15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Multiscale Two-view Stereo using Convolutional Neural Networks for Unrectified Images

Pramod Yadati, Anoop M Namboodiri
CVIT, KCIS , IIIT Hyderabad
pramod.yadati@research.iiit.ac.in, anoop@iiit.ac.in

Abstract

Two-view stereo problem is a well researched problem in 3D computer vision. Algorithms proposed in the past have focussed on rectified stereo images where the epipolar lines are parallel to the horizontal axis. The general problem of computing stereo correspondences for unrectified images without any knowledge of calibration parameters is an important problem but unexplored as yet. Our idea in this paper is to predict depth maps from two unrectified stereo images using a modified Flownet architecture. Since, datasets for depth map reconstruction for unrectified stereo images for deep learning do not exist, we have created a dataset of turn table sequences of 3D models from Google 3D warehouse. Following the concepts of Attention modelling, we implement an architecture for combining correlations computed at multiple resolutions using a simple element-wise multiplication of the correlations to aid the architecture to resolve correspondences for textureless and repeated textured surfaces. Our experiments show both qualitaitve and quantitative improvements of depth maps over the original Flownet architecture.

1 Introduction

Multi-view stereo problem is a well researched and an important part of 3D computer vision. It is one of the most common techniques to obtain 3D geometry of a scene. Two view stereo problem is a special case of Multi-View Stereo problem (MVS), wherein two views of the scene are provided to infer the 3D geometry of the scene. In a stereo vision system, when the cameras are calibrated, i.e the intrinsic and extrinsic matrices are known, the images can be rectified and computing correspondences under the epipolar geometry constraint becomes a 1-D search problem. In an alternative setup wherein only the intrinsic parameters are known, the problem is posed as Structure from Motion problem (SfM) which is a well researched area in 3D computer vision. Incase of both intrinsic and extrinsic parameters being unknown, the problem is of Fundamental matrix computation, which is a challenging problem in computer vision. It is important to note that all these problems become hard when only two-views of the scene are available.

Two-view stereo for depth map prediction is a tougher problem to solve as compared to MVS because of redundancy in MVS i.e, there are many images to infer 3D geometry and aggregation of depth maps inferred from groups of images is sufficient [3]. On the other hand, two-view stereo algorithms require rectified images. In the absence of calibration parameters, rectification will require computation of Fundamental matrix which will be subjected to errors. These errors will cause faulty correspondences between the rectified images. An end-to-end solution using deep networks holds promise in such situations.

Figure 1: Example of 'unrectified' stereo images. Our model takes images which are turn table sequences with arbitrary rotation angle $\leq 15°$, and predicts the depth map with respect to the left view.

Our main contributions are two fold :-

1. We propose an end-to-end deep learning solution to predict depth maps for unrectified stereo images using a modified Flownet architecture. We propose an Attention modelling approach to combine correlations at multiple scales to disambiguate textureless and repeated textured surfaces.

2. Due to unavailability of datasets for unrectified stereo for deep learning, we have created a rich dataset of turn table sequences of 3D models with varying textures and repeated textured surfaces using the Google warehouse 3D models.

1.1 Related Work

In general, Two-view stereo correspondence algorithms assume camera calibration and epipolar geometry and surfaces in the scene to be lambertian surfaces. Hence, the input to these algorithms are typically rectified images.

Stereo algorithms: In general, Stereo algorithms can be classified into broad categories namely - Global, Local and Hierarchical methods. Local algorithms compute disparity at a given pixel based on intensity values within a finite window, usually making implicit smoothness assumptions by aggregating support. Global methods on the other hand make explicit smoothness assumptions and solve an optimization problem that minimizes a global cost function as function of the intensity values in a window surrounding the pixel and smoothness term. Hierarchical methods operate on an image pyramid where results from coarser levels are used to constrain a more local search at finer levels. A comprehensive comparative study of some of the algorithms was performed by Scharstein and Szeliski [6].

Deep learning based stereo algorithms : Convolutional neural networks have been applied to low-level vision tasks such as predicting optical flow . A

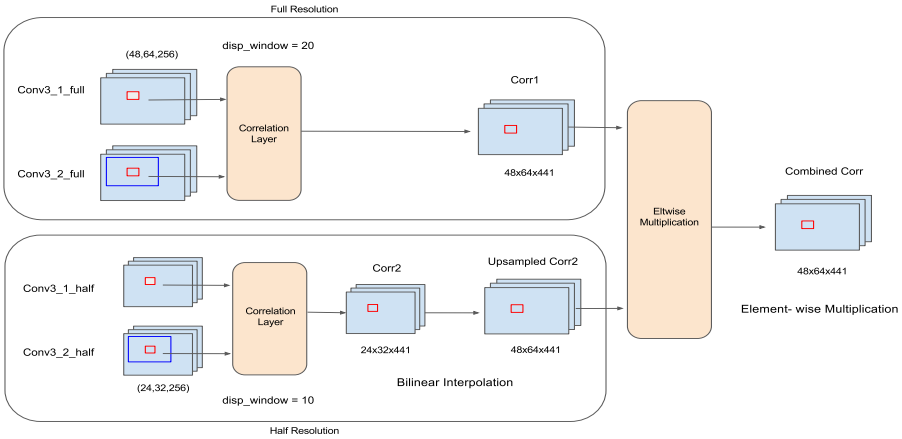


Figure 2: Architecture of Multi-scale correlation: Conv3 outputs from the Flownet architecture of image1 (left image) and image2 (right image) are fed into the correlation layer. For the half resolution correlations, the Conv3 outputs are down sampled by a factor of two using a max pool layer and fed to the correlation layer. The correlation layer for the halved resolution would compute correlations in a displacement window half the size used for full resolution.

siamese based fully convolutional network was introduced by Dosovitskiy et al. [2] which implements a correlation layer that combines the convolution outputs of consecutive frames followed by a expansion network which predicts optical flow between the frames. This work was extended by N.Mayer et al. [5] to predict disparity by modifying the correlation layer to search in single direction to compute correspondences for rectified stereo images. Žbontar and LeCun [7] propose siamese based CNN architectures to predict matching cost between image patches of rectified stereo images by minimizing the binary cross entropy loss. Zagoruyko and Komodakis [8] explore various architectures namely siamese, psuedo-siamese and 2-stream architectures for computing similarity score between 64×64 grayscale image patches. Luo et al. [4] aim to learn a probability distribution over all disparity values using a smooth target probability distribution. They achieve quick computation speeds by incorporating a dot-product layer similar to Žbontar and LeCun [7] and Chen et al. [1]. Chen et al. [1] introduce a deep embedding learning model which can extract discriminative features from stereo patches. The matching cost thus computed is used to build a MRF-based stereo framework for predicting disparity map.

We propose an end to end deep learning solution to predict depth maps for unrectified images using a modified Flownet architecture to deal with textureless and repeated textured surfaces. Following the concepts of Attention modeling of neural networks, we combine correlations at multiple scales to disambiguate textureless and repeated textured surfaces for computing correspondences.

2 Method

A common practice of stereo correspondence algorithms is to rectify input images using the 3D geometry of the setup as the problem reduces to 1D search problem owing to epipolar geometry constraints. A sub-problem to the unrectified stereo problem is a constrained setup of turn-table sequences. The attempt of this work is to learn to predict depth maps from two image sequences of such a setup. The learnable parts of the MVS pipeline are feature extraction and finding correlations between the stereo images. The flownet architecture implements these modules of the pipeline for predicting the optical flow in a video sequence.

We would like to focus on learning correlations for unrectified images. The Flownet-Corr architecture [2] implements a data-data correlation layer which basically computes a dot product of convolution features in a sliding window fashion. It relies on first three convolution layers to compute task specific features which can be used to compute correlations and finally predict optical flow for each pixel. The correlation layer computes a dot product of a 3×3 window of convolutional features of stereo images in a sliding window fashion. For computational tractability, the sliding window is constrained to slide in a area around the target pixel called the displacement window. Thus the context captured which would help in finding correlations is essentially limited by the static window size of the sliding window and the size of displacement window which is in a way inescapable.

The side effects of this setup is particularly prominent in finding correlations for textureless and repeated texture surfaces. The ambiguity in finding correlations at a certain context (window size) can be resolved by examining larger context (larger window size). Attention models in neural networks are loosely based on human attention mechanism which essentially is able to focus on certain region of an image at high resolution based on the surrounding region at low resolution and adjusting focal point over time. We incorporate this concept to the Flownet architecture to account for the ambiguity in estimating correspondences for textureless and repeated textured regions in the image.

In the Flownet architecture, the correlation layer output is a vector of N channels, each of which represent the strength of correlation between the corresponding 3×3 windows of the stereo image convolutions. The idea is to combine the correlations obtained at multiple resolutions in a way that resembles attention mechanism. Correlations are computed at full and half resolutions keeping the feature extraction

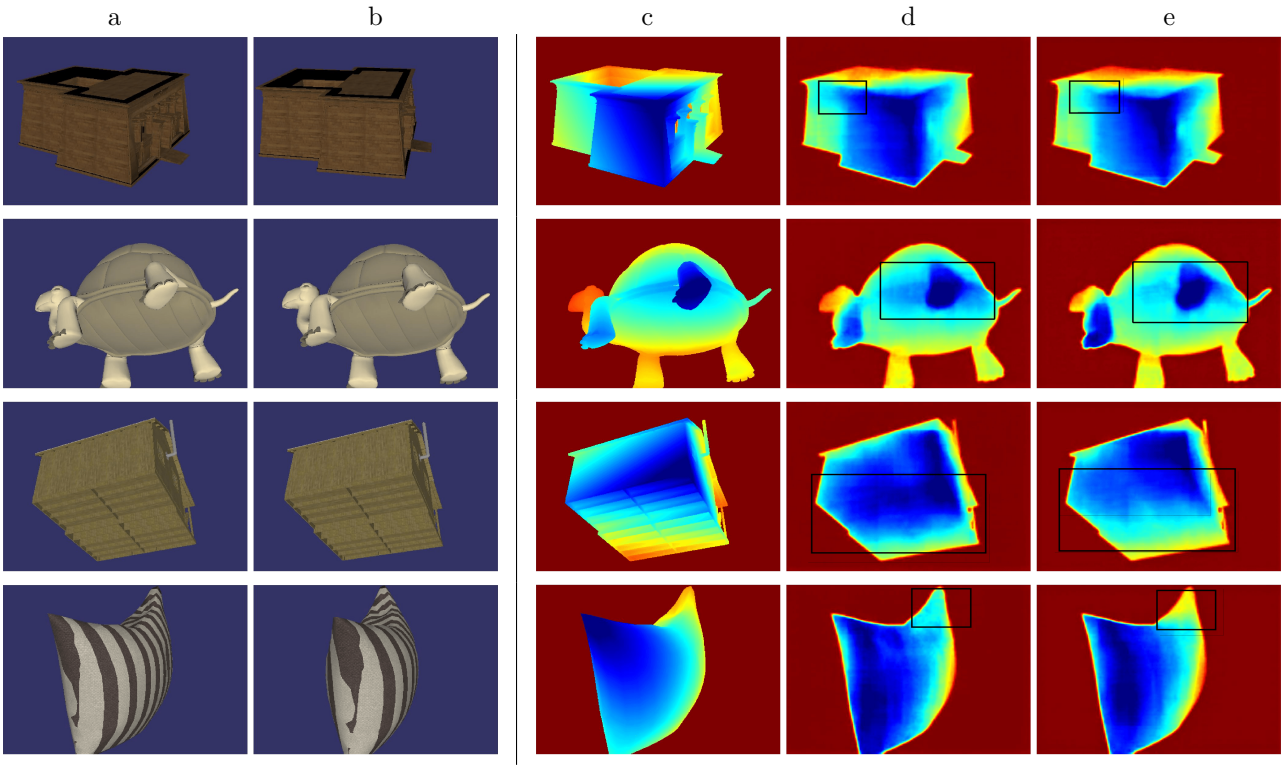


Figure 3: Results: (a) Left image; (b) Right image; (c) Ground truth depth map; (d) FlownetCorr architecture; (e) Our architecture. This figure shows models with good distribution of textureless and repeated-textured surfaces. Rows 1, 2 are results on image sequences whose models were part of the training data, but the orientations are not. These typically show that the model can generalize over random orientations. Rows 3 and 4 show results on models which are not part of the training data. These results show model performance on random orientations and different surface textures.

part i.e the conv layers common for both resolutions. After normalizing the correlations of both resolutions, we perform a simple element-wise multiplication of the correlations. The correlations at the coarser scale (half resolution) act as weights to the correlations at the finer scale (original resolution).

3 Experiments

A Unrectified image dataset for deep learning does not exist till date. A synthetic dataset of 3D models was created from Google 3D Data Warehouse. Using Opengl, turn table sequences of various 3D models at different orientations was created. Around 60,000 pairs of training images and depth maps of the left images were used to train this Multiscale Flownet network. The flownet architecture was used to predict disparity maps for rectified stereo images [5]. The model was trained on a GPU using Caffe using a simple perpixel L2 loss function untill significant decrease in average error was not found. Graphs in figure 5 show the performances of our model as compared to the FlownetCorr model on validation and test models respectively.

As we can see from Figure 3, Our model shows slight improvement in capturing correspondences which is evident from the distribution of depth levels. The model essentially captures how corresponding pixels move across the image sequence and render them as depth values. The pixels which have moved more would be much closer to the camera than the pixels which move less. Pixel correspondence in a hierarchical context is always more accurate and robust to local repeated textures and textureless surfaces. Evidence of this is quite clearly visible in the images. Using the correlations computed at a coarser level, our model is able to produce better depth resolution at finer scales as intended.

The ambiguities in computing textureless and repeated textured surfaces is better resolved by our model. Our model captures the relative depths between pixels consistently well across different textures and object shapes as is evident from Figure 3. Figure 4 shows instances where both models fail owing to small orientation changes. Even for humans, the relative depths are ambiguous given two views with small orientation changes.

4 Conclusions and Future work

Two-view stereo problem for unrectified images is difficult yet interesting problem to solve. Given its practical uses, A constrained setup of the general problem needs to be explored. Multi-view stereo algorithms which require multiple images for computing correspondences would not work given only two images with unknown calibration parameters. On the other hand, Two-view stereo algorithms proposed till now compute correspondences for rectified images. Rectification of two images with unknown geometry will result in errors which can affect the correspondences computed after rectification. We have attempted to predict depth maps from turntable sequences. We show how we can resolve correspondences for textureless and repeated textured surfaces by combining correlations computed

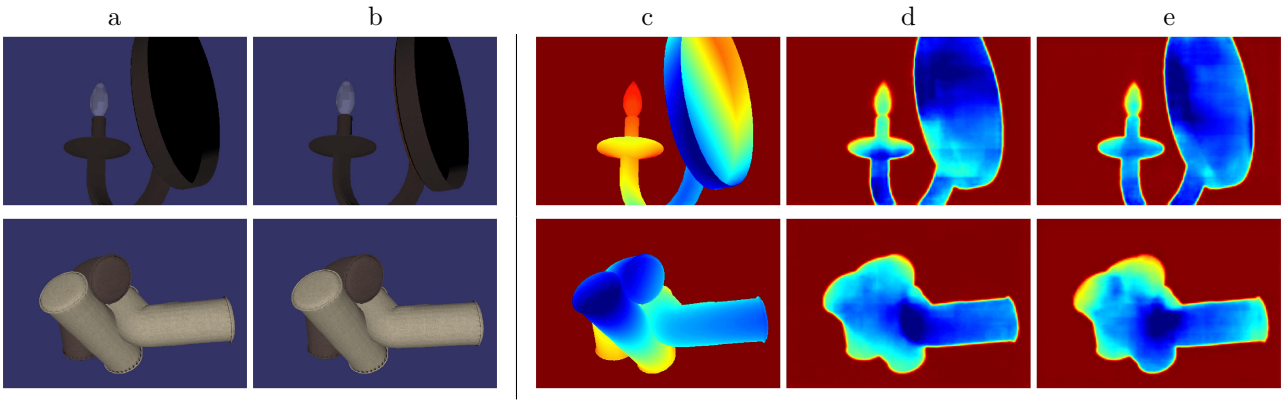
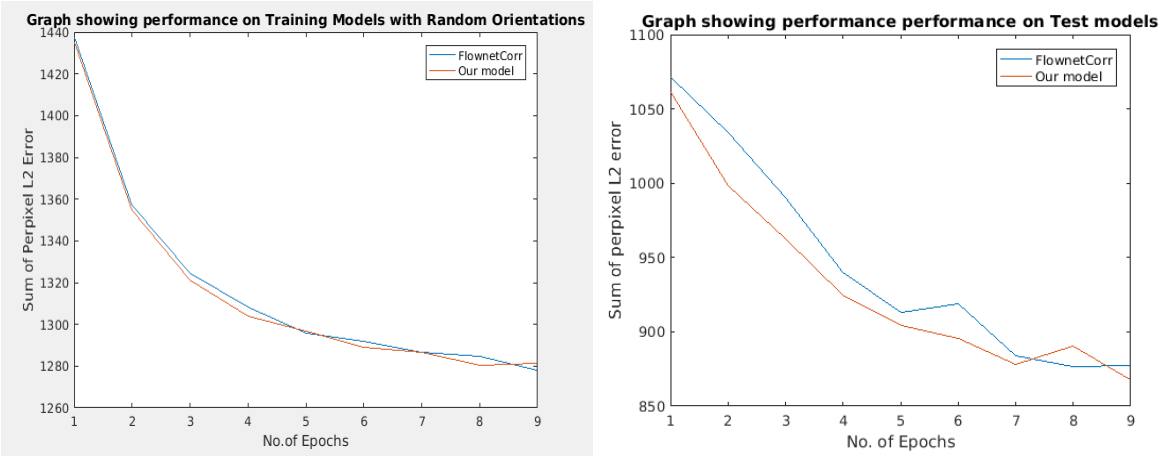


Figure 4: Failure Cases: (a) Left image; (b) Right image; (c) Ground truth depth map; (d) FlownetCorr architecture; (e) Our architecture. Rows 1 and 2 show poor correspondence computations owing to small rotation angles. Visually, even for human eye the depth levels are ambiguous.



(a) Validation performance

(b) Testset performance

Figure 5: Graph (a) shows the performances of our model and FlownetCorr model on Training models with random orientations. Though the performance of both models based on per-pixel squared error on an average is similar, the difference is visually evident in the previous figures. Graph (b) shows the performances of our model and FlownetCorr model on Test models with random orientations. Our model performs slightly better quantitatively as compared to the original FlownetCorr architecture on unseen models.

at multiple resolutions using a modified flownet architecture. We would like to incorporate more varied 3D models to our training dataset including real world models to improve the testset performances.

References

[1] Zhuoyuan Chen, Xun Sun, Liang Wang, Yinan Yu, and Chang Huang. A deep visual correspondence embedding model for stereo matching costs. In *ICCV*, 2015.

[2] Alexey Dosovitskiy, Philipp Fischery, Eddy Ilg, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, Thomas Brox, et al. Flownet: Learning optical flow with convolutional networks. In *ICCV*. IEEE, 2015.

[3] Yasutaka Furukawa and Carlos Hernndez. Multi-view stereo: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 9(1-2):1–148, 2015. ISSN 1572-2740. doi: 10.1561/0600000052.

[4] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *CVPR*, 2016.

[5] N.Mayer, E.Ilg, P.Häusser, P.Fischer, D.Cremers, A.Dosovitskiy, and T.Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016.

[6] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision*, 47(1-3), apr 2002. ISSN 0920-5691.

[7] Jure Žbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.*, 17(1), jan 2016. ISSN 1532-4435.

[8] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *CVPR*, 2015.