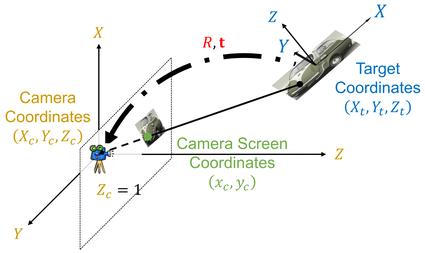
Direct 3D Pose Estimation of a Planar Target

Hung-Yu Tseng[1], Po-Chen Wu[1], Ming-Hsuan Yang[2], Shao-Yi Chien[1]
[1]National Taiwan University
[2]University of California, Merced

IEEE 2016 Winter conference on Applications of Computer Vision

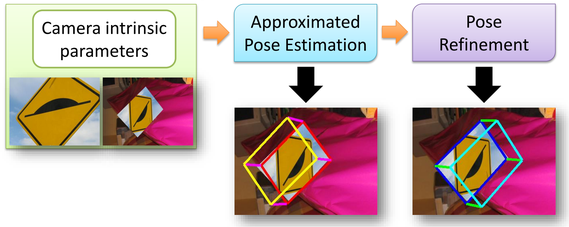
Introduction

- Robust 3D pose estimation on both texture and textureless planar target.



Estimate **R** and **t**, which is six degree of freedom.

- Two-step algorithm. Estimate the pose in the first stage, refine and disambiguate the pose in refinement stage.



Experiment I

- Compared algorithms

SIFT / ASIFT → RANSAC → OPnP

- Evaluation
  - Rotation error (degree) : $E_{\mathbf{R}} = acosd((Tr(\mathbf{R}^{\mathrm{T}} \cdot \hat{\mathbf{R}}) - 1)/2)$
  - Translation error (%) : $E_{\mathbf{t}} = \|\hat{\mathbf{t}} - \mathbf{t}\| / \|\hat{\mathbf{t}}\| \times 100$
  - Success rate (%) : SR indicates the percentage of poses that $E_{\mathbf{R}} < 20^{\circ}$ and $E_{\mathbf{t}} < 10\%$
- Real dataset from [5]
  - Compute all candidate poses with 4 corners using OPnP.
  - Select the correct pose from all candidates.

Proposed Algorithm

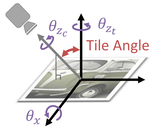
Approximated Pose Estimation

- Find the pose with minimum appearance distance.

$$E_a(\mathbf{p}) = \frac{1}{n_t}\sum_{i=1}^{n_t} |I_c(\mathbf{u}_i) - I_t(\mathbf{x}_i)| \quad (1) \qquad \mathbf{x} = (X_t, Y_t),\ \mathbf{u} = (x_c, y_c)$$

I_c: Camera image, I_t: Target image, n_t: Number of sample points

- Theorem
  - It is shown in [1] that (Mean variation of I_t)
    $|E_a(\mathbf{p}_1) - E_a(\mathbf{p}_2)| = O(\varepsilon \bar{V})$, if (2)
    $\forall \mathbf{x}_i \in I_t : d\left(T_{\mathbf{p}_1}(\mathbf{x}_i) - T_{\mathbf{p}_2}(\mathbf{x}_i)\right) = O(\varepsilon)$. (3)



- ε-Covering Set Construction
  - Factorize the rotation matrix $\mathbf{R}$ as $\mathbf{R}_{\mathbf{z}}(\theta_{z_c})\mathbf{R}_{\mathbf{x}}(\theta_x)\mathbf{R}_{\mathbf{z}}(\theta_{z_t})$.
  - Any two consecutive poses $\mathbf{p}_k$ and $\mathbf{p}_k + \Delta\mathbf{p}_k$ on each dimension in the set must satisfy (3).
  - Minimum appearance distance $E_a(\mathbf{p}_b) = O(\varepsilon)$.
- Coase-to-fine search Estimation
  1. Construct pose set S with coarser ε.
  2. Evaluating E_a of each pose and obtain best pose $\mathbf{p}_b$ associated with $E_a(\mathbf{p}_b)$.
  3. Select the poses within a threshold.

     $S_L = \{\mathbf{p}_L \mid E_a(\mathbf{p}_L) < E_a(\mathbf{p}_b) + L\}$ (4)
  4. Expand S_L with finer ε'.

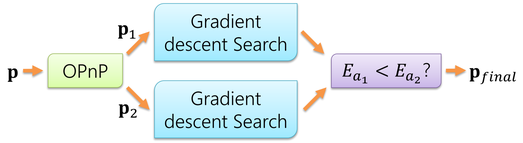
     $S' = \{\mathbf{p}' \mid \exists \mathbf{p}_L \in S_L : (3) \text{ holds for } \mathbf{p}', \mathbf{p}_L, \varepsilon'\}$ (5)
  5. Repeat 2, 3 and 4 until reaching wanted ε^*.

Pose Refinement

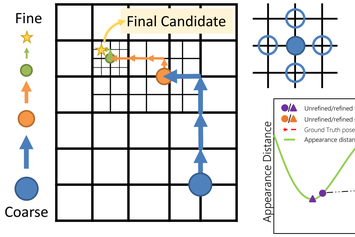
- Solve pose ambiguity problem and further refine the estimated pose.

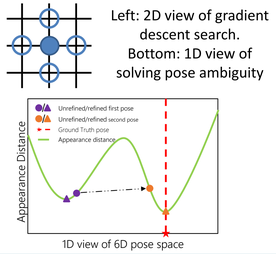
Pose Ambiguity : Different poses with similar appearance

- Refinement with two candidate poses.



1. Apply OPnP[2] to compute all stationary points, select two points with smallest error as the candidate poses.
2. Use gradient descent search to further refine each poses.
3. Choose the one with smaller E_a to be the final pose.



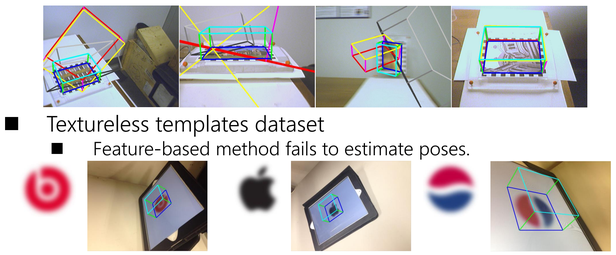


Left: 2D view of gradient descent search.
Bottom: 1D view of solving pose ambiguity

- Real dataset from [3]
  - 6 templates, each contains 16 video sequences.

video	Unconstraint			Perspective Distortion			Motion 9			Dynamic Lighting		
	$E_{\mathbf{R}}(^{\circ})$	$E_{\mathbf{t}}(\%)$	SR(%)	$E_{\mathbf{R}}(^{\circ})$	$E_{\mathbf{t}}(\%)$	SR(%)	$E_{\mathbf{R}}(^{\circ})$	$E_{\mathbf{t}}(\%)$	SR(%)	$E_{\mathbf{R}}(^{\circ})$	$E_{\mathbf{t}}(\%)$	SR(%)
SIFT	89.4	167	26.6	69.0	134	41	115	107	7.33	105	136	16.3
ASIFT	60.2	**17.9**	46.9	39.3	**12.1**	70.7	95.7	41.6	6.14	38.5	**13.1**	63.8
Direct	**29.1**	18.7	**65.7**	**27.7**	23.3	**77**	**50.4**	**6.58**	**26.7**	**27.5**	22.7	**71.7**

- Textureless templates dataset
  - Feature-based method fails to estimate poses.

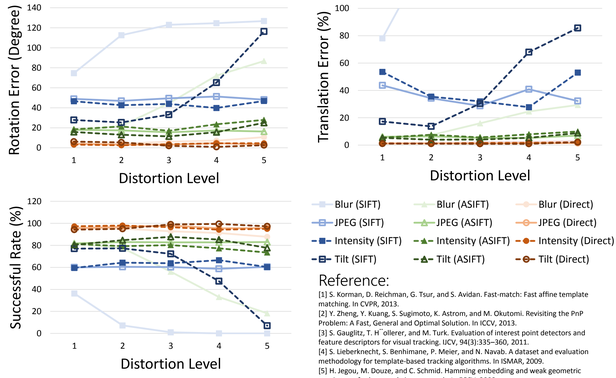


Experiment II

- Synthetic dataset
  - 8 template images from [4], 100 background images from [5].
  - Totally 8400 test images, each image is generated according to the random generated pose.
  - 4 varying conditions : blur, JPEG, intensity, tilt angle.

	Bump Sign			Stop Sign			Lucent			MacMini Board		
	$E_{\mathbf{R}}(^{\circ})$	$E_{\mathbf{t}}(\%)$	SR(%)	$E_{\mathbf{R}}(^{\circ})$	$E_{\mathbf{t}}(\%)$	SR(%)	$E_{\mathbf{R}}(^{\circ})$	$E_{\mathbf{t}}(\%)$	SR(%)	$E_{\mathbf{R}}(^{\circ})$	$E_{\mathbf{t}}(\%)$	SR(%)
SIFT	100	54.8	10	69.2	35.3	38	30.1	16.5	72	28.0	13.4	78
ASIFT	72.1	24.3	22	5.07	**0.74**	96	**1.90**	**0.38**	**100**	6.37	2.59	**96**
Direct	**3.84**	**0.80**	**96**	**2.81**	0.91	**98**	2.62	1.06	98	**5.37**	**2.56**	**96**

	Isetta			Philadelphia			Grass			Wall		
	$E_{\mathbf{R}}(^{\circ})$	$E_{\mathbf{t}}(\%)$	SR(%)	$E_{\mathbf{R}}(^{\circ})$	$E_{\mathbf{t}}(\%)$	SR(%)	$E_{\mathbf{R}}(^{\circ})$	$E_{\mathbf{t}}(\%)$	SR(%)	$E_{\mathbf{R}}(^{\circ})$	$E_{\mathbf{t}}(\%)$	SR(%)
SIFT	20.6	16.2	82	13.5	4.57	90	97.3	212	24	17.1	27.3	86
ASIFT	2.08	**0.49**	98	**1.16**	**0.35**	**100**	51.2	16.7	52	**2.76**	**0.36**	**96**
Direct	**1.49**	0.87	**100**	1.95	0.87	**100**	**2.87**	**1.06**	**96**	6.68	1.64	94



Reference:
[1] S. Korman, D. Reichman, G. Tsur, and S. Avidan. Fast-match: Fast affine template matching. In CVPR, 2013.
[2] Y. Zheng, Y. Kuang, S. Sugimoto, K. Astrom, and M. Okutomi. Revisiting the PnP Problem: A Fast, General and Optimal Solution. In ICCV, 2013.
[3] S. Gauglitz, T. H¨ollerer, and M. Turk. Evaluation of interest point detectors and feature descriptors for visual tracking. IJCV, 94(3):335–360, 2011.
[4] S. Lieberknecht, S. Benhimane, P. Meier, and N. Navab. A dataset and evaluation methodology for template-based tracking algorithms. In ISMAR, 2009.
[5] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In ECCV, 2008.

Conclusion

The proposed direct pose estimation performs favorably in terms of accuracy and robustness against state-of-the-art feature-based approaches on both synthetic and real dataset.