Dynamic object detection and reconstruction using multi-view bootstrapping

Abstract

Reconstruction of dynamic scenes from multiple unsynchronized cameras in the wild is a challenging problem. Most of the research in the multi camera based reconstruction has been focused in two aspects. With multiple cameras being stationary with dynamic objects in the scene or static scene with cameras moving. We propose a algorithm to solve the problem of dynamic scene reconstruction from multiple moving cameras. We also show the advantages of our method by proposing a improvement in detection using multi-view bootstrapping. We use semantic keypoints and spatio-temporal bundle adjustment to densily reconstruct the dynamic scenes. Using the accurate reconstructions, we improve the accuracy of the keypoint detections. The novel method of using multiple camera based method for training the detections can be applied to any task in computer vision.

1. INTRODUCTION

Reconstruction and infering about a scene is the most important problem in computer vision. Although there have been plethora of work in scene understanding and reconstruction from images and videos. Most of the methods cannot completely infer the dynamics of a scene because of bottleneck of the number of views. We believe this problem can be solved using multiple cameras recording the scene. Most of the research in this direction has been from multiple static cameras. We propose reconstruction of dynamic objects captured from multiple unsynchronized cameras. We believe accurate reconstruction of scene in the wild can be used for training wide range of computer vision tasks like detection, semantic segmentation etc. Further reconstructions from multi camera systems have wide applications in areas like robot navigation, virtual reality and Autonomous driving.

With the advent of deep learning most of the research has focused on end-to-end learning of most of the tasks in computer vision. We believe that geometric constraints impose a very strong prior on the scene and exploiting these constraints help in learning the reconstruction of a scene. Our method proposes a reconstruction pipeline using features from the deep learning based keypoint detection algorithm. We further use geometry based multi-view constraint to improve the detection accuracy in the images. We show results on keypoint based improvement algorithm but the multi view information can be used to improve and learn any vision task like detection segmentation etc.

Dynamic objects reconstruction is a well studied problem due to its applications in all the computer vision problems. Most of the earlier works have used sparse features to segment and reconstruct moving and stationary objects simultaneously. We interpolate the idea to semantic keypoints i.e. wheels, head lights etc and densely reconstruct the enviroment. We add multiple constraints using bundle adjustment like motion prior, semantic cues for improving the reconstructions.

The main contributions of the paper are outlined:

1. Time synchronization of multiple videos by optimizing for motion of moving objects.

2. Constraint Bundle adjustment for reconstruction of dynamic moving objects viewed from multiple cameras.

3. Improving the detection accuracy using reconstruction of dynamic scenes in the wild.

2. RELATED WORK

2.1. Multi-Camera systems

2.2. 3D reconstruction

2.3. Geometry for learning

3. PROBLEM STATEMENT

Lets assume C video cameras observing M moving objects with N 3d points on each object over time. The problem of 3D reconstruction can be posed as estimating the evolving 3D points $X_m^n(t)$, where m is the 3D point belonging to a moving object. The 2D point $x_n^c(f)$ of the moving 3D points on camera n can be given as:

$$\begin{bmatrix} x_n^p(f) \\ 1 \end{bmatrix} = K_c(f)[R_c(f) \;\; T_c(f)] \begin{bmatrix} X_n(f) \\ 1 \end{bmatrix}$$

where $K_c(f)$ is the intinsic camera matrix, $R_c(f)$ and $T_c(f)$ are the relative camera rotation and translation respectively of frame f. We denote this transformations $x_c^n(f) = P_n(f, X^n(t))$. The time correspoding to frame f is related to the continous global time t linearly: $f = \alpha_c t + \beta$, where α_c is the frame rate and time offset respectively. T

3.1. RANSAC Based Object Fit

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 any time instant 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:

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

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 joints of the object for accurate fitting of the rigid object in 3D. A joint in 3d can be defined as the difference of the 3D points between the centers of the joint. We define a joint as $J(p1, p2) = X^{p1}(t) - X^{p2}(t)$ where p1 and p2 correspond to different parts of the rigid object which are linked.

$$E_l(t) = \sum_{r=1}^{R} ||J(r) - E(J(r))||^2 \quad (2)$$

where $E(J(r))$ is the mean length of joint r. We model the symmetry constraint in the bundle adjustment for better reconstruction of rigid objects.

$$E_j(t) = \sum_{q=1}^{Q} ||J(q) - J(S(q))||^2 \quad (3)$$

where $S(q)$gives the corresponding symmetric joint of q i.e. if q is the joint connecting the left wheels of a car S(q)is the joint connecting the right wheels of the car. All the above constraints are enough to fit a object from multiple views we minimize the object specific error for finding the tracks of object accross views in stead of triangulation methods. The energy can be defined as:

$$E_o(t) = E_r(t) + E_l(t) + E_j(t) \quad (4)$$

The following energy is minimized for all the views in a ransac based formulation to find the correspondece across views.

Algorithm 1: Object specific Reconstruction

Input: $\{x_c^m(t)\}, \{K', R', T'\}, \beta'$
Output: $\{X_p^m(t)\}, \{K, R, T\}, \beta$

```
n ← Total object Hypothesis in C views (Sec. 4.1);
while Object Visibility < Min object views do
    repeat
        h = Sample(n);
        solve eq. 4 for h ;
        s = Reprojectfit(K', R', T', x_c^m(t), n) ;
        loop++;
        if s<a then
            Object Visibility = s;
        end
    until loop < MaxIteration;
end Sec 3.1
Track keypoints for s objects (Sec 4.2);
solve Eq.6 for {X_p^m(t)}, {K, R, T}, β;
Reproject and retrain the network(sec 3.4);
```

3.2. Multi-Video time synchronization

All the videos captured are taken from unsynchronized cameras. The first problem is to solve the synchronization problem by observing motion of objects from multiple views. we propose a multi camera video synchronization method for reconstructing from multiple videos. We assume that the rotation and translation of the object to be constant over time. We follow the work of [1], which synchronizes the time offset between two video frames assuming motion of a rigid object from non overlapping views. We extend the formulation further to synchronize multiple videos observing a moving rigid objects. The time offset δ with respect to global time can be computed using a addition constraint in the bundle adjustment. which is the as rigid as possible constraint which imrpoves the reconstruction as well as gives an accurate time offset. The error term is given as:

whereE_R is the rigidity constraint on the 3d points belonging to the object. It is given by

$$E_R(t) = \sum_{i=1}^{T^n-1} \sum_{p=1}^{P} ||(X'_{p1} - X'_{p2}) - R*(X_{p1} - X_{p2})||^2 \quad (5)$$

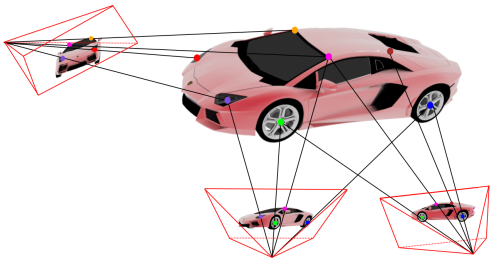

Figure 1. The picture depicts the projection of 3D keypoints of car onto thier respective videos

3.3. Constrained 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:

$$E = arg \min_{X(t),\{K,R,T\},\alpha,\delta} \sum_{t=1}^{T} E_o(t) + E_m(t) + E_R(t) \quad (6)$$

where w represents the weights of each term in the bundle adjustment. E_m is defined as the motion prior over the trajectory of the points.

$$E_m(t) = \sum_{p=1}^{P} \sum_{i=1}^{T^n-1} ||\frac{X^p(t^{i+1}) - X^p(t^i)}{t^{i+1} - t^i}||^2 (t^{i+1} - t^i) \quad (7)$$

and Although the reocnstructions from the constrained bundle adjustment works well. It is highly dependent on the keypoint detector. We propose to improve the reconstruction by using local features like sift around the keypoints. Since the motion of the object is continuous, we enforce the E_r, E_m, E_h on these points improve the localization of the keypoints.

3.4. Finetune Network

We believe that having a good 3D reconstruction from multiple views we can improve the detectors which gave the initial object keypoints. Since we use some geometric feature to improve the reconstruction the localization of the keypoints is improved. We reproject the reconstructed keypoints and finetune the detector for a better detector.

4. Implementation Details

4.1. 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.

4.2. 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.

References

[1] T. Gaspar, P. Oliveira, and P. Favaro. Synchronization of two independently moving cameras without feature correspondences. In *European Conference on Computer Vision*, pages 189–204. Springer, 2014.