A View of Computer Vision

David McAllester

Toyota Technologial Institute at Chicago (TTI-Chicago)

Joint work with

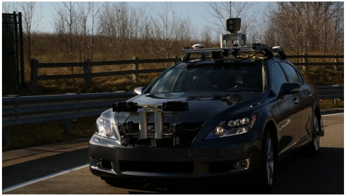
Koichiro Yamaguchi (TCRDL)

Michael James (TRINA)

Harry Yang (TTIC)

Raquel Urtasun (University of Toronto)

Do We Need Vision for Robotics?



Toyota Autonomous Car

Velodyne (Lidar)

The Value of Vision

- Vision systems can provide depth information that very dense in space and time.
- Vision systems can provide accurate motion information.
- Vision presents an important scientific challenge.

Geometric vs. Semantic Vision

Rendering — converting scene geometry to images.

Inverse Rendering (low level vision) — converting images to scene geometry (e.g., stereo, structure from motion, shape from shading).

Semantic Vision (high level vision) — vision involving conceptual labels such as “car”, “bicycle”or “building” (e.g., image classification, object detection, semantic segmentation).

Both are important and should be done jointly.

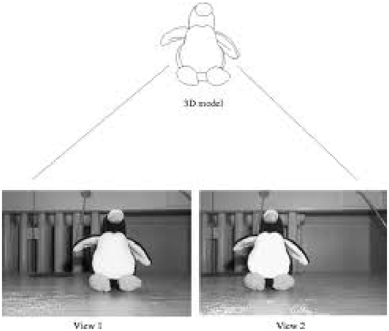
Why is Low Level Vision Important?

Autonomous Learning — learning from simply being in the world. Learning without labeled data.

Human babies learn autonomously.

Stereo and motion seems central to autonomous learning.

Stereo



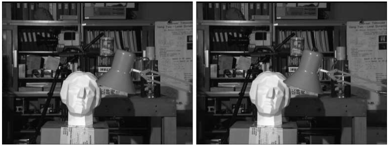
Correspondences between pixels from two cameras determine depth.

Unfortunately, the correspondences are surprisingly difficult for computers to see.

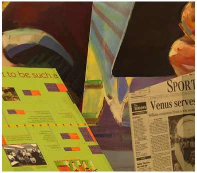
Motion Stereo

Given two images of a **Static Scene**, with known camera positions, the correspondence of a given pixel lies along a known epi-polar line. The position on the epi-polar line corresponds to depth in the image.

The Middlebury Sudio-Image Stereo Benchmark (2001)



Tsukuba



Venus

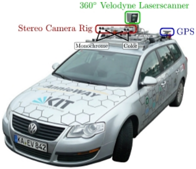
Cones

Teddy

The KITTI Driving-Scene Benchmark (2012)

Karlsruhe Institute of Technology (KIT)
Toyota Technological Institute at Chicago (TTI)

Andreas Geiger, Philip Lenz, Christoph Stiller, Raquel Urtasun



KITTI Stereo Leader Board, November 11, 2014

Team	Data*	Date	Error Rate	Time	Proc.
Zbontar and LeCun. NYU	S	August 2014	2.61%	100s	GPU
Yamaguchi et al. TCRDL, TTIC	S, MS	Feb. 2014	2.83%	35s	1 core
Vogel et al. ETH Zurich TU Darmstadt	S, MS	March 2014	3.05%	300s	1 core
Anonymous (TTIC)	S	November 2014	3.30%	6s	6 cores
Yamaguchi et al.	S	Sept 2014	3.39%	2s	1 core
Yamaguchi et al.	S	March 2013	3.40%	300s	4 cores
⋮ 64 entries total	⋮	⋮	⋮	⋮	⋮

* S means stereo pair, MS means motion stereo pair.

KITTI Motion Stereo Leader Board, November 11, 2014

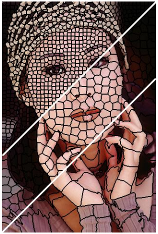
Team	Data*	Date	Error Rate	Time	Proc.
Vogel at al. ETH Zurich TU Darmstadt	S, MS	March 2014	2.72%	300s	1 core
Yamaguchi et al. TCRDL, TTIC	S, MS	Feb. 2014	2.82%	35s	1 core
Yamaguchi et al.	MS	Sept 2014	3.39%	11s	1 core
Vogel at al.	S, MS	April 2013	3.56%	200s	4 cores
⋮ 54 entries total	⋮	⋮	⋮	⋮	⋮

* S means stereo pair, MS means motion stereo pair.

TCRDL &TTIC System: Slanted Plane Model

Birchfield and Tomasi, ICCV, 1999

Create superpixels on the left image.



Shown are SLIC superpixels, Achanta et al., 2010

Assume that in each superpixel the scene is a plane.

The problem is then to estimate the plane for each superpixel using the stereo data.

Slanted Plane models have dominated the Middelbury benchmark.

TCRDL & TTIC System: Overview

- Slanted Plane Model.
- Discrete edge labels — each boundary between two superpixels is labeled as occlusion, hinge or coplanar.
- Segmentation, plane parameters and edge labels are inferred by joint energy minization. The energy involves the following terms.
  - The match energy for a setting of plane parameters is computed from SGM semi-sparse disparity values computed by the SGM stereo vision algorithm (Hirschmuller, 2005).
  - The smoothness energy between adjacent planes depends on the edge label between them.
  - SLIC-like segmentation energy.

Stereo Pairs and Motion Stereo Pairs

The system can be either with a single stereo pair (S), a single motion stereo pair (MS), or both (S, MS).

In all three settings we use a variant of semi-global matching (SGM) to produce an initial semi-dense depth estimate. Joint SGM for both S and MS is nontrivial.

The semi-dense SGM estimate is then smoothed with a slanted-plane model.

The smoothing algorithm is the same in all three cases.

Slanted-Plane Smoothing by Alternating Optimization

We alternate between optimizing:

- The segmentation.
- The segment plane parameters.
- The boundary labels.

Depth-Aware SLIC Segmentation

We want to minize $E_{\mathrm{seg}} = \sum_{p \in \mathrm{pixels}} E_{\mathrm{seg}}(p, s_p)$.

$E_{\mathrm{seg}}(p, s_p)$ involves the position, color and depth of p and the mean position, mean color and plane parameters of s_p.

We use alternating minimization:

```
Initialize a set of segments (typically a grid).
repeat
    Compute segment parameters so as to minimize $E_{\mathrm{seg}}$ holding the segmentation fixed.
    Compute the segmentation $s_p$ so as to minimize $E_{\mathrm{seg}}$ holding the plane parameters fixed.
until Convergence
```

Problems with Alternating-Minimization Segmentation

Individual segments become disconnected and must be converted to connected segments in a post-processing phase.

The final segmentation is then not a local minimizer of the energy.

Topology-Preserving Segmentation (ECCV 2014)

```
Initialize the stack to contain all boundary pixels
while the stack is not empty do
    Take pixel $\mathbf{p}$ off the stack.
    if $\mathbf{valid\_connectivity}(\mathbf{p}) = 0$ then
        continue
    end if
    $s_p = \mathrm{argmin}_{\hat{s}_p \in \{S_q,\, q \in \mathcal{N}_4(p)\}} \; E_{\mathrm{seg}}(p, \hat{s}_p)$
    if $s_p$ was changed then
        Update Segment Parameters.
        Push pixels in $\mathcal{N}_4(\mathbf{p})$ onto the stack.
    end if
end while
```

KITTI Stereo Leader Board, November 11, 2014

Team	Data*	Date	Error Rate	Time	Proc.
Zbontar and LeCun. NYU	S	August 2014	2.61%	100s	GPU
Yamaguchi et al. TCRDL, TTIC	S, MS	Feb. 2014	2.83%	35s	1 core
Vogel et al. ETH Zurich TU Darmstadt	S, MS	March 2014	3.05%	300s	1 core
Anonymous (TTIC)	S	November 2014	3.30%	6s	6 cores
Yamaguchi et al.	S	Sept 2014	3.39%	2s	1 core
Yamaguchi et al.	S	March 2013	3.40%	300s	4 cores
⋮ 64 entries total	⋮	⋮	⋮	⋮	⋮

* S means stereo pair, MS means motion stereo pair.

Moving Objects — Scene Flow

Two stereo pairs taken at different times should determine both the depth and the 3D-motion of each point yielding a scene flow image (depth plus motion at each pixel).

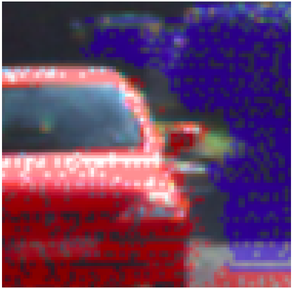
Problems with KITTI

The KITTI benchmark does not provide gound truth data for scenes with moving objects.

This is becasue ground-truth depth must be constructed from Lidar points taken at a time different from the time the images were taken and this can only be done reliably for static scenes.

Problems Even for Static Scenes

Even for static scenes the KITTI ground truth is not perfect:

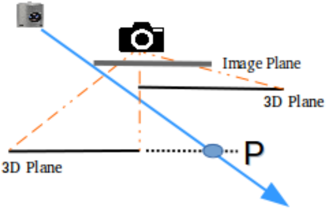


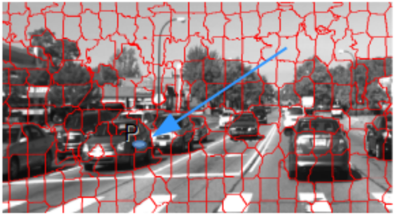
The KITTI ground truth consists of (quite accurate) *inferred* labels.

Inferring ground truth for scene flow is too difficult — scene flow is typically evaluated on synthetic data (e.g. MPI Sintel, 2012).

Avoiding Inferred Labels: Direct Lidar Evaluation

A system is scored by its ability to predict the flight time of a lidar probe.





For scenfow evaluation the lidar probe is at a time different from the image data.

Evaluating Existing Software Packages

We have implemented a “predictor” for converting disparity images to lidar predictions.

Stereo	Non-Occ	All	KITTI-All
SPS-St	4.55%	5.79%	4.41%
rSGM	5.84%	7.33%	6.60%
opencv-BM	6.61%	7.88%	13.76%
opencv-SGBM	9.54%	10.32%	9.13%
ELAS	10.80%	11.09%	9.96%

Other Evaluations are Possible

We will eventually implement predictors for converting the output of each of the following to lidar predictions.

- Image Segmentation.
- Optical Flow.
- Scene Flow.
- Monocular Depth Estimation.

Teams will be free to write their own predictors or select from a list of default predictors provided by the evaluation site.

Lidar Evaluation and Autonomous Learning

People can confirm vision by touching.

An autonomous car can confirm vision with lidar.

Direct lidar evaluation should allow autonomous learning of vision systems.

Semantic vision should improve the accuracy of inverse rendering.

Direct lidar evaluation should therefore also support autonomous learning of semantic vision.

Deep Neural Networks (DNNs)

Over the last three years there has been enormous progress in applications of DNN research.

The best systems (by a large margin) in speech recognition and image classification are now based on DNNs.

There is now active research on the use of DNNs in viturally all applications of machine learning.

DNNs in ImageNet Image Classification

Year	Team	Error Rate
2012	U of Tokyo	26.1%
2012	U of Toronto (DNN)	15.3%
2013	NYU (DNN)	11.2%
2014	Google (DNN)	6.7%

PASCAL Object Detection: A Moore's Law of AI?

	bicycle	bus	car	motorbike	person	20 class average
2007	36.9	23.2	34.6	27.6	21.3	**17.1**
2008	**42.0**	23.2	32.0	38.6	**42.0**	**22.9**
2009	**46.8**	**43.8**	**37.2**	42.0	**41.5**	**27.9**
2010	54.3	54.2	**49.1**	51.6	**47.5**	33.7 (**36.8**)
2011	(**58.1**)	(**57.6**)	(**54.4**)	(**58.3**)	(**51.6**)	(**40.9**)
2012	(**54.5**)	(**57.1**)	(**49.3**)	(**59.4**)	(**46.1**)	(**41.1**)
2013 (DNN)	(**56.3**)	(**51.4**)	(**48.7**)	(**59.8**)	(**44.4**)	(**43.2**)
2014 (DNN)						(**63.8**)

KITTI Stereo Leader Board, November 11, 2014

Team	Data*	Date	Error Rate	Time	Proc.
Zbontar and LeCun. NYU (DNN)	S	August 2014	2.61%	100s	GPU
Yamaguchi et al. TCRDL, TTIC	S, MS	Feb. 2014	2.83%	35s	1 core
Vogel et al. ETH Zurich TU Darmstadt	S, MS	March 2014	3.05%	300s	1 core
Anonymous (TTIC)	S	November 2014	3.30%	6s	6 cores
Yamaguchi et al.	S	Sept 2014	3.39%	2s	1 core
Yamaguchi et al.	S	March 2013	3.40%	300s	4 cores
⋮ 64 entries total	⋮	⋮	⋮	⋮	⋮

* S means stereo pair, MS means motion stereo pair.

Summary

- It seems likely that future inverse rendering systems will base both the match energies and the image smoothing on DNN methods.

- It also seems likey that direct lidar evaluation will allow autonomous learning of inverse rendering DNN systems.

- Autonomous learning of inverse rendering may lead to autonomous learning of semantic vision.

- I believe that there is still room for new theory in the area of autonomous learning.