Spatio-Temporal reconstruction of Rigid objects from multiple cameras

Abstract

Multi camera based understanding of dynamic scenes from unsynchronized images is a challening problem in reconstruction. We propose a solution to the problem of 3D reconstruction of multiple moving rigid objects from unsynchronized multiple cameras. Most of the reconstruction piplines leverage reconstruction using reprojection error of features, we propose a novel semantic features pipeline and object based feature tracking for reconstruction of dynamic objects. We leverage multiple rigid object constraints in a bundle adjustment framework to improve the reconstuction.

1. METHOD

We Follow the basic pipeline of reconstcruion i.e. feature detection and tracking, correspodence of the keypoint across views and then bundle adjustment for accurate reconstruction of the 3d points. In each of the segments of the pipeline we incorporate geometric and semantic information of the objects for better reconstruction. We propose instance keypoint segmentation for better feature detection of important points on the rigid object. We first formulate the problem statement for the problem. Lets assume N video cameras observing P 3d points over time $X^p(t)$ and their 2D projection $x_n^p(f)$ on camera p and frame f is given as:

$$\begin{bmatrix} x_n^p(f) \\ 1 \end{bmatrix} = K_n(f)[R_n(f)|T_n(f)] \begin{bmatrix} X_n(f) \\ 1 \end{bmatrix}$$

where $K_n(f)$ is the intrisic camera matrix, $R_n(f)$ and $T_n(f)$ are the relative camera rotation and translation, respectively. We denote this transformations $x_c^n(f) = P_n(f, X^n(t))$

1.1. Instance Aware Keypoint detection

Given a images with multiple moving objects. The amount of objects which can move can be sorted into a small set of categories like cars and humans. We propose a instance keypoint detection on these categories for improving the reconstruction of these objects. We leverage deep learning framwork for fast and accurate detection of semantic features in images. We pass the input image through a detection pipeline and each detected bounding box is passed through a object specific keypoint detection pipeline for computation of semantic features in an image.

1.2. Object aware keypoint tracking

Tracking of keypoints has generally been considered as a search problem over the image space for features of similar shape in adjustcent images. Semantic aware tracking correspondences to tracking of the important parts using the knowledge of the object i.e. the keypoints on the wheels are tracked by finding the correspondences of cars overtime. This gives a better tracks of the keypoints and longer correpondeces compared to the feature based tracking methods.

1.3. Correspondence across views

We currently have tracks of objects and thier respective keypoints in each individual video for all videos of the scene. For accurate reconstruction of objects from multiple views we need to find corrospondences across views. We propose a object specific bundle adjustment for finding correspondes across views. Lets assume we have h object proposals in N videos at time t. The problem of correspondence is finding g, where $g \subset h$ object bounding boxes which belong to the same object. Since there is a spatial constraint on the bounding boxes over multiple views, Fig 1 shows the bundle adjustment we use to find correspondes across views. We propose a ransac based methodology for finding correspondenses across views and minimize the following reprojection error:

$$E_r = \sum_{n=1}^{N} \sum_{p=1}^{P} V_c^n(f)\sigma_c^n(F)||P_n(f, X^p(t)) - x_n^p(f)||^2 \quad (1)$$

where E_r is the image reprojection error, $V_n^p(f)$ is a binary indicator of the visibilty of the keypoint. We add additional constraints like symmetry and the length of different parts of the object for accurate fitting of the rigid object in 3D.

$$E_j = \quad (2)$$

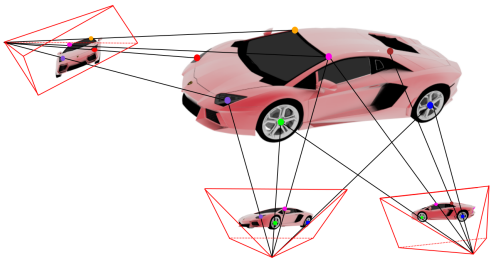

Figure 1. The picture depicts the projection of 3D keypoints of car onto thier respective videos

$$E_l = \sum_{n=1}^{N} \sum_{p=1}^{P} V_c^n(f)\sigma_c^n(F)||P_n(f, X^p(t)) - x_n^p(f)||^2 \quad (3)$$

1.4. Spatio-Temporal Bundle Adjustment

Once we have the correspondences across cameras and views, We get the fit a motion model based rigid object over time. The formulation is given by:

(4)