Discussing: 3D Reconstruction from Accidental Motion

Fisher Yu[1] David Gallup[2]

[1]Prinston University

[2]Google Inc.

Conference on Computer Vision and Pattern Recognition, 2014

Idea

What Perform 3D reconstruction from the accidental motion by photographer trying to take a still image. Create a depth map corresponding to the taken image.

Why Perspective change, simulated aperture, object segmentation etc.

How

- Structure from motion
- Dense reconstruction based on that structure

Background on SfM

The goal: To calculate the structure of a scene (and the cameras used to record it) from a sequence of images.
Typical approach:

1. Detect features in each image
2. Match features across images
3. Estimate relative camera and feature positions from shared features

Problems:

- Small baseline in accidental motion (angle between images usually < 0.2 degrees)
- Small baseline means depth uncertainty is large

Feature detection

- A feature window is a small area surrounding a feature point.
- Goal is to find a location for which the minimum change caused by shifting the window is large.
- Let $I(x, y)$ be the energy of image I at location (x, y), and W be the feature window. We want to maximize

$$E(u, v) = \sum_{(x,y) \in W} [I(x + u, y + v) - I(x, y)]^2$$

  for some small displacement (u, v).
- Since motion is small, first order taylor approximation is good enough, and we get the following, where $I_x = \frac{\delta I}{\delta x}$. Let the term in the parens be H.

$$(u, v) = \sum_{(x,y) \in W} \left[\begin{bmatrix} I_x & I_y \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \right]^2 = \begin{bmatrix} u & v \end{bmatrix} \left(\sum_{(x,y) \in W} \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \right) \begin{bmatrix} u \\ v \end{bmatrix}$$

Feature detection: Algorithm

- We can identify feature type by analyzing the eigenvalues of H. In particular, corners are features where the eigenvalues λ_+, λ_- are both large.

The algorithm thus becomes:

1. Compute the gradiant at each point of the image.
2. Compute the H matrix from the gradients.
3. Compute the eigenvalues of H.
4. Find points where eigenvalues are large; $\lambda_- > threshold$ (λ_- is only large for corners, λ_+ is also large for edges).
5. Choose those points where λ_- is a local maximum as feature points.

Problem This is invariant in rotation and mostly in magnitude, but not in scale.

Solution Find the scale for which the Harris operator, $f_{Harris} = det(H) - k trace(H)^2$, is locally maximized. Feature points are local maxima in both position and scale.

Feature matching

- Extract the features from all images independentely.
- Match features extracted from different images by comparing descriptors.
- Multiple approaches:
  - Naive: Match pixels within feature windows within some threshold.
  - MOPS: Detect features at many scales using a Gaussian pyramid (this is the solution from last slide). Extract window, scale down to $\approx 20\%$, rotate to horizontal and normalize the window by subtracting the mean, dividing by its stdev. Store/Compare these results.
  - SIFT: Find the *"best scale"* to represent each feature. Divide window (at that scale) into cells, and for each cell, compute edge orientations (gradiant-90°), throws out weak edges, create histogram from remaining edges.

Feature tracking

The general idea:

- Extract the features from first image.
- Attempt to find the same features in the next image.

Advantage Takes advantage of small changes between subsequent frames, not affected by potentially large changes over full sequence.

Disatvantage Large changes between subsequent frame are not handled well.

Translational Motion Assume small displacement between subsequent frames (≈ 1 px).

Constant Flow Assume pixels in the feature window have the same displacement.

Feature tracking

Kanade-Lucas-Tomasi (KLT)

The KLT algorithm:

For every pair of consecutive images I, J and window W:

- The displacement of a feature, $\mathbf{d}$ is given by $G\mathbf{d} = \mathbf{e}$, where $G = \int_W \mathbf{g}\mathbf{g}^T w\delta A$ and $\mathbf{e} = \int_W (I - J)\mathbf{g} w\delta A$.
- G,a 2×2 symmetric coefficient matrix, can be calculated from the second-order moments of gradient estimates of one frame.
- $\mathbf{e}$, a 2D vector, can be calculated from the difference between the two frames along with the gradients from above.
- After a few iterations of the above, the displacement estimate will stabilize. The original KLT article states *"Each feature required typically fewer than five iterations . . . to stabilize . . . within* $\frac{1}{100}$*th of a pixel"*.

KLT is used in this paper due to the nature of the image sequence:

- Small motions, between frames as well as overall
- They track from reference image to all others, *not* between sequential pairs due to this, which reduces the accumulative localization error of feature tracking.

Bundle Adjustment

- Based on an error metric, iteratively refine feature coordinates and the relative motion and optical characteristics of camera.
- Error is usually Euclidian distance from estimated projection of the features onto a reference image to their actual position.
- Boils down to a minimization of this *reprojection error*, and we can use a nonlinear least-squares algorithm (usually Levenberg-Marquardt).
- Finds local minimum; needs good initialization to find global minimum. Usually initialized with 2-view reconstruction.

  Problem Small baseline makes 2-view reconstruction for initialization undesirable.

  Break Camera characteristics for their image sequence is likely to be known and constant.

Bundle Ajustment

The mathmatical definition:

$$\min_{\mathbf{a}_j, \mathbf{b}_i} \sum_{i=1}^{n} \sum_{j=1}^{m} v_{ij} d(\mathbf{Q}(\mathbf{a}_j, \mathbf{b}_i), \mathbf{x}_{ij})^2$$

where $Q(\mathbf{a}_j, \mathbf{b}_i)$ is the predicted projection of point i on image j, and $d(\mathbf{x}, \mathbf{y})$ denotes the Euclidian distance between the image points epresented by vectors $\mathbf{x}$ and $\mathbf{y}$.
It is by definition tolerant to missing image projections.

Solution: SfM

But first, some definitions

- Assume they have N_c images and N_p features.
- Let the first image be the reference image.
- The i-th camera is related to the reference camera by a rotation matrix $\mathbf{R}_i$ and a relative translation $\mathbf{T}_i = \begin{bmatrix} T_i^x & T_i^y & T_i^z \end{bmatrix}^T$.
- Assume $\mathbf{P}_j$ is the position of the j-th feature in the coordinate system of the *reference* camera.
- The position of the j-th feature in the i-th camera's coordinate space is then $\mathbf{R}_i\mathbf{P}_j + \mathbf{T}_i$.
- Let $\mathbf{\Theta} = [\theta_i^x, \theta_i^y, \theta_i^z]$ be the rotation angles of the i-th camera. Since the angles are small, we can approximate $\mathbf{R}_i$ to

$$\mathbf{R}_i = \begin{bmatrix} 1 & -\theta_i^z & \theta_i^y \\ \theta_i^z & 1 & -\theta_i^x \\ -\theta_i^y & \theta_i^x & 1 \end{bmatrix}$$

Solution: SfM

More definitions

- To make the optimization problem easier, parameterize features by their inverse depth, $\mathbf{P}_j = \frac{1}{w_j} \begin{bmatrix} x_j & y_j & 1 \end{bmatrix}^T$ where (x_j, y_j) is the projection of $\mathbf{P}_j$ in the reference image.
- The projection of $\mathbf{P}_j$ onto the i-th image is $\mathbf{p}_{ij} = \begin{bmatrix} p_{ij}^x & p_{ij}^y \end{bmatrix}^T$
- They define their projection function $\pi : \mathbb{R}^3 \rightarrow \mathbb{R}^2$, that is $\pi(\begin{bmatrix} x & y & z \end{bmatrix}^T) = \begin{bmatrix} \frac{x}{z} & \frac{y}{z} \end{bmatrix}^T$

Solution: SfM

Analysis

- They use the L_2 norm, defined as i $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$ over an $m \times n$ matrix $\mathbf{A}$, to measure the reprojection error, as it has nice a statistical interpretation and can be robustified.
- Based on the above definitions, they present this cost function of bundle adjustment in the retina plane:

$$F = \sum_{i=1}^{N_c}\sum_{j=1}^{N_p}\|p_{ij} - \pi(\mathbf{R}_i\mathbf{P}_j + \mathbf{T}_i)\|^2$$

- Now, assume the camera poses are fixed and given. Finding the depth estimation of a feature point is minimizing:

$$F_i(w_j) = \sum_{i=1}^{N_c}(f_j^x(w_j) + f_j^y(w_j))$$

Solution: SfM

Analysis: Depth estimation

- Expanding $F_i(w_j)$, we get:

$$F_i(w_j) = \sum_{i=1}^{N_c}((\frac{p_{ij}^x c_{ij} - a_{ij}^x + w_j(p_{ij}^x d_{ij} - b_{ij}^x)}{-\theta_i^y x_j + \theta_i^x y_j + 1 + w_j T_i^z})^2$$

$$+(\frac{p_{ij}^y c_{ij} - a_{ij}^y + w_j(p_{ij}^y d_{ij} - b_{ij}^y)}{-\theta_i^y x_j + \theta_i^x y_j + 1 + w_j T_i^z})^2)$$

- The paper showes that $F(w_j)$ is convex in $\langle 0, min_i|\frac{-\theta_i^y x_j + \theta_i^x y_j + 1}{2T_i^z}|\rangle$, the upper bound of which is supposed to be far greater than reasonable values for w_j, so one can easily optimze w_j for the reprojection error in $F_i(w_j)$.
- It is also noted that noise in $\mathbf{p}_{ij}$ does not alter this convex interval. It also does not depend on the rotation matrix; it is an exact property of depth estimation with small motion.

Solution: SfM

Analysis: Depth estimation & uncertainty

Points at infinity If feature points are approximately at infinity, the cost function can be approximated to

$$F \approx \sum_{i=1}^{N_c} \sum_{j=1}^{N_p} [(p_{ij}^x c_{ij} - a_{ij}^x)^2 + (p_{ij}^y c_{ij} - a_{ij}^y)^2]$$

and is a convex function of the camera rotation angles on the domain around zero.

Depth uncertainty The relationship between inverse depth w, baseline b, focal length f and disparity d for two stereo images is $w = \frac{d}{fb}$. Ignoring qunatization errors and mismatches, the inverse depth estimation at a given pixel is

$$Var[\hat{w}] = \mathbf{E}[(\frac{d+\epsilon}{fb})^2] = \frac{Var[\epsilon]}{f^2 b^2}$$

where ϵ is the feature localization error.

Solutino: SfM

Analysis: Depth uncertainty continued

The paper goes on to show that if the feature detection errors are independant, the stdev of the inverse depth estimation, $Var[\hat{w}]$, deacreses linearly with $\sqrt{n}$, where n is the number of observations of the feature point.
Note that if the errors are fully correlated, multiple observations do not help reduce uncertainty.
Similar conclusions can be draw for depth.

Solution: SfM

Initialization & features

- It follows from this analysis that random structure might be a good initialization.
- Given a sequence of images, they select a reference view, and initialize all cameras to zero relative rotation or translation.
- They detect corner features in the reference image and track these from the reference image to the other images with the Kanade-Lucas-Tomasi feature tracker.
  - They require that all features can be tracked to all images, and they set a threshold for the maximum color gradiant difference per pixel between two patches.
  - KLT provides subpixel accuracy which is important due to the small baseline.

Solution: SfM

Optimization

- They use Ceres Solver an open source c++ library for solving large, compilacted nonlinear least squares problems developed at google.
- The reference view camera is fixed at the coordinate space origin.
- Usually, outliers can be neglected after the feature tracking and selection initialization, but in some cases robustifiers applied to the cost function can improve the reconstructin results.
- After each optimization, they remove negative depth features and reoptimize with the remaining ones.

Solution: Dense Reconstruction

Structure estimated, what now?

- Now that they have a scene structure, they want to estimate the depth of the images to reconstruct the 3D scene.
- Because they have captures from a similar viewpoint, they can only reconstruct a 3D scene from that viewpoint.
- Thus, they aim to construct a depth map for the reference image as the 3D reconstruction output.

  Problem At the pixel level, the depth signal tends to be noisy.

  Solution They adopt plane sweeping and conditional random fields (CRF) to solve a smooth depth map.

  Because The confidence of depth minima is low in general (not just for textureless areas) due to the small baseline, so details can be easily smoothed. To retain detail, they propose using long-range pixel connections in the CRF energy function.

Solution: Dense Reconstruction

Formulation

- Let $\mathcal{I}$ be the index set of the pixels in the reference image I, and $I(i), i \in \mathcal{I}$ be the color of the i-th pixel.
- The goal is to determing D, a dense depth map of I.
- Let $\mathcal{L}$ map the pixel indices in $\mathcal{I}$ to 2D locations in I.
- Let P be the photo-consistency function such that $P(i, d)$ is the consistency score of the i-th pixel at distance d.
- They then intend to mimize the energy $E(D) = E_p(D) + \alpha E_s(D)$, where E_p is the standard photo-consistency term $E_p(D) = \sum_{i \in \mathcal{I}} P(i, D(i))$, and α is the data term.
- E_s is a smoothness term used to regularize the depth estimation. It often represents first- or second-order CRF.

  Problem The adjecent connected model doesn't effectively regularize the noisy depth data.

  Solution Long range pixel connectivity.

Solution: Dense Reconstruction

Formulation, smoothness

They introduce $C(i, j, I, \mathcal{L}, D)$, which gives a score for depth assignment of the i-th and j-th pixels based on the color intensities and their location in the reference image. This gives us

$$E_s(D) = \sum_{i \in \mathcal{I}, j \in \mathcal{I}, i \neq j} C(i, j, I, \mathcal{L}, D)$$

where

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

in which p_c is a robust measurment of depth difference, and θ_c and θ_p are parameters to control the connection strength and range.
They choose p_c to be the truncated linear function,
$p_c = min(t, |D(i) - D(j)|)$, where t is a threshold.
This aims to connect pixes in an area of similar colors since they are likely to belong to the same object.

Solution: Dense Reconstruction

Energy function, justified

Data term If we only optimize the data term of the energy function, a very noisy result is found. Observe the WTA-result in fig 3d of the paper.

1st Order Smoothness See fig 2d-f, where they show the result of several levels of First Order Smoothness. Some areas are still noisy, while others are oversmoothed, leading to surfaces being reconstructed in layers. This motivates long range connectivity.

Long Range Connectivity Connecting pixels at a longer range instead of adjacent ones. They first smooth the reference image with mean shift before using its color to compute C. As resolution increases, θ_p should be chosen to grow, and due to efficient mean field inference the running time doesn't change with the values of θ_c and θ_p.

Experiments

User behavior

They conducted a study in user behavior to determine the magnitude of accidental translation during still photography.

Method Users theyre asked to record a calibration pattern (0.5m away) as if they theyre photgraphing it, and theyre instructed to hold the camera still for 5s.
9 participants and two cameras theyre evaluated: a camera phone, and a point-and-shoot.

Results

All users	Speed (mm/s)	Stdev. (mm) after 1s	2s	3s
Phone: Mean	18.07	2.18	3.35	3.81
Phone: Stdev.	6.67	1.11	1.99	2.31
P&S: Mean	9.23	1.71	3.02	3.99
P&S: Stdev.	2.10	0.65	1.23	1.88

Conclusion At 3s, a stdev. of 3.9mm translation of camera center gives a sufficient baseline for a good reconstruction under reasonable conditions.

Experiments

SfM

Method They use the method described earlier. They remove feature tracking outliers by average pixel difference in a patch (≈ 6 for an 8-bit encoded gray image). They initialize the features' depth to a uniformly random value between 2m and 4m.

Results See figure 3 in the paper.
Due to the avoidance of two-view reconstruction for each image pair, SfM is fast. 1000 points and a 100 cameras *"usually takes several seconds on a modern desktop"*.

Conclusion It is observed that feature tracking outliers are inevitable, but usually don't affect the result, hotheyver a robustifier can help in the cases they do (as mentioned). If one is not needed, the result is usually *better* without.

Experiments

SfM, multiple images

Image count Figure 6 in the paper shows baseline between a camera and the reference camera in the model of all reconstructed image (gree), and the structure error as a sum of square differences between the reconstruction at a given number of images vs. all images.

Observation While a spike can be observed at around 65 images, this is likely to be due to outliers in the feature tracking. The curve with robustness in the tracking (right) has no such spike. They observe that the depth estimation uncertainty decreases with an increase in image count (fig 5e), and from fig 6 that more images help reduce the effect of outliers.

Experiments

Dense Reconstruction

In the experiments run, the following values for the Long Range Connectivity function theyre used:

- Image size: 480×270 px
- θ_c (connection strenght): 20-30 px
- θ_p (connection range): 5-9 px
- t (connectiviy threshold): 15% of the total label number.

Conclusion

- For small motion image sequences, random feature depth relative to the reference image and identical camera poses are good initialization for the bundle adjustment cost function.
- While the 3D features at the background have very high uncertainty, the foreground features clearly show the 3D structure.
- Based on the noisy nature of the photo consistency measurement, long range connectivity between pixels to regularize the depth map is proposed.
- The reulting depth map is shown to be of high enoughn quality to make "*perceptually plausible refocused images*", as seen in fig 3f of the paper.

Questions?