International Institute of Information Technology, Hyderabad

Computer Vision

Undergraduate Course in

Computer Science

3D Reconstruction from Accidental Motion

Instructor:

Prof. Avinash Sharma

Teaching Assistant:

Gurkirat Singh Chauhan

Candidates:

Kartik Garg

Pravalika Mukkiri

Ashish Gupta

Spring'23

« Reality must take precedence over public relations,
for nature cannot be fooled. »

Richard P. Feynman

« Vision is a picture of the future,
that produces passion. »

Bill Hybels

Abstract

The paper and us tackle with the problem of Dense Reconstruction of the scene from accidental motion. By accidental motion we refer to the slight movements which happen due to metabolism processes in our body. So we aim to record a small video before and after taking a still. Then our primary goal is to construct a dense depth map using bundle adjustment. We then use CRF Framework to regularize depth to get smooth depth maps. Although the motions seem to be small leading to a small baseline which further leads to unreliable depth estimates. But in theory we can make use of many such measurements to make reliable depth estimates. The results lead to the possibility that depth estimates thus depth maps of sufficient quality can come "for free"

Contents

Introduction

3D-Reconstruction from Accidental Motion, the title itself can be broken into two tasks primarily that is "3D Reconstruction" and "Accidental Motion"
3D Reconstruction is a field in which a lot of research has already gone and we know it is related to Bundle Adjustment. In our use case we use CERES SOLVER provided by google to solve our problem of bundle adjustment. Accidental Motion on the other hand is a new Buzzword here. What we mean by accidental motion is actually inevitable motion. While clicking a picture, if we don't use the auto stabilization feature then there is always some motion which cannot be evaded. It usually takes a couple of seconds to take a picture. And during this short span of time we are bound to have the following

- Hand Shaking maybe due to surrounding air or respiration.
- Metabolism processes such as heart beats etc.

So we are bound to have such small motion. So we ponder upon the question **"If we were to capture a short video before and/or after the capture of a still, would it be possible to use the baseline from accidental motion to reconstruct the scene"** Here the baseline arises from the translation.

Challenges Arising

There are primarily two challenges arising which are listed below

- Common approaches of 3D Reconstruction end up giving good results since they assume an adequate baseline for the algebraic methods they adopt. And algebraic methods tend to be reliable when the conditions meet, but in our case when we make a claim of having small baseline then the algebraic methods tend to become vulnerable.
- Again the depth estimates have a lot of uncertainty owing to a small baseline.
- Therefore previous stereo methods produce artifacts.

1. Solution

1.1 Vague Solution

Paper suggests we can use multiple images together to do SFM directly. Due to accidental motion, they suggest to parameterize the 3D points using inverse depth relative to a reference view. Motivation behind such parameterization if that **It helps regularize the bundle adjustment.**
Moreover, in the paper the authors have suggested the following

- It is good to initialize with random depth and identical camera poses. It has been verified experimentally by the authors.
- Having multiple images, in our case either a high frame rate or a little longer duration, can help reduce uncertainty.

1.2 Important Details

- Since the depth map is weak and noisy owing to small baseline in our method, the popular first order CRF is not very effective and results in over smoothed depth maps that is **low order connections cannot regularize depth effectively**
- So, the paper proposes to use long range connections and it is found that direct connections between a pixel and its bigger neighbourhood can improve the dense reconstruction in the current case.

More Intrinsic details can be found in the Discussions Section.

1.3 Observations From Analysis

- When Camera poses are fixed, it is convex to get the depth of a **feature** relative to reference view.
  - What Features?, Harris/Shi-Tomasi ones

- It is convex to optimize the rotation for the points at infinity when an approximation is used.

2. Discussion

2.1 Definitions

Formulation of the problem is as follows

- Let us assume we have N_c images and N_p points in 3D. Here N_c is the number of images which we will extract from the video sequence and N_p is the number of shi tomasi features which can be tracked in all the images. Here, we have already removed the outliers using Homography
- We will assume the first image to be our reference image, and for all other images we assume the camera has been rotated with reference to the **reference image aka first image**
- Let us use P_j to denote the j^{th} point in the frame of reference camera, Now this point can be writtes as $R_i P_j + T_i$ in the frame of the i^{th} image's camera where R_i and T_i is the rotation and translation of the i^{th} camera with respect to the reference frame.
- Since we are assuming samll motion, so it can be assumed that the roation angles are small. Thus we can approximate the 3D Rotation matrix as follows (We use the approximations that sin(θ) -> 0 and cos(θ) -> 1 for small θ)

$$\begin{bmatrix} 1 & -\theta_i^z & \theta_i^y \\ \theta_i^z & 1 & -\theta_i^x \\ -\theta_i^y & \theta_i^x & 1 \end{bmatrix}$$

- The **inverse depth parameterization** is done as following, $P_j = \frac{1}{w_j}[x_j, y_j, 1]^T$. Here x_j and y_j is the projection of j^{th} point in the reference frame.
- However, the projection of j^{th} point on the i^{th} image is denoted by p_{ij} and is given by $[p_{ij}^x, p_{ij}^y]^T$

- We denote the projection function as π and is given as $\pi([x, y, z]^T) = [\frac{x}{z}, \frac{y}{z}]^T$

2.2 Cost Function Analysis

We make use of L_2 norm to calculate the re projection error and our ultimate goal is to minimize this re projection error. The cost function can be defined as follows

$$F = \sum_{i=1}^{N_c} \sum_{j=1}^{N_p} ||p_{ij} - \pi(R_i P_j + T_i)||^2$$

After following the calculation shown **here**, the cost function can be simplified as follows

$$F = \sum_{i=1}^{N_c} \sum_{j=1}^{N_p} (\frac{e_{ij}^x + f_{ij}^x w_j}{c_{ij} + d_{ij} w_j})^2 + (\frac{e_{ij}^y + f_{ij}^y w_j}{c_{ij} + d_{ij} w_j})^2$$

where ,

$$\begin{aligned}
a_{ij}^x &= x_j - \theta_i^x y_j + \theta_i^y, \\
b_{ij}^x &= T_i^x, \\
a_{ij}^y &= y_i - \theta_i^x + \theta_i^z x_j, \\
b_{ij}^y &= T_i^y, \\
c_{ij} &= -\theta_i^y x_j + \theta_i^x y_j + 1, \\
d_{ij} &= T_i^z, \\
e_{ij}^x &= p_{ij}^x c_{ij} - a_{ij}^x, \\
f_{ij}^x &= p_{ij}^x d_{ij} - b_{ij}^x, \\
e_{ij}^y &= p_{ij}^y c_{ij} - a_{ij}^y, \\
f_{ij}^y &= p_{ij}^y d_{ij} - b_{ij}^y,
\end{aligned}$$

2.3 Revisiting 1.2

If we assume that the correct camera poses are given that is we know the rotation and translation matrix (We will get these details after bundle adjustment). The depth estimation problem reduces to find only w_j in the above formula for the cost function and everything else is known to us now.

Let us consider the general form of each of the individual terms in the function, it comes out to be $(\frac{x-a}{x-b})^2$ where a and b are the zero/root and pole(Asymptotic point) of the function. When $a > b$, the function is convex in $(b, \frac{3a}{2} - \frac{b}{2})$. When $a < b$, the interval is reversed and is given by $(\frac{3a}{2} - \frac{b}{2}, b)$. These details are given by basic mathematics and calculus. On further analysis, we find that the function becomes convex as far as the reasonable values of w_j are considered.

2.4 Dense Reconstruction

Since we aim to get a depth map of a reference view as a 3D reconstruction output. Because the depth signal at each pixel tends to be noisy in this case (Due to small baseline). So we adopt CRF with plane sweeping algorithm to get smoothed depth maps. To preserve the details while smoothing the depth map, it has been proposed to use long range connection between pixels in the CRF energy function, which can pass information to a pixel effectively.

2.4.1 Formulation

The input is a set of images and pixels. In addition we have maps which maintain the actual pixel values along with the positions at which they are in the images. Let P denote the photo-consistency function such that P(i,d) is the photo consistency score of the i^{th} pixel at distance d.

The energy we want to minimize is given by

$$E(D) = E_p(D) + \alpha E_s(D)$$

where $E_p(D)$ is the standard photoconsistency term of the form and is obtained using the **plane sweeping algorithm**

$$E_p(D) = \sum_i P(i, D(i))$$

E_s is the smoothness term to regularize. In general it is first or second order CRF (Conditional Random Field) Model which allows the passage of information between adjacently connected pixels. But for our case, **we need more than just adjacently connected pixels**
So the paper proposes to connect pixels with longer range so that the photo consistency can be effectively aggregated from an area to a pixel in it. To build such connections we define a term

$$C(i, j, I, L, D)$$

which gives a score for the depth assignment of the i^{th} and j^{th} pixels based on color intensities (given by I) and their locations (given by L) and D is the depth assignment in context. The score is given as

$$C(i, j, I, L, D) = \rho_c(D(i), D(j)) \times exp(-\frac{||I(i) - I(j)||^2}{\theta_c} - \frac{||L(i) - L(j)||^2}{\theta_p})$$

and the smoothness term E_s is now a summation over all C's. ρ_c is just the difference between the depth of the two pixels in context.
The purpose of all this is to connect pixels in a region with similar colors such that they have consistent depth beacuse they should be more likely to belong to the same object.

3. Algorithm

- Select a reference view and initialize all camera poses with zero rotation and translation.
- Parameterize the points by inverse depth
- Get points of interest
  - We use shi tomasi corner detection to find points of interest
  - One corners are found in the reference image we use KLT Tracking to track them over the whole set of iamaes
  - But once we use KLT Feature tracking there is still a chance of getting false matches resulting in outliers. So we make use of Homography to handle that.
- Once we have the above things ready, we move onto the bundle adjustment
  - Export bundle file
  - Create initial point cloud
  - Use CERES solver with appropriate params
  - CERES will return is final point cloud and the final bundle file
- Once we have the 3D structure from bundle adjustment, we now need to smooth the depth map as explained earlier
- We adopt plane sweep algorithm with the suggested changes in the cost function so as to get good smoothed maps.

Conclusions

RESULTS ARE PRESENTED/DISCUSSED IN SLIDES

The paper claims that they have proposed the first practical system to reconstruct 3D structure from small motion image sequences. We discover that in the case of small motion, random point depth relative to a reference view and identical camera poses are good initialization for the bundle adjustment problem.
Further based on the noisy depth maps, the paper has suggested to use long range connections and improvise upon the regularization terms which resuls in better depth maps.