RESEARCH STATEMENT

BUILDING A VIRTUAL TIME MACHINE

Minh P. Vo
The Robotics Institute, Carnegie Mellon University
Email: mpvo@cs.cmu.edu Web: http://www.cs.cmu.edu/~mvo

Imagine if we could go back in time all along the memory lane to revisit the fine moments of our life such as our first birthday, our wedding ceremony, or the moment our first kid was born, etc., like everything is happening around us. While building a *real* time machine may require some more lifetimes, the advent of affordable and high quality smartphone cameras has presented a great opportunity to build a *virtual* time machine in a near future. Any significant event is now captured from multiple perspectives. These collections of social video data provide a unique opportunity for rich explorations of the scenes, far exceeding what is possible with a single camera.

Despite great potentials, automatically fusing such rich visual data into a single comprehensive model that could facilitate content browsing is challenging. Those social videos could be acquired from different locations, angles, zoom settings, with varying focus, different types of cameras (high quality, low quality), and at different times. This is a new form of visual data. Due to the dynamics and diversity of the data, it is difficult to correspond images of the same object in the videos. Additionally, the mathematics of reconstructing moving objects from unsynchronized cameras is not well-understood. Consequently, current techniques on large scale analytics of community visual data are limited to static scenes and commercially available applications for event browsing either discards content of certain clips to create predefined multi-angle videos without interactive scene exploration (CrowdFilk) or requires expensive hardware with elaborated setup (Intel FreeD).

My research seeks to **automatically organize the massive visual data captured in unconstrained settings into a comprehensive 4D environment along with its semantics for scene browsing and editing**. Overcoming this challenge unlocks strong potential applications in Virtual Reality, Augmented Reality, surveillance system, statistical understanding of crowd psychology, and even the chance of reconstructing historic events from the Internet videos. As an example application, since the video are captured from different perspectives and at slightly different times, combining their information in 3D space effectively re-creates the scene at higher spatial and temporal resolution than what can be obtained from individual videos. Imagine we could playback the moment when your wife entered your wedding ceremony, when you poured the champagne into the stack of champagne glasses, or when your wife threw the bouquet to the crowd at slow motion and from all possible angles. Such precious moments, lying deep in your memory, magically come back to us immersively and realistically. The same system can also be used for surveillance or statistically study of crowd psychology as the unconstrained visual data has been automatically organized, reconstructed, and tracked in both space and time. As another example, consider a historical tragedy such as The Boston Bombing in 2013. This event was captured by infrastructure cameras and civilian personal cameras from many different angles. If the information was automatically fused from the vast amount of visual data, both the suspects and the explosive devices could be quickly identified. More importantly, the civilian could be evacuated along the save escape routes that were planned according to the reconstructed dynamic scene model. The pattern of people running away from the event could also be studied for better construction of emergency evacuation routes.

4D EVENT RECONSTRUCTION PIPELINE

Compared to static scene reconstruction pipeline from community images, reconstructing social dynamic event from casual videos possesses distinctive differences in the input (images vs. videos) and the expected output (spatially coherent 3D reconstruction vs. spatiotemporally consistent 4D reconstruction). The consequences of these differences are great challenges in the correspondence estimation (as we no longer have luxury of time to image the scene as much as we want), and the reconstruction of moving points (as the classical triangulation constraint is inapplicable for asynchronous videos). Thus, the problem of 4D organization and reconstruction of social dynamic event from videos in unconstrained settings requires a ground-up re-building of previous static scene reconstruction pipelines.

My Ph.D. research proposes **multiple novel core components for 4D reconstruction pipeline of dynamic scenes captured by multiple unsynchronized video cameras in an unconstrained crowd-captured setting**. Figure 1 shows the proposed pipeline. The key is to exploit additional cues from the scene semantics, the physics of motion dynamics, and the transient behavior of rigid shape deformation for the shape estimation problem. These cues help us to solve problem hierarchically in three stages: coarse human reconstruction using semantic priors, sparse but accurate trajectory reconstruction of salient features using motion prior, and densify the reconstruction using shape deformation prior

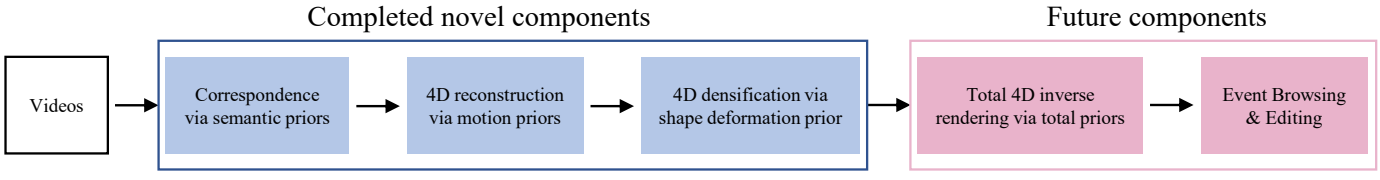


Figure 1: *A dynamic social event reconstruction from unconstrained videos pipeline.*

Spatiotemporal People Tracking in Complex Group Activity [1]

Consider an interesting dynamic event that involves multiple people actively interacting with each other, e.g. a surprised birthday party. The static objects in the background such as buildings and walls are mostly occluded by humans. Since the number of static features observed from multiple views is significantly smaller than the number of dynamic features, camera localization is very challenging. There are few visual

features of static objects available to build a reliable model of the scene. As the number of people and the complexity of the activity increases, both camera localization and human tracking become considerably more challenging due to frequent occlusions and strong viewpoint and appearance variations. Existing algorithms rely on expensive, calibrated, and synchronized multi-camera systems for reliable people tracking for such scenario. However, from a high-level reasoning perspective, a different approach to understand the geometry of the scene from the same set of visual data is to associate the same person across different cameras. Together with semantic 2D human pose estimation, such association can provide constrains for rough camera localization with respect to the scene.

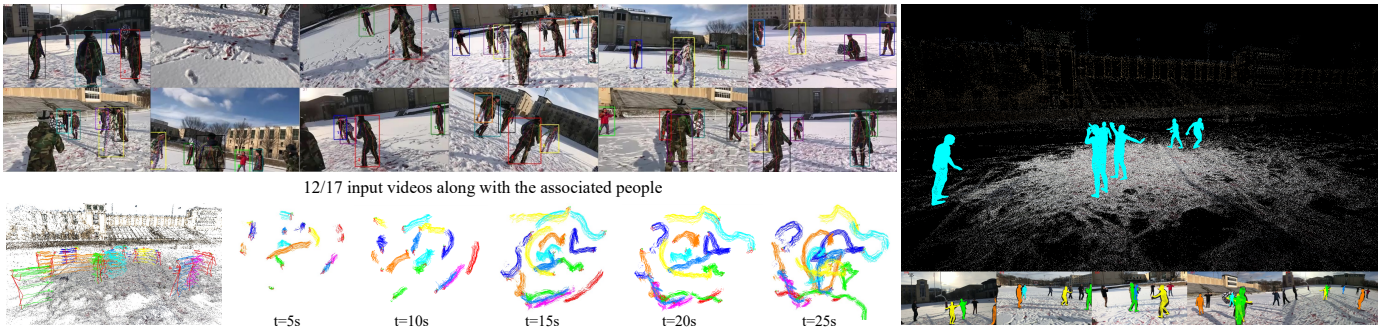


Figure 2: *3D tracks of complex group activity from hand-held smartphone and head-mounted GoPro cameras. A total of* 14 *people, each associated with a bounding box of unique color, are tracked over time. My method gives smooth and clean trajectories despite strong occlusion, similar people appearance, and complex motion pattern. The human mesh model is fitted to the estimated 3D skeleton, providing a coarse approximation of the human body shape for later refinement.*

To address the problem of people association across multiple viewpoints and time instances, I advocate **the use self-adaptive learning of strong appearance descriptor specifically for the domain videos for people matching**. We combine motion tracking, mutual exclusion constraints, and multi-view geometry in a multi-task learning framework to automatically adapt a generic person appearance descriptor to the domain videos. A discriminative person descriptor enables the use of clustering for tracking individual persons. Since the association is solved globally, there is no tracking drift. As a bi-product of the association, the videos are also temporally aligned (up to the speed of the observed motion). To estimate spatially stable and temporally coherent 3D skeleton for each person, we exploit human physical prior on limb length and symmetry to constrain the reconstruction. We validate the proposed approach on challenging sequences of people involved in complex and highly dynamic group activity captured by at least 15 hand-held smartphones and head-mounted cameras without any constrains on the capture or scene behavior (see Figure 2 for one of those sequences [1]). Despite difficulties, our method shows significant improvement in association accuracy (18%) and 3D human skeleton reconstruction (5 to 10 times) over the baselines. The existence of spatiotemporally coherent 3D skeleton eases the human morphable mesh model fitting step providing a good initialization for further body shape refinement.

Spatiotemporal Bundle Adjustment for Dynamic Scene Reconstruction [2]

The estimated temporal alignments and camera poses in the previous stage are refined using a novel spatiotemporal bundle adjustment algorithm. The classical bundle adjustment is built on the ray triangulation constraint. However, this constraint is invalid for moving points captured in multiple unsynchronized videos. No moving 3D point is seen from at least to asynchronous cameras at the same time. Currently, there is no consumer mechanism to ensure that multiple personal cameras, i.e., smartphones or consumer camcorders, are simultaneously triggered.

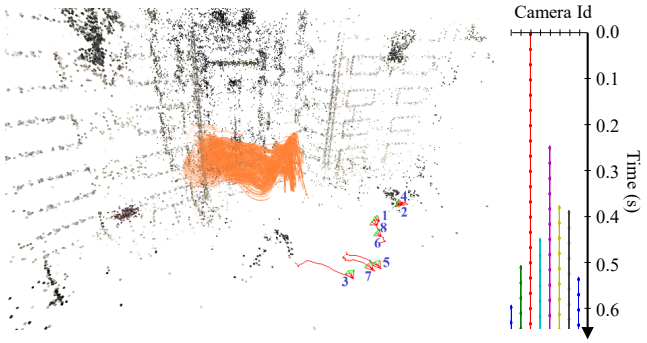
To optimally solve the dynamic 3D reconstruction problem, we must first recognize all the constituent sub-problems that exist. The classic problems of point triangulation and camera resectioning in the static case are subsumed. Two new problems arise: reconstructing 3D trajectories of moving points and estimating the temporal alignment of each camera. Second, we must recognize that the sub-problems are tightly coupled and must be solved jointly. As an example, consider the problem of estimating 3D camera pose. While segmenting out stationary points and using them to estimate camera pose is a common strategy, it ignores evidence from moving points that are often closer to the cameras and therefore provide tighter constraints for precise camera calibration. Imprecise camera calibration and quantization errors in estimating discrete temporal offsets result in significant errors in the reconstruction of moving points. None has considered the interactions and solving these four problems jointly in the past.

Since there is only one observation of the point at any time instances, the reconstruction of its motion/trajectory is ill-posed. Clearly, there are infinitely many such trajectories and each of these paths corresponds to a different temporal sequencing of the rays. Yet, the true trajectory must also correctly align all the cameras. Thus, to sufficiently constraint the solution, I recognize that **both the geometry and dynamics of the point have to be considered**. Ideally, among all possible the solutions satisfying the image observation, a dynamically plausible trajectory corresponds to the correct temporal alignment. We analyze several physical motion priors, i.e. least kinetic energy, least force, and least action, on a large motion capture corpus of CMU MoCap database, and carefully integrate of these priors within the reconstruction pipeline. This formulation not only allows sub-frame temporal alignment but also fully exploits the asynchronous video streams to recover sparse but accurate 3D trajectories

[1] See `https://www.youtube.com/watch?v=ZDuaJzcLcdE` for other results.

Time-aligned estimated from 8 sub-sampled videos (30fps) showed on the original (120fps) videos

Insets of the time-aligned images



4D trajectory reconstruction (240fps)

Figure 3: *The aligned images, estimated from temporally down sampled video at 30 fps, are shown for the original video captured at 120fps. As showed in the inset of aligned images, the shadow casted by the folding cloth are well temporally aligned across images. Our motion prior based approach produces plausible reconstruction for the entire course of the action even with relatively low frame-rate cameras (30fps).*

at much high temporal resolution than the framerate of the input videos. As a demonstration, we reconstruct 3D trajectories of dynamic actions captured outdoor at 240fps from 8 hand-held cameras captured at 30fps (see Figure 3)[2]. This algorithm is being taught by Prof. Martial Hebert in his graduate-level course, Geometry-based Methods of Computer Vision, at CMU.

Shape Triangulation: Densify 3D Trajectories without Dense Multiview Correspondences [3]

For an immersive event browsing experience, the scene must be densely reconstructed. This requires finding dense correspondence across multiple cameras. However, this is challenging because the video data could be acquired from very different perspectives from camera of different quality, or may suffer from realistic optics artifacts such as motion blur, exposure saturation, or de-focusing. Additionally, some dynamic objects may have homogeneous appearance. This is why previous work only produces a sparse set of 3D trajectory of moving points [2].

The key observation is that it is usually easier to perform dense and accurate 2D tracking (event transient points appeared due to shading) within same the video than finding reliable dense correspondences cross different videos. Furthermore, if the object is rigid, it can be reconstructed using classical structure from motion algorithms. While it is unlike for human motion to be rigid, rigidity motion may be plausible for a transient amount of time, say 2 consecutive frames given the current framerate of consumer camera. Built upon these intuition, I formulate a shape triangulation to recovery denser 3D trajectory of dynamic points. This formulation uses sparse multi-view correspondence to identify corresponding region undergoing a same motion from different cameras and exploits in-view motion tracking for 3D motion estimation. **No spatial constraints are needed for single-view tracking points. No temporal constraints are needed for multi-view matching points. No scale ambiguity occurs in merging the motion estimation across cameras.**

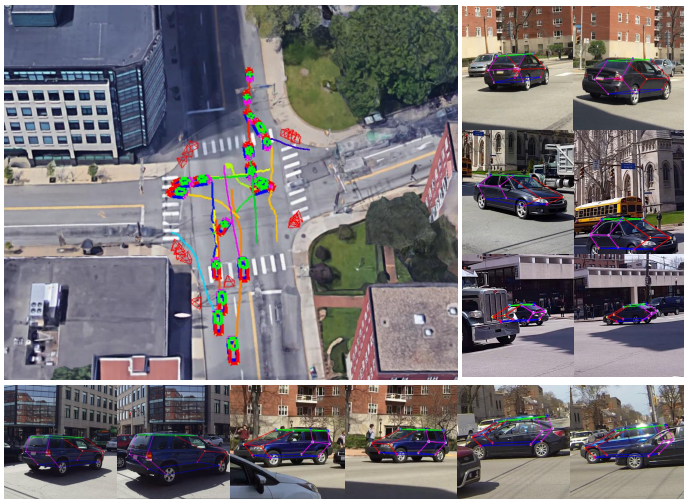


Figure 4: *Reconstruction of vehicles crossing a busy intersection, making turns, going straight and changing lanes. A subset of vehicle skeletons (3D detector locations) and their 3D trajectories are augmented within the Google Earth view of the intersection. The reconstructions are reprojected into multiple views of two cars (a sedan and an SUV) demonstrating good performance under partial occlusions.*

Together with a junior graduate student, we apply the idea to estimate the motion of vehicles at a busy intersection from 21 hand-held cameras (see Figure) [3]. We demonstrate that incomplete and imprecise semantic keypoint detection across multiple views can be fused with precise but sparse single-view tracks of Harris feature to reconstruct moving vehicles even in severe occluded scenarios. The shapes (even sparse) and motions of the vehicles recovered using our approach can be invaluable to traffic analysis, including vehicle type, speed, density, trajectory, and frequency of events such as near-accidents or for (semi-)autonomous vehicles approaching the intersection.

[2] See `http://www.cs.cmu.edu/~ILIM/projects/IM/STBA/` for our 960fps motion reconstruction from the original 120fps videos and other results

[3] See `https://youtu.be/RUtK7xRtyns` for more analysis

FUTURE RESEARCH THRUSTS

Building a complete engine of event browsing and editing requires great expansion of my research to many interesting directions. The goal is to make the system robust, efficient, and as automatic as possible for real world capture scenarios.

Reflectance and Illumination Estimation

Estimating the intrinsic appearance of the object and the underlying scene illumination from images lies at the heart of many important visions and graphics problems such as appearance editing, virtual object inserting, or shape recovery. Currently, most algorithms assume diffuse reflectance and distant gray-scale point light sources, both approximated using low dimensional basis functions such as Spherical Harmonics. More sophisticated representations such as wavelets or DSBRDF are rarely used due to computational tractability issues. For community causal videos, the inputs are often radiometrically uncalibrated and have limited dynamic range, leading to significant model deterioration if applied naively. My interest in this domain are three-folds: (1) compact statistical modeling of arbitrary lightings and surface reflectance properties from casual videos; (2) automatic adaptation of learned priors to the specific domain settings via learning algorithms; (3) efficient framework powered by the learned priors for reflectance and illumination estimation.

Priors-based Event Reconstruction

Despite significant progress in camera calibration and scene reconstruction, they often fail when used in casual dynamic scene settings due to camera shaking, motion blur, degenerated configurations (no motions, pure rotation, or forward motion), frequent occlusion, or textureless scenes (human clothing or man-made backgrounds). Thus, the camera poses may not be accurately estimated at all time, the reconstructed environment are coarse and full of artifacts. Even well-studied tasks such as camera localization could be computationally expensive due to all-camera-pair matching for correspondence search as the cameras baseline is typically quite small for reliable per-camera SLAM-based localization.

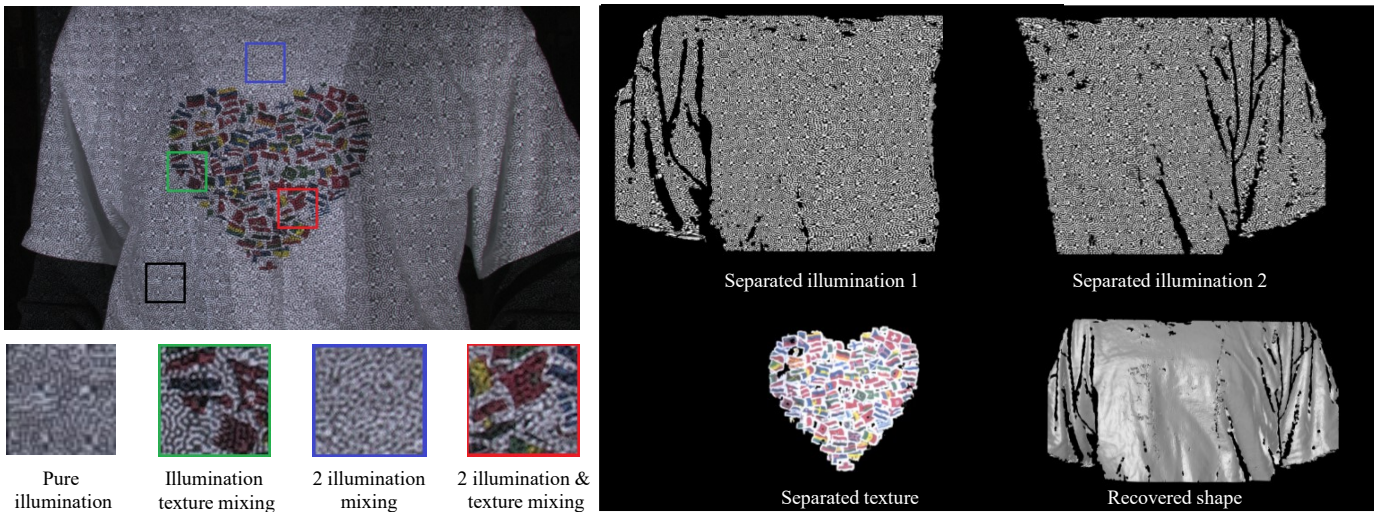


Figure 5: *When a highly texture shirt illuminated by 2 projectors, the mixture of surface color and the illuminations makes it difficult to establish reliable and dense camera-projector correspondences. By exploiting known illumination patterns and a partial observation of the surface color, we can decompose the mixed appearances into its original illuminations and the surface texture. The decomposed illumination facilitated dense camera-projector correspondences for detailed shape reconstruction and the decomposed surface texture allows dense surface motion tracking.*

I strongly advocate the use of priors from semantics (segmentation, surface normal), motion dynamics, and photometry (surface reflectance and environment illumination), and the exploitation of advances in consumer cameras (GPS, IMU, network-based time synchronization) to efficiently solve the scene reconstruction problem. As an example, while strong interpretation of their activities can be observed from fitting statistical human mesh model to the scene (sparse trajectories vs. coarse template shape fitting), the visualization can be photorealistic and immersive if the true shape and motion of the human bodies can be recovered via an inverse rendering process where different body parts (face, arms, and clothing) modeled by their best reflectance priors, appeared under the same illumination partially captured by the videos, are jointly optimized for reflectance constancy and motion coherency over a window of time (see Figure 5[4] for one early result in this direction). As another example, while traditional multi-view stereo algorithms are unreliable in reconstructing textureless regions such as the planar wall, it is relatively easy to segment and further deduce that pixels within that region share the same surface normal. Together with the sparse set of reliable 3D points triangulated from the correspondence search, this feedback information constrains and remove spurious reconstruction noise.

[4]See `http://www.cs.cmu.edu/~ILIM/projects/IL/TextIllumSep/` for more results

Event Browsing and Editing

Despite significant efforts in event reconstruction, artifacts in both shape and motion may still occur, sometimes simply due to the lack of observations (insufficient number of cameras). Unfortunately, human eyes are sensitive to those artifacts, especially for scenes familiar with themselves. Approaches that texturemap the reconstruction for visualization may not look unpleasant to unprofessional consumers. Additionally, it is not clear that a perfect inverse rendering algorithm is required for immersive visual perception. I plan to investigate hybrid approaches of 3D reconstruction and Image-based rendering to address those artifacts. Both approaches have complementary strength and should be combined. For examples, textureless regions which are hard to reconstruction may be well-represented using homography warping and Poisson blending.

Besides being able to visualize our personal events immersive, the browsing can be more joyful if one can edit the details (frowning face to smiling face), insert funny objects or delete annoying ones in a spatiotemporally coherent manner. To remove unwanted objects, besides inpainting techniques, which could be slow and produce incoherent results, one interesting way to exploit the rich information available from multi-video setting is the synthetic aperture approach, which allows seeing through occlusion. The availability of the shape reconstruction, reflection, and lighting estimation could make the filled in regions indistinguishable from the real ones. I am interested in exploring the interactions between semantics, surface reflectance, lighting, geometry, and the abundance of pixels data for novel scene browsing and editing experience.

References

[1] **M. Vo**, E. Yumer, K. Sunkavalli, S. Hadap, Y. Sheikh, and S. Narasimhan, "3D Tracking of Complex Group Activity", *CVPR*, 2018, (Submitted).

[2] **M. Vo**, S. Narasimhan, and Y. Sheikh, "Spatiotemporal Bundle Adjustment for Dynamic Scenes Reconstruction", *CVPR*, 2016.

[3] D. Reddy, **M. Vo**, and S. Narasimhan, "CarFusion: Incorporating Point Tracking and Part Detection for Dynamic 3D Reconstruction of Vehicles", *CVPR*, 2018, (Submitted).

[4] **M. Vo**, S. Narasimhan, and Y. Sheikh, "Texture Illumination Separation for Single-shot Structured Light Reconstruction", *IEEE TPAMI*, 2016.