Comparing the Robustness of different depth map algorithms- findings of the paper

Wednesday, September 4, 2024 4:54 PM

Light field imaging processing is a new and upcoming technique to use for finding the depth maps and for reconstruction.

Three depth maps are evaluated here. They are:
- Least squares gradient (LSG)
- Plane sweeping
- Epipolar plane strategies

The metrics used to compare are:
- Depth estimation accuracy
- Computation time

Simulated and real scenes were used.

What is Depth mapping?
- Depth maps contain information about the distance of objects from a specific perspective or reference point. Each pixel is assigned a value which represents the distance of that pixel from the reference point. It is a 3D matrix representation of the scene.
- And the process of creating this depth map is called depth mapping.
- The current approach to create a depth map is to use specialized time of flight or depth cameras.
- These cameras measures the depth by emitting an infrared beam of light and measuring the time it returns.

The problem in this approach of creating a depth map is that it does work for coarse representations but it is troublesome for the features which are below the resolution of the depth camera being used. It also has range limitations because of the scattering effect of the infrared light. So creating depth maps in outdoor or vast areas is difficult and or very accurate.
This method is also computationally expensive as it requires multiple views of the same scene.

The dataset used in this paper is the 4D Light Field Benchmark Dataset and Stanford Light Field Archive.

In the 4d Light Field Benchmark dataset each light field has 9x9 views which has a resolution of 512x512 and in the Stanford Light Field archive each light field has 17x17 views which has a resolution of 960x1280.

- **LEAST SQUARES GRADIENT METHOD**
  - Here light field imaging is done for a single scene in various viewpoints.
  - Light field imaging captures the intensity of light at each point and its direction as well. It is represented as a 4D function as L(x,y,u,v), where,
  (x,y) are the spatial coordinates and (u,v) are the angular coordinates.
  - When the viewpoint is changed, objects at different depths will appear to shift. This shift is called disparity.
  - And LSG aims to minimize this disparity.
  - The displacements of the shift are named as d_x and d_y signifying the displacements in the x and y direction respectively.
  - So, the light field function at a point (x,y) from a viewpoint (u,v) can be written as:

$$L(x, y, u, v) = L(x - d\Delta_x, y - d\Delta_y, u + \Delta_x, v + \Delta_y)$$

- This method now looks for the displacement d such that it makes the difference between the images from different viewpoints as possible.
- The error E is defined as:

$$E = \int_\alpha \sum_p L(x, y, u, v) - L(x - d\Delta_x, y - d\Delta_y, u + \Delta_x, v + \Delta_y)$$

$$d^* = \frac{\sum_p (L_x L_u + L_y L_v)}{\sum_p (L_x^2 + L_y^2)}$$

where, L_x and L_y are the spatial derivatives of the light field function in the x and y directions. They represent how quickly light intensity changes as we move across the image.
And, L_u and L_v are the angular derivatives with respect to the viewpoint position. They represent how the light intensity changes when the viewpoint changes.

A larger value of d represents that the object is closer to the camera and a smaller value of d represents that the object is far from the camera.

- **PLANE SWEEPING METHOD**
  - The method used in this paper is slightly modified from the original method where only 2 views of the light field were considered but here the entire light field is considered.
  - Initially, light field of different viewpoints are refocused to a common center view. The spatial (x,y) and angular (u,v) coordinates and a hypothesized disparity d are adjusted. This process is called 4D shearing which is carried out using the following formula:

$$L_d(x, y, u, v) = L(x + ud, y + vd, u, v)$$

Where the LHS represents the refocused light field view and L(x,y,u,v) represents the original light field data.
This operation aligns the light rays as if the scene were in focus at a particular depth(disparity).
- After refocusing, the new refocused view is evaluated as to how well it aligns with the actual observed data. This is known as Cost Volume.
- Cost Volume is calculated by stacking the refocused images for all possible disparities and computing a cost for each pixel.
- The cost C(x,y,d) at each pixel for a particular disparity is calculated as:

$$\bar{L_d}(x, y, u, v) = \frac{1}{|U||V|} \sum_{u \exists U} \sum_{v \exists V} L_d(x, y, u, v)$$

$$C(x, y, d) = \frac{1}{|U||V|} \sum_{u \in U} \sum_{v \in V} (L_d(x, y, u, v) - \bar{L_d}(x, y, u, v))^2$$

- This cost represents the variance across all viewpoints for the hypothetical depth.
- Low cost indicates that the image is in focus for that disparity.
- To reduce noise and to produce a cleaner depth map, the cost volume is filtered using a box filter of size 3x3.
- This smoothing operation helps in creating a more consistent depth map by averaging out small variations in cost.

- After calculating the cost volume and filtering it, the optimal value for depth is found and this is used as the depth for that pixel.

- **EPIPOLAR-PLANE AND FINE-TO-COARSE REFINEMENT METHOD**
  - This method is proven to be robust even against occlusion.
  - It uses parallel processing.
  - It combines edge detection, radiance sampling and iterative refinement to produce a high-quality depth map.
  - Epipolar-plane image(EPI) is a 2D slice of 4D light field data. When one of the spatial and angular dimensions are fixed it results in epipolar image.
  - EPIs help in handling occlusions because, occlusions manifest as discontinuities in the EPI, they can be detected and managed effectively.
  - Fine-to-coarse refinement is the process used to iteratively refine a solution starting from a high-resolution level and working down to a lower-resolution(coarse) level or vice-versa.
  - In this method, the first step is to compute the edge confidence for each pixel in the central view image I(x,y). This will help us identify the strong edges in the central view image.

$$C_e(x,y) = \sum_{(x',y') \in N(x,y)} \| I(x,y) - I(x',y') \|$$

  - It uses a sliding window of size 3x7 to calculate the confidence based on the intensity difference.
  - The threshold is set to 0.005 at the initial level (level 0) and 0.01 for the subsequent levels in the fine-to-coarse procedure.
  - Then for each pixel in the central view, they sample the light field data at different disparities d across different views.

$$R(x,y,u,v,d) = L(x+(\hat{u}-u)d, y+(\hat{v}-v)d, s, t) | s = 1..n,\ t = 1..m$$

  - Then using this radiance we can compute a score of color density S as

$$S(x,y,d) = \frac{1}{R(x,y,u,v,d)} \sum_{r \in R(x,y,u,v,d)} K(r-\bar{r})$$

  - Then they evaluated the consistency of radiance across views at each disparity d. Then, they applied a kernel function to smooth the scores.
  - The radiance value is updated using the mean shift algorithm which improved the robustness of the score of color density.

$$\bar{r} \leftarrow \frac{K(r-\bar{r})r}{K(r-\bar{r})}$$

  - For whichever value the color density score S(x,y,d) was maximum is selected as the optimal disparity. Then , depth confidence $C_d(x,y)$ is calculated based on edge confidence and color density.

$$C_d(x,y) = C_e(x,y) \| S_{max} - \bar{S} \|$$

  - Then using these values an initial depth map D(x,y) is created keeping only the disparities with high depth confidence. Then a median filter of window size 3x3 is applied to this map to reduce noise.
  - This disparity map is saved for further processing steps in fine-to-coarse. But this map has many gaps or areas with uncertain disparity values.
  - Since there are many uncertain disparity values, the fine-to-coarse refinement method updates

the range of possible disparity values also known as disparity bounds. This helps in narrowing the search space for disparity during the refinement process.

- The method is as follows:
  - First step is to apply Gaussian filter with a kernel size of 7x7 and a standard deviation of 0.5 is applied to the central view image.
  - After performing Gaussian blurring, they down-sample the disparity map with a factor of 0.5 and then compute Ce again.
  - These steps will be repeated till the dimension of I is less than 10 pixels.
  - Then the disparity map is up-sampled from the coarsest level to fill-up every pixel without changing the d that was obtained from finer levels and then combined to form the final disparity map D.

The depth Z is calculated through the Matlab tool Heidelberg Dataset provided which is based on the equation:

$$Z = \frac{fb}{d}$$

Where f = focal length, b = baseline, which is the distance between adjacent views in the light field, d = disparity.

They also want to show the effectiveness of the refinement process in the accuracy of the depth map.

- **Analysis:**
  - **LSG Method:** This method was accurate in reconstructions of the depth map and it took less time. But it did not work well on untextured background as it did on foreground.

  - **Plane Sweeping Method:** This method was tested with different numbers of depth planes. The results showed that as the number of depth planes increased, the computation time also increased. This result was already expected as more depths require more computation to find the optimal disparity.
  The MSE (Mean-squared error) decreased as the number of depths increased but the decrease was minimal, order of 10^-4. This indicates that after a certain point increasing depth planes offers diminishing returns in accuracy.
  It also showed that this method could not handle occlusions well.
  And it was taking longer time to run, almost 3 times longer than the Epipolar Plane method and 18 times longer than the LSG method.

  - **Epipolar-Plane and Fine-to-Coarse Refinement Method:** This peak signal-to-noise ratio of the depth maps generated by this method was higher than LSG method but is lower than the Plane sweeping method.
  The runtime was about ten times higher than LSG but nearly half of the plane sweeping method.
  It possess a good balance between efficiency and image quality and requires further optimization to handle noise and low-texture areas effectively.