Master Thesis Proposal

Disparate Lens-Stiching

Author:
Max Geiselbrechtinger, BSc.
Mat.Nr.:01609418

Supervisors:
Univ.Prof. Dipl.-Ing. Dr.techn. Axel Jantsch
Dipl.-Ing. Martin Lechner, BSc.

April 11, 2022

1 Motivation and Problem Statement

Image stitching describes the process of aligning images from different perspectives to form seamless photo-mosaics. This procedure has a multitude of use cases, from the creation of ultra wide-angle panoramas on smartphones to high-resolution satellite maps [Sze22].

Another application can be found in the emerging field of autonomous vehicles or driver assistant systems as demonstrated in [WYL20]. These tasks comprise of embedded systems that provide ubiquitous and consistent surveillance of the vehicles surrounding to enable safe and reliable operation. In order to gain a wide-area of observation a variety of different technologies can be utilized and even combined together (eg. LiDAR, RADAR, etc.). Digital cameras present a cheap and flexible solution as they are able to convey huge amounts of information and can therefore be utilized for various tasks in the area of autonomous locomotion. However a major drawback of cameras is their limited field of view (FOV). This issue is usually handled by distributing multiple cameras across the vehicle. To form a continuous and consistent view given a set of partially overlapping camera images a stitching pipeline can be used.

In case the camera positions are precisely calibrated the stitching task can be efficiently solved by exploiting epipolar geometry constraints [Sze22]. However for wide baselines this requires robust and usually bulky mounting frames that would increase both the vehicles weight and manufacturing costs. To provide large FOV surveillance without the structural requirements of calibrated camera setups different image registration techniques can be deployed. In particular the registration process for partially overlapping images can be approached in two ways. Either by local feature detection and matching followed by the estimation

of the transformation parameters. Or by a more global approach in which a pixel-based loss function that relates the images according to the parametric model is directly minimized. Modern systems usually rely on the feature-based variants as they can be designed to work robust and efficient even over wide view-point changes and substantial occlusion where global pixel-to-pixel based methods mostly fail [Sze22].

As reported by Jin et al. [JMM+21] the development of feature detectors has come a long way. Today there exists a whole catalog of methods on how to reliably detect and describe local features. With the advent of deep learning further approaches which utilize Convolutional Neural Networks (CNN) to learn features and their descriptors have been developed. Balntas et al. show in [BLVM17] that it is difficult to establish a general performance ranking given the large quantity of methods and the absence of defacto standards for evaluation metrics and datasets. These circumstances impede the decision process for feature matchers in image stitching pipelines.

In general image matching suffers from problems such as texture less scenes, insufficient regions of overlap, strong changes in illumination, ghosting, etc. In the context of autonomous locomotion effects like ghosting or occlusion of features are not really severe as the image acquisition usually is synchronous and the view point changes between cameras are not too drastic. However given the immense sensory input required for autonomously acting systems modern vehicles cannot simply add a dedicated camera systems for each new task. Therefore the same cameras may have to serve as input for multiple different processes. Thus require the image stitching pipeline to deal with the alignment of an inhomogeneous pool of images. Such images may have varying FOVs, resolution differences of ratios up to 1:7 and possibly only small regions of overlap. All these properties can have a detrimental effect on the performance of image stitching algorithms.

2 Aim of the Work

In this thesis a sample of prominent classical feature detection and matching algorithms will be analyzed according to their performance in real-world applications such as a driver-assistant system. For this a specific dataset has to be compiled which mimics the fundamental aspects of this use case. To do so we will use consecutive frames of the CMU Seasons dataset [SMT+18] and establish ground truth through algorithmic refinement of hand labeled image correspondences. To address issues of the inhomogeneous camera setup we will augment images of the dataset with transformations that mimic the setup. Besides evaluating classic feature matching approaches on this dataset, the utilization of novel deep learning based feature matching and description methods is explored. As a further step the feature matchings performance deterioration due to inhomogeneous camera setups is addressed by preprocessing images according to task specific information which should assist state-of-the-art algorithms. In addition a selection of promising feature detection methods will be evaluated on an NVIDIA Jetson™ platform to asses their performance under the computational

constraints and power limitations of embedded hardware. By these means a reliable image stitching pipeline for autonomous vehicles shall be established.

3 Methodological Approach

1. **Literature Review:**
   An extensive survey of the vast quantity of existing feature detection and matching algorithms is performed. Furthermore a qualitative understanding of all steps underlying an image stitching pipeline has to be gained in order to asses required design decisions.

2. **Model and Data Preparations:**
   A process to generate a representative dataset with accurate ground truth is implemented and executed. Also a model of a realistic camera setup is derived to augment the dataset with synthetically transformed images.

3. **Evaluation of State-of-the-Art:**
   A selection of state-of-the-art feature matching and detection algorithms, hand crafted as well as learned ones, are evaluated on the novel dataset to asses their utility in a real-world autonomous driving context. Furthermore promising algorithms are implemented on an embedded hardware platform to asses their applicability under realistic constraints.

4. **Algorithmic Adaptation:**
   Additional measures are implemented to tackle the peculiarities of the disparate lens camera setup to derive a reliably working image stitching pipeline.

4 State-of-the-Art

Szeliski provides a very good overview of the image stitching problem as well as descriptions to some classical approaches along with the necessary background knowledge in his Computer Vision and Applications book [Sze22]. The classical image stitching pipeline consists of first identifying repeatable keypoints in the images followed by the construction of local descriptors for all keypoints. The obtained descriptors of two or more images are then matched by minimizing a distance metric. It is crucial that the descriptors are designed to be invariant to certain scene and view-point changes in order for the matching to be effective. To evaluate the geometric transformation that is described by the resulting correspondences regression method, that is insensitive to mismatches, is performed. Finally the resulting transformation is applied to map/warp the images onto a composition surface to form a consistent view.

The advancements of CNNs as powerful feature extractors transferred the spark of deep learning to the field of computer vision. While the structure of classical image stitching pipelines has not fundamentally changed over the past the debate on how/where neural networks shall contribute to this process is

not settled yet. Research thereby pursues three main paths, whether to learn detection only, description only or both tasks jointly. Tian et al. [TBN+20] claim that the separated structure of is beneficial for enhanced generalization. Also the results of Jin et al. in [JMM+21] show that the holistic (end-to-end) learning approaches are not yet able to achieve the accuracy of hand-crafted or partially learned feature detectors.

Handcrafted Features

One of the most prominent feature detection and matching algorithms for classic image stitching pipelines is SIFT [Low04]. Although it has shown very reliable in a lot of applications and benchmarks it requires large amounts of computation and memory. This inspired the development of more resource efficient derivatives such as for example SURF [BETVG08] or move to binary descriptors as in ORB [RRKB11]. In [AZ12] methods for boosting feature detector performance are introduced (RootSIFT, etc.). The comparisons in [JMM+21] show that also AKAZE [AS11] performs reasonably good and additionally uses a sophisticated method to determine the scale space which might be relevant to alleviate the issues due to the large resolution discrepancies of inhomogeneous multi-camera setups.

Learned Features

As already mentioned there exist end-to-end learning approaches such as Key.Net [BLRPM19] but we refrain from investigating further on them as they have not yet proven ready for field deployment as assessed in [JMM+21]. Therefore we focus on methods that learn descriptors from patches around keypoints that have been extracted by localizing extrema in Difference of Gaussian (DoG) pyramids. Many of the descriptors are learned using CNNs based on the L2-Net [TFW17], for example HardNet [MMRM17], which performed good in the benchmarks [JMM+21] and [BLV+20]. TFeat [BRPM16] is based on a shallow CNN which should reduce the computation required for the inference process which is beneficial if run on an embedded platform. There also exist inverted approaches such as D2D from Tian et al. [TBN+20] which learns a dense set of descriptors and selects keypoints among them using a saliency metric.

As described by Szeliski [Sze22] the matching process for image stitching is usually based on finding the nearest neighbors of descriptors to form correspondences between keypoint. However there exists a broad spectrum of heuristics and approximation methods to speed up the search process. Also some descriptors, for example binary descriptors as used by ORB, can leverage special hardware accelerated operations to speed up the matching.

The estimation of the motion models parameters has to take care of possible mismatches from the image registration step. This can be achieved by using RANdom Sample Consensus (RANSAC) to determine a good set of inliers which serve as the input for a Least Squares (LS) estimator [Sze22]. Another approach

is to directly use LS with a robust objective function such as Least Median of Squares. Jin et. al [JMM+21] provide an extensive summary of parameters and extension to RANSAC methods used in image registration.

To evaluate the resulting image stitching procedures data with corresponding ground truth is needed. A dataset for image matching usually consists of image pairs or sets that contain the same scene captured from different viewpoints. The ground truth is provided as parameters to the relating motion model. In the most general form this constitutes a projective transformation matrix [MS04]. Another type of dataset consists of a collection of image patches together with a listing of correspondences in this patch-set [BLVM17]. This approach is well suited for the training of CNN based descriptors but leaves out the step of determining the keypoints. Most of the popular datasets for image matching contain buildings and sights (eg. Photo Tourism [WB07], Oxford-Affine [MS05], Rome Patches [PDH+15]) which is not really representative for autonomous driving tasks. The car-centric datasets (KITTI [GLU12], RobotCar [MPLN17], CMU Seasons [SMT+18]), however are usually focused on different task an therefore don't contain appropriate ground truth or have the cameras set up without overlapping FOVs. Which renders them also unsuitable for our purposes. Wang et al. [WYL20] simulate a driving environment using the CARLA simulator [DRC+17] and stitch renderings with different lighting and weather conditions. Although the renderings are quite realistic most of the scenes are certainly not as detail rich as real-world photographs are.

5 Relevance to the Curricula of Computer Engineering

This thesis explores classical as well as machine-learning based computer vision algorithms. These algorithms are to be deployed in novel Cyber-Physical Systems (eg. autonomous vehicles). These systems require robust and efficient algorithms to process visual inputs that enable them to interact with their environment. The fundamental knowledge to understand and design such algorithms is taught in the Computer Engineering curriculum. Besides the numerous courses that provide necessary mathematical and computer scientific skills the following courses can be closely linked to the subject of this thesis.

- 182.763 Stochastic Foundations of Cyber-Physical Systems
- 376.054 Machine Vision and Cognitive Robotics
- 389.166 Signal Processing 1
- 389.170 Signal Processing 2
- 389.119 Parameter Estimation Methods

References

[AS11] Pablo F Alcantarilla and T Solutions. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE Trans. Patt. Anal. Mach. Intell*, 34(7):1281–1298, 2011.

[AZ12] Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2911–2918, 2012.

[BETVG08] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[BLRPM19] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key.net: Keypoint detection by handcrafted and learned cnn filters. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5836–5844, 2019.

[BLV+20] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, Tinne Tuytelaars, Jiri Matas, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2825–2841, 2020.

[BLVM17] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5173–5182, 2017.

[BRPM16] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Bmvc*, volume 1, page 3, 2016.

[DRC+17] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[GLU12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[JMM+21] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2):517–547, 2021.

[Low04] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[MMRM17] Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. *Advances in neural information processing systems*, 30, 2017.

[MPLN17] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017.

[MS04] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International journal of computer vision*, 60(1):63–86, 2004.

[MS05] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 27(10):1615–1630, 2005.

[PDH+15] Mattis Paulin, Matthijs Douze, Zaid Harchaoui, Julien Mairal, Florent Perronin, and Cordelia Schmid. Local convolutional features with unsupervised training for image retrieval. In *Proceedings of the IEEE international conference on computer vision*, pages 91–99, 2015.

[RRKB11] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. IEEE, 2011.

[SMT+18] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8601–8610, 2018.

[Sze22] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer, 2nd edition, 2022.

[TBN+20] Yurun Tian, Vassileios Balntas, Tony Ng, Axel Barroso-Laguna, Yiannis Demiris, and Krystian Mikolajczyk. D2d: Keypoint extraction with describe to detect approach. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[TFW17] Yurun Tian, Bin Fan, and Fuchao Wu. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 661–669, 2017.

[WB07] Simon AJ Winder and Matthew Brown. Learning local image descriptors. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[WYL20] Lang Wang, Wen Yu, and Bao Li. Multi-scenes image stitching based on autonomous driving. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, volume 1, pages 694–698, 2020.