Learning to Detect Light Field Features

Kelly Guan
Energy Resources Engineering
Stanford University
kmguan@stanford.edu

Farah Memon
Bioengineering
Stanford University
farahm@stanford.edu

Lars Jebe
Electrical Engineering
Stanford University
larsjebe@stanford.edu

***Abstract*—Feature detection and description is the basis for much of computer vision. The ability to detect features in images quickly and reliably is necessary for a wide range of applications, including object recognition, change detection and 3D reconstruction. This project will focus on a learning-based approach to detecting features in order to obtain Structure from Motion (SfM). This 3D-reconstruction technique is used for example in robots that have to navigate autonomously through unknown and quickly changing environments, such as self-driving cars.**

I. Motivation

Conventional methods for feature detection on 2D-images such as SIFT [1] or learning-based methods [2–4] work well in environments with sufficient illumination and without reflective or refractive surfaces. They do not work well in challenging environments that e.g. contain many occlusions or are low-light. Light field images contain a lot of information about the 3D scene they captured and about surface textures. If we can find a way to leverage this information that is contained in the third and fourth dimension of the light field image, feature detection has the potential to work much more reliably.

II. Dataset

The dataset contains 4251 light field images indoor and outdoor scenes in 31 different categories. Each scene is captured from 3-6 different viewpoints. The complete dataset has a size of approximately 200 GB.

(a) Raw light field (b) Rendered image

Fig. 1: Rendered and raw light field image

III. Task Definition and Scope

The class project is part of a larger research project ongoing at Stanford University on learning feature detection, description and matching from light field images. The goal and milestones below reflect only a subset of the complete project. Depending on how fast the project progresses, additional milestones might be added throughout the course of the quarter.

A. Goal

This project will focus on the detection of features only. The goal of this project is the development of a convolutional neural network that classifies input patches from light field images as good or bad features. The closely related task of feature description and matching lies outside the scope of this project.

B. Milestones

1) **Proof of Concept.** Comparison between a 2D and a 3D convolutional autoencoder for light field image slices to show that prior information about the third light field dimension can help the network to understand the inherent structure of a light field.
2) **Ground truth sampling.** The dataset is unlabeled. To train the network, examples of 'good' and 'bad' patches are needed. Bad patches will be sampled randomly from the data; good patches are obtained by using 2D SIFT and rejecting all features that are not used by COLMAP 3D-scene matching [5].
3) **Baseline 2D detector network**. Design of a CNN that uses the labeled patches to learn to classify input patches as 'good' or 'bad', similar to the approaches in [3, 4], and evaluation of the results against state-of-the-art 2D learned features [6].
4) **(optional) Pseudo-4D Convolution.** Implementation of a CNN that uses light field-specific higher-dimensional custom convolutions that can be calculated much faster than full 4D convolutions and can capture more relevant information about the image than a 2D network.

REFERENCES

[1] David G. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints". In: *International Journal of Computer Vision* 60.2 (2004), pp. 91–110.

[2] Xufeng Han et al. "Matchnet: Unifying feature and metric learning for patch-based matching". In: *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE. 2015, pp. 3279–3286.

[3] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. "SuperPoint: Self-Supervised Interest Point Detection and Description". In: *arXiv preprint arXiv:1712.07629* (2017).

[4] H. Altwaijry et al. "Learning to Match Aerial Images with Deep Attentive Architectures". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 3539–3547.

[5] Johannes Lutz Schönberger and Jan-Michael Frahm. "Structure-from-Motion Revisited". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[6] Johannes Lutz Schönberger et al. "Comparative Evaluation of Hand-Crafted and Learned Local Features". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.