Developments of 3D Computer Vision Since 2017

Yihong Wu
Institute of Automation
Chinese Academy of Sciences
ACCV, Perth, 2018.12
Contents

➢ Preface

➢ Image matching

➢ Visual localization: PnP, SLAM

➢ 3D reconstruction: SfM, learning, RGBD

➢ Trends

There are various academic awards and prizes named in his honour, Marr Prize

**Three stages:**

- **A primal sketch**
  - Feature extraction:
    - Points, edges, regions

- **A 2.5D sketch**
  - Viewer-centered three dimensional view of the environment

- **A 3D model**
  - A continuous, 3-dimensional map
3D vision is very important in computer vision

AR、VR、Robotics


• 2016: AR、VR
• 2017: Driverless car、robot、AGV、3D camera
  • 2017. 6. 5, Apple ARKit
  • 2017. 8, Google ARCore
• 2018: Driverless car、robot、AGV、3D camera
  • Boston Dynamics: Spotmini, 3D visual navigation
SpotMini is a small four-legged robot. It weighs 25 kg (30 kg if you include the arm). SpotMini is all-electric and can go for about 90 minutes on a charge, depending on what it is doing. The sensor suite includes stereo cameras, depth cameras, an IMU, and position/force sensors in the limbs. These sensors help with navigation and mobile manipulation.
Similar but Different: projective geometry
Main tasks

• Image matching
• Localization
• 3D reconstruction
Image matching

\[ \text{minimize} \quad g(R, C, F, X) \]
Localization and 3D reconstruction

Images captured by cameras

To compute camera pose
To model the 3D environment structure
Contents

➢ Preface

➢ Image matching

➢ Visual localization: PnP, SLAM

➢ 3D reconstruction: SfM, learning, RGBD

➢ Trends
Image matching

- Feature detection
- Descriptor extraction
- Matching
- Evaluation and dataset
● Overview
  • Traditional designed descriptors methods
  Learning descriptor methods: deep learning
  • Feature detection: deep learning
  • In practice: Traditional methods

● Learning feature detection
  • **CovDet**: CNN learning Covariant feature, Zhang, Yu, Kumar, Chang, CVPR2017
  • **AffNet**: CNN learning affine regions, Radenovic, Matas, arXiv2017
  • Haoliang Li, Sinno Jialin Pan, Shiqi Wang, Alex C. Kot. Domain Generalization with Adversarial Feature Learning. CVPR 2018.
● **Learning descriptors:**

- **L2Net:** progressive sampling strategy, relative distance between descriptors and extra supervision. Tian, Fan, Wu, CVPR 2017
- **HardNet:** improve L2Net, Mishchuk, Mishkin, Radenovic, Matas, NIPS 2017
- **DeepCD:** learns a pair of complementary descriptors of binary and float, Yang, Hsu, Lin, Chuang, ICCV 2017
- **Spread-out:** regularization term to maximize the spread in feature descriptor inspired by the property of uniform distribution, Zhang, Yu, Kumar, Chang, ICCV 2017 (pairwise and triplet losses + regularization technique)
- **PPFNet:** Global Context Aware Local Features for Robust 3D Point Matching. Haowen Deng, Tolga Birdal, Slobodan Ilic. CVPR 2018 (N-tuple loss, 3D point cloud)
- Georgios Georgakis, Srikrishna Karanam, Ziyuan Wu. End-to-End Learning of Keypoint Detector and Descriptor for Pose Invariant 3D Matching. CVPR 2018 (depth image)
L2Net:

- Input: 32*32 patch    Output: 128D, directly matched by L2 distance

- Network structure:
  - 6 number of 3*3 convolutional layers;
  - 1 number of 8*8 convolutional layers;
  - CIC (Convolution in Convolution) only for training, similar features are matched, different features aren’t matched;
  - LRN (Local Response Normalization) to normalize the network output

- Training:
  - Cost function with multiple errors:
    - Similar error: relative distance
    - Error term for descriptor compactness: decrease redundancy among different dimensions for decreasing overfitting
    - Error term for intermediate feature maps: for generalization ability without more parameters
  - 2-4 hours, GPU

- Speed of descriptor extractions: 21.3 k patch/s
Hpatches dataset

ECCV 2016 workshop “Local Features: state of the art, open problems and performance evaluation”, L2 net ranked No.1 in all tasks: image matching; image retrieval; image classification
Matching

1. J. Bian, W. Lin, Y. Matsushita, S. Yeung, T. Nguyen, M. Cheng. GMS: Grid-based Motion Statistics for Fast, Ultra-robust Feature Correspondence. CVPR2017 (GMS (Grid-based Motion Statistics), grids, encapsulating motion smoothness as the statistical likelihood of a certain number of matches in a region, real time) NDL


7. Fudong Wang, Nan Xue, Yipeng Zhang, Xiang Bai, and Gui-Song Xia. Adaptively Transforming Graph Matching. ECCV 2018 (With a linear representation map of the transformation, the pairwise edge attributes of graphs $\rightarrow$ unary node attributes, which enables us to reduce the space and time complexity significantly.)

8. Mohammed E. Fathy, Quoc-Huy Tran, M. Zeeshan Zia, Paul Vernaza, and Manmohan Chandraker. Hierarchical Metric Learning and Matching for 2D and 3D Geometric Correspondences. ECCV 2018. (While a metric loss applied to the deepest layer of a CNN, is often expected to yield ideal features irrespective of the task, in fact the growing receptive field as well as striding effects cause shallower features to be better at high precision matching tasks)

9. Yiran Zhong, Hongdong Li, Yuchao Dai. Open-World Stereo Video Matching with Deep RNN. ECCV 2018. (RNN, a continuous stereo video, 1) Feature-Net(CNN), (2) Match-Net(EncoDeco) $\rightarrow$ a depth-map, without a pre-training process, and without the need of ground-truth depth-maps as supervision)
Evaluation and dataset


- **Dataset Hpatches**
  
  Brown
  Hpatches: data quality, evaluation methods

Contents

➢ Preface

➢ Image matching

➢ Visual localization: PnP, SLAM

➢ 3D reconstruction: SfM, learning, RGBD

➢ Trends
Visual localization

- Known 3D knowledge
  - PnP, SLAM relocalization
  - RGBD SLAM
  - SLAM/Pose with learning
  - Marker SLAM

- Unknown 3D knowledge
  - General SLAM or Point\line\edge\plane SLAM
  - Camera+IMU SLAM
  - Semantic SLAM
  - Event camera SLAM
  - Relative pose
• **Known 3D knowledge**
  (2D-based methods achieve the lowest localization accuracy, 3D-based methods offer more precise pose)
  PnP, SLAM relocalization: 2D to 3D matching
  RGBD SLAM: 3D to 3D matching
  Learning SLAM; Marker SLAM

• **Unknown 3D knowledge**
  SLAM, real-time and online

**Complete overview for visual localization:**
**Known 3D knowledge:**

**PnP, SLAM relocalization**

**Large scale environment, heterogeneous data**

1. Nathan Piasco, Désiré Sidibé, Cédric Demonceaux, Valérie Gouet-Brunet. A survey on Visual-Based Localization: On the benefit of heterogeneous data. Pattern Recognition, 2018. (known environment, two distinct families: indirect(retrieval) and direct(6D))

2. Dylan Campbell, Lars Petersson, Laurent Kneip and Hongdong Li. Globally-Optimal Inlier Set Maximisation for Simultaneous Camera Pose and Feature Correspondence, ICCV 2017. (Marr Prize Honorable Mention)


Since a large proportion of outliers are common for this problem, we instead propose a globally-optimal inlier set cardinality maximisation approach which jointly estimates optimal camera pose and optimal correspondences.

$$\nu^* = \max_{\mathbf{R}, \mathbf{t}} f(\mathbf{R}, \mathbf{t})$$

$$f(\mathbf{R}, \mathbf{t}) = \sum_{\mathbf{f} \in \mathcal{F}} \max_{\mathbf{p} \in \mathcal{P}} 1(\theta - \angle(\mathbf{f}, \mathbf{R}(\mathbf{p} - \mathbf{t})))$$

Branch-and-bound to search the 6D space of camera poses, guaranteeing global optimality without requiring a pose prior to accelerate convergence.

4 Lemmas, 3 Theorems
Step 1. Build a Map-Graph

• $G(V, E)$: $V$ graph nodes, $E$ graph edges

Each edge is assigned a weight $c_{ij}$

For the $i$-th 3D point, denote the set of database images that contain this point as $A_i$. The state transition matrix is given by:

$$c_{ij} = \frac{|A_i \cap A_j|}{|A_j|}$$

$$C = [c_{ij}]$$
Step 2. Compute query vector

• Given a 2D image, compute:

\[
q_i = \sum_{f \in \mathcal{O}(i)} \frac{\sqrt{w_{fi}}}{N_i} \cdot \log \left( \frac{N}{N_f} \right)
\]

\[
w_{fi} = \exp\left(-H^2(f, i)/\sigma^2\right), \forall i \in [1..N].
\]

a 2D query feature f and a 3D map point i

\[
q_i \leftarrow \left( \frac{q_i}{\sum_{i=1}^{N} q_i} \right), \forall i \in [1..N].
\]
Step 3. Random walk on map-graph

Given a map-graph \( G(V,E) \) along with a state transition matrix \( C \): a Markov Network or Markov Random Field

\[
p(t + 1) = \alpha C p(t) + (1 - \alpha) q.
\]

Once the iteration converges, they sort this steady-state probability vector in descending order, which gives the final “matchability” of every 3D point to the set of 2D query features.


12. Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, Akihiko Torii. InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. CVPR 2018. (CNN based 2D image retrieval, CNN features and P3P combination to match between 3D and 2D, which can deal with textureless indoor scenes to some extent, indoor dataset)

RGBD SLAM


• Pyojin Kim, Brian Coltin, and H. Jin Kim. Linear RGB-D SLAM for Planar Environments. ECCV 2018. (jointly estimates camera position and planar landmarks in the map within a linear Kalman filter framework, and solve for the rotational motion of the camera using structural regularities in the Manhattan world)

Deep learning can improve camera localization robustness, having better performance in weak textures, illumination changes, season changes.


18. Eric Brachmann and Carsten Rother. Learning Less is More – 6D Camera Localization via 3D Surface Regression. CVPR 2018. (fully convolutional neural network for densely regressing scene coordinates, defining the correspondence between the input image and the 3D scene space.)


22. Peng Wang, Ruigang Yang, Binbin Cao, Wei Xu, Yuanqing Lin. DeLS-3D: Deep Localization and Segmentation with a 3D Semantic Map. CVPR 2018. (Initial coarse camera pose from GPS/IMU → a label 3D semantic map, 3D semantic map+the RGB image→ pose CNN, to realize camera pose determination)

Marker SLAM

<table>
<thead>
<tr>
<th>AR Toolkit (IWAR1999)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARTag (CVPR2005)</td>
</tr>
<tr>
<td>AprilTag (ICRA2011)</td>
</tr>
<tr>
<td>AprilTag 2 (IROS2016)</td>
</tr>
<tr>
<td>ChromaTag (ICCV2017)</td>
</tr>
</tbody>
</table>

3D matching  →  PnP, RANSAC
The problem:
Are correspondences necessary?
Can we localize a camera without correspondences?

The proposed method:
Circular marker SLAM
\[ r_1 = \frac{(m_0 \times m_1) \times l_\infty}{\left\| (m_0 \times m_1) \times l_\infty \right\|} \]
\[ t = s_0 m_0 \]
\[ r_3 = \pm \frac{1}{\sqrt{l_\infty^T l_\infty}} l_\infty = s_3 \frac{1}{\sqrt{l_\infty^T l_\infty}} l_\infty \]
\[ r_2 = r_3 \times r_1 \]

\[ f_u = \sqrt{\frac{a_1 b_1 (a_2^2 - b_2^2) - a_2 b_2 (a_1^2 - b_1^2)}{a_3 b_3 (a_1^2 - b_1^2) - a_1 b_1 (a_3^2 - b_3^2)}} \]
\[ \cdots \]
\[ f_v = \sqrt{\frac{a_1 b_1 (a_2^2 - b_2^2) - a_2 b_2 (a_1^2 - b_1^2)}{a_3 b_3 (a_2^2 - b_2^2) - a_2 b_2 (a_3^2 - b_3^2)}} \]
• AR/VR
• Robot grabbing at texture less circular objects
• Machine docking
Visual localization

- Known 3D knowledge
  - PnP, SLAM relocalization
  - RGBD SLAM
  - SLAM/Pose with learning
  - Marker SLAM
- Unknown 3D knowledge
  - General SLAM or Point\line\edge\plane SLAM
  - Camera+IMU SLAM
  - Semantic SLAM
  - Event camera SLAM
  - Relative pose
Unknown 3D knowledge, SLAM

Reviews:


Point\line\edge\plane fusion SLAM/General SLAM

Weak texture, illumination changes, empty corridor etc:
Point, line, edge, plane fusion to improve camera localization robustness

8. Y. Ling and S. Shen. Building maps for autonomous navigation using sparse visual SLAM features. IROS 2017. (incremental SLAM, real-time dense mapping, and free space extraction)
A novel framework fuses the advantages of direct and feature methods. Both accuracy and speed are considered.
## FMD Stereo SLAM: Fusing MVG and Direct Formulation Towards Accurate and Fast Stereo SLAM

Public dataset EuRoc

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Length /Duration</th>
<th>OURS</th>
<th>SVO stereo</th>
<th>ORBSLAM stereo without loop</th>
</tr>
</thead>
<tbody>
<tr>
<td>MH_01_easy</td>
<td>80.6m/182s</td>
<td><strong>3.80</strong></td>
<td>8.00</td>
<td>4.03</td>
</tr>
<tr>
<td>MH_02_easy</td>
<td>73.5m/150s</td>
<td><strong>3.76</strong></td>
<td>8.00</td>
<td>4.16</td>
</tr>
<tr>
<td>MH_03_medium</td>
<td>130m/132s</td>
<td>5.36</td>
<td>29.00</td>
<td><strong>4.78</strong></td>
</tr>
<tr>
<td>MH_04_difficult</td>
<td>91.7m/99s</td>
<td><strong>9.20</strong></td>
<td>267.00</td>
<td>40.49</td>
</tr>
<tr>
<td>MH_05_difficult</td>
<td>97.6m/111s</td>
<td><strong>9.30</strong></td>
<td>43.00</td>
<td>10.27</td>
</tr>
<tr>
<td>V1_01_easy</td>
<td>58.6m/144s</td>
<td>8.72</td>
<td><strong>5.00</strong></td>
<td>8.85</td>
</tr>
<tr>
<td>V1_02_medium</td>
<td>75.9m/83.5s</td>
<td>20.11</td>
<td><strong>9.00</strong></td>
<td>9.75</td>
</tr>
<tr>
<td>V1_03_difficult</td>
<td>79.0m/105s</td>
<td>53.28</td>
<td>36.00</td>
<td><strong>16.44</strong></td>
</tr>
<tr>
<td>V2_01_easy</td>
<td>36.5m/112s</td>
<td>8.85</td>
<td>9.00</td>
<td><strong>6.21</strong></td>
</tr>
<tr>
<td>V2_02_medium</td>
<td>83.2m/115s</td>
<td><strong>7.67</strong></td>
<td>52.00</td>
<td>7.96</td>
</tr>
<tr>
<td>V2_03_difficult</td>
<td>86.1m/115s</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

FMD: 109 Hz
ORB SLAM: 15 Hz
Stereo DSO: Large-Scale Direct Sparse Visual Odometry with Stereo Cameras
<table>
<thead>
<tr>
<th>Seq.</th>
<th>( t_{rel} )</th>
<th>( r_{rel} )</th>
<th>( t_{rel} )</th>
<th>( r_{rel} )</th>
<th>( t_{rel} )</th>
<th>( r_{rel} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>0.84</td>
<td>0.26</td>
<td>0.83</td>
<td>0.29</td>
<td>1.09</td>
<td>0.42</td>
</tr>
<tr>
<td>01</td>
<td>1.43</td>
<td>0.09</td>
<td>1.38</td>
<td>0.20</td>
<td>2.13</td>
<td>0.37</td>
</tr>
<tr>
<td>02</td>
<td>0.78</td>
<td>0.21</td>
<td>0.81</td>
<td>0.28</td>
<td>1.09</td>
<td>0.37</td>
</tr>
<tr>
<td>03</td>
<td>0.92</td>
<td>0.16</td>
<td>0.71</td>
<td>0.17</td>
<td>1.16</td>
<td>0.32</td>
</tr>
<tr>
<td>04</td>
<td>0.65</td>
<td>0.15</td>
<td>0.45</td>
<td>0.18</td>
<td>\textbf{0.42}</td>
<td>0.34</td>
</tr>
<tr>
<td>05</td>
<td>0.68</td>
<td>0.19</td>
<td>0.64</td>
<td>0.26</td>
<td>0.90</td>
<td>0.34</td>
</tr>
<tr>
<td>06</td>
<td>\textbf{0.67}</td>
<td>0.20</td>
<td>0.82</td>
<td>0.25</td>
<td>1.28</td>
<td>0.43</td>
</tr>
<tr>
<td>07</td>
<td>0.83</td>
<td>0.36</td>
<td>\textbf{0.78}</td>
<td>0.42</td>
<td>1.25</td>
<td>0.79</td>
</tr>
<tr>
<td>08</td>
<td>\textbf{0.98}</td>
<td>0.25</td>
<td>1.07</td>
<td>0.31</td>
<td>1.24</td>
<td>0.38</td>
</tr>
<tr>
<td>09</td>
<td>0.98</td>
<td>0.18</td>
<td>\textbf{0.82}</td>
<td>0.25</td>
<td>1.22</td>
<td>0.28</td>
</tr>
<tr>
<td>10</td>
<td>\textbf{0.49}</td>
<td>0.18</td>
<td>0.58</td>
<td>0.28</td>
<td>0.75</td>
<td>0.34</td>
</tr>
</tbody>
</table>

| mean | 0.84          | 0.20          | 0.81          | 0.26          | 1.14          | 0.40          |

Table 1. Comparison of accuracy on KITTI training set. \( t_{rel} \) translational RMSE (\%), \( r_{rel} \) rotational RMSE (degree per 100\( m \)). Both are average over 100\( m \) to 800\( m \) intervals. Best results are shown as bold numbers.
Camera+IMU, other sensor SLAM

Multi cameras, multi kinds of sensors fusion to improve camera localization robustness

Semantic SLAM

Semantic SLAM: geometry with contents to improve SLAM accuracy and simultaneously to understand environments.

Event Camera SLAM

**Event camera**: capture light intensity changes for each pixel.

Relative pose


Deep learning methods are very active, but are less accurate or have lower generalization ability to larger scenes outside training data.

Traditional methods are still the main used methods in practice.
Contents

➢ Preface

➢ Image matching

➢ Visual localization: PnP, SLAM

➢ 3D reconstruction: SfM, learning, RGBD

➢ Trends
3D reconstruction

- Structure from motion
  - Single image: learning depth
  - Multi images: Learning matching/ disparity

From depth camera: limited visual scope; for non-rigid objects; for large environments or an entire object by RGBD SLAM

Others
Structure from motion

- Point detection and matching
- Epipolar geometry computation
- Camera pose and structure computation
- Bundle adjustment
- Point cloud processing

- Incremental
- Global
- Hybrid

Onur Ozyesil, Vladislav Voroninski, Ronen Basri, Amit Singer
Survey on Structure from Motion
Acta Numerica, 2017
**Incremental**

1. Hainan Cui, Shuhan Shen, Xiang Gao, Zhanyi Hu. Batched Incremental Structure-from-Motion. 3DV 2017 (Speed up, selecting 3D points and verifying camera poses)
2. Joseph DeGol1, Timothy Bretl1, Derek Hoiem. Improved Structure from Motion Using Fiducial Marker Matching. ECCV 2018. (Put markers in the scene, Higher accuracy, limiting matches, changing order of images, enforcing new bundle adjustment constraints)

**Global**

3. Anders Eriksson, Carl Olsson, Fredrik Kahl, Tat-Jun Chin. Rotation Averaging and Strong Duality. CVPR 2018. (the role of duality principles within the problem of rotation averaging, no duality gap)
Translation and rotation are estimated individually


(rotation solving is transformed as a linear L1 norm problems that is robust to outliers. After rotations are solved, the parameter number decrease greatly. Then to solve only translations by incremental methods make errors decrease too.

Cameras are grouped, each group are reconstructed incrementally, then all groups are reconstructed globally

3. Siyu Zhu, Runze Zhang, Lei Zhou, Tianwei Shen, Tian Fang, Ping Tan, and Long Quan. Very Large-Scale Global SfM by Distributed Motion Averaging. CVPR 2018. (one PC, 1.2 m)
1. Runze Zhang, Siyu Zhu, Tian Fang, Long Quan. Distributed Very Large Scale Bundle Adjustment by Global Camera Consensus. 29-38, ICCV 2017. (global consensus based on ADMM a general consensus framework regardless of the number of parameters of camera )

2. Hainan Cui, Shuhan Shen, Zhanyi Hu. Tracks Selection for Robust, Efficient and Scalable Large-Scale Structure from Motion. PR 2017. (Formulate the tracks selection task as finding a subset of tracks to cover multiple spanning trees of epipolar geometry graph, decrease bundle adjustment constraint numbers and speed up bundle adjustment greatly)

Fusion from large view angles


Point cloud/3D point processing

Mesh/texture mapping; complete; fuse/match

1. Nan. et al., PolyFit: Polygonal Surface Reconstruction from Point Clouds, ICCV 2017
2. Kelly. et al., BigSUR: Large-scale Structured Urban Reconstruction, TOG 2017
3. Zhu. et al., Variational Building Modeling from Urban MVS Meshes, 3DV 2017
6. Angela Dai, Daniel Ritchie, Martin Bokeloh, Scott Reed, Jurgen Sturm, Matthias Nießner. ScanComplete: Large-Scale Scene Completion and Semantic Segmentation for 3D Scans, CVPR 2018. (completion)
7. Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, Martial Hebert. PCN: Point Completion Network, 3DV 2018, Best Honorable Mention. (completion)


11. A. Parra Bustos and T.-J. Chin. Guaranteed outlier removal for point cloud registration with correspondences. PAMI, 2018. (reduce the input to a smaller set, which can reduce outlier quickly and reliably)
Deep learning

Single image


A end-to-end DNN learns to directly infer a set of plane parameters and corresponding plane segmentation masks from a single RGB image. More than 50,000 piece-wise planar depthmaps for training and testing from ScanNet, a largescale RGBD video database is generated.
Stereo/binocular: matching, disparity

Model accuracy: repeated patterns, occlusion areas, textureless regions, reflective surfaces, weak light ----→ challenge problems

Model Generalization

Model speed

1. A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach and A. Bry. End-to-End Learning of Geometry and Context for Deep Stereo Regression. ICCV 2017. (1. employ 3-D convolutions to regularize the cost volume; 2. use a differentiable soft argmin function to regress sub-pixel disparity)

2. J. Pang, W. Sun, J. Ren, C. Yang and Q. Yan: Cascade residual learning: A two-stage convolutional neural network for stereo matching. ICCV 2017 (use color image and residual network to refine the disparity estimation)

3. Z. Jie, P. Wang, Y. Ling, B. Zhao, Y. Wei, J. Feng and W. Liu: Left-Right Comparative Recurrent Model for Stereo Matching. CVPR 2018 (introduce an soft attention mechanism accompanying recurrent learning to simultaneously check consistency and select proper regions for refinement)
Stereo

4. J. Chang and Y. Chen. Pyramid Stereo Matching Network. CVPR 2018 (1. introduce a pyramid pooling module to incorporate hierarchical context information; 2. stacked hourglass 3D-CNN to regularization)

5. Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, Jianfeng Zhang. Learning for Disparity Estimation through Feature Constancy. CVPR 2018 (use feature constancy to refine the initial disparity)


16. S. Khamis, S. Fanello, C. Rhemann, A. Kowdle, J. Valentin, S. Izadi. StereoNet: Guided Hierarchical Refinement for Real-Time Edge-Aware Depth Prediction. ECCV 2018 (60fps on an Nvidia Titan X; gain the initial disparity at a very low resolution cost volume and refine the disparity by a learned edge-aware upsampling function)


**Tuple**

Nonrigid or from RGBD:
1. Suryansh Kumar, Yuchao Dai, Hongdong Li. The 1st Winner of “Non-Rigid Structure from Motion Challenge 2017” @ CVPR-2017
2. Suryansh Kumar, Yuchao Dai, Hongdong Li. Spatial-temporal union of subspaces for multi-body non-rigid structure-from-motion. Pattern Recognition, 2017. (An unified framework to jointly segment and reconstruct multiple non-rigid objects, along both temporal direction and spatial direction)


Others:

10. Qilin Sun, Xiong Dun, Yifan Peng, Wolfgang Heidrich. Depth and Transient Imaging With Compressive SPAD Array Cameras. CVPR, 2018. (to overcome the spatial resolution limit of SPAD arrays by employing a compressive sensing camera design. And then the depths are reconstructed by a method of TVAL3.)
Development trends

• Traditional computation and learning fusion:

• **Multi sensor fusion:** camera, IMU etc.

• **Combination with hardware:** depth camera, 3D camera, event camera, light field camera...


• **Combination with applications:**
  AGV, driverless car, robotics, AR, VR

• **Brain inspired:**
  Semantic+, Matching Localization, 3D Map

  but far away to simulate human intelligence
**Traditional computation and learning fusion**


<table>
<thead>
<tr>
<th>A primal sketch</th>
<th>A 2.5D sketch</th>
<th>A 3D model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature extraction:</td>
<td>Viewer-centered three dimensional view of the environment</td>
<td>A continuous, 3-dimensional map</td>
</tr>
<tr>
<td>Points, edges, regions</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Not limited in the higher level, fusing learning into the computation of each level is a trend. Marr’s vision frame is still the main stream.

--- Yihong Wu, NLPR of IA of CAS

After thirty years, Tomaso Poggio adds one higher level beyond the computational level, that is the learning.

I am not sure that Marr would agree, but I am tempted to add learning as the very top level of understanding, above the computational level. Only then may we be able to build intelligent machines that could learn to see—and think—without the need to be programmed to do it.

Thanks
yhwu@nlpr.ia.ac.cn
http://vision.ia.ac.cn/Faculty/yhwu/index.htm