3D Semantic Trajectory Reconstruction from 3D Pixel Continuum

Jae Shin Yoon
University of Minnesota
jsyoon@umn.edu

Ziwei Li
University of Minnesota
lixx3686@umn.edu

Hyun Soo Park
University of Minnesota
hspark@umn.edu

Abstract

This paper presents a method to assign a semantic label to a 3D reconstructed trajectory from multiview image streams. The key challenge of the semantic labeling lies in the self-occlusion and photometric inconsistency caused by object and social interactions, resulting in highly fragmented trajectory reconstruction with noisy semantic labels. We address this challenge by introducing a new representation called 3D semantic map—a probability distribution over labels per 3D trajectory constructed by a set of semantic recognition across multiple views. Our conjecture is that among many views, there exist a set of views that are more informative than the others. We build the 3D semantic map based on a likelihood of visibility and 2D recognition confidence and identify the view that best represents the semantics of the trajectory. We use this 3D semantic map and trajectory affinity computed by local rigid transformation to precisely infer labels as a whole. This global inference quantitatively outperforms the baseline approaches in terms of predictive validity, representation robustness, and affinity effectiveness. We demonstrate that our algorithm can robustly compute the semantic labels of a large scale trajectory set (e.g., millions of trajectories) involving real-world human interactions with object, scenes, and people.

1. Introduction

Now cameras are deeply integrated in our daily lives, e.g., Amazon Cloud Cam and Nest Cam, reaching soon towards *3D pixel continuum*—every 3D point in our space is observed by a network of ubiquitous cameras. Such cameras open up a unique opportunity to quantitatively analyze our detailed interactions with scenes, objects, and people continuously, which will facilitate behavioral monitoring for the elderly, human-robot collaboration, and social telepresence. A 3D trajectory representation of human interactions [8, 19, 25, 41, 42] is a viable computational model that measures microscopic actions at high spatial resolution without prior scene assumptions. Unfortunately, the representation is lacking semantics, i.e., it is important to know not only where a 3D point is but also *what it means* and *how associated with other points*. For instance, Figure 1 illustrates semantic labeling of dense 3D trajectories that can computationally describe the spatial and temporal relationship between basketball player's hand and ball.

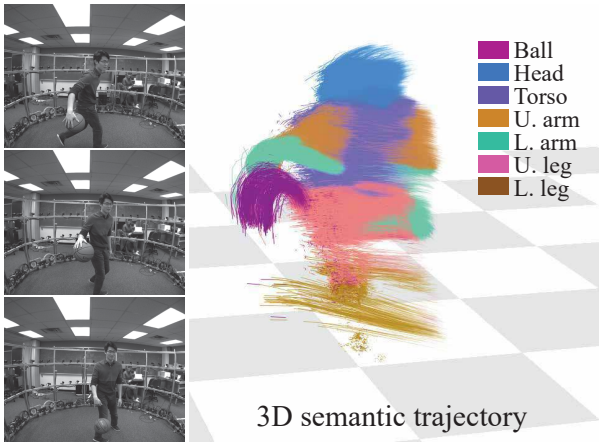


Figure 1. Given 3D dense reconstructed trajectories, we assign their semantic meaning using multiple view image streams. Each trajectory is associated with semantic labels such as body parts and objects (basketball). For illustrative purpose, the last 10 frames of trajectories are visualized.

However, assigning a semantic label to each trajectory in real-world scenes involves with two principal challenges. (1) Missing data: interactions with objects and people inherently introduce self-occlusion, resulting in 3D reconstruction of highly fragmented trajectories, i.e., each trajectory emerges and dissolves in different time instances where existing approaches of global spatial reasoning such as articulated body [42] and shape basis [8, 41] are not applicable. (2) Noisy and coarse recognition: existing visual recognition systems were largely built on single view images, which are often fragile to heavy background clutter, self-occlusion, and non-iconic object pose. This issue further escalates when coarse recognition models such as a bounding box representation are used, i.e., not all pixels in a detection window belong to the same object class.

In this paper, we present a method to precisely assign the

semantic label on dense 3D trajectory stream reconstructed by a large scale multi-camera system that emulates the 3D pixel continuum. Our semantic labeling method leverages two cues. (a) 2D visual cue: albeit noisy, it is possible to geometrically consolidate the outputs of 2D image recognition across multiview images. We conjecture that among many views, there exists a set of views that can confidently label a 3D trajectory. We introduce a new representation called *3D semantic map*—a probability distribution over semantic labels per 3D trajectory constructed by a probability of visibility and recognition confidence. (b) 3D spatial cue: a set of trajectories that belong to the same objects can be expressed by local rigid transformation. We use the local rigid transformation to compute the trajectory affinity that can link long-range fragmented trajectories.

Our system takes a set of synchronized multiview image streams captured by 69 HD cameras[1]. Given 3D reconstructed trajectories from image streams, we build the 3D semantic map to find the view that best represents the semantics of a 3D trajectory. We use the 3D semantic map and trajectory affinity computed by local rigid transformation to precisely infer labels as a whole. This global inference is conducted via multi-class graph-cuts in Markov Random Field (MRF).

The core contributions of this paper include: (1) 3D semantic map: we introduce a novel concept for trajectory semantics encoding the distribution over labels computed by view-pooling; (2) Long range affinity: estimation of local rigid transformation around a trajectory allows relating with distant trajectories; (3) Multiple view human interaction dataset: we collect 9 new datasets involving in various human interactions including pet/social interactions, dance, sports, and object manipulations; (4) A modular design of 3D pixel continuum: we design a space that can densely measure human interactions from nearly exhaustive views by modularizing commodity parts, which is scalable and customizable.

2. Related Work

Humans can effortlessly *read* the intent of others through subtle behavioral cues in a fraction of second [4], and high resolution videos are now able to capture such cues via our interactions with surrounding environments. The pixels in the videos can be tracked to form long term trajectories to encode the interactions both in 2D and 3D.

2D trajectory As many objects are roughly rigid and move independently, motion provides a strong discriminative cue to group pixels and recognize occluding boundary, precisely. A core challenge of motion segmentation lies in fragmented nature of trajectories caused by tracking failure (occlusion, drifting, and motion blur). Embedding trajectories into low dimensional space has been used to robustly measure trajectory distance in the presence of missing data without pre-trained models [9,13,16,30], and 2D trajectories can be decomposed into 3D camera motion and deformable object models [28,35,40]. Visual semantics learned by object recognition frameworks provide stronger cues to cluster trajectories [21,22,38].

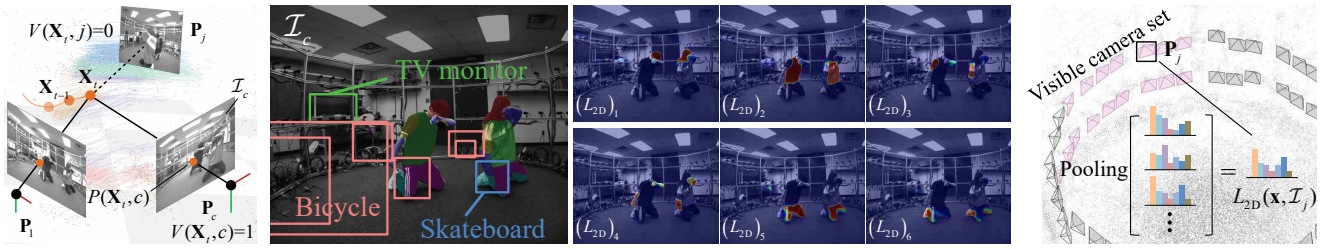
3D trajectories Due to dimensional loss in the process of 2D projection, reconstructing 3D motion from a monocular camera is an ill-posed problem in general, i.e., the number of variables (3D motion parameters) is greater than the number equations (projections). However, when an object undergoes constrained deformation such as face, its 3D shape can be recovered by enforcing spatial regularity, e.g., shape basis [8, 34, 41, 42], template [33], and mesh [39]. A key challenge of this approach is to learn a shape prior that can express general deformation, often requiring an instance specific pre-trained model, or inherent rank minimization where the global solution is difficult to be achieved [1, 10]. A trajectory based representation directly addresses this challenge. Motion is described by a set of trajectory stream where generic temporal regularity is applied through DCT trajectory basis [2,27], polynomial basis [5,20], and linear dynamical model [36]. A spatiotemporal constraint can further reduce dimensionality, resulting in robust 3D reconstruction [3,26,40]. When multiple view images are used, it is possible to represent general motion with topological change without any spatial and temporal prior [18,19].

Unlike 2D trajectories, *semantic labeling* of 3D trajectories is largely uncharted research area. Notably, Yan and Pollefeys [42] presented a trajectory clustering algorithm based on articulated body structure, i.e., an object is composed of a kinematic chain of rigid bodies where the articulated joint and its rotational axis lie in the intersection of two shape subspaces. Later, image segmentation cues have been incorporated to recognize a scene topology, i.e., preclustering object instances, to reconstruct dynamics scenes from videos in the wild [11,15,32]. Note that none of these work has addressed semantics. The work by Joo et al. [18] is closest to our approach where the trajectory clustering is based on 3D rigid transformation of human anatomical keypoints. Our method is not limited to human bodies, which enables modeling general human interactions with scenes, objects, and other people.

3. System Overview

Our system takes 69 synchronized image streams at 30 Hz from a multi-camera system (Section 7). We use the standard structure from motion pipeline [17,37] to calibrate the camera and reconstruct trajectory stream in 3D as described in Section 6. These 3D reconstructed trajectories

[1]Our system reaches average 6.4 pixels/cm^3, resulting in the most dense 3D pixel continuum. cf) 0.44 pixels/cm^3 for the Panoptic Studio at CMU [18,19]



(a) Geometry (b) 2D multiple view recognition (c) View-pooling

Figure 2. (a) A 3D point $\mathbf{X}_t$ at the t time instant is observed by multiple cameras $\{\mathbf{P}_c\}_{c\in\mathcal{C}}$ where the point is fully visible to the c^{th} camera if $V(\mathbf{X}_t, c) = 1$, and zero otherwise. We denote the 2D projection of the 3D point onto the camera as $P(\mathbf{X}_t, c)$. (b) For each image $\mathcal{I}_c$, we use the recognition confidence (body segmentation [23]/object bounding box [29]) to build $L_{2D}(\mathbf{x}|\mathcal{I}_c)$ at each pixel $\mathbf{x}$ where the i^{th} element of L_{2D} is the likelihood (confidence) of the recognition for the i^{th} object class as shown on the right. For the illustration purpose, we only visualize the likelihood of body segments overlaid with the image while L_{2D} also includes object classes. (c) We construct the 3D semantic map $L_{3D}(\mathcal{X})$ via pooling L_{2D} over multiple views (view-pooling) by reasoning about visibility. The magenta camera is the visible camera set, and the bar graphs represent L_{2D}. The figures are best seen in color.

are used to reason about their semantic labels by consolidating 2D recognition confidence in multiple view images: 3D semantic map is constructed using view-pooling (Section 5.1), and affinity between long range fragmented trajectories is measured by computing local rigid transformation (Section 5.2).

4. Notation

We represent a fragmented trajectory with a time series of 3D points: $\mathcal{X} = \{\mathbf{X}_t \in \mathbb{R}^3\}_{t=T_e}^{T_d}$ where $\mathbf{X}_t$ is the 3D point in the trajectory at the t time instant, and T_e and T_d are emerging and dissolving moments of the trajectory, respectively.

The 3D point $\mathbf{X}_t$ is projected onto the visible c^{th} camera projection matrix, $\mathbf{P}_c = \mathbf{K}_c\mathbf{R}_c \begin{bmatrix} \mathbf{I}_3 & -\mathbf{C}_c \end{bmatrix} \in \mathbb{R}^{3\times4}$ to form the 2D projection, $P(\mathbf{X}_t, c) \in \mathbb{R}^2$ where $\mathbf{K}_c$ is the intrinsic parameter of the camera encoding focal length and principal points, and $\mathbf{R}_c \in SO(3)$ and $\mathbf{C}_c \in \mathbb{R}^3$ are the extrinsic parameters (rotation and camera center), i.e., $P(\mathbf{X}_t, c) = \begin{bmatrix} \mathbf{P}_c^1\widetilde{\mathbf{X}}_t/\mathbf{P}_c^3\widetilde{\mathbf{X}}_t & \mathbf{P}_c^2\widetilde{\mathbf{X}}_t/\mathbf{P}_c^3\widetilde{\mathbf{X}}_t \end{bmatrix}^{\mathsf{T}}$ where $\widetilde{\mathbf{X}}$ is the homogeneous representation of $\mathbf{X}$, and $\mathbf{P}_c^i$ indicates the i^{th} row of $\mathbf{P}_c$. We assume the camera extrinsic and intrinsic parameters are pre-calibrated and constant across time (no time index).

A probability of point visibility at c^{th} camera is represented as $V(\mathbf{X}_t, c) \in [0, 1]$ as shown in Figure 2(a). The c^{th} camera produces the image at the t time instant $\mathcal{I}_t^c$. Each pixel $\mathbf{x}$ is associated with the confidence of semantic labels, i.e., $L_{2D}\left(\mathbf{x} \in \mathbb{R}^2|\mathcal{I}_c\right) \in [0, 1]^N$ where N is the number of object classes[2]. For instance, L_{2D} can be approximated by the last layers of a convolutional neural network as shown in Figure 2(b). Our framework can build on general 2D recognition framework that can produce a confidence map while in this paper, we focus on two main pre-trained models: body semantic segmentation [23] and bounding box object recognition [29] trained with COCO [24] and ImageNet [31] datasets.

[2]The object classes include objects, body parts, and independent instances.

5. Semantic Trajectory Labeling

Given 3D reconstructed trajectories, we present a method to precisely infer their semantic labels. A fundamental challenge lies in the fragmented and noisy nature of the 3D reconstruction and image semantic labeling. A key innovation is the *3D semantic map* that can encode the visual semantics of a 3D trajectory by consolidating the 2D recognition confidence across multiple view image streams. We integrate the 3D semantic map in conjunction with long term trajectory affinity into a graph-cut formulation to infer the semantic labels jointly.

5.1. 3D Semantic Map

We define the 3D semantic map, $L_{3D} \in [0, 1]^N$, a probability distribution over semantic labels per 3D based on a probability of visibility and 2D recognition confidence measured at the 2D projections of the trajectory onto all multiple view images:

$$L_{3D}\left(\mathcal{X}\right) = \frac{1}{\Delta T}\sum_{t=T_e}^{T_d} \underset{c\in\mathcal{C}}{\text{Pool}}\left(L_{2D}\left(P\left(\mathbf{X}_t, c\right)|\mathcal{I}_c\right)\right), \quad (1)$$

where $\Delta T = T_d - T_e$ is the life span of the trajectory, $c \in \mathcal{C}$ is the camera index, and $\mathcal{C}$ is the camera index set, i.e., $|\mathcal{C}|$ is the number of cameras. The 3D trajectory label is evaluated at the 2D projection $P\left(\mathbf{X}_t, c\right)$ across all cameras over the trajectory life span.

To alleviate noisy and coarse 2D recognition results, we

introduce a view-pooling operation:

$$L_{c^*} = \underset{c \in \mathcal{C}}{\text{Pool}}\left(L_c\right) \quad \text{s.t. } c^* = \underset{c \in \mathcal{C}}{\text{argmin}} \sum_{j=1}^{C} V_c \|L_c - L_j\|^2,$$

where we denote $L_{2D}\left(P\left(\mathbf{X}_t, c\right)|\mathcal{I}_c\right)$ as L_c, and $V(\mathbf{X}_t, c)$ as V_c by an abuse of notation. The view-pooling operation finds the best view among the visible cameras that is consistent with other view predictions (the weighted median of $\{L_c\}_{c \in \mathcal{C}}$).

The view-pooling operation is based on our conjecture that among many views, there exist a few views that can confidently predict an object label. It is robust to noisy recognition outputs as shown in Figure 2(b) where many false positive bounding boxes are detected. The visibility based confidence measure can suppress inconsistent detection across views, and weighted median pooling can prevent from a view biased L_{3D}. This allows the pooled L_{2D} temporally consistent, which makes averaging over time meaningful.

Figure 2(c) illustrates the view-pooling operation over all multiview image streams. A set of L_c (bar graphs) at the projected locations $\{P(\mathbf{X}, c)\}_{c \in \mathcal{C}}$ are used for the view-pooling that finds the L_{c^*} that best represents the distribution of L_c. For an illustrative purpose, we highlight the cameras that have high visibility with magenta color, i.e., $V(\mathbf{X}, c) > \epsilon_e$.

5.2. 3D Trajectory Affinity

An object that undergoes locally rigid motion provides a spatial cue to identify the affinity between fragmented trajectories. Consider two trajectories $\mathcal{X}_i$ and $\mathcal{X}_j$ that have overlapping lifetime, $\emptyset \neq \mathcal{S} = [T_e^i, T_d^i] \cap [T_e^j, T_d^j]$ where the superscript in T_e and T_d indicates the index of the trajectory. We measure the affinity of the trajectories as follow:

$$A(i,j) = \exp\left(-\left(\|\mathbf{e}_i^j\|/\tau\right)^2\right) \tag{2}$$

where $A \in \mathbb{R}^{M \times M}$ is an affinity matrix whose (i,j) entry measures the reconstruction error:

$$\mathbf{e}_i^j = \max_{t-1, t \in \mathcal{S}} \left\|\mathbf{X}_t^j - \mathbf{R}_t^i \mathbf{X}_{t-1}^j - \mathbf{t}_t^i\right\|.$$

$\mathbf{e}_i^j$ is the Euclidean distance between $\mathbf{X}_t^j$ and the predicted point by its emerging location $\mathbf{X}_{T_e}^j$ via its local transformation $(\mathbf{R}_t^i, \mathbf{t}_t^i) \in SE(3)$ (rotation and translation) learned by the i^{th} trajectory $\mathcal{X}_i$. This measure can be applied to long range trajectories, which establish a strong connection across an object, e.g., left hand to left elbow trajectories. $i, j \in \mathcal{T} = \{1, \cdots, M\}$ where M is the number of trajectories. Unlike difference of pairwise point distance measure that has been used for trajectory clustering [18], our affinity takes into account general Euclidean transformation ($SE(3)$) that directly measures rigidity.

We learn the local transformation $(\mathbf{R}_t^i, \mathbf{t}_t^i)$ of the i^{th} trajectory at each time instant, given a set of neighbors:

$$\mathbf{R}_t^i = \Delta\mathbf{X}_t^{\mathcal{N}_i}\left(\Delta\mathbf{X}_{t-1}^{\mathcal{N}_i}\right)^{-1}, \quad \mathbf{t}_t^i = \mathbf{R}_t^i \mathbf{X}_{t-1}^i - \mathbf{X}_t^i \tag{3}$$

where $\Delta\mathbf{X}_t^{\mathcal{N}_i}$ is a matrix whose columns are made of relative displacement vectors of neighboring trajectories with respect to $\mathcal{X}_i$, i.e., $\Delta\mathbf{X}_t^j = \mathbf{X}_t^j - \mathbf{X}_t^i$ where $j \in \mathcal{N}_i$ is the index of neighboring trajectories. The set of neighbors are chosen as

$$\mathcal{N}_i = \left\{ j \,\middle|\, \max_{t \in \mathcal{S}} \left\|\mathbf{X}_t^j - \mathbf{X}_t^i\right\| < \epsilon \right\},$$

where ϵ is the radius of a 3D Euclidean ball. Note that not all ϵ-neighbors belong to the same object which requires to evaluate the trajectory with Equation (2).

In practice, evaluating Equation (2) for all trajectories are computationally prohibitive. For example, it requires 10^{10} evaluations are needed for 100,000 trajectories[3] to fill in all entries in the affinity matrix A. Since it is unlikely that far distance trajectories belong to the same object class, we restrict the evaluations only for ϵ_a-neighbors ($\mathcal{N}_i^a$) that are sufficient to cover a large portion of objects and greater ϵ, e.g., $\epsilon_a = 30$cm and $\epsilon = 5$cm. Further, we randomly drop-out connections between neighboring trajectories for computational efficiency. This also increases the robustness of trajectory affinity that is often biased by the density of trajectories. When computing the local transformation in Equation (3), we embed RANSAC [14]: choosing random three trajectories from ϵ-neighbors and finding the local transformation that produces the maximum number of inliers.

5.3. Trajectory Label Inference

Inspired by multi-class pixel labeling using α-expansion [7], we infer the trajectory labels $U : \mathcal{T} \rightarrow \mathcal{L}$ where $\mathcal{L} = \{1, \cdots, N\}$ is the index set of object classes, by minimizing the following cost:

$$C(U) = \sum_{i \in \mathcal{T}} \phi(l_i, U(i)) + \lambda \sum_{i \in \mathcal{T}} \sum_{j \in \mathcal{N}_i^a} \psi(U(i), U(j)) \tag{4}$$

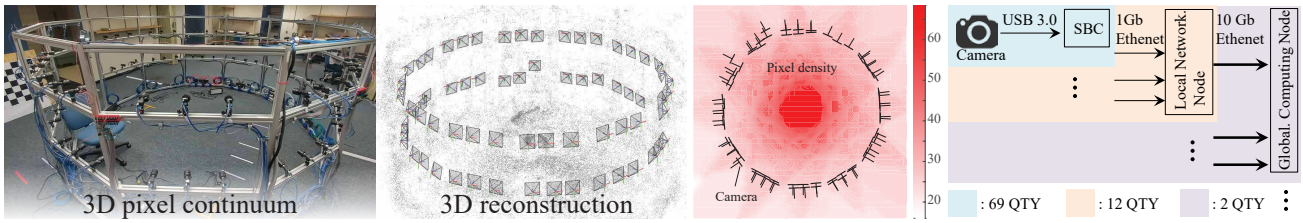
where λ is a hyper-parameter that control the weight between data ϕ and smoothness ψ costs.

The data cost can be written as:

$$\phi(l_i, U(i)) = \begin{cases} 0 & \text{if } l_i = U(i) \\ L_{3D}\left(\mathcal{X}_i\right)_{l_i} & \text{if } l_i \neq U(i) \end{cases},$$

where it penalizes the discrepancy between the 3D semantic map predicted by a series of 2D recognitions and assigned

[3] In our experiments, the number of trajectories is order of $10^4 \sim 10^6$.



(a) System geometry (b) Pixel density (c) System architecture

Figure 3. (a) We build a multi-camera system composed of 69 cameras running at 30 Hz. (b) The multi-camera system creates the 3D pixel continuum where all 3D points in the enclosed space are measured by multiple images. We visualize the pixel density using maximum intensity projection seen from top view. At the center of the stage, more than 60 pixels can measure a unit cm^3 cubic. (c) The system architecture is designed using modular units, which makes the system highly scalable.

label. $L_{3D}\left(\mathcal{X}_i\right)_{l_i}$ is the l_i^{th} entry of L_{3D} that measures the likelihood of $\mathcal{X}_i$ being class l_i.

The smoothness cost can be described by the trajectory affinity:

$$\psi(U(i),U(j)) = \begin{cases} 0 & \text{if } U(i)=U(j) \\ A(i,j) & \text{if } U(i)\neq U(j) \end{cases},$$

where it penalizes the label difference between trajectories that undergo the same local rigid transformation. l_i is the label index computed from L_{3D}:

$$l_i = \underset{l\in\mathcal{L}}{\text{argmax}}\ L_{3D}\left(\mathcal{X}_i|\{\mathbf{P}_c,\mathcal{I}_c\}_{c\in\mathcal{C}}\right).$$

Due to multi-class labeling, minimization of Equation (4) is highly nonlinear while the iterative α-expansion algorithm has been shown a strong convergence towards the global minimum [7, 12].

6. 3D Trajectory Reconstruction

We reconstruct 3D trajectory stream by leveraging the era system described in Section 7. In this section, we describe the procedure of the 3D trajectory reconstruction algorithm modified from Joo et al. [19] to produce denser and more accurate trajectories. **(1) Camera calibration** We calibrate the intrinsic parameter of each camera (focal length, principal points, and radial lens distortion), independently, and use standard structure from motion to calibrate extrinsic parameters (relative rotation and translation). In the bundle adjustment, the extrinsic and intrinsic parameters are jointly refined. To accelerate further image based matching, we learn the image connectivity graph [37] $\mathcal{G}_m = (\mathcal{V}_m, \mathcal{E}_m)$ through exhaustive pairwise image matching, e.g., two cameras that have more than 90 degree apart are unlikely to match to each other. **(2) Point cloud triangulation** At each time instant, we find dense feature correspondences using grid-based motion statistics (GMS) [6] among $\mathcal{G}_m$ and triangulate each 3D point $\mathbf{X}$ with RANSAC. The initial visibility for the c^{th} camera is set to $V(\mathbf{X},c) = \exp(-\left(\|P(\mathbf{X},c) - x(c)\|/\sigma\right)^2)$ where the σ is the tolerance of the reprojection error and $x(c)$ is the corerspondecne point at camera c. **(3) 3D point tracking** The triangulated points are used for build trajectory stream. For each point $\mathbf{X}_t$ at the t time instant, we project the point onto the visible set of cameras, i.e., $P(\mathbf{X}_t, c \in \mathcal{V})$ where $\mathcal{V} = \{j|V(\mathbf{X}_{t-1},c) > \epsilon_s\}$ where ϵ_s is the threshold for the probability of visibility. These projected points are tracked in 2D using optical flow and triangulated with RANSAC to form $\mathbf{X}_{t+1}$. Similar to the visibility initialization, the probability of visibility $V(\mathbf{X}_{t+1}, c)$ is updated using reprojection error. We iterate this process (tracking→triangulation→visibility update) until the average reprojection is higher than 2 pixels or the number of visible cameras $|\mathcal{V}|$ is less than 2.

7. 3D Pixel Continuum Design

To demonstrate the 3D pixel continuum where every 3D point is observed by multiple cameras, we build a large scale multi-camera system composed of 69 cameras as shown in Figure 3(a). Two rows of the cameras enclose cylindrical space (3m diameter $\times$ 2.5m height) that facilitates capturing diverse human interactions. A camera produces a HD resolution image (1280$\times$1024) where the maximum pixel density per unit cm^3 reaches to more than 60 pixels. It runs at 30 Hz precisely triggered by a master camera node: the master camera sends PWM signal through General Purpose Input/Output (GPIO) port when its shutter opens, which triggers the rest 68 slave cameras, achieving sub-nano second accuracy. To alleviate the trigger signal attenuation due to a number of camera connections, we design a signal amplifier that can feed the targeted electric current.

All cameras produce a sheer amount of visual data at each second (280 GB/s), which introduces severe data traffic in the global computing node. Instead, we modularize the image processing using a single board computer (SBC): the image data stream from each camera is transferred through USB 3.0 to its own SBC that is dedicated to JPEG image compression, resulting in $\sim$ 400 KB/image with minimal loss of image quality. This compressed data is transferred to two global computing nodes through multi-

ples of 10 Gb Ethernet network switches. The global computing nodes write the data into designated PCIe interfaced solid state drives (SSD). The architecture is summarized in Figure 3(c).

The key features of the system design is scalability and cost effectiveness. The modularized system design allows increasing the number of cameras and size of the system without introducing system complexity: the module of camera-SBC-Network switch can be augmented in the existing system. Also the hardware frame is build on modular T-slotted aluminum frame where the modification of geometric camera placement can be easily customizable. All parts including hardware, electronic devices, and cameras are commodity items where no system specific design is needed.

8. Results

To validate our semantic trajectory reconstruction algorithm, we evaluate on real-world datasets collected by the 3D pixel continuum described in Section 7.

8.1. Human Interaction Dataset

9 new vignettes that include diverse human interactions are captured: **Pet interaction**: A dog owner naturally interacts with her dog: ask him to sit, turn around and jump. The dog also plays with his doll and seek snack while walking around with the owner. This pet interaction demonstrates strength of our system, i.e., reconstructing fine detailed interactions, not limited to humans [18]; **International Latin ballroom dance**: Two sport dancers practice for Cha-cha style dance competition where the physical interactions between them are highly stylized. The dancers wear textureless black suit and skirt where semantic labeling is likely noisy; **K-Pop group dance**: Two experienced K-Pop dancers perform the group break dance. The dances are designed to be synchronized, jerky, and fast; **Object manipulation**: Two students manipulate various objects such as doll, flowerpot, monitor, umbrella, and hair drier in a cluttered environments. This vignette demonstrates that the system is able to handle multiple objects; **Bicycle riding**: A person rides a bicycle that induces large displacement. This interaction introduces significant occlusion, i.e., the person is a part of the bicycle; **Tennis swing**: A person practices fore- and back-hand strokes with a tennis racket. The tennis racket is often difficult to detect as the racket head is mostly transparent; **Basketball I**: A student player practices dribbling which includes fast ball motion; **Basketball II**: An other player tries to block the opponent's motion that includes severe occlusion between players. We make these data including images, calibration, 3D trajectories, and their semantic labels, publicly available through the following website: http://www-users.cs.umn.edu/~jsyoon/Semantic_trajectory/

8.2. Quantitative Evaluation

We quantitatively evaluate our representation and algorithm in terms of three criteria: (1) robustness of 3D semantic map (view-pooling); (2) effectiveness of the affinity measure; and (3) predictive validity of semantic labels where all datasets are used for the evaluations. Note that as no ground truth data or benchmark dataset is available, we conduct ablation studies to validate our methods.

Robustness of 3D semantic map We introduce the view-pooling operation that takes the weighted median of recognition confidence based on visibility. This operation allows robustly predicting the 3D semantic map L_{3D} as it is not sensitive to erroneous detection. To evaluate its robustness, we measure the temporal consistency of the view-pooling operation along a trajectory. Ideally, the view-pooled recognition confidence should remain constant across time as it belongs to the trajectory of the same object. We compare the view-pooling with average-pooling across randomly all cameras using normalized correlation measure across time, i.e., $NC(L^0_{\text{vp}}, L^t_{\text{vp}})$ where L^t_{vp} is the view-pooled recognition confidence at the t time instant. We summarize the results on all sequences in Table 1. Our method shows a graceful degradation as time progress up to 15 seconds while the average-pooling is highly biased by noisy recognition, which produces drastic performance gradation (no temporal coherence).

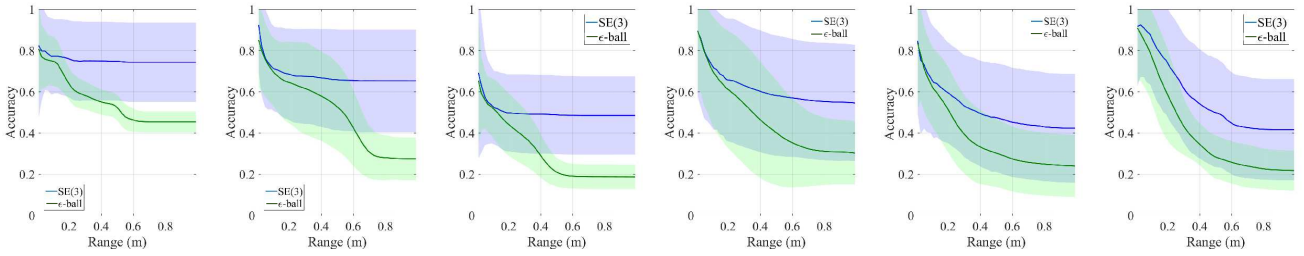
Time (second)	1s	3s	5s	7s
View pool	**0.96**±0.01	**0.90**±0.02	**0.89**±0.03	**0.88**±0.02
Ave. pool	0.43±0.10	0.44±0.10	0.43±0.10	0.48±0.09
Time (second)	9s	11s	13s	15s
View pool	**0.89**±0.02	**0.88**±0.03	**0.87**±0.05	**0.79**±0.08
Ave. pool	0.44±0.09	0.43±0.10	0.42±0.10	0.37±0.10

Table 1. Time consistency of 3D semantic map

Effectiveness of affinity measure We compute the affinity based on local transformation per trajectory. This method is highly effective to relate with long term fragmented trajectories. We compare the validity of our affinity measure with that of ϵ_s-neighbors ($\mathcal{N}_s$), i.e., the distance between trajectories over time remains less than ϵ_s. To evaluate, two neighboring trajectories for both methods are randomly chosen and projected onto cameras. Concretely, we measure $\sum_{j\in\mathcal{N}_s} E(i,j)$ where

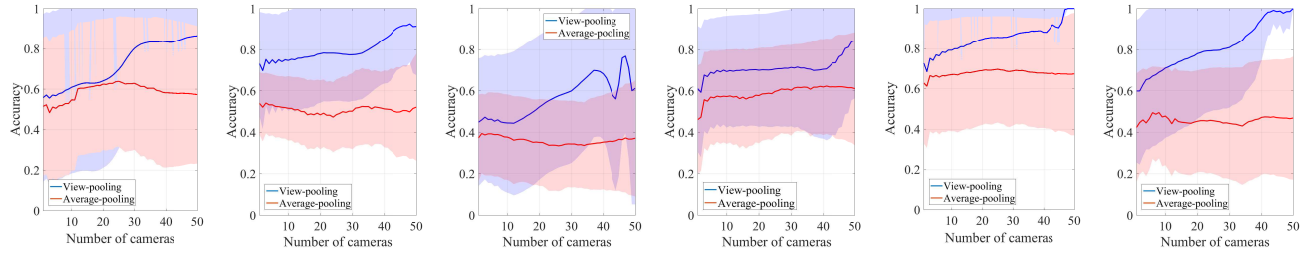
$$E(i,j) = \begin{cases} 0 & \text{if } L(P(\mathbf{X}^i_t, c)|\mathcal{I}_c) = L(P(\mathbf{X}^j_t, c)|\mathcal{I}_c) \\ 1 & \text{otherwise} \end{cases}.$$

$L : \mathbb{R}^2 \rightarrow \mathcal{L}$ outputs the semantic label index given the 2D projection. If the measure is small, it indicates that the neighbors are correctly identified. Figure 4 illustrates the comparison over 6 different sequences. Each one has different global and local motion. If the motion is largely global, the affinity measure can confuse as multibody motion is identified as a rigid body motion as shown in Basketball



(a) Pet int. (b) L. dance (c) K-Pop (d) Bike (e) Tennis (f) Basketball II

Figure 4. We evaluate the effectiveness of our affinity map computed by estimating local Euclidean transformation SE(3). While the effectiveness of ϵ_s-neighbors diminishes rapidly after 10 cm, our method still holds for longer range, e.g., 1 m.



(a) Pet int. (b) L. dance (c) K-Pop (d) Tennis (e) Basketball I (f) Basketball II

Figure 5. We evaluate semantic label prediction via an ablation study: to use a subset of cameras to assign the semantic labels to the trajectories and validate the labels by comparing the labels of projections with the held-out images. Our view-pooling method outperforms the average-pooling with large margin for all sequences.

II. Nonetheless, our method outperforms the ϵ_s-neighbors for all sequences. In particular, it shows much stronger performance at long range trajectories (0.6-1 m), which makes the large scale label inference possible.

Predictive validity of 3D semantic label We evaluate the semantic label inference via cross validation scheme. We label a 3D trajectory with a subset of cameras and project onto the held-out camera to evaluate the predictive validity. Ideally, the trajectory label should be consistent with any view as visibility is considered, and therefore, the projected label must agree with the recognition result. As we infer the semantic labels of the trajectories jointly by consolidating multiple view recognition, the number of cameras plays a key role in the inference. We test the predictive validity by changing the number of cameras to label trajectories as shown in Figure 5. When the number of cameras is few, e.g., 1-5, our method using view-pooling performs similarly with average-pooling. However, the performance quickly is boosted as the number of camera increases, i.e., in most cases, it produces more than 0.6 accuracy at 20 cameras for inference. In Table 2, we further compare our method with the approach from Joo et al. [18], where the semantic label on the trajectory is inferred by 3D human body anatomical key-points. As highlighted in Figure 6, our method outperforms [18] in all possible scenarios (e.g. occlusion, dynamic deformation, object interaction, multiple people).

	R.motion	B.ball I	Latin	K-Pop	Pet	Bike	Tennis
Joo et al. [18]	0.8547	0.8862	0.7532	0.5019	0.4819	0.5307	0.7317
AP(1)	0.7532	0.6271	0.5388	0.3730	0.5145	0.5297	0.4607
AP(30)	0.8578	0.6879	0.5014	0.3431	0.6276	0.6341	0.6029
AP(69)	0.8584	0.7309	0.7769	0.5706	0.6018	0.6162	0.6691
VP(1)	0.8403	0.7259	0.7307	0.4485	0.5755	0.7432	0.6099
VP(30)	0.9092	0.8650	0.7753	0.5992	0.8015	0.7064	0.7133
VP(69) [**Ours**]	**0.9326**	**0.9572**	**0.8753**	**0.6985**	**0.8132**	**0.8394**	**0.8438**

Table 2. We compare our method with multiple baselines in terms of accuracy. AP(x) and VP(x) refer to average-pooling and view-pooling, respectively where x is the maximum number of visible cameras.

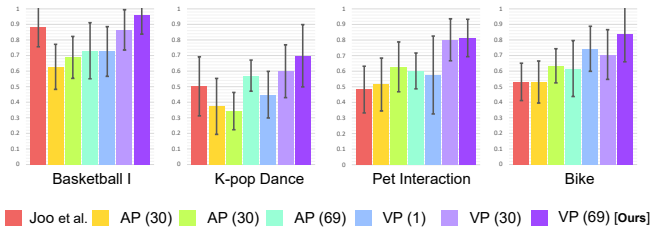
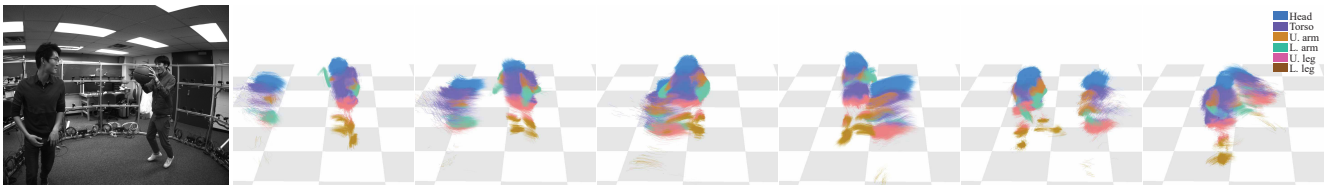


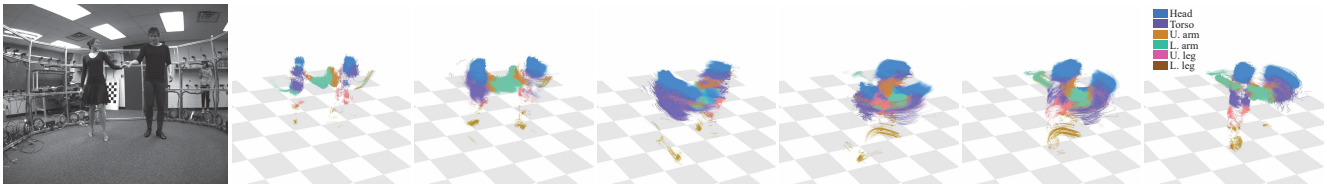
Figure 6. Our method outperforms all baselines. The notation, AP(x) and VP(x) are consistent with in Table 2

8.3. Qualitative Evaluation

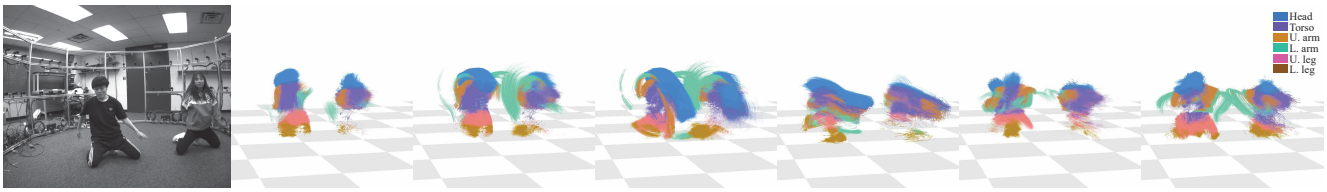
We apply our method to reconstruct dense semantic trajectories in 3D as shown in Figure 1, 7, 8, and 9. The colors of the trajectories are associated with the semantic labels.



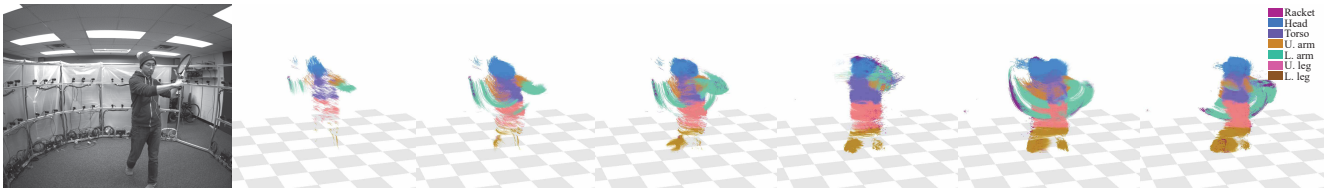
(a) Basketball II



(b) Latin dance



(c) K-Pop



(d) Tennis

Figure 7. Qualitative evaluation. Best seen in color. For an illustrative purpose, the last 30 frames of the trajectories are visualized.

Figure 8. Pet interaction

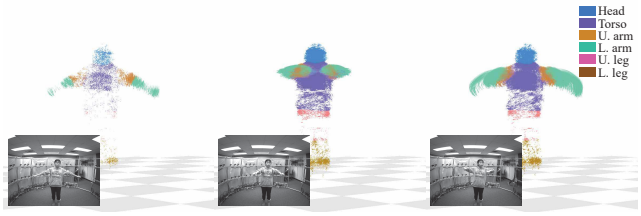


Figure 9. Range of motion

9. Discussion

We present an algorithm to reconstruct semantic trajectories in 3D using a large scale multi-camera system. This problem is challenging because of fragmented trajectories and noisy/coarse recognition in 2D. We introduce a new representation to encode the visual semantics to each trajectory called 3D semantic map that allows us to consolidate multiple view noisy recognition results by leveraging view pooling based on their visibility and recognition confidence. 3D spatial relationship between fragmented trajectories is modeled by local rigid transformation that can establish the connection between long range trajectories. These two cues are integrated into a graph-cut formulation to infer precise labeling of the trajectories. Note that Our framework is not specific to the choice of the 2D recognition models.

The first wave of the optic technology enabled cameras to be emerged and embedded in our space. The second wave will be *connectedness*: multiple cameras will measure our interactions and cooperatively understand their semantic meaning. This paper takes the first bold step towards establishing a computational basis for understanding 3D semantics at fine scale.

References

[1] I. Akhter, Y. Sheikh, and S. Khan. In defense of orthonormality constraints for nonrigid structure from motion. In *CVPR*, 2009. 2

[2] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Nonrigid structure from motion in trajectory space. In *NIPS*, 2008. 2

[3] I. Akhter, T. Simon, S. Khan, I. Matthews, and Y. Sheikh. Bilinear spatiotemporal basis models. *SIGGRAPH*, 2012. 2

[4] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *IJCV*, 1992. 2

[5] S. Avidan and A. Shashua. Trajectory triangulation: 3D reconstruction of moving points from a monocular image sequence. *PAMI*, 2000. 2

[6] J. Bian, W.-Y. Lin, Y. Matsushita, S.-K. Yeung, T. D. Nguyen, and M.-M. Cheng. Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. In *CVPR*, 2017. 5

[7] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 2001. 4, 5

[8] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *CVPR*, 1999. 1, 2

[9] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010. 2

[10] Y. Dai, H. Li, and M. He. A simple prior-free method for non-rigid structure-from-motion factorization. In *CVPR*, 2012. 2

[11] A. Del Bue, X. Lladó, and L. Agapito. Segmentation of rigid motion from non-rigid 2d trajectories. *Pattern Recognition and Image Analysis*, 2007. 2

[12] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov. Fast approximate energy minimization with label costs. *IJCV*, 2012. 5

[13] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. In *CVPR*, 2009. 2

[14] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *ACM Communications*, 1981. 4

[15] K. Fragkiadaki, M. Salas, P. Arbelaez, and J. Malik. Grouping-based low-rank trajectory completion and 3d reconstruction. In *NIPS*, 2014. 2

[16] K. Fragkiadaki, G. Zhang, and J. Shi. Video segmentation by tracing discontinuities in a trajectory embedding. In *CVPR*, 2012. 2

[17] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004. 2

[18] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015. 2, 4, 6, 7

[19] H. Joo, H. S. Park, and Y. Sheikh. Map visibility estimation for large-scale dynamic 3d reconstruction. In *CVPR*, 2014. 1, 2, 5

[20] J. Y. Kaminski and M. Teicher. A general framework for trajectory triangulation. *Journal of Mathematical Imaging and Vision*, 2004. 2

[21] A. Kundu, Y. Li, F. Daellert, F. Li, and J. M. Rehg. Joint semantic segmentation and 3d reconstruction from monocular video. In *ECCV*, 2014. 2

[22] A. Kundu, V. Vineet, and V. Koltun. Feature space optimization for semantic video segmentation. In *CVPR*, 2016. 2

[23] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017. 3

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[25] K. E. Ozden, K. Cornelis, L. V. Eychen, and L. V. Gool. Reconstructing 3D trajectories of independently moving objects using generic constraints. *CVIU*, 2004. 1

[26] H. S. Park and Y. Sheikh. 3d reconstruction of a smooth articulated trajectory from a monocular image sequence. In *ICCV*, 2011. 2

[27] H. S. Park, T. Shiratori, I. Matthews, and Y. Sheikh. 3D reconstruction of a moving point from a series of 2D projections. In *ECCV*, 2010. 2

[28] S. Rao, R. Tron, R. Vidal, and Y. Ma. Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. *PAMI*, 2010. 2

[29] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In *CVPR*, 2017. 3

[30] S. Ricco and C. Tomasi. Video motion for every visible point. In *ICCV*, 2013. 2

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 3

[32] C. Russell, R. Yu, and L. Agapito. Video pop-up: Monocular 3d reconstruction of dynamic scenes. In *NIPS*, 2014. 2

[33] M. Salzmann, J. Pilet, S. Ilic, and P. Fua. Surface deformation models for nonrigid 3D shape recovery. *PAMI*, 2007. 2

[34] A. Shaji, A. Varol, L. Torresani, and P. Fua. Simulataneous point matching and 3D deformable surface reconstruction. In *CVPR*, 2010. 2

[35] Y. Sheikh, O. Javed, and T. Kanade. Background subtraction for freely moving cameras. In *ICCV*, 2009. 2

[36] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3d human figures using 2D image motion. In *ECCV*, 2000. 2

[37] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from Internet photo collections. *IJCV*, 2008. 2, 5

[38] B. Taylor, A. Ayvaci, A. Ravichandran, and S. Soatto. Semantic video segmentation from occlusion relations within a convex optimization framework. In *CVPR Workshop*, 2013. 2

[39] J. Taylor, A. D. Jepson, and K. N. Kutulakos. Non-rigid structure from locally-rigid motion. In *CVPR*, 2010. 2

[40] L. Torresani and C. Bregler. Space-time tracking. In *ECCV*, 2002. 2

[41] L. Torresani, D. Yang, G. Alexander, and C. Bregler. Tracking and modeling non-rigid objects with rank constraints. In *CVPR*, 2001. 1, 2

[42] J. Yan and M. Pollefeys. Automatic kinematic chain building from feature trajectories of articulated objects. In *CVPR*, 2006. 1, 2