Understanding real-world scenes for human-like machine perception*

Armin Mustafa[1][0000−0002−1779−2775] and Adrian Hilton[1][0000−0003−4223−238X]

University of Surrey, Guildford, GU27XH, UK
a.mustafa, a.hilton@surrey.ac.uk

Abstract. The rise of autonomous machines in our day-to-day lives has led to an increasing demand for machine perception of real-world to be more robust, accurate and human-like. The research in visual scene understanding over the past two decades has focused on machine perception in controlled environments such as indoor, static and rigid objects. There is a gap in literature for machine perception in general complex scenes (outdoor with multiple interacting people). The proposed research addresses the limitations of existing methods by proposing an unsupervised framework to simultaneously model, semantically segment and estimate motion for general dynamic scenes captured from multiple view videos with a network of static or moving cameras. In this talk I will explain the proposed joint framework to understand general dynamic scenes for machine perception; give a comprehensive performance evaluation against state-of-the-art techniques on challenging indoor and outdoor sequences; and demonstrate applications such as virtual, augmented, mixed reality (VR/AR/MR) and broadcast production (Free-view point video - FVV).

Keywords: Scene Understanding · Reconstruction · Semantic Segmentation.

1 Introduction

Existing state-of-the-art methods to reconstruct and understand general dynamic scenes [2] have several limitations: ambiguities in appearance between objects degrades performance; a large number of inter-occluding dynamic objects and complex outdoor scenes are not handled; the approach requires sufficient wide-baseline cameras (static/moving) to cover the scene; changing scene illumination leads to errors in the reconstruction; non-lambertian surface reflectance; and the quality of output is limited. The goal of the proposed research is to introduce a framework for simultaneous 4D reconstruction and scene understanding of highly complex natural scenes such as crowds and complex social human interactions from video, as shown in Figure 1 for publicly available Juggler dataset [3] captured with 6 moving cameras. Joint semantically coherent object-based long-term 4D scene flow estimation, co-segmentation and reconstruction is proposed

* Supported by RAEng and EPSRC

exploiting the coherence in semantic class labels both spatially, between views at a single time instant, and temporally, between widely spaced time instants of dynamic objects with similar shape and appearance. Semantic tracklets are introduced to robustly initialize the scene flow in the joint estimation and enforce temporal coherence in 4D flow, semantic labelling and reconstruction between widely spaced instances of dynamic objects. This research is an enabling technology for autonomous operation of machines to operate safely alongside people in our complex ever-changing everyday environments (home or workplace).

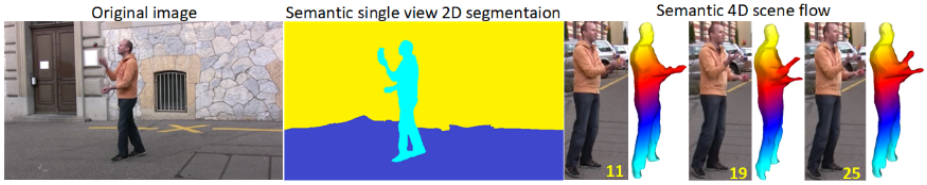

Fig. 1. Example of proposed framework resulting in an accurately labeled segmentation, 4D reconstruction and scene flow (represented by color mask propagation in dynamic object of the scene).

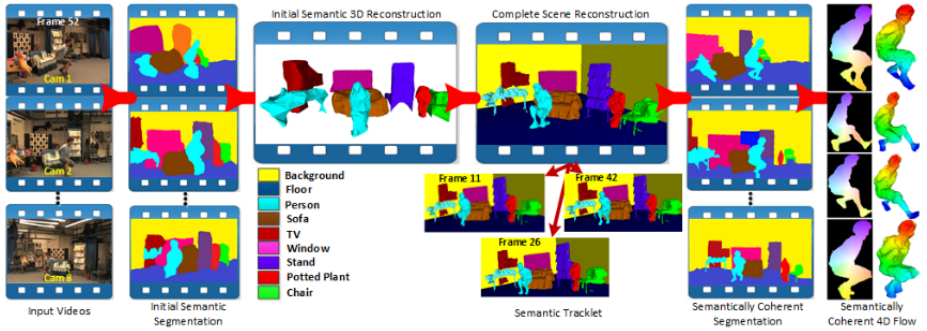

Fig. 2. Joint semantic segmentation, reconstruction and flow framework.

2 Methodology

The proposed framework for semantic temporal coherence, illustrated in Figure 2, comprises the following stages:

Initial Semantic Segmentation: Initial semantic labels are estimated for each pixel in the image using Mask RCNN [4].

Initial Semantic Reconstruction: Semantic information for each view is combined with sparse 3D feature correspondence between views to obtain an initial semantic 3D reconstruction [1]. This initial reconstruction combines semantic information across views but results in inconsistency due to inaccuracies in the initial per-view segmentation.

Semantic Tracklets: To enforce long-term semantic coherence temporally we propose semantic tracklets that identify a set of similar frames for each dynamic object. Similarity between any pair of frames is estimated from the per-view semantic labels, appearance, shape and motion information. Semantic trackets provide a prior for the joint space-time semantic co-segmentation and reconstruction to enforce temporal coherence.

Joint Semantic Flow, Co-segmentation and Reconstruction: The initial semantic segmentation and reconstruction is refined per-view for each dynamic object through joint optimisation of flow, segmentation and shape across multiple views and over time using the semantic tracklets. Per-view information is merged into a single 3D model using Poisson surface reconstruction. The process is repeated for the entire sequence to obtain semantically coherent long-term dense 4D scene flow, co-segmentation, and reconstruction for the complete scene for human-like machine perception.

Temporal and semantic coherence results using proposed approach are demonstrated in Figure 3. The first frame is color coded using unique color gradient and colors are reliably propagated to the sequence using proposed 4D scene flow demonstrating it's usefulness.

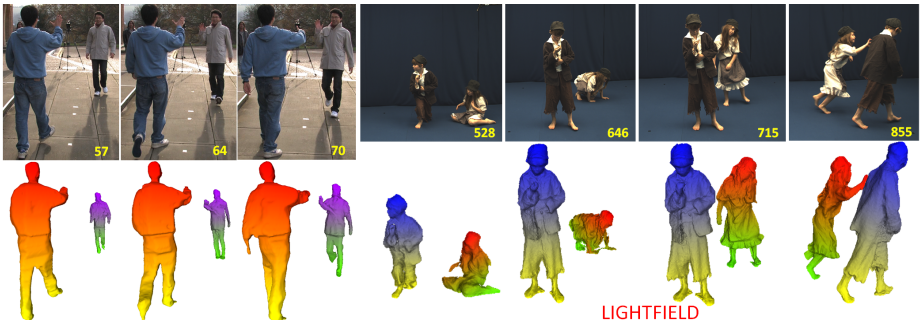

Fig. 3. 4D semantic scene flow results demonstrated on two datasets: Cathedral outdoor dataset on the left from [2] and Lightfield indoor dataset on the right from [5]

3 Conclusion

This paper proposes a novel approach to joint semantic 4D scene flow, multi-view co-segmentation and reconstruction of complex dynamic scenes for applications in VR and FVV. Temporal and semantic coherence is enforced over long-time frames by semantic. Comparative evaluation demonstrates that enforcing semantic coherence achieves significant improvement in scene flow and segmentation of general dynamic indoor and outdoor scenes captured with multiple hand-held cameras. Introduction of space-time semantic coherence in the proposed framework achieves better reconstruction and flow estimation against state-of-the-art methods. Results and applications will be shown in the talk.

References

1. Mustafa, A., and Hilton, A.: Semantically coherent co-segmentation and reconstruction of dynamic scenes. In: CVPR 2017
2. Kim, H., Guillemaut, J., Takai, T., Sarim, M., and Hilton, A.: Outdoor Dynamic 3D Scene Reconstruction. In: IEEE CSVT 2012
3. Ballan, L., Brostow, G.J., Puwein, J., and Pollefeys, M.: Unstructured video-based rendering: Interactive exploration of casually captured videos. In: ACM TOG 2010
4. He, K., Gkioxari, G., Dollar, P., and Girshick, R.B.: Mask RCNN. In: CVPR 2017.
5. Mustafa, A., Volino, M., Guillemaut, J., and Hilton, A.: Semantically coherent co-segmentation and reconstruction of dynamic scenes. In: 3DV 2017