CV PROJECT REPORT

-DEPTH ESTIMATION IN VIDEO

MEMBERS and Contribution:

Kolla Varuna krishna, S20190010093: Training code(Hourglass model implementation), Transfer learning and testing on Davis dataset.

K.venkatesh, S20190010090: getting the pre-trained models and loading them. And model Options. creating training and testing data of dataset TUM

G.vikram, S20190010055: Testing the video on tum dataset, creating them into training and testing data on dataset DAVIS

D.karthik, S20190010042: Evaluating the Result, Data loaders, getting datasets, making the frames into videos.

<u>ABSTRACT:</u>

Depth Estimation in Video is very important for many use-cases. In this report, we present the survey and methodology we implemented. We may not even be knowing that our regular day-to-day appliances use depth. As much as it is important it still has challenges.

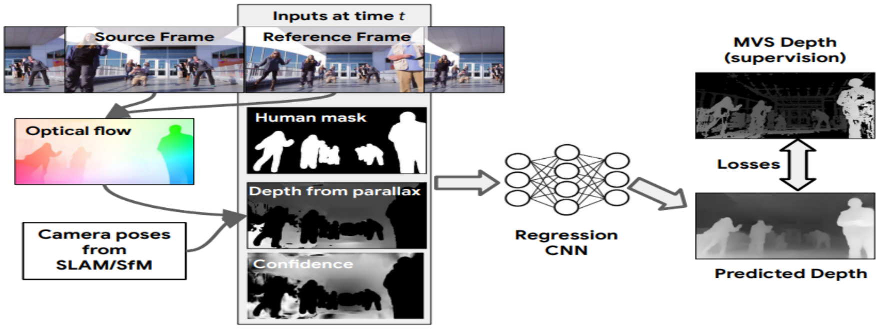
<u>Literature Survey:</u>

[1]. COLMAP:

- Uses SFM, and MVS to find the depth.
- Structure From Motion(SFM) is generally used to structure images. This means it estimates photo location, orientation, and Camera parameters.
- Multi-view stereo(MVS) takes these camera parameters, location, orientation, etc information from SFM and makes a 3D dense point cloud, which can get depth information.
- As we are looking into videos it will consider only neighboring frames for construction.
- Sfm – for 3d reconstruction, camera params - if used for depth estimation can find only for key points(points in 3d reconstruction)
- Mvs – takes op from sfm and uses them to find the depth at every pixel

[2]. Learning the Depths of Moving People by Watching Frozen People:

- Trained on Mannequin challenge dataset.
- And fine-tuned on DAVIS and TUM DATASET
- Can predict depth even when object and camera is moving.
- Uses a data-driven approach and learns human depth priors from a new source of data.
- At inference time, uses motion parallax cues from the static areas of the scene.
- The below picture shows the architecture the paper proposed. Uses optical flow, camera parameter predictions, and a human mask.
- The evaluation metric used is Si-RMSE, RMSE.



[3].CONSISTENT VIDEO DEPTH ESTIMATION

- They used a conventional structure-from-motion reconstruction to establish geometric constraints on pixels in the video.
- Unlike the ad-hoc priors in classical reconstruction, they used a learning-based prior, i.e., a convolutional neural network trained for single-image depth estimation.
- At test time, we fine-tune this network to satisfy the geometric constraints of a particular input video, while retaining its ability to synthesize plausible depth details in parts of the video that are less constrained.
- Visually, their results appear more stable.
- LIMITATION: The time taken while testing is a lot, 3 sec video takes 47 mins to find the depth map.

CHALLENGES:

Poses: Our method currently relies on COLMAP to estimate the camera pose from a monocular video. In challenging scenarios, COLMAP may not be able to produce reliable sparse reconstruction and camera pose estimation

Dynamic motion: Our method supports videos containing moderate object motion. It breaks for extreme object motion

Speed: we do not support online processing. For example, our test-time training step takes about 40 minutes for a video of 244 frames and 708 sampled flow pairs

Flow: Unreliable flow is filtered through forward-backward consistency checks, but it might be by chance erroneous in a consistent way. In this case, our method will fail to produce correct depth

MOTIVATION:

- **Augmented reality**
  - A fundamental problem in AR is to place an object in 3D space such that its orientation, scale, and perspective are properly calibrated. For this, we need depth information.
- **3d reconstruction:**
  - To reconstruct a 3d image/object we need depth information and multiple frames of the object.
- **Robotics and object trajectory estimation:**
  - **In autonomous vehicles:** With depth information, we can estimate the trajectory along the third dimension. the distance, velocity, and acceleration values of the object within reasonable accuracy.
- **Haze and Fog removal:**
  - Haze and Fog are natural phenomena that are a function of depth. Distant objects are obscured to a greater extent.
- **Portrait mode:**
  - Blur applied as a function of depth creates a much more appealing image than using just uniform blur.

METHODOLOGY:

CV PROJECT REPORT
-DEPTH ESTIMATION IN VIDEO

IMPLEMENTED PAPER: *Learning the Depths of Moving People by Watching Frozen People*

DATASET: DAVIS dataset.

EXPLAINING PAPER METHODOLOGY :

Steps:

— A sequence of images is given as input.

— Take two consecutive images in this sequence.

— Find the Human Mask using Mask RCNN(Pretrained Model) which detects the human mask.

— using FlowNet 2.0(Pretrained Model) and the two consecutive images, it will find the optical flow.

— Using ORB-SLAM2 (Pretrained Model) and the two consecutive images, it will find the Camera Parameters.

— With these two, optical flow and camera parameters, it will find the depth from parallax and confidence.

—> Now, We pass these 3 inputs (depth from parallax, confidence, human mask) to Regression CNN, and that will be trained.

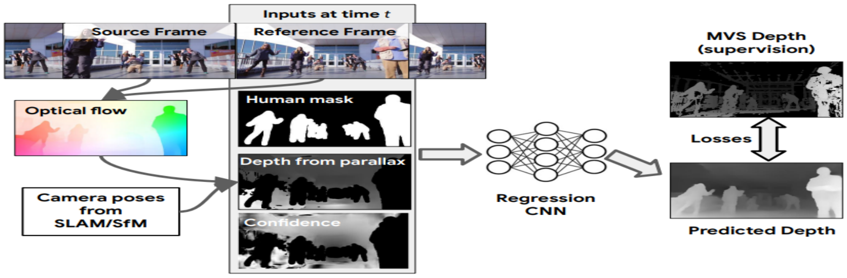
— While Testing we need to pass an image and it will automatically create depth from parallax, confidence, human mask and pass to our trained model and it will predict the depth map.

HOW TO RUN:

- Run the cv_project.ipynb file.
- First, get all the pre-trained models, and training should be done using Davis dataset.
- Then training can be done using “train_davis_videos.py”
- Then testing can be done using “test_davis_videos.py”
- The output will be stored in ”test_data/viz_prediction/”; the output will be a sequence of frames.
- To make it to a video, you need to run the cell below of the testing code.
- For the Evaluation score, We are using RMSE.

EXPLAIN MODELS :

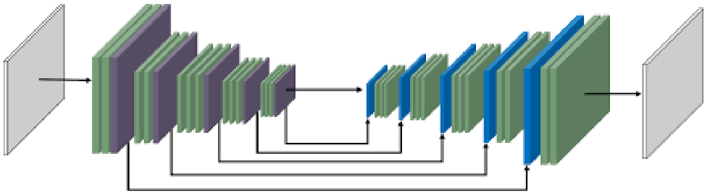
- We use the pixel2pixel model:
    - Model structure: At its core uses the Hourglass model
        - {
            - Takes 3 inputs, convolve it
            - Do batch normalization and apply ReLU activation function
            - This will first convolve down and then upsample.
        - }



Hourglass model :

CV PROJECT REPORT

-DEPTH ESTIMATION IN VIDEO



EXPERIMENTAL RESULTS:

METHOD	RMSE(DATASET - DAVIS)	RMSE(DATASET - TUM)
Learning the Depths of Moving People by Watching Frozen People	0.70	0.80

EVALUATION METRIC:

- *Si*-RMSE: scale-invariant Root Mean Square Error. Can also use RMSE, L1 relative error.

QUANTITATIVE ANALYSIS:

- The results show that it is not up to human-level depth estimated depth maps and needs to be improved a lot. But human error can also be considered. and humans can't predict the depth at the pixel level. So, finding the score based on pixel-level may not be optimal.

CONCLUSION:

- Finding DEPTH is very important , A lot of applications are using depth. Scope of improvement is there , like no discontinuity in depths, less time to produce results.As much as it is important still there are challenges.

REFERENCES:

[1] . Schönberger, J. L., Zheng, E., Frahm, J. M., & Pollefeys, M. (2016, October). Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision* (pp. 501-518). Springer, Cham.

[2]. Li, Z., Dekel, T., Cole, F., Tucker, R., Snavely, N., Liu, C., & Freeman, W. T. (2019). Learning the depths of moving people by watching frozen people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4521-4530).

[3]. Luo, X., Huang, J. B., Szeliski, R., Matzen, K., & Kopf, J. (2020). Consistent video depth estimation. *ACM Transactions on Graphics (ToG)*, *39*(4), 71-1.