Master kickoff presentation

Natural marker extraction and learning for binocular images and 6D pose estimation

Mahdi Saleh
12.08.2016

Supervisors
Benjamin Busam
Federico Tombari

FRAMOS

Technische Universität München

+

Johns Hopkins
Whiting School of Engineering

Overview

- Problem statement
- Challenges
- Related works
- Initial solutions
- Early results
- Further plans

FRAMOS

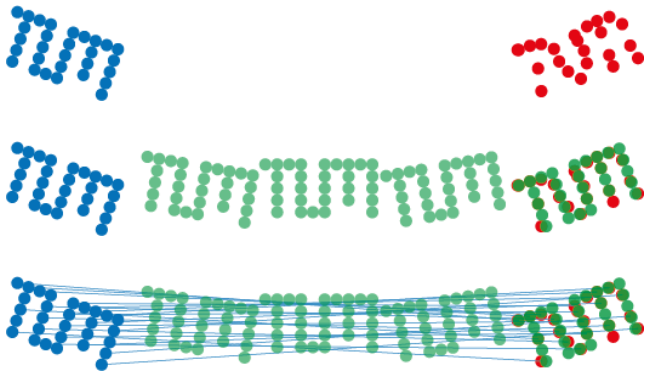
Technische Universität München

+

Johns Hopkins
Whiting School of Engineering

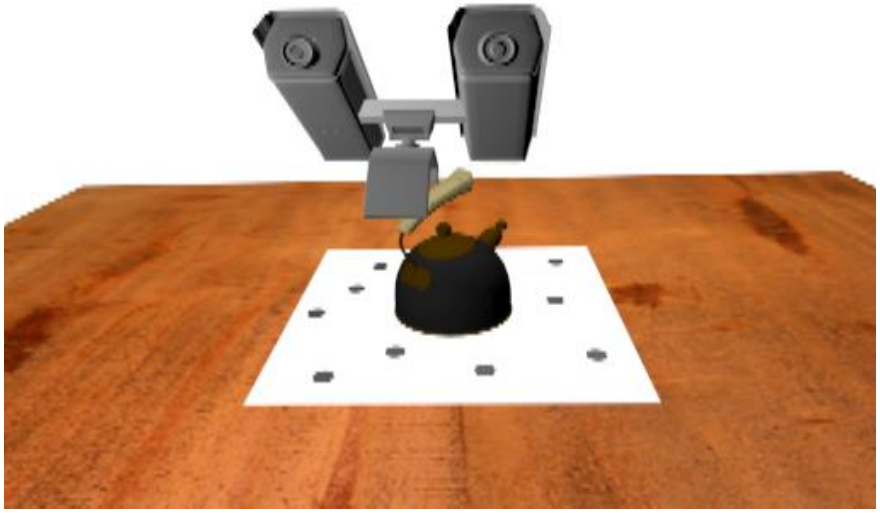
Problem statement

- Framos OTS detects markers on an object and outputs 3D Pose accurately and fast.
  - Busam et al. ***A Stereo Vision Approach for Cooperative Robotic Movement Therapy*** *ICCVW, Santiago, Chile, December 2015*
- Learn some natural markers
- Video input with known 6DOF pose > See slide 04

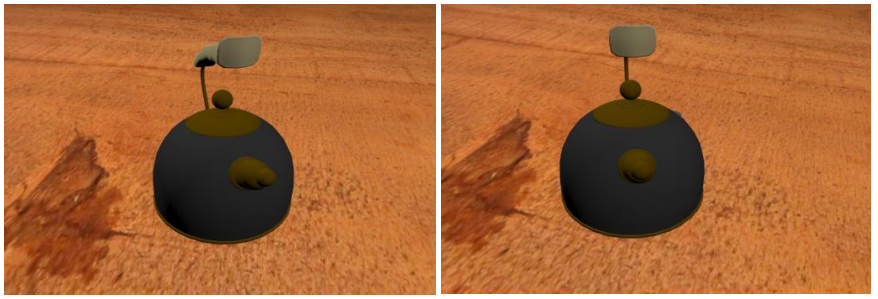


Problem statement

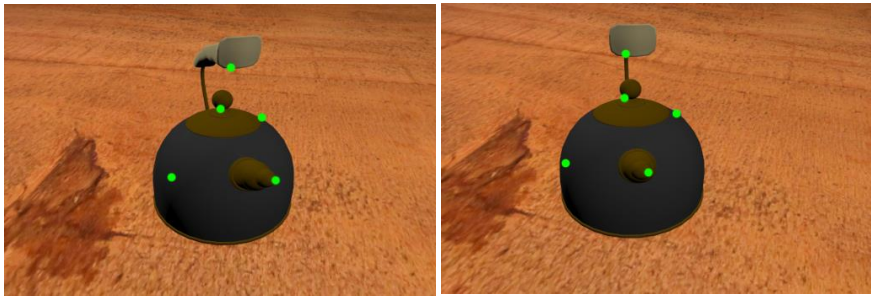
Training video



Input stereo for evaluation



What we want



Challenges

- Robust, fast calculation
- Discriminative keypoints
- Symmetry
- Highly textured to texture less
- Not direct 3D pose estimation

Image taken from hema.nl, August 2016

Image taken from living4me.de, August 2016

Image taken from pixabay.com, August 2016

Related Works

Early handcrafted features used for pose estimation

- Harris Corner Detector
  - Harris et al. **A combined corner and edge detector** . Proceedings of the 4th Alvey Vision Conference. 1988.
- SIFT
  - Lowe et al. **Distinctive image features from scale-invariant keypoints.** ICCV 2004, Springer.
- HOG
  - Dalal et al. **Histograms of Oriented Gradients for Human Detection.** *CVPR*, 2005.
- SURF
  - Bay et al. **Speeded Up Robust Features**, ECCV 2006
- BRISK
  - Leutenegger et al., **BRISK: Binary Robust Invariant Scalable Keypoints**. *ICCV* 2011.
- LINEMOD
  - Hinterstoisser et al., **Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes**. *ACCV* 2012.

Related Works

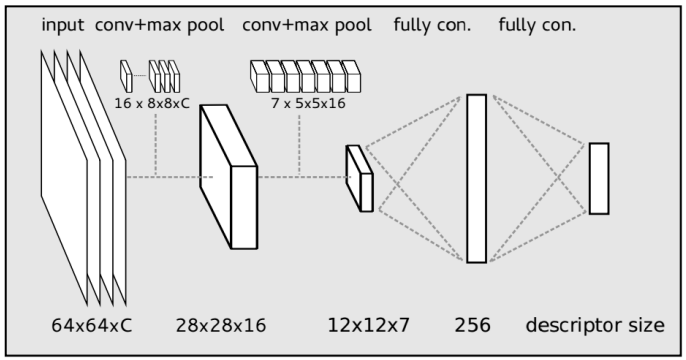
Training keypoints by SVM or boosting

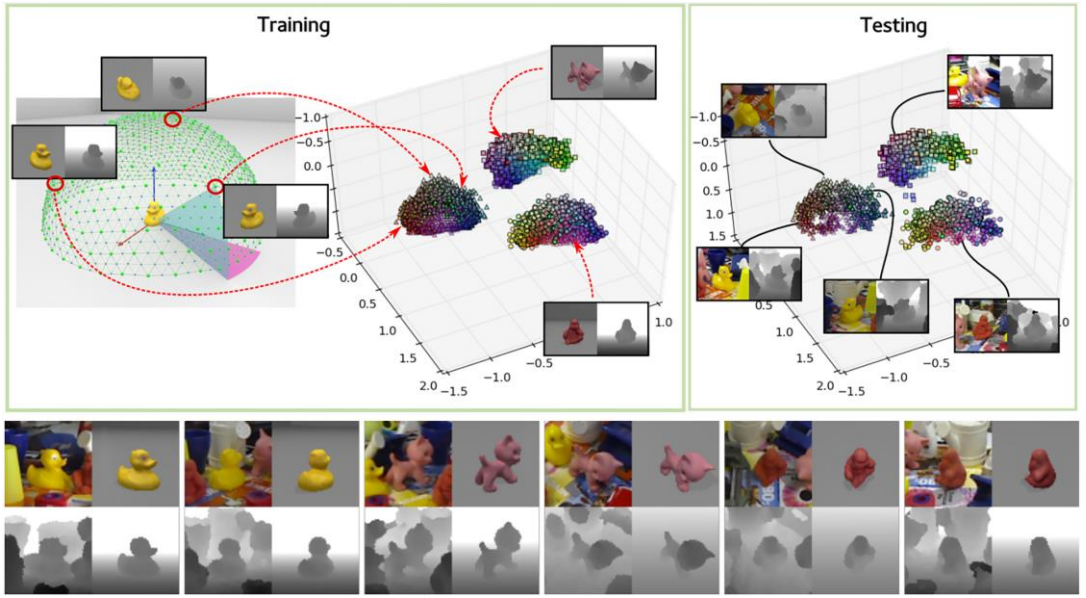
- Learning HOG models for viewport classification
  - Gu et al. **Discriminative Mixture-of-Templates for Viewpoint Classification.** In ECCV, 2010.
- Does not consider the pose estimation problem
  - Malisiewicz et al. **Ensemble of Exemplar-SVMs for Object Detection and Beyond.** ICCV, 2011.
- Exemplars were recently used for 3D object detection and pose estimation, but still rely on a handcrafted representation.
  - Aubry et al. **Seeing 3D Chairs: Exemplar Part-Based 2D-3D alignement Using a Large Dataset of CAD Models.** CVPR, 2014.

Related Works

Wohlhart et al. "**Learning descriptors for object recognition and 3d pose estimation**." *CVPR* 2015.

- Trains RGB-D objects from different poses on a CNN network
  - Finds object class and its 3D pose
  - K-nearest neighbor search in descriptor space





Related Works

Verdie, Yannick, et al. "**TILDE: A Temporally Invariant Learned DEtector**."*CVPR 2015*

- Finding keypoints under drastic changes
  - Extracts keypoints on all image
  - If the batch of images has frequent keypoints in a point, the point is a positive location
  - Uses regression to train patches
  - Tries different regressions

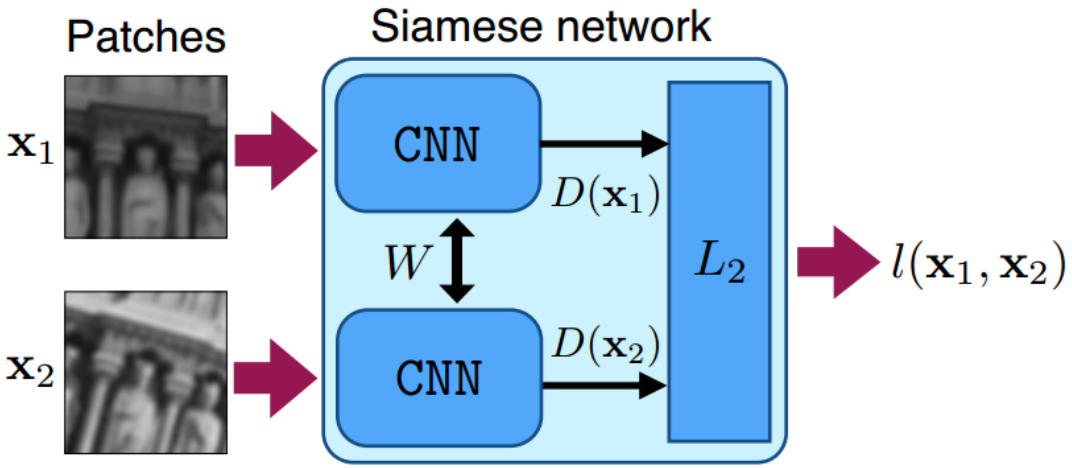
(a) Original images (b) SIFT (c) SURF (d) FAST-9 (e) Our keypoints

#keypoints	*Webcam* (2%)	*Oxford* Stand.	*Oxford* (2%)	*EF* Stand.	*EF* (2%)
TILDE-GB	33.3	54.5	32.8	43.1	16.2
TILDE-CNN	36.8	51.8	49.3	43.2	27.6
TILDE-P24	40.7	**58.7**	**59.1**	**46.3**	**33.0**
TILDE-P	**48.3**	58.1	55.9	45.1	31.6
FAST-9	26.4	53.8	47.9	39.0	28.0
SFOP	22.9	51.3	39.3	42.2	21.2
SIFER	25.7	45.1	40.1	27.4	17.6
SIFT	20.7	46.5	43.6	32.2	23.0
SURF	29.9	56.9	57.6	43.6	28.7
TaSK	14.5	25.7	15.7	22.8	10.0
WADE	27.5	44.3	51.0	25.6	28.6
MSER	22.3	51.5	35.9	38.9	23.9
LCF	30.9	55.0	40.1	41.6	23.1
EdgeFoci	30.0	54.9	47.5	46.2	31.0

Related Works

Simo-Serra et al. "**Discriminative learning of deep convolutional feature point descriptors**." *ICCV* 2015.

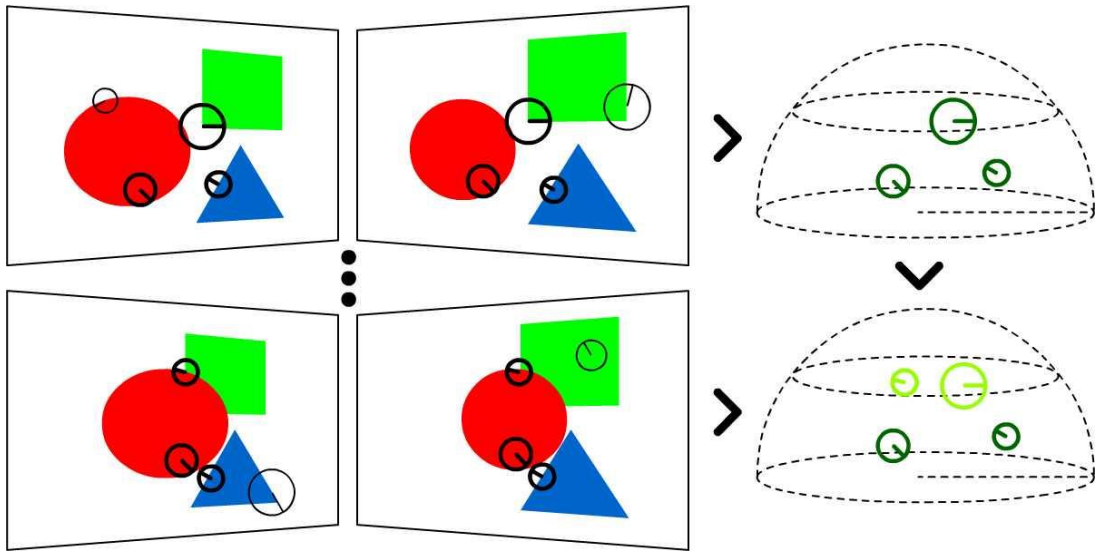
- Trains the network on sets of matching and unmatching patches
  - The CNN outputs are 128-D descriptor
  - Some tied weights between two networks
  - Easy matching using Euclidean distance



Initial Solutions

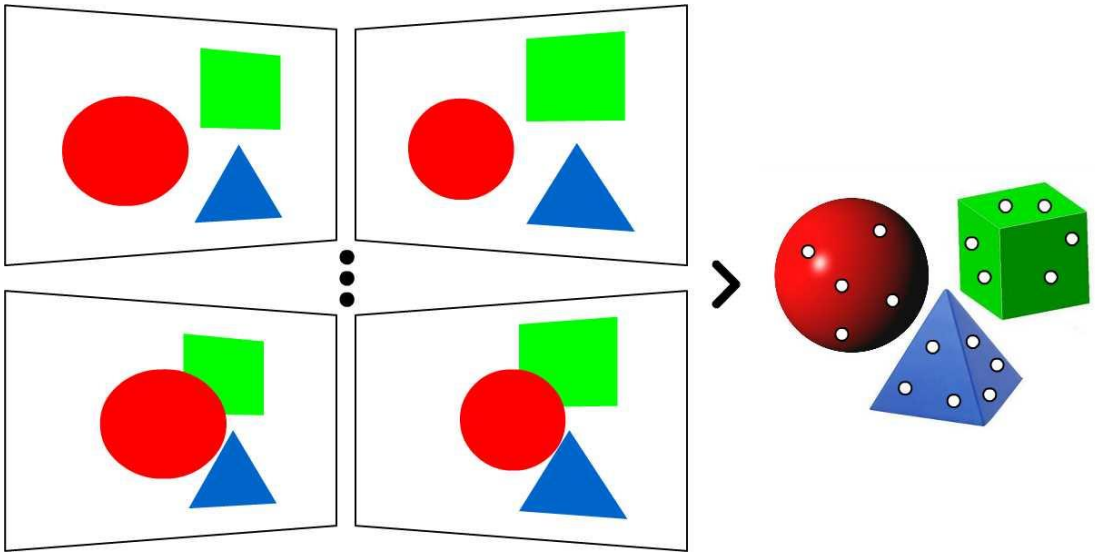
1. Train a CNN on the pairs of images with their corresponding pose and try to get the pose directly similar to "*Learning Descriptors for Object Recognition and 3D Pose Estimation*"

2. Finding keypoints/corners based on a conventional keypoint detectors. Filter and project them in 3D using stereo. train a network on best keypoints



Initial Solutions

3. Finding the rough 3D surface of the object from stereo matching and chose some points uniformly around it
   - Need for a reliable stereo matching algorithm
   - No descriptor

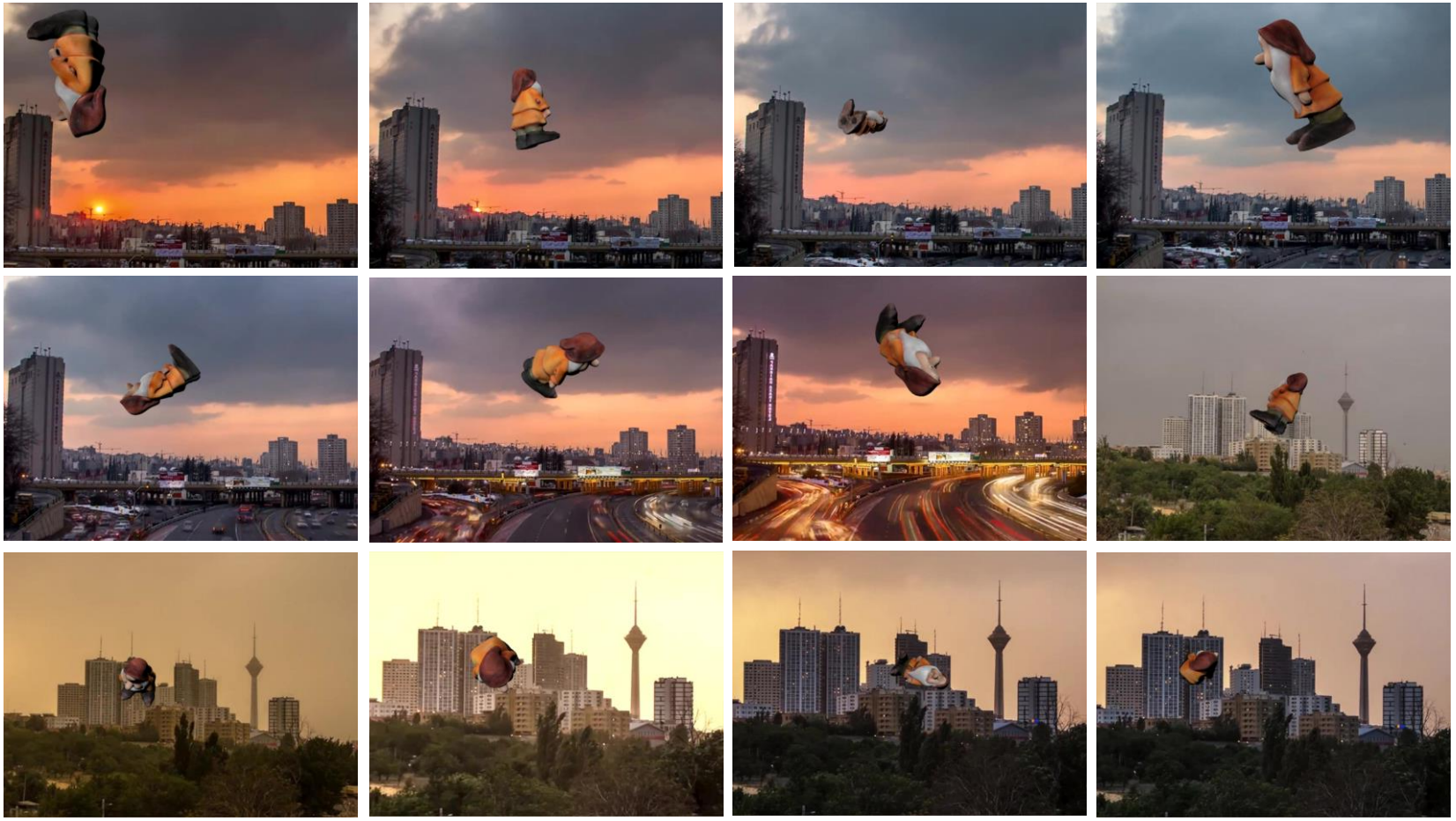


Early Results

Dataset Creation

- Initially 6 models from the website www.turbosquid.com
- The objects are different from textured to textureless
- Objects are rendered in 500 frames from a list of different poses in 640x480
- A set is also rendered with checkerboard plane for stereo camera calibration
- Maxscript used for reading input, transformation and automatic rendering
- Time-lapse video on background

Early Results



Right camera sample images from dwarf object

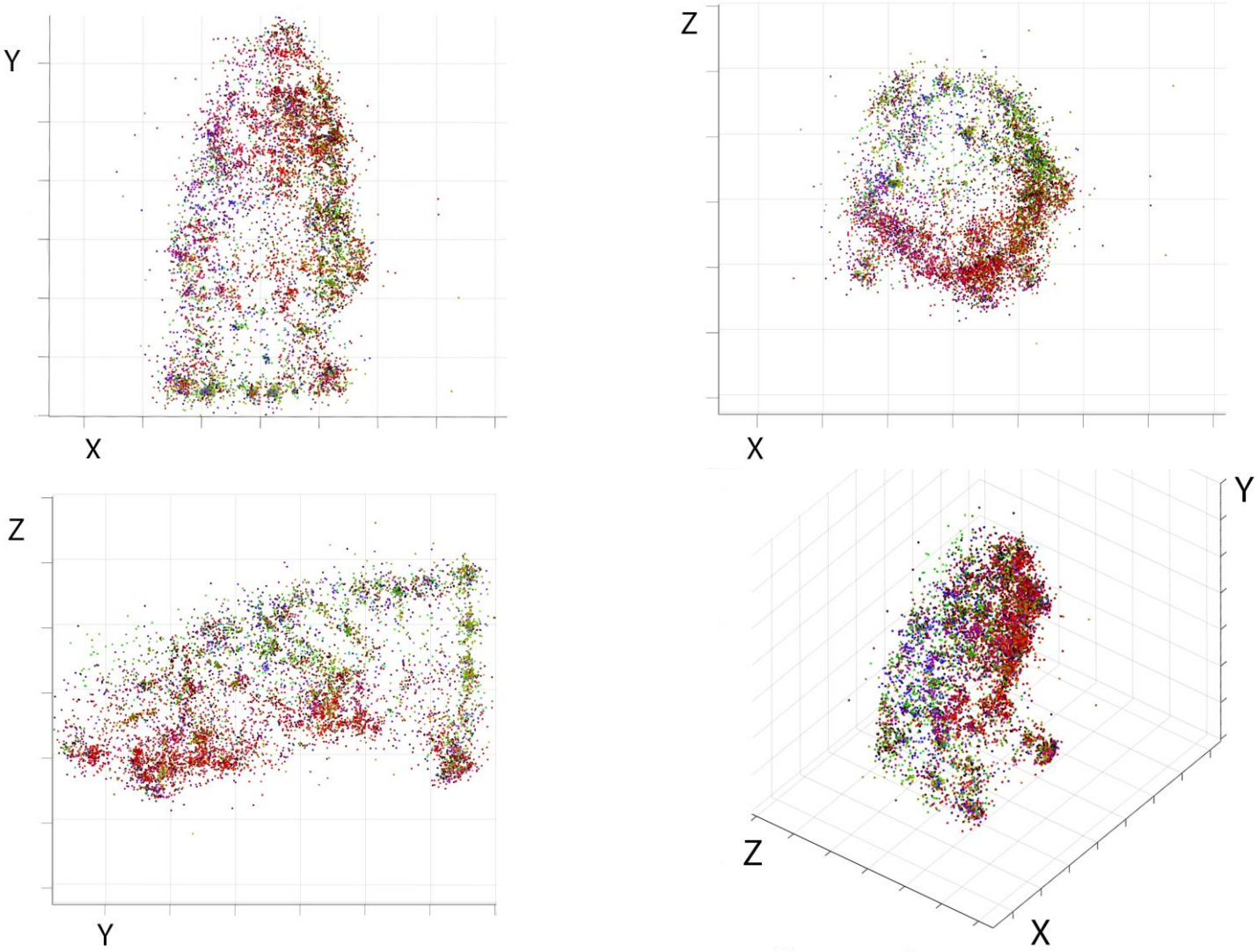
Early Results

2D Keypoint matching

- A pool of SIFT, SURF, MSER and Harris is created
- Then we do stereo triangulation and 3D points transformation

Keypoint matching samples of stereo image pairs

Early Results

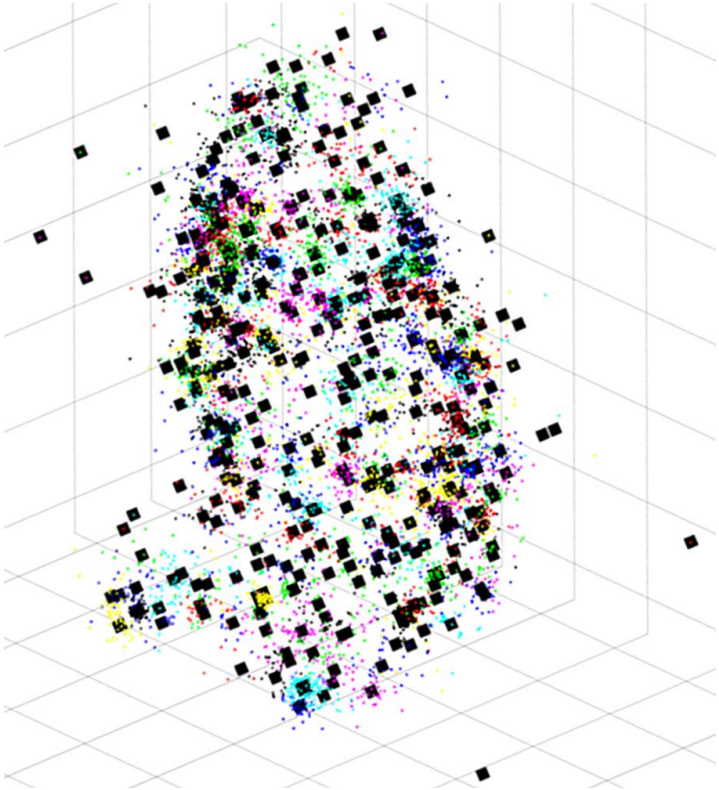


Over 8100 projection of extracted keypoints for the dwarf object

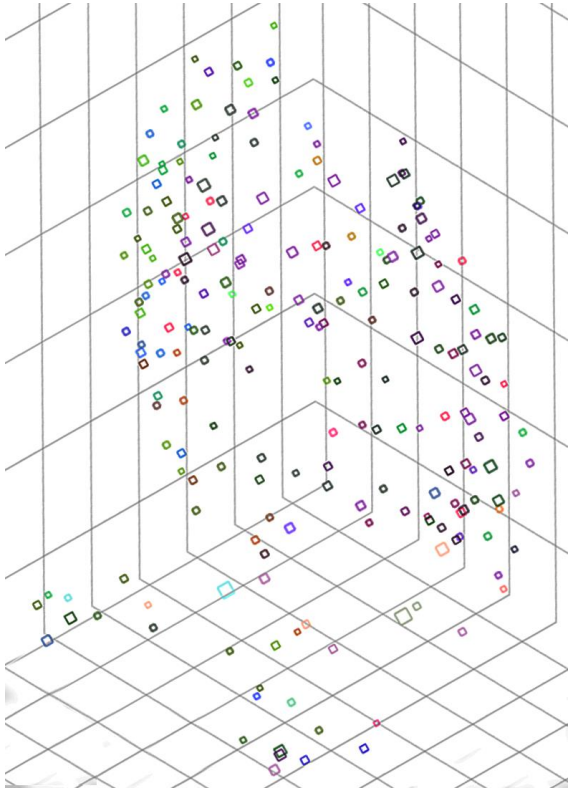
Early Results

K-means for keypoint candidates

- Up to 400 labels based on the locations
- The classes with less than a threshold (here 15) population are filtered
- Keypoints repetition and mean of quaternion pose shown with size and color



K-means output



Keypoint candidates and their pose and repetition

Early Results

Keypoint reprojection

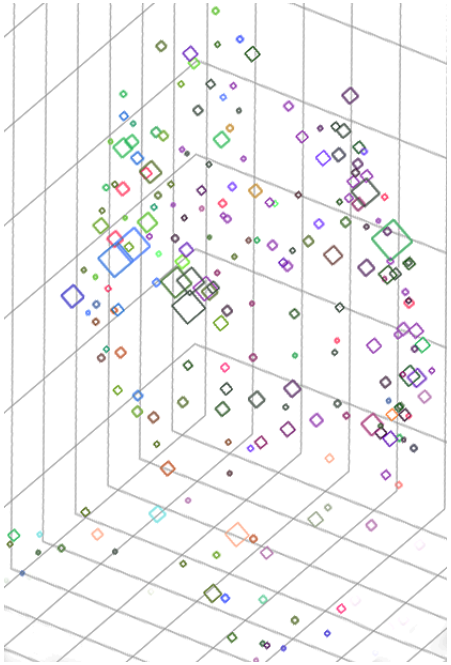
- Backprojection of all keypoints on stereo pairs
- Big uncertainty whether a keypoint is back of the object and therefore not seen
- Neighborhood, depth and poses are checked to overcome this

Keypoint reprojection samples on left images

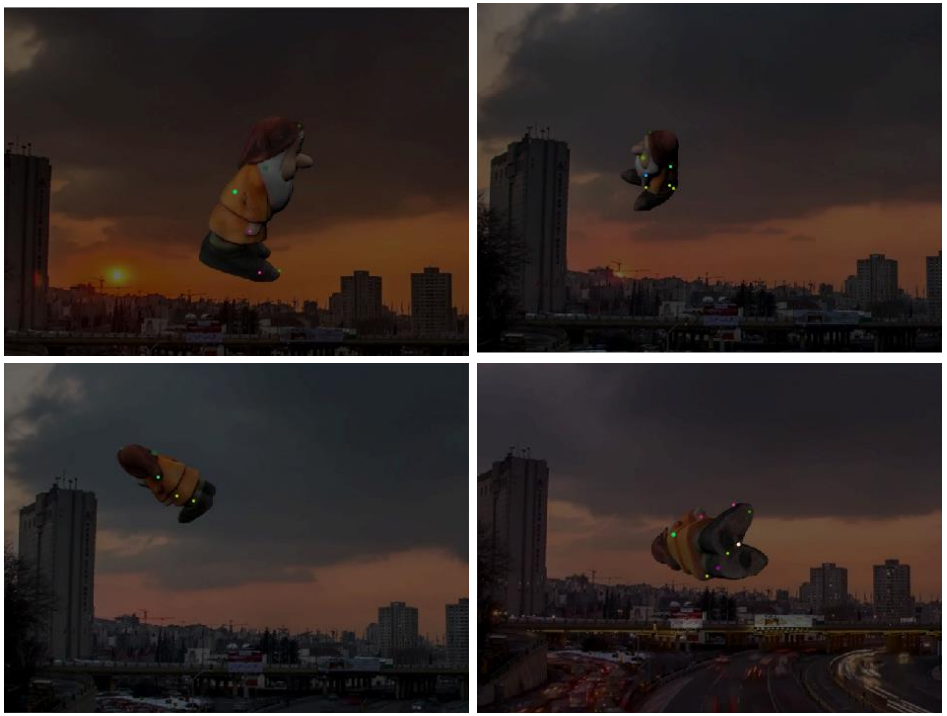
Early Results

Best keypoint selection

- Keypoint are scored based on some promoting criteria
- **Discrimination**: based on distance of the mean of descriptors of a keypoint to all the means
- **Each frame has a single vote**: The summation of discriminations for every frame is the same
- **Repetitiveness**: based on number of frames a keypoint is seen
- **Widely distributed**: in a 3D neighborhood the best rated keypoint is selected
- selected keypoints are reprojected again
- A 40x40 patch is taken from around the keypoints



Keypoint shown with scores

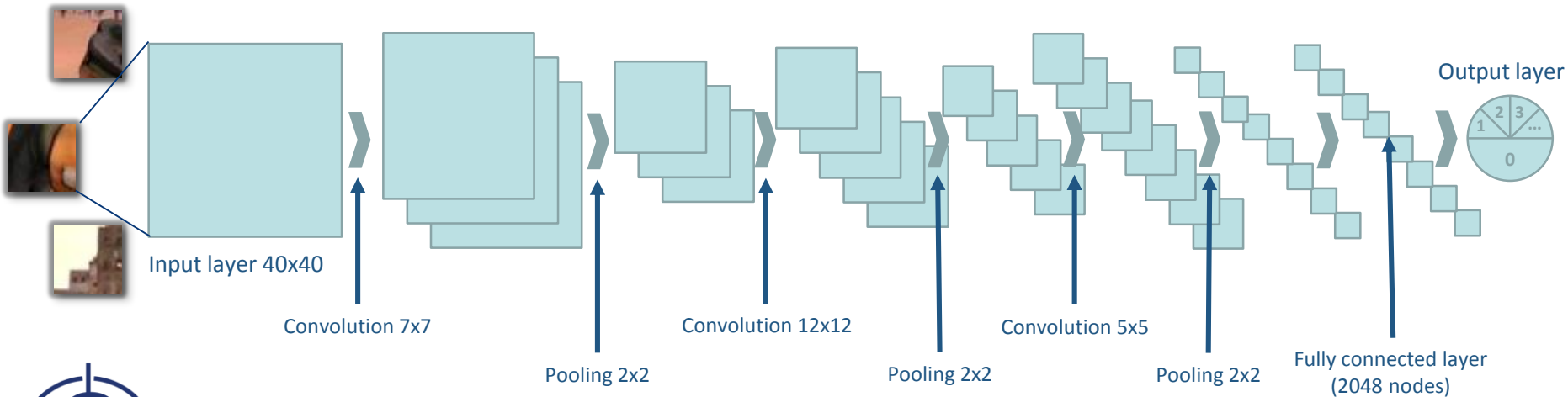


Selected Keypoint reprojection samples on left images

Early Results

Deep learning

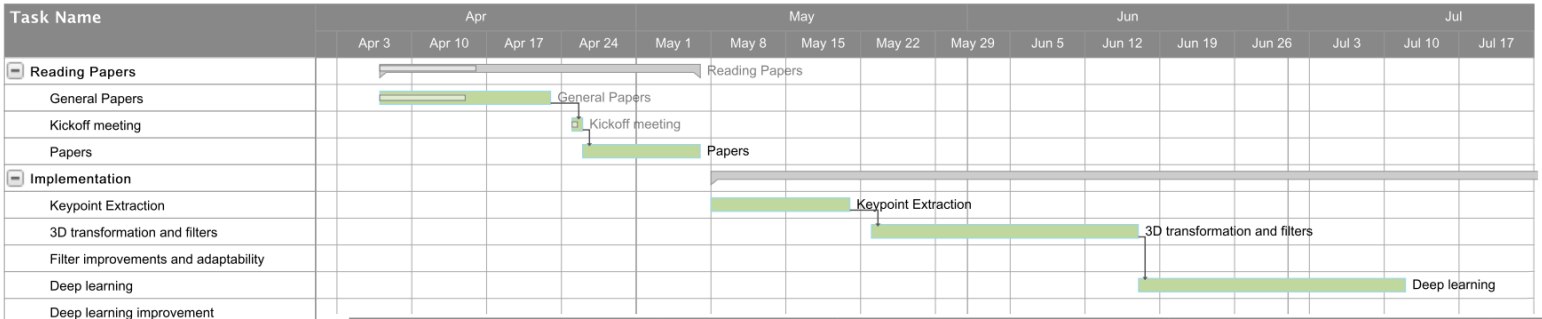
- Keypoint patches + false patches (50% object region, 50% surroundings)
- For accuracy improvement and ease of marker registration task binary classification is turned to marker classification
- Microsoft CNTK has been used for training
- 80% of dataset(~140K images) for training, 20% for test
- Data augmentation using Matlab and some using CNTK are applied. An SGD with drop out is used, with a decreasing learning rate
- Architecture inspired from TILDE paper with some changes
- Initial training error around 2% and test error around 5%

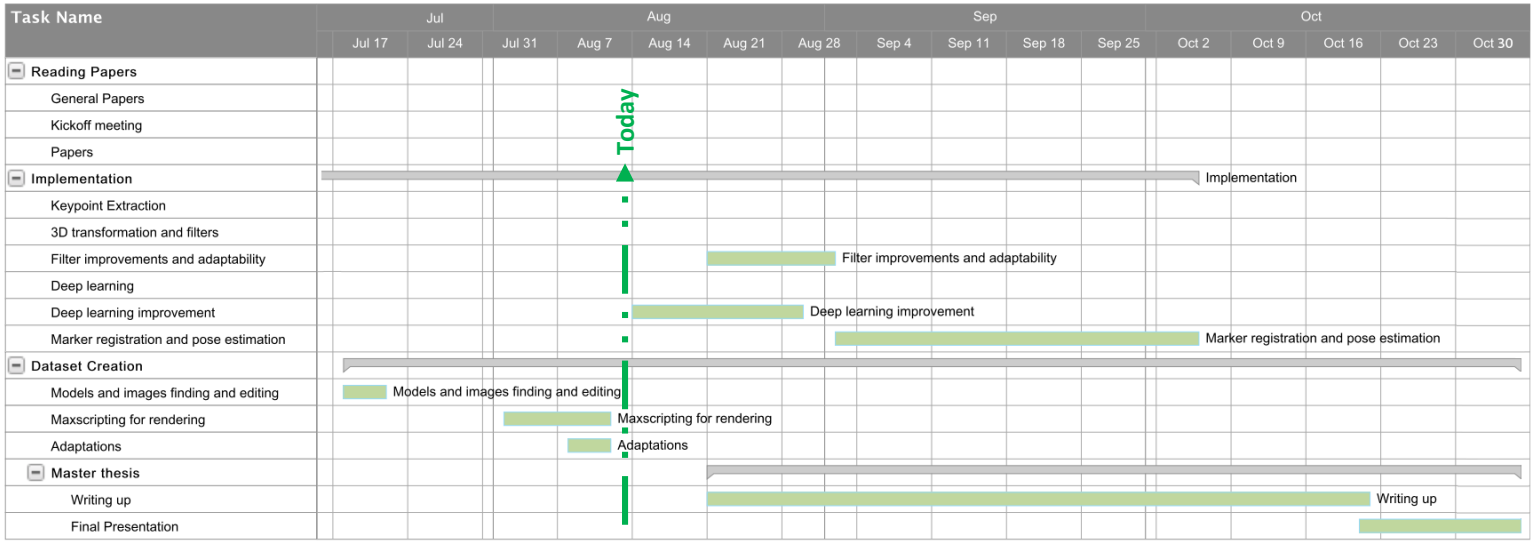


Deep network layers

Further Plans

- Marker registration using Hungarian method should be applied
- Machine learning part still needs some work
- Reprojection filters needs some adaptation for different objects
- Bug fixes
- Writing up the thesis





Gant Chart for the plans

THE END

Questions?