseline algorithm [11]. The images are grayscale, and the camera pose estimation is difficult to perform due to the homogeneity of the pedestrians and to the repeating architectural patterns. For the continuous calibration, one hundred image pairs are used, sampled at an interval of 3 seconds in order to allow pedestrians to move significantly. On these images, the pedestrian detector, outputting a list of bounding boxes for each frame, is used. Surprisingly, YOLOv3 is still able to detect most of the separate persons in the images, even those with small sizes at a large distance to the cameras. However, the crowd of people sitting and praying on the lower left corner are mostly missed (see Figure 2), probably as a result of the cluttered black-white and less discriminative textures. Lastly, the re-identification algorithm cannot be applied due to the low resolution of the pedestrians and due to the missing RGB information.

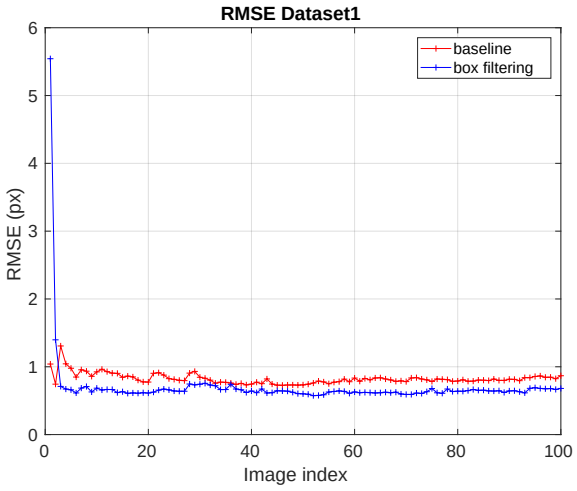
The second dataset, denoted in the following as Dataset2 or D2, contains four volunteers following free trajectories in an open planar area, with very few static elements on the sides in the fields of view of the cameras to be calibrated. The difficulty in this case is to obtain a calibration with a good local precision across all the field of view, due to the scarcity of visually salient features. The images are RGB with a good resolution and obviously, the applied re-id model works consistently well on almost all images of the dataset, independent of the size and body pose, while being robust to occlusion and crossing as well. A sample image pair from both datasets is presented in Figure 1.

The implementation is based on the open source code of [11], which is available online[2]. Not taking into account the detection/re-identification algorithm running time, the filtering strategy is called during the guided matching, and the additional computational cost is negligible with respect to the baseline algorithm.

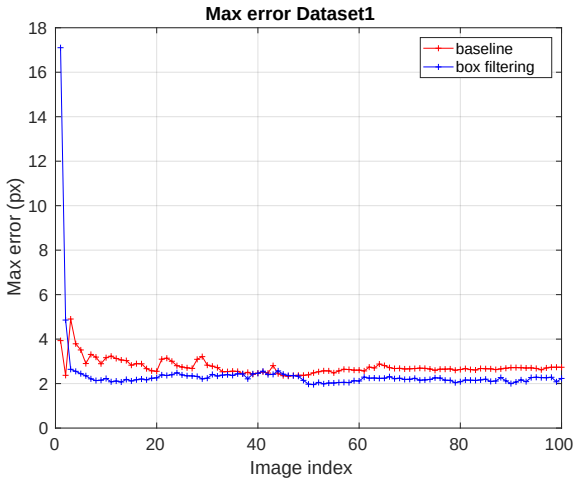
4.2. Quantitative evaluation – D1

We present in Figure 3 the evolution of the RMSE and of the Max geometric error for Dataset1. The additional check introduced is based on the detection boxes associated to the pedestrians. The RMSE is evaluated across all the common field of view and, as a global measure of quality, does not underline the final difference introduced by the additional box filtering. However, the final RMSE value is improved across all the image from 0.87 to 0.68 px. The impact is more visible for the ME, which is reduced from 2.73 to 2.23 px, *i.e.*, over 18%. These gains have a significant positive impact for these such small targets, for applications which perform tasks in multiple cameras such as multi-view detection, accurate localization or gait analysis.

Note also that for filtering strategies the initial estimation errors may be higher than the ones of the baseline algorithm, due to the fact that fewer matches are used in the beginning. However, since the matches are more consistent in terms of associating pedestrians correctly to pedestrians, the benefits become visible progressively during the estimation. An additional drawback of the large initial values is that the plots are squished toward the final part of the graphic; for exact numerical values, the reader is referred to Table 1.

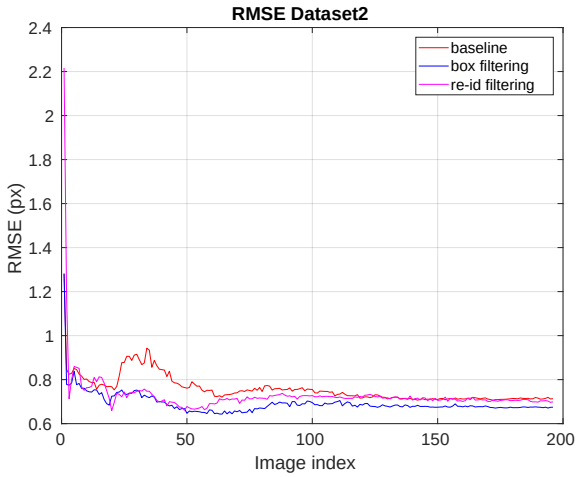


(a)

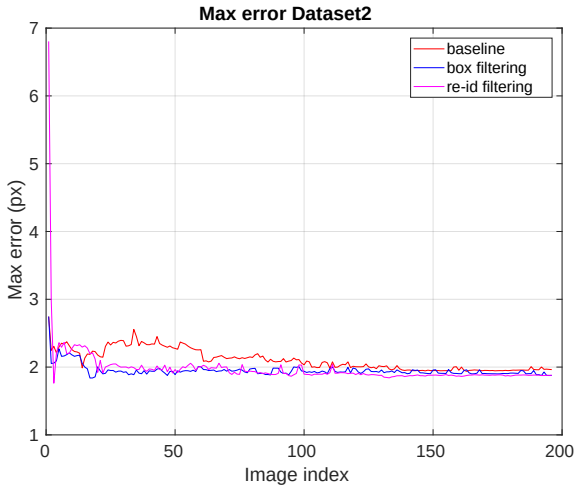


(b)

Figure 3: (a) RMSE and (b) Max error for Dataset1. The baseline estimation algorithm and the box filtering strategy are compared.



(a)



(b)

Figure 4: (a) RMSE and (b) Max error for Dataset2. The baseline estimation algorithm, the box filtering and the re-id filtering strategies are compared.

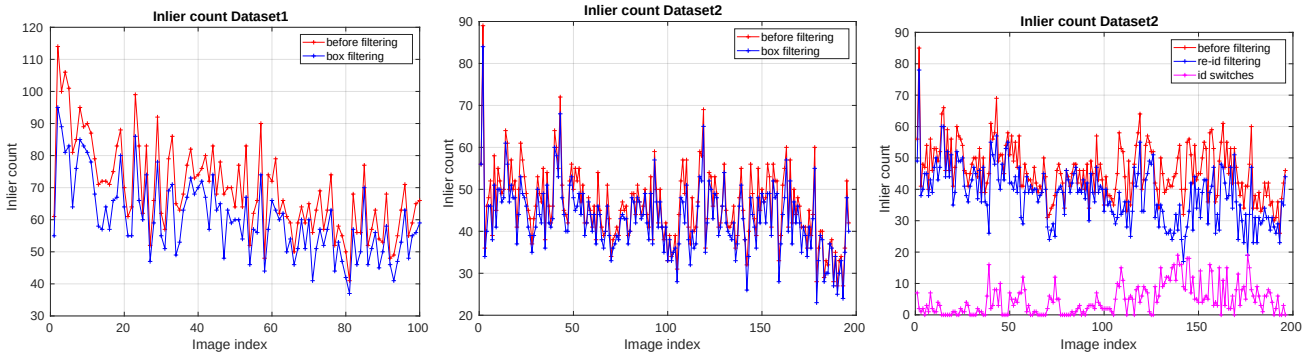
4.3. Quantitative evaluation – D2

We present in Figure 4 the evolution of the RMSE and of the Max geometric error for Dataset2. In addition to the pedestrian detection box filtering, we run on this dataset a re-identification check in order to reject pedestrian-to-pedestrian matches which are not validated by an identical id. For this dataset, the difference regarding the global quality measure (RMSE) are less significant, but in terms of Max geometric error, both methods achieve a better performance than the baseline. Overall, we can also notice from Figures 4a and that the two filtering strategies manage to stabilize the solution faster, due to the fact that they reject more aggressively the spurious associations. The explanation for which the effectiveness of the re-id filtering is limited compared to the one of the simpler box check is twofold. First of all, the accumulation of observations for the second filtering strategy is slower, and it would probably benefit more from additional observations. Secondly, the presence of a higher number of pedestrians would also increase the error rate for the two other algorithms and degrade their performance relatively to this last one.

4.4. Summary and discussion

Beyond the relative improvement in accuracy of the new poses, it is also important to visualize to what extent filtering is used actively during pose estimation. In order to present the frequency of the the match rejections performed by the two filtering strategies we employed, we refer the

[2] https://github.com/MOHICANS-project/fundvid



(a) (b) (c)

Figure 5: (a) Filtering for Dataset1 (box detections) (b) Filtering for Dataset2 (box detections) (c) Filtering for Dataset2 (re-id information)

reader to Figure 5. In all the graphs, the red plot represents for each frame the number of new matches which were validated by [11] as inliers and which would be directly integrated into the global inlier set. In Figures 5a and 5b, we present in blue the number of matches which are still accepted after the box filtering step, for Dataset1 and Dataset2 respectively, and we notice that a significant ratio (approximately 10%) of matches are rejected. As expected, the number of rejected matches decreases, as the global solution is guided towards the correct one. The magenta plot in Figure 5c shows the *subset* of rejected matches due to the re-identification check, which is not present in the previous two Figures. One may thus notice that even in the presence of a limited number of pedestrians, the confusions due to the association of wrong ids are quite frequent if only a local appearance similarity is employed.

An additional remark related to the magenta plot in Figure 5c is that it sums at the same time the erroneous matches rejected by the re-identification and the potential good matches rejected due to re-identification errors. Overall, the rejection of good matches due to re-identification errors is preferable as it will just slow down the pose estimation convergence, but it will not cause the algorithm to drift towards a wrong pose.

Finally, Table 1 presents the values of the RMSE and ME for the considered strategies and the two datasets, at the beginning and at the end of the video sequence based estimation. Relatively to the initial errors, the improvements obtained using a relatively straightforward match filtering strategy are very interesting, as subpixel improvements in the image domain translate for 3D related tasks in significant improvements of the metric scale estimations for urban scene dimensions.

Figure 5 and Table 1 underline that for calibrating cameras with overlapping fields of view in complex urban scenes, the pedestrian detectors or, when applicable, the re-identification algorithms may guide effectively the relative pose estimation.

Algorithm	**RMSE**$_i$	**RMSE**$_f$	**ME**$_i$	**ME**$_f$
Baseline D1	1.04	0.87	3.93	2.73
Box D1	5.54	**0.68**	17.10	**2.23**
Baseline D2	1.28	0.71	2.75	1.96
Box D2	1.28	**0.67**	2.75	**1.88**
Re-id D2	2.21	0.70	6.80	**1.88**

Table 1: Comparison between the initial (columns 2 and 4) and final (columns 3 and 5) performance indicators, for the two datasets (D1 and D2) and for the three algorithms used: the baseline estimator, the detection box filtering and the re-identification based filtering.

5. Conclusion

In this paper, we propose an additional filtering step for relative camera pose estimation in urban scenes, derived from the presence of moving pedestrians which are used as calibration instruments. The additional level of match consistency is enforced either based on a pedestrian detector, or when possible on a more accurate re-identification algorithm. The results show that global geometric consistency and local visual consistency algorithms may be assisted significantly by these additional checks.

In future work, we intend to explore further the link between the re-identification and pose estimation in order to improve jointly the performance of these two different components of our proposed algorithm. We also intend to study the influence on our algorithm of different families of detectors and re-identification algorithms at medium and high

densities.

Acknowledgment

This study was partially supported by the S^{2}UCRE project, co-funded by the German Federal Ministry of Education and Research (BMBF) under grant 13N14463 and by the French National Research Agency (ANR) under grant ANR-16-SEBM-0001.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 1

[2] Y. Caspi, D. Simakov, and M. Irani. Feature-based sequence-to-sequence matching. *Int. J. Comp. Vis.*, 68(1):53–64, 2006. 1

[3] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996. 2

[4] 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*, 24(6):381–395, 1981. 2

[5] R. Girshick. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 2

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 2

[7] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 3

[8] D. Martinec and T. Pajdla. Robust rotation and translation estimation in multiview reconstruction. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. 1

[9] L. Moisan and B. Stival. A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix. *Int. J. Comp. Vis.*, 57(3):201–218, 2004. 2

[10] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 1

[11] N. Pellicanò, E. Aldea, and S. L. Hégarat-Mascle. Robust wide baseline pose estimation from video. In *23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, December 4-8, 2016*, pages 3820–3825, 2016. 1, 2, 4, 6

[12] N. Pellicanò, E. Aldea, and S. Le Hégarat-Mascle. Wide baseline pose estimation from video with a density-based uncertainty model. *Machine Vision and Applications*, Jun 2019. 1

[13] A. Ravichandran and R. Vidal. Video registration using dynamic textures. *Patt. Anal. Mach. Intell.*, 33(1):158–171, 2011. 1

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 2

[15] J. Redmon and A. Farhadi. YOLO9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017. 2

[16] J. Redmon and A. Farhadi. YOLOv3: An incremental improvement. In *arXiv:1804.02767*, 2018. 2

[17] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 2

[18] F. Sur, N. Noury, and M.-O. Berger. Computing the uncertainty of the 8 point algorithm for fundamental matrix estimation. In *19th British Machine Vision Conference-BMVC 2008*, page 10, 2008. 1

[19] X. Tan, C. Sun, X. Sirault, R. Furbank, and T. D. Pham. Feature matching in stereo images encouraging uniform spatial distribution. *Pattern Recognition*, 48(8):2530–2542, 2015. 1