Spatio-Temporal Feature Matching for Time-Varying Structure from Motion

Keywords: structure from motion, temporal reasoning, feature matching

Introduction

The ability to build three-dimensional structure from visual information is a goal that spans a multitude of domains, both inside and outside computer vision; building architects are hired to create accurate real-world representations of their designs, 3D printing, and figurine manufacturers aim to do this as well. In computer vision, this desire is realized by 3D reconstruction, Structure from Motion (SfM), and simultaneous localization and mapping (SLAM). While many advancements in computer vision have been made in recreating the geometry of a scene based on spatial features, the ability to recreate dynamic scenes has been challenging. There has been significant development in matching viewpoints from multiple cameras to the same position in the world, which has yielded improvements in many applications: panorama stitching, 3D reconstruction, SLAM, SfM, and more. However, despite these improvements, reconstruction of dynamic scenes relies not only on spatial signals but temporal signals as well. Most of the approaches to spatial feature matching rely on being invariant to multiple factors such as illumination, scale, and affine transformations. However, some of these factors could be useful in finding correspondence between images that exist in the same spatio-temporal span, which makes matching these features challenging. While feature matching in this context may be hard for humans to solve, I propose a neural network-based approach to tease out the invariances that are desirable for spatio-temporal feature matching.

Background

SfM is able to create three-dimensional structures from large collections of two-dimensional images by creating point clouds of common points from an image captured by multiple camera angles. SfM has gone through multiple variations since its inception, including non-rigid SfM, time-varying SfM, and SLAM; these variations aim to create structure from deformable objects, track appearance changes in a structure across time, and create maps of unknown environments in real-time, respectively. While all of these implementations differ in their applications, they share a need for efficient feature matching algorithms. There have been successful attempts at matching features in the spatial domain, both in neural-network-based approaches [1] and traditional multiple-view stereo pipelines [2]. Despite this, matching points between images in the same world location and in the same timeframe have been challenging. Previous attempts at spatio-temporal feature matching have mainly been applied to matching features within consecutive video frames [3], or have been treated as a binary classification (either the image was in a specified time frame, or it was not) [4]. Time-varying SfM has traditionally relied on the latter approach to combat temporal misregistration in image metadata, resulting in the potential loss of feature-rich matches due to misclassification.

Proposal

With time-varying SfM as the motivator, I propose using modern deep convolutional neural network (CNN) architectures to find correspondence between images that not only exist at the same point in the world, but also in the same specified temporal range. Determining correspondence between spatial image features usually requires some sort of invariance, such as illumination, scale, or affine transformations. To address this, I plan on expanding upon the D2D [1] architecture by feeding the output of the decoder into a multi-layer perceptron (MLP) to generate a temporal classification. The D2D architecture determines correspondence by matching all points between images and then selecting the best matches based on a learned distinctiveness score. Given the complexity of spatio-temporal feature matching, I plan to start with relatively wide temporal spans (season, month, etc.) and work towards shorter spans. The loss function will be defined by taking into consideration the correspondence and distinctiveness scores proposed in D2D, along with the mean-squared error of the temporal classification, measured in a discrete time step. The correspondence loss uses a standard Euclidian distance metric to accomplish feature-wise comparisons between two image descriptor maps. The distinctiveness loss is an estimate of how often a feature descriptor is mismatched with another feature calculated by an MLP regressor on the descriptor maps. Attaching an MLP to the output of the decoder will allow the CNN to learn the invariances necessary to determine spatio-temporal feature matches. In the case of the MLP not learning desirable features, I will explore modifying other architectures such as ResNet-* to use as a head for the proposed D2D backbone.

To train this model we need ground truth for spatial matches as well as ground truth for temporal matches. Our spatial ground truth will come from the MegaDepth [5] dataset, a large scale dataset composed of internet images with ground truth depth maps. However, the timestamps associated with images from MegaDepth are often misregistered. Our temporal ground truth will come from webcam streams, which have accurate timestamps but lack in viewpoint variation. By training our model on both MegaDepth and webcam streams, there is potential for accurate spatio-temporal matches.

Intellectual Merit

If this approach is successful, the ability to match features in a spatio-temporal domain will allow for a more robust set of features that can be used in a variety of applications. The inability to match features in this domain has hindered progress in time-varying SfM as it has only been able to use planar geometry in its reconstructions. This can be partially attributed to the temporal constraints of extracted features. By solving the problem of spatio-temporal matching, more progress will be made towards creating non-planar geometric structures in time-varying SfM.

Broader Impacts

Having the ability to find spatio-temporal correspondence between images not only has the potential to impact computer vision problems, but interdisciplinary problems as well. Given enough data, one could imagine preserving cultural landmarks by way of time-varying SfM. Consider a popular, well-photographed structure that collapsed and was then rebuilt into something still culturally meaningful; time-varying SfM, along with these proposed features, could create a three-dimensional representation that demonstrates the transformation that this structure had gone through. Researchers working in remote sensing would also benefit from these proposed features by tracking area changes over time in a more accurate manner provided my proposal could be generalized to overhead imagery.

References

[1] Olivia Wiles, Sebastien Ehrhardt, & Andrew Zisserman. (2020). D2D: Learning to find good correspondences for image matching and manipulation.

[2] Lowe, D. (2004). Distinctive Image Features from Scale-Invariant Keypoints *Int. J. Comput. Vision, 60*(2), 91–110.

[3] Patrick Ruhkamp, Ruiqi Gong, Nassir Navab, & Benjamin Busam. (2020). DynaMiTe: A Dynamic Local Motion Model with Temporal Constraints for Robust Real-Time Feature Matching.

[4] Matzen, K., & Snavely, N. (2014). Scene Chronology. In *Proc. European Conf. on Computer Vision*.

[5] Zhengqi Li, & Noah Snavely (2018). MegaDepth: Learning Single-View Depth Prediction from Internet Photos. In *Computer Vision and Pattern Recognition (CVPR)*.