Course Notes 5: Multi-View Stereo

1 Introduction

Reconstructing 3D structure from 2D images is a key challenge in computer vision. Multi-View Stereo (MVS) addresses this by using images from different viewpoints to create dense 3D reconstructions. It starts with stereo matching—finding corresponding points to estimate depth. Traditional methods use hand-crafted similarity metrics and geometric constraints but struggle with issues like textureless regions and occlusions. Learning-based approaches, using deep neural networks, improve robustness and accuracy. This lecture covers both classical methods like PMVS and modern approaches such as MVSNet.

2 Stereo Matching

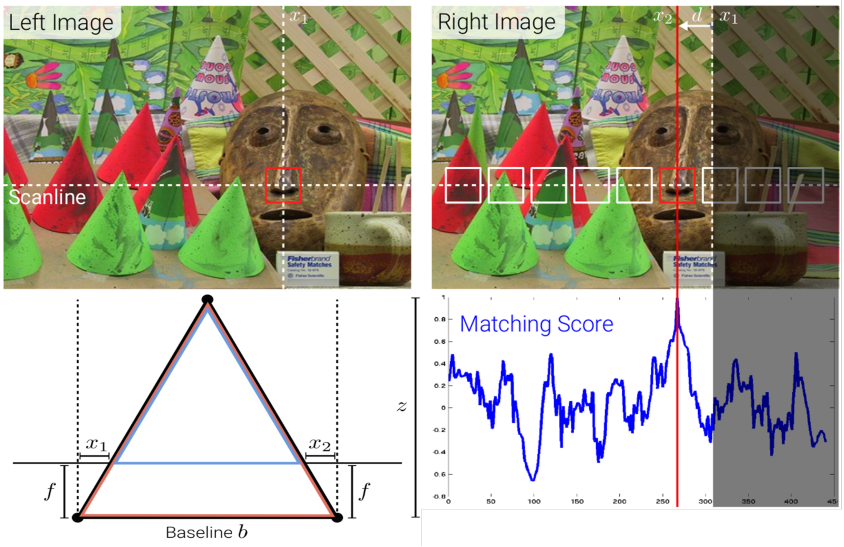

Figure 1: Stereo matching (Image credit: Andreas Geiger).

Stereo matching is a technique in computer vision used to estimate the depth of a scene by finding corresponding points between two images captured from slightly different viewpoints. The process involves several steps:

- **Rectification:** Align the stereo images so that corresponding points lie on the same horizontal line.
- **Matching:** For each pixel in one image, search along the corresponding epipolar line in the other image to find the best match.

- **Disparity Computation:** Calculate the horizontal difference (disparity) between matched pixels.
- **Depth Estimation:** Use the disparity to estimate depth, typically via the relation

$$z = \frac{f\,b}{d},$$

where f is the focal length, b is the baseline (distance between cameras), and d is the disparity.

2.1 Similarity Measures

The quality of the match is determined by comparing small windows (patches) around the pixels in both images. Common similarity measures include:

Sum of Squared Differences (SSD) sums the squared differences in intensity between corresponding pixels:

$$\mathrm{SSD}(u,v) = \sum_{(i,j)\in\mathcal{W}} \left[I_L(u+i,v+j) - I_R(u+i-d,v+j)\right]^2.$$

Lower SSD values indicate a better match.

Sum of Absolute Differences (SAD) is similar to SSD but uses absolute differences instead of squared differences:

$$\mathrm{SAD}(u,v) = \sum_{(i,j)\in\mathcal{W}} \left|I_L(u+i,v+j) - I_R(u+i-d,v+j)\right|.$$

Like SSD, lower SAD values indicate a closer match between patches, and it is often preferred for its computational simplicity.

Normalized Cross-Correlation (NCC) measures the similarity between patches after normalizing for local brightness variations:

$$\mathrm{NCC}(u,v) = \frac{\sum_{(i,j)\in\mathcal{W}} \left[I_L(u+i,v+j) - \bar{I}_L\right]\left[I_R(u+i-d,v+j) - \bar{I}_R\right]}{\sqrt{\sum_{(i,j)\in\mathcal{W}} \left[I_L(u+i,v+j) - \bar{I}_L\right]^2 \sum_{(i,j)\in\mathcal{W}} \left[I_R(u+i-d,v+j) - \bar{I}_R\right]^2}},$$

where $\bar{I}_L$ and $\bar{I}_R$ are the mean intensities of the patches in the left and right images, respectively. A value close to 1 indicates a strong match.

2.2 Failure cases of stereo matching

Traditional stereo reconstruction approaches use hand-crafted similarity metrics (e.g., NCC) and regularization techniques such as Semi-Global Matching (SGM) to recover 3D points. Recent stereo benchmarks have reported that although traditional algorithms achieve high accuracy, there remains significant room for improvement in reconstruction completeness. The primary reason for this limitation is that the hand-crafted similarity measures and block matching methods perform well mainly with Lambertian surfaces and tend to fail in the following scenarios:

- **Textureless Surfaces:** It is difficult to infer geometry from textureless surfaces (e.g., a white wall) because they appear similar from different viewpoints.

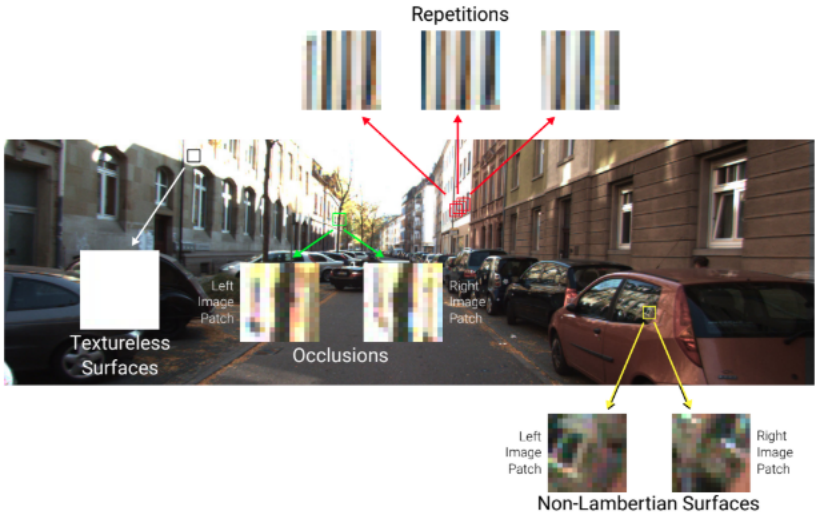

Figure 2: Stereo matching failure cases (Image credit: Andreas Geiger).

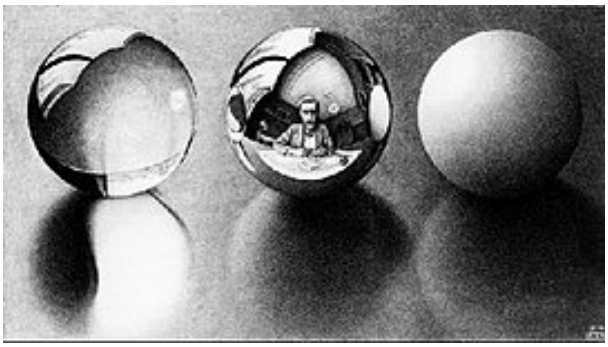

Figure 3: Three geometrically identical spheres with different material property (by M. C. Escher, 1746).

- **Occlusions:** Parts of scene objects may be partially or completely hidden in certain views due to occlusions.
- **Repetitions:** Block matching techniques can yield similar responses for different surfaces when geometric and photometric patterns are repetitive.
- **Non-Lambertian Surfaces:** Surfaces that deviate from Lambertian reflectance exhibit different appearances across viewpoints.
- **Other Non-Geometric Variations:** Factors such as image noise, vignetting, exposure changes, and lighting variations can further degrade performance.

2.3 Learning-based Approaches for Stereo Matching

To tackle failure cases of traditional stereo-matching methods, such as textureless surfaces, occlusions, repetitive patterns, and non-Lambertian effects, learning-based approaches have been introduced. These methods leverage large datasets and deep network architectures to learn feature spaces that are more robust to stereo matching than raw RGB representation.

2.3.1 Siamese Networks for Stereo Matching

Zbontar et al. [5] proposed an early method that employs a *siamese network* architecture. In this approach, two identical subnetworks process patches from the left and right images independently. Each subnetwork, typically implemented as a multilayer perceptron (MLP) or a convolutional neural network (CNN), extracts a compact feature vector from the input patch.

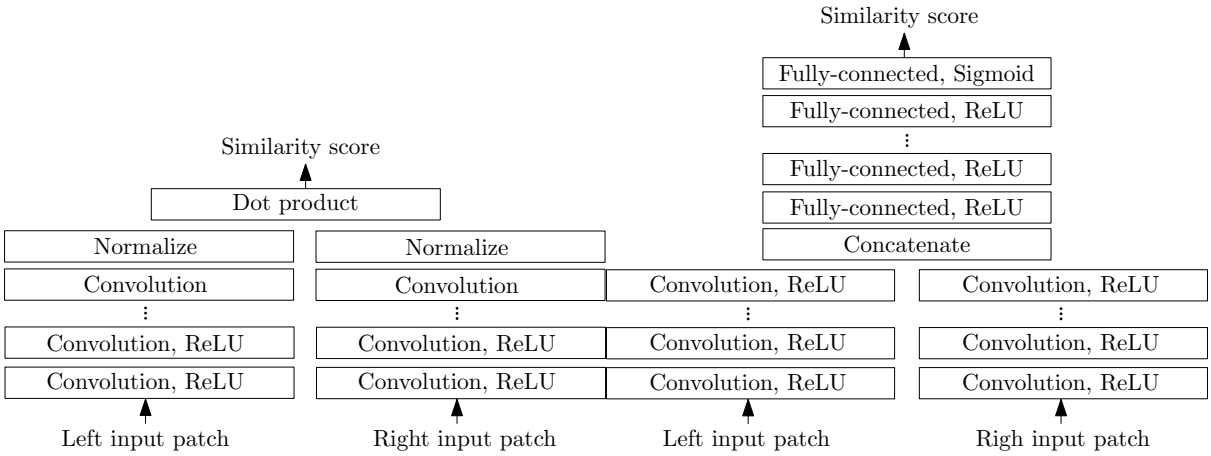

Figure 4: MLP-based siamese network for stereo matching

The network is trained with a contrastive loss using pairs of matching and non-matching patches, ensuring that corresponding patches yield higher similarity score while non-corresponding ones estimate the lower score. The resulting similarity scores, computed over a range of disparities, is integrated into a stereo matching pipeline to recover the disparity map.

2.3.2 DispNet: Disparity Estimation Network Architecture

DispNet [3] is one of the pioneering works that utilizes an end-to-end trained deep neural network for stereo matching. The network takes a pair of rectified stereo images (left and right views) as input and estimates the disparity map directly. The architecture is inspired by U-Net, featuring an encoder-decoder structure with convolutional layers, downsampling, and skip connections that help preserve spatial details when restoring the original resolution.

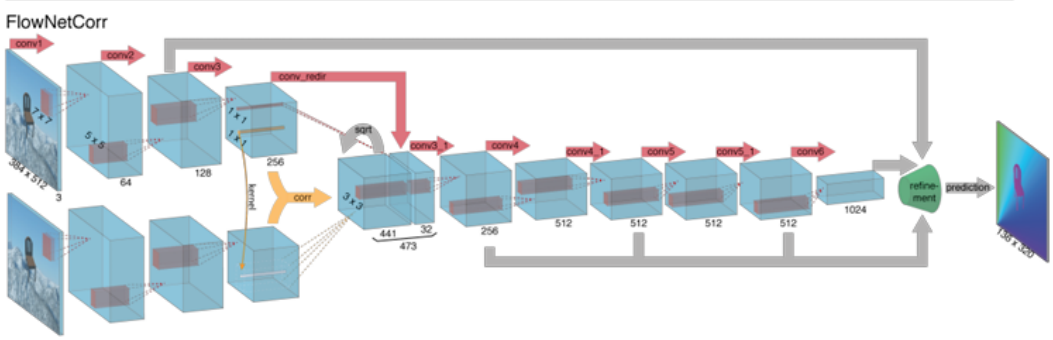

Figure 5: Disparity matching network architecture of DispNet.

Key components of the DispNet architecture include:

- **Correlation Layer:** After several convolutional and pooling layers, a specialized correlation layer is introduced to compute the matching cost between the feature maps from the left and right images. This mimics traditional block matching but operates on learned feature representations.
- **Multi-Scale Loss:** To guide the network during training, a multi-scale loss function is employed. Disparity predictions are generated at multiple scales within the decoder, and each prediction is compared with a downscaled version of the ground truth disparity. The total loss is a weighted sum of disparity errors across these scales, improving convergence and accuracy at different resolutions.
- **Curriculum Learning Strategy:** Training is performed in a progressive manner. Initially, the network is trained on synthetically generated simple scenes at low resolutions. As training progresses, the complexity of the scenes and the resolution of the inputs are gradually increased. This curriculum learning strategy helps the model to learn robust features before being fine-tuned on more challenging real-world datasets.

2.3.3 GC-Net for Deep Stereo Regression

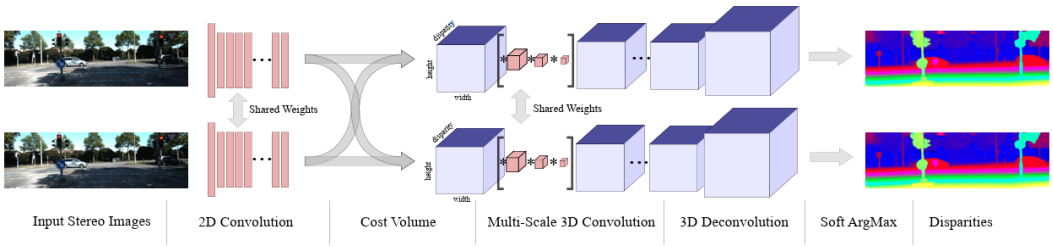

Figure 6: End-to-end deep stereo regression architecture, GC-Net.

GC-Net[2] further improved previous methods by explicitly integrating geometric reasoning into the stereo-matching process. The main components of GC-Net are:

- **Cost Volume Construction:** Features are extracted from both images and a 4D cost volume is constructed by concatenating feature maps across candidate disparities.
- **3D Convolutional Cost Aggregation:** The constructed cost volume is processed with 3D convolutional layers. This allows the network to aggregate contextual and geometric information over both spatial and disparity dimensions, enforcing smoothness and consistency in the predictions.
- **Differentiable Disparity Regression:** After aggregation, a softmax function is applied along the disparity dimension to convert costs into a probability distribution. The final disparity is then computed as a weighted sum:

$$d = \sum_{d'} d' \cdot \text{softmax}(-C(d')), \tag{2.1}$$

where $C(d')$ denotes the cost associated with disparity d'.

3 Multi-view Stereo

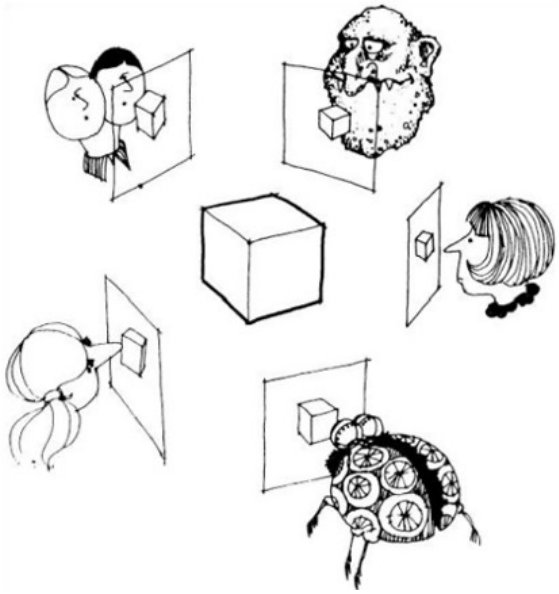

Figure 7: MVS aims to find a 3D shape that explains the images (Image credit: Svetlana Lazebnik).

Multi-view Stereo (MVS) aims to recover a dense 3D reconstruction of a scene from multiple images taken at different viewpoints. Unlike traditional stereo matching that relies on just two views to compute disparity maps, MVS exploits the redundancy in multiple views to produce highly detailed and robust 3D models. In this lecture notes, we introduce two prominent approaches to MVS: a classical method known as PMVS and modern learning-based methods that leverage differentiable homography- MVSNet.

3.1 PMVS: Patch-based Multi-view Stereo

Figure 8: Overall approach of PMVS. From left to right: A sample input image, detected features, reconstructed patches after the initial matching, final patches after expansion and filtering, and the mesh model.

PMVS[1] is a classical approach for dense 3D reconstruction. It builds a point cloud by reconstructing small surface patches and then expanding these patches to cover the scene. The main steps of PMVS include:

- **Patch Detection:** Reliable feature points are detected across the available views.
- **Patch Expansion:** Starting from these feature points, local planar patches are generated and iteratively expanded to densely cover the scene.
- **Consistency Checks:** Each patch is evaluated for both photometric consistency (similar appearance across views) and geometric consistency (alignment with epipolar constraints) to ensure accurate reconstruction.

By enforcing these constraints, PMVS effectively filters out incorrect matches and produces a robust, dense 3D point cloud representing the scene's surfaces.

3.2 Learning-based Methods: Differentiable Homography and MVSNet

Recent developments in deep learning have led to the emergence of learning-based MVS methods that integrate the entire reconstruction pipeline into an end-to-end differentiable framework.

3.2.1 Differentiable Homography

Differentiable homography warping is a key component in modern learning-based multi-view stereo (MVS) methods. It aligns feature maps from different views onto a common reference view, which is crucial for constructing the cost volume used in depth estimation. Given a homogeneous pixel coordinate $\mathbf{p}$ in the reference image and a depth hypothesis d, the corresponding pixel coordinate $\mathbf{p}'$ in a source view is computed as:

$$\mathbf{p}' \sim \mathbf{H}(d, p) = \mathbf{K}\left(\mathbf{R}\left(d\,\mathbf{K}^{-1}\mathbf{p}\right) + \mathbf{t}\right), \tag{3.1}$$

where:

- $\mathbf{p}$ is the homogeneous pixel coordinate in the reference image,
- d is the depth candidate (hypothesis),
- $\mathbf{K}$ is the intrinsic calibration matrix,
- $\mathbf{R}$ and $\mathbf{t}$ represent the rotation and translation from the reference camera to the source camera.

This equation can be understood in three sequential steps:

1. **Back-projection:** The pixel $\mathbf{p}$ is back-projected into the 3D space of the reference camera as $d\,\mathbf{K}^{-1}\mathbf{p}$, assuming the point lies at depth d.
2. **Transformation:** This 3D point is then transformed to the source camera coordinate system using the rotation $\mathbf{R}$ and translation $\mathbf{t}$.
3. **Projection:** Finally, the transformed 3D point is projected back onto the source image plane by applying the intrinsic matrix $\mathbf{K}$, resulting in the pixel coordinate $\mathbf{p}'$.

Feature Warping. With the homography defined in Equation 3.1, feature maps can be warped from the source view to the reference view. Let $\mathbf{F}_s$ denote the feature map of the source image. The warped feature map at the reference view for a depth hypothesis d is given by:

$$\mathbf{F}_s^{\text{warp}}(x, y; d) = \mathbf{F}_s \left(\mathbf{H}(d, [x, y, 1]^\top) \right), \tag{3.2}$$

where $\mathbf{H}(d)$ is derived from the differentiable homography equation.

Figure 9: Hans Holbein's "The Ambassadors" shows a perspective effect akin to homography, where the skull appears more natural from a slanted view than from a straight-on view.

This warping operation is typically implemented with bilinear interpolation, ensuring it remains fully differentiable with respect to both the depth d and the network parameters. Repeating this process for multiple depth hypotheses allows the construction of a cost volume for depth estimation.

Intuition. Differentiable homography allows the network to "simulate" different views of the scene by warping features as if they were captured from various depths. This process aids in identifying the depth that best aligns features across multiple views, which is essential for accurate multi-view depth estimation.

3.2.2 MVSNet

MVSNet[4] is a representative deep learning-based MVS method that leverages a cost volume approach for depth estimation. The overall approach is taking to GC-Net. Its pipeline can be summarized in the following steps:

- **Feature Extraction:** A deep convolutional neural network extracts high-level features from each input image.
- **Cost Volume Construction:** Differentiable homography warps the extracted features from multiple views onto a common reference view over a range of depth hypotheses, forming a 3D cost volume.

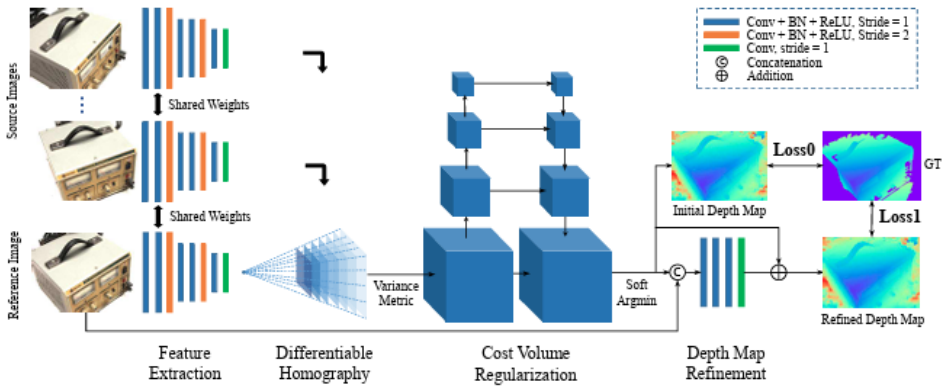

Figure 10: MVSNet architecture.

- **3D CNN Regularization:** A 3D convolutional neural network processes the cost volume to regularize the matching costs, yielding refined depth probability distributions.
- **Depth Estimation:** The final depth map is obtained either by selecting the depth hypothesis with the highest probability or by applying a soft-argmin operation over the cost volume.

The end-to-end differentiable design of MVSNet enables simultaneous training of all stages, from feature extraction and cost volume construction to depth estimation, on large datasets. This integrated approach has resulted in state-of-the-art performance in reconstructing detailed 3D structures, even in challenging scenes.

References

[1] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 2009.

[2] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, 2017.

[3] 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.

[4] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[5] Jure Žbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 2016.