DescriptorEnsemble: An Unsupervised Approach to Image Matching and Alignment with Multiple Descriptors

林彥宇 副研究員
Yen-Yu Lin, Associate Research Fellow
中央研究院 資訊科技創新研究中心
Research Center for IT Innovation, Academia Sinica

About myself

- Yen-Yu Lin, *Ph.D., Associate research fellow, CITI, Academia Sinica*
  - Education

    Ph.D. in CSIE, National Taiwan University, 2005-2010

    M.S. in CSIE, National Taiwan University, 2001-2003

    B.B.A. in IM, National Taiwan University, 1997-2001
  - Work experience

    Assistant research fellow, 2011-2015

    Research assistant , IIS, Academia Sinica, 2004-2010
- Research interests
  - Computer vision

    Object recognition, image segmentation, action recognition, people counting, …
  - Machine learning

    Kernel methods, boosting, transfer learning, graphical model, deep learning, …

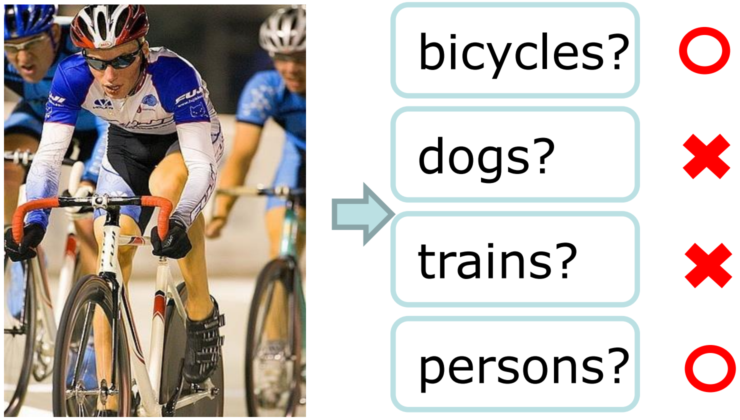
Computer vision and machine learning

- Research interests:
  - Computer Vision (CV):

    Let computers see, recognize, and interpret the world like humans
  - Machine Learning (ML):

    A statistical way to learn how human visual system works
  - Goal: *Design **ML** methods to facilitate **CV** applications*

Research topics 1/3

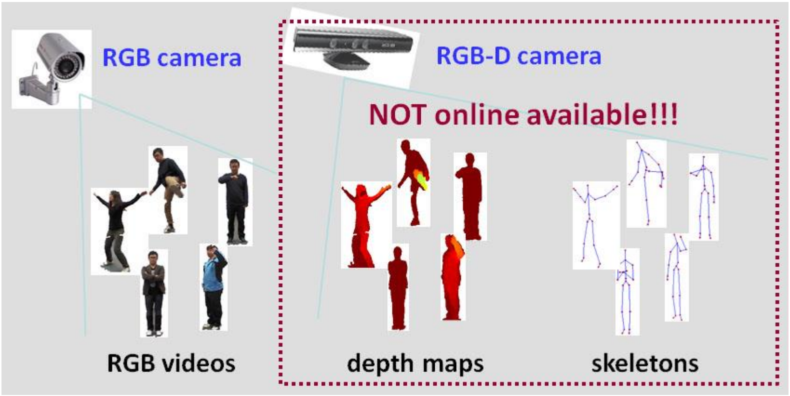


CV: *object recognition*
ML: *multiple kernel learning*
TPAMI'11, ICCV'09, NIPS'08

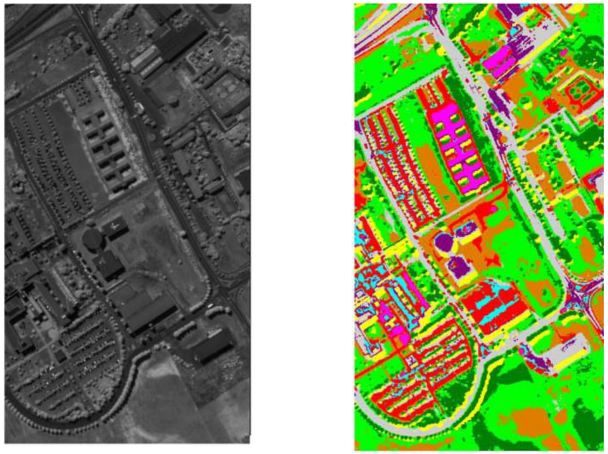
CV: *image segmentation*
ML: *graphical model*
CVPR'14, TIP'14, ACCV'12

CV: *face detection* **ML**: *vector-valued boosting*
US Patent'07, CVPR'05, ECCV'04

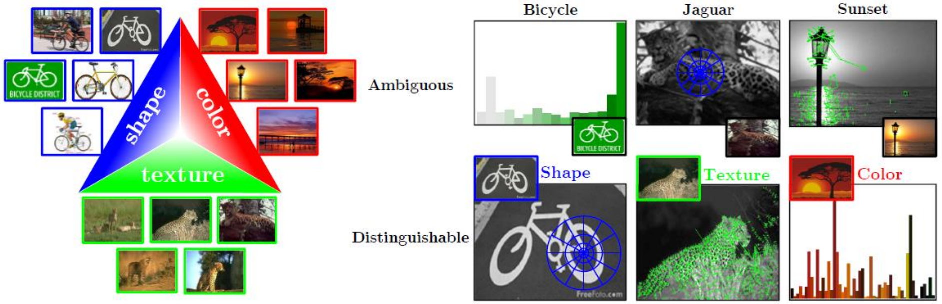
Research topics 2/3



CV: *action recognition*
ML: *low-rank reconstruction*
TIP'15, CVPR'14, ICASSP'14

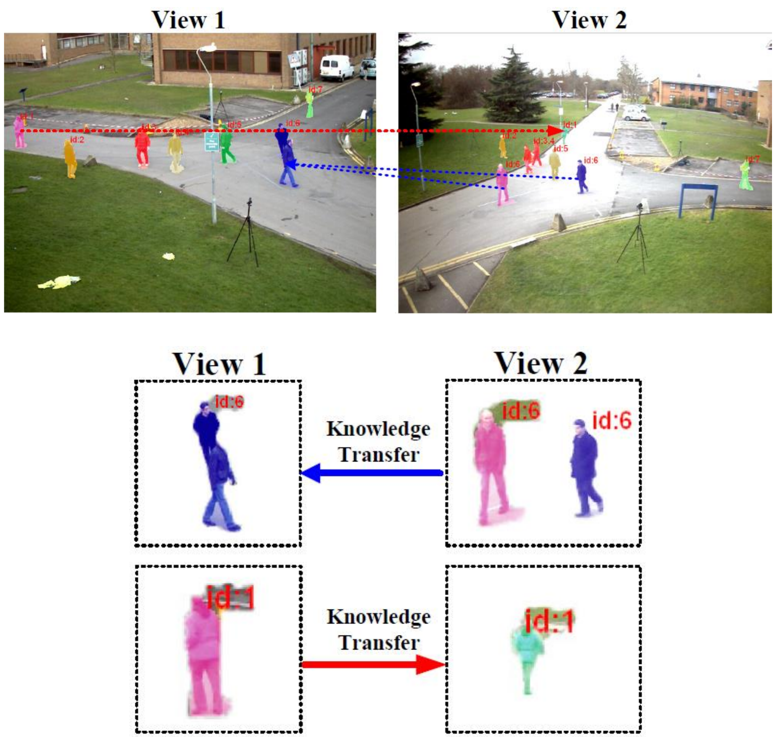


CV: *hyperspectral classification*
ML: *multiple kernel learning*
TGRS'15



CV: *image clustering* **ML**: *self-organizing map*
TMM'14, ICPR'12, ECCV'10

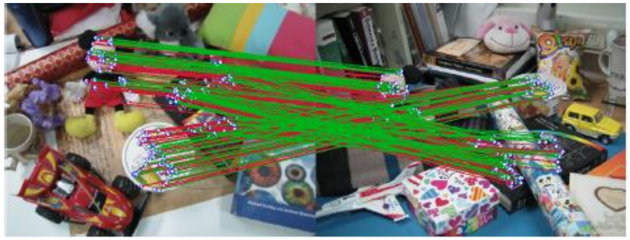
Research topics 3/3



CV: *multi-view people counting*
ML: *transfer learning*
TIP'15, ACM MM'12, WIFS'11

Input images

Hough voting: 207/222



Inverted Hough voting: 337/369

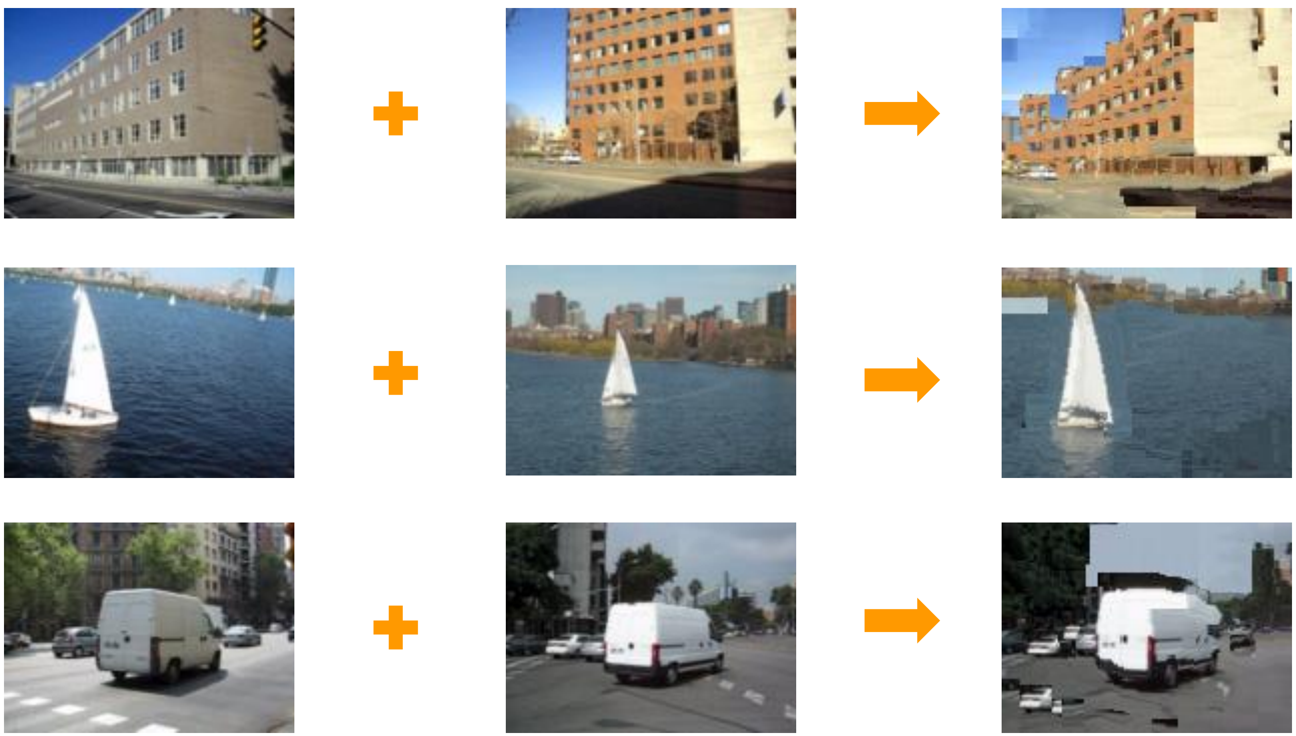
CV: *image matching*
ML: *kernel density estimation*
TIP'16, TPAMI'15, CVPR'15, CVPR'13

Image matching

- Seek common objects and regions between images
- Salient point detection + Description + Matching

Image alignment

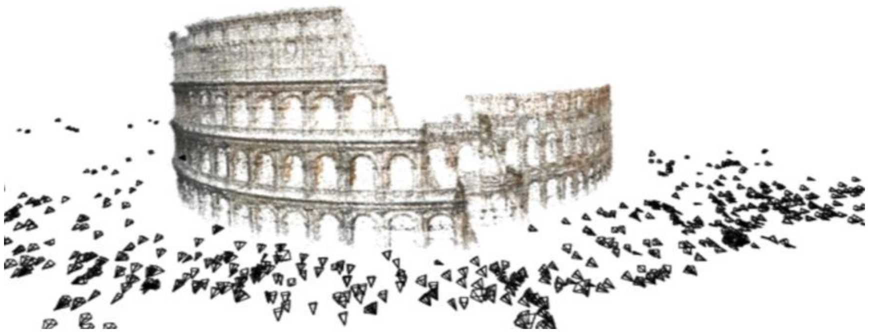
- Image matching: sparse matching on interest points
- Image alignment: ***dense*** matching on pixels



Liu et al. 2011

Fundamental research topics

3D Reconstruction

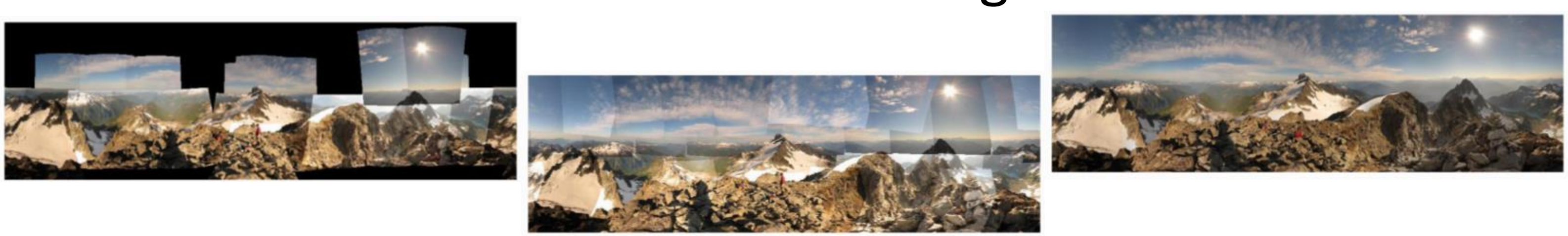


Agarwal et al. 2009

Image Retrieval

Zheng et al. 2015

Panorama Stitching

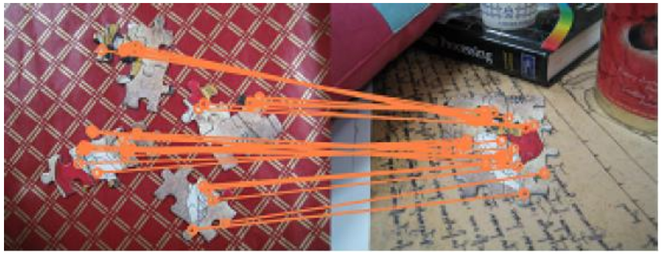


Brown and Lowe 2007

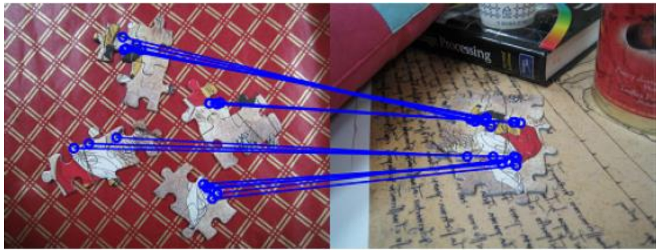
Motivation

- Three different descriptors on two image pairs

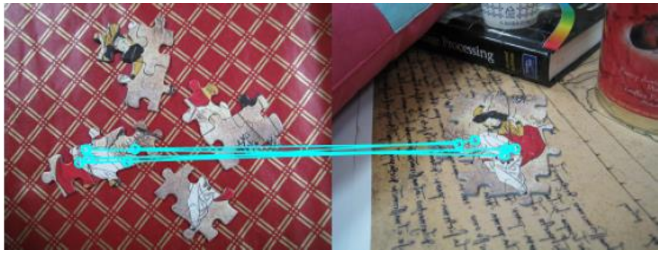
SIFT

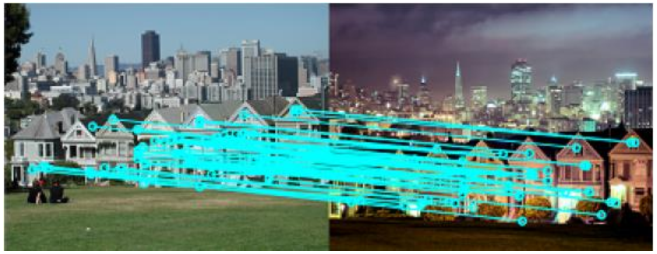


LIOP



GB





Motivation

2/2

- The optimal descriptor for matching varies from image to image
  - The effectiveness of a descriptor is image dependent

- **Descriptor fusion** may be a feasible way for improving the performance of both sparse and dense matching

- Two difficulties arise
  - Descriptor variations: different dimensions, scales, distances, …

    A common domain for fusing descriptors
  - Unsupervised nature of image matching/alignment

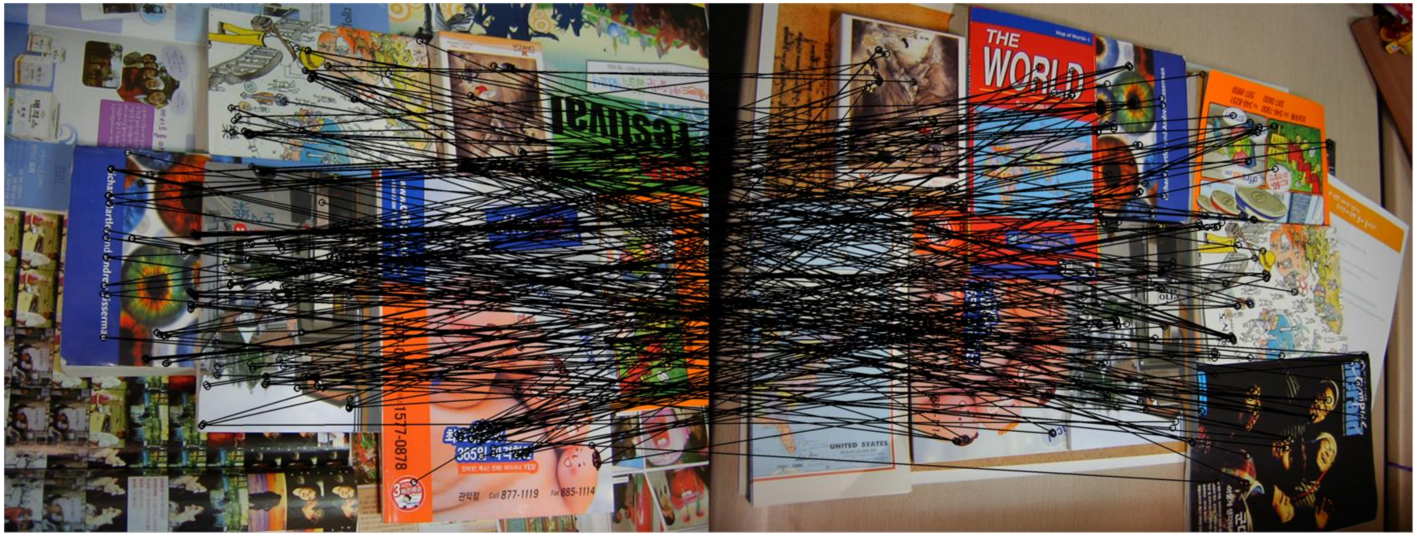
    A measure for estimating the goodness of descriptors

Observation

- A pair of images with the detected points

Observation

- A pair of images with the detected points
  - Wrong matches



Observation

- A pair of images with the detected points
  - Correct matches
  - High **geometric** and **spatial** consistence

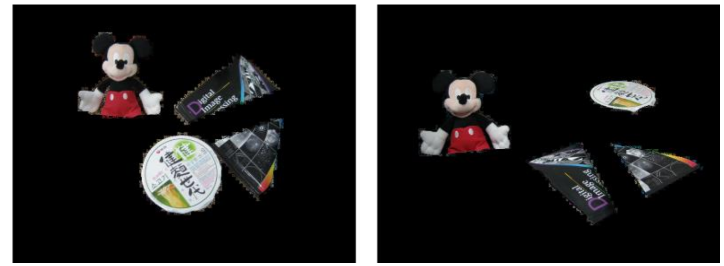
DescriptorEnsemble for Unsupervised descriptor fusion

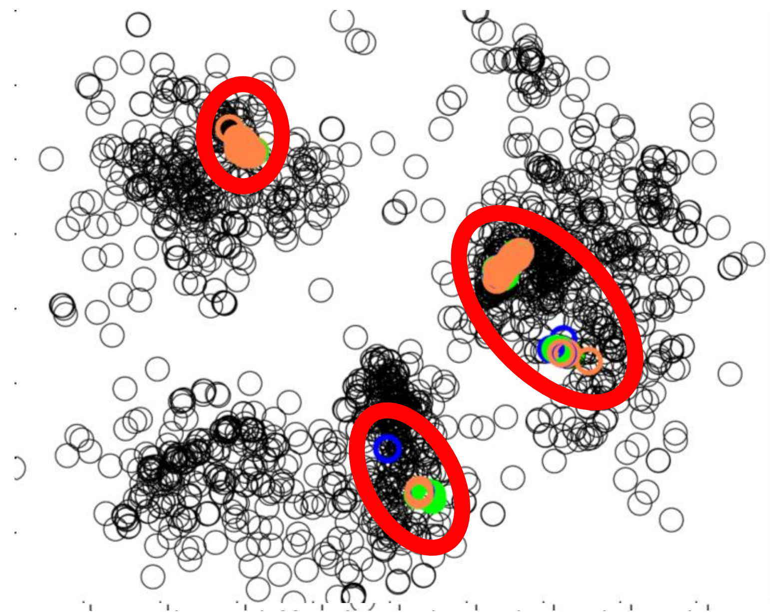
GB

DAISY

LIOP

SIFT





Affine Transformation Space

DescriptorEnsemble for image matching

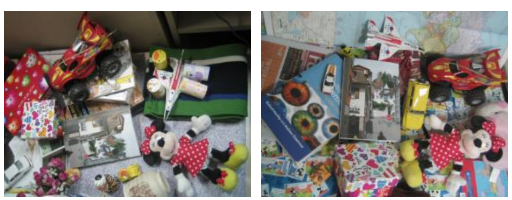
- Collect a set of plausible match candidates
  - For each point and each descriptor, find its k NNs in the opposite image
  - Total $N \times k \times M$ match candidates
- Construct a graph
  - Create a node for each match candidate
  - Add an edge between two neighboring candidates (**spatial** consistence)
  - Set the edge weight as the re-projection error (**geometric** consistence)
- Compute the **geodesic distance** between each candidate pair
- **One-class SVM** for candidate ranking
  - Positive data gather together (Correct matches have high density)
  - Negative data are negative in their own ways (Wrong matches irregularly distribute)

Experiments: Dataset

- Three datasets:
  - Co-recognition Dataset (**Co-reg**)
    Cho et al. 2009
  - Symmetric Dataset (**SYM**)
    Hauagge and Snavely 2012
  - Oxford VGG Dataset (**VGG**)
    Mikolajczyk and Schmid 2004

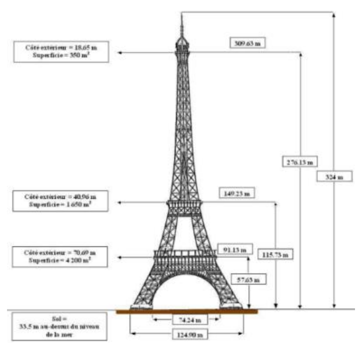
Experiments: Dataset

- Co-reg dataset:
  - 6 image pairs
  - Multiple common objects with cluttered backgrounds



Experiments: Dataset

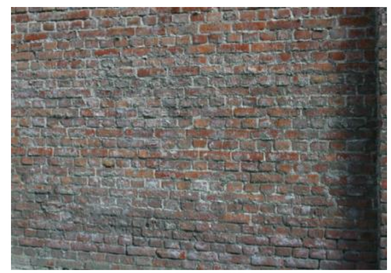
- SYM dataset:
  - 43 image pairs
  - Various variations: ages, rendering styles, day/night

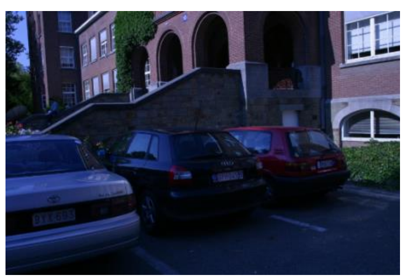


Experiments: Dataset

- VGG dataset:
  - 8 image pairs
  - Each pair contains only one or two transformation

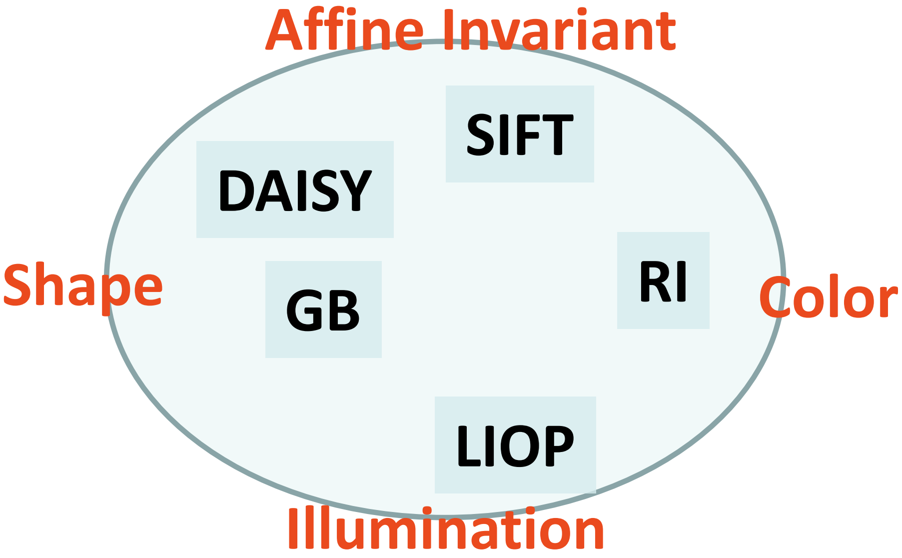




Experiments: Descriptor

- **SIFT** [Lowe 2004]
- **LIOP** [Wang et al. 2011]
- **DAISY** [Tola et al. 2010]
- Raw Intensities(**RI**)
- Geometric Blur(**GB**) [Berg and Malik 2001]



Experiments: Competing methods

- Image matching baselines
  - ACC [Cho et al. 2009]
  - HV [Chen et al. 2013]
  - SM [Leordeanu and Hebert 2005]
  - VFC [Ma et al. 2014]
- Fusion baselines
  - Concatenation (Conca.)
  - Conca. + HV
  - Ranking
  - Ratio

Experiments: Quantitative results

- Performance in mAP (mean of Average Precision)

method	descriptor	dataset		
		Co-reg	SYM	VGG
SM	SIFT	55.30	18.92	70.73
	LIOP	38.74	16.79	87.71
	DAISY	46.71	22.72	78.62
	RI	34.57	7.99	59.55
	GB	12.13	32.57	62.96
	Average	37.49	19.80	71.92
ACC	SIFT	*60.28**	26.74	82.53
	LIOP	29.83	19.49	91.76
	DAISY	36.49	29.97	86.33
	RI	15.10	10.70	67.98
	GB	8.88	29.28	64.36
	Average	30.12	23.57	78.59
HV	SIFT	60.12	22.21	82.31
	LIOP	43.97	18.69	91.78
	DAISY	50.06	26.92	88.14
	RI	37.14	11.88	70.33
	GB	12.08	*38.28**	71.84
	Average	40.67	23.60	80.88
VFC	SIFT	31.11	20.76	82.24
	LIOP	11.79	14.12	*93.51**
	DAISY	16.29	21.41	85.03
	RI	4.51	4.63	47.13
	GB	1.77	28.82	64.01
	Average	13.09	17.95	74.38

method	descriptor	dataset		
		Co-reg	SYM	VGG
CAT	All	19.62	7.90	61.27
CAT+HV	All	39.70	13.36	75.13
Ranking	All	48.48	27.99	85.68
Ratio	All	53.61	28.75	90.47
Ours	All	**79.59**	**46.83**	**93.81**

Summary

- ***DescriptorEnsemble*** for image matching: A general approach to matching images with multiple, complementary descriptors
- It works with any elliptical interest region detectors
  - Hessian-affine, Harris-affine, MSER, or their combinations
- It makes no assumption about the adopted descriptors, and fuses them in the domain of affine transformations
  - SIFT, geometric blur, LIOP, DAYSI, self-similarity descriptor, …
- Off-the-shelf solvers
  - Dijkstra’s algorithm for computing geometric distances
  - LibSVM for one-class SVM

DescriptorEnsemble for image alignment

- An approach that leverages **multiple complementary descriptors** and **matching-guided neighborhoods** to enhance descriptor-based image alignment
  - Multiple descriptors: the optimal descriptor typically varies from image to image, or even pixel to pixel
  - Matching guided neighborhood: object boundaries better reveal during the process of alignment

Challenges

(a) different scales and rotations

(b) different lighting

(c) parital occulsion

(d) complex objects and clutter background

Observation 1

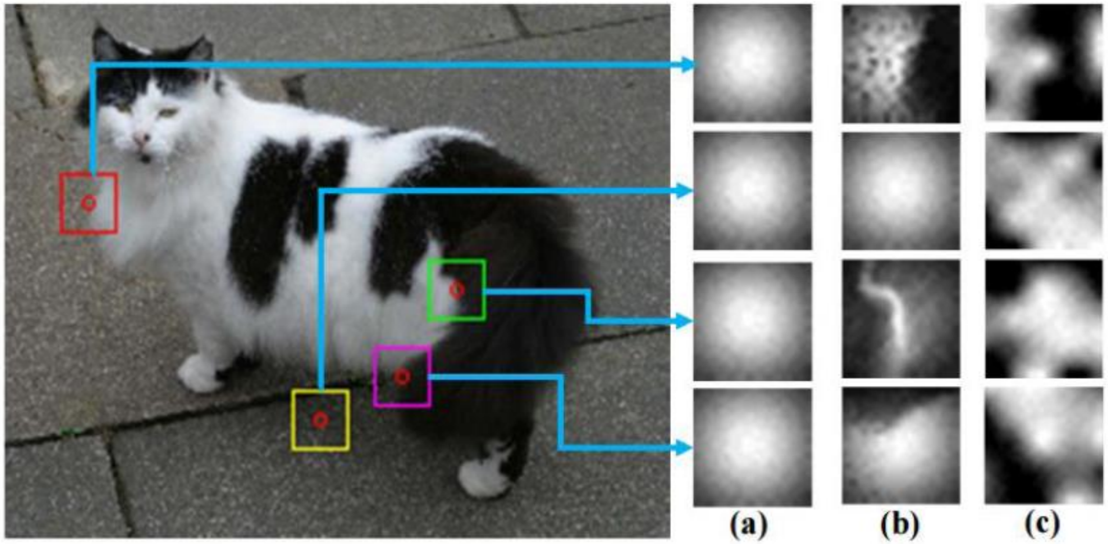
- The optimal descriptor typically varies from image to image, or even pixel to pixel

(a) I_1 (b) I_2 (c) SIFT (d) GB [1] (e) LIOP [2]

[1] Berg and Malik, "Geometric blur for template matching," *CVPR*, 2001
[2] Wang, et al., "Local intensity order pattern for feature description," *ICCV*, 2011

Observation 2

- Neighborhood for smoothness enforcement should be consistent with object boundaries
- Object boundaries can be better discovered during the process of alignment



(a) Gaussian weight. (b) Bilateral weight. (c) Ours.

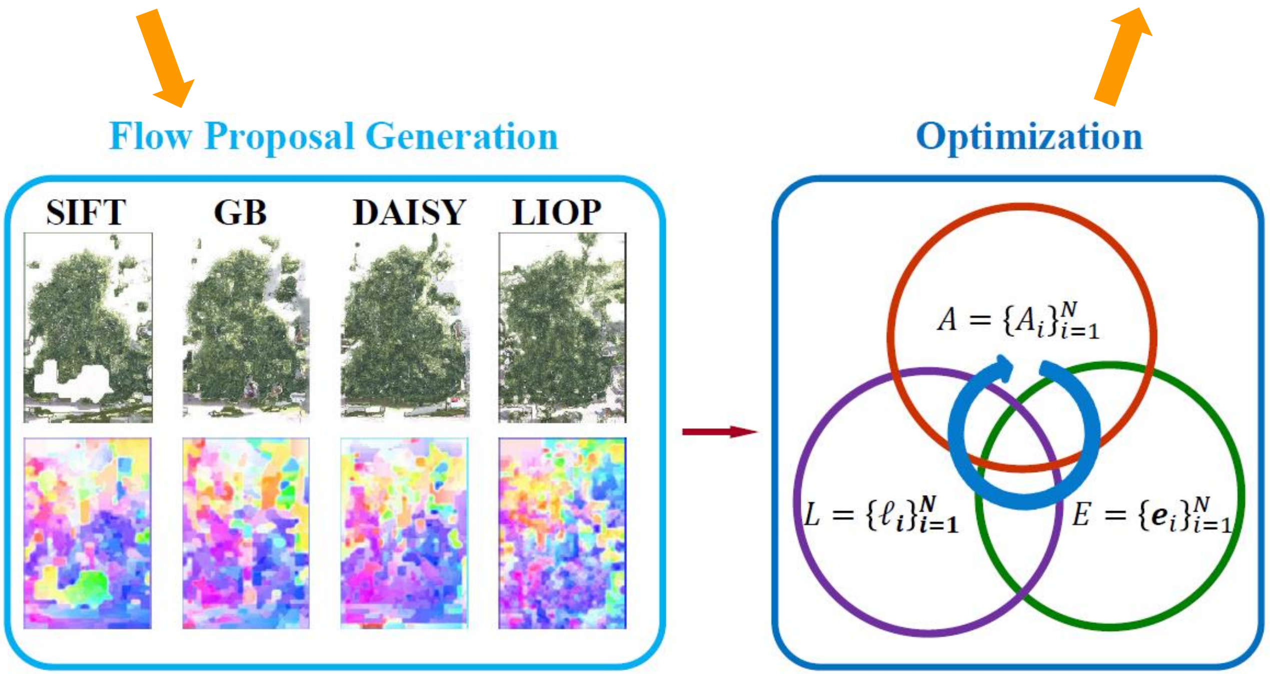
Our idea

- Associate each pixel to be aligned with an affine matrix
  - The affine matrix approximates the local flow map centered on that pixel
  - Affine matrices serve as the common domain for fusion
- Associate each pixel with a learnable neighborhood
  - Better neighborhood facilitates image alignment
  - Better alignment results help reveal the underlying neighborhood
- Integrate the learning of pixel-specific affine matrix and neighborhood into image alignment

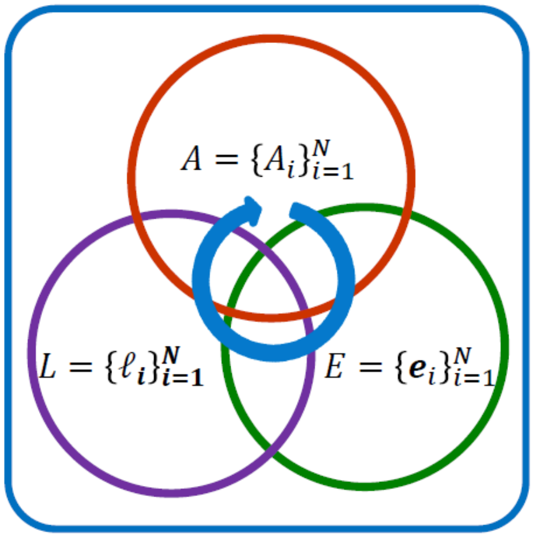
Our approach: an overview

input

output



The optimization problem



- Alignment objective:

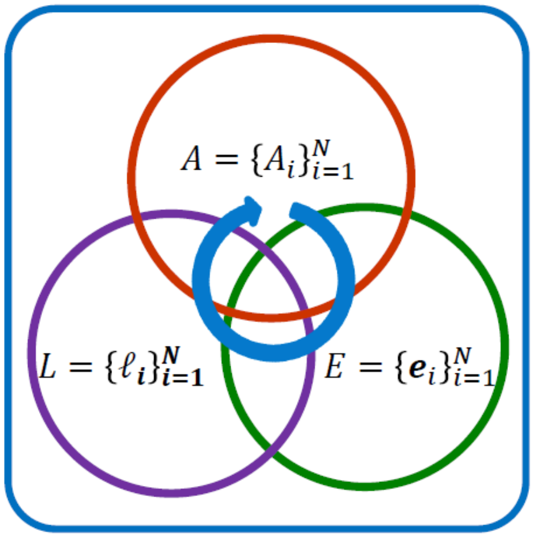
$$\min_{E,A,L} \quad \sum_{i=1}^{N} J(\ell_i, A_i, \boldsymbol{e}_i)$$
$$\text{s.t.} \quad \boldsymbol{e}_i \succcurlyeq 0, \boldsymbol{e}_i^T \mathbf{1} = 1, \text{for i} = 1,2,\ldots,N,$$

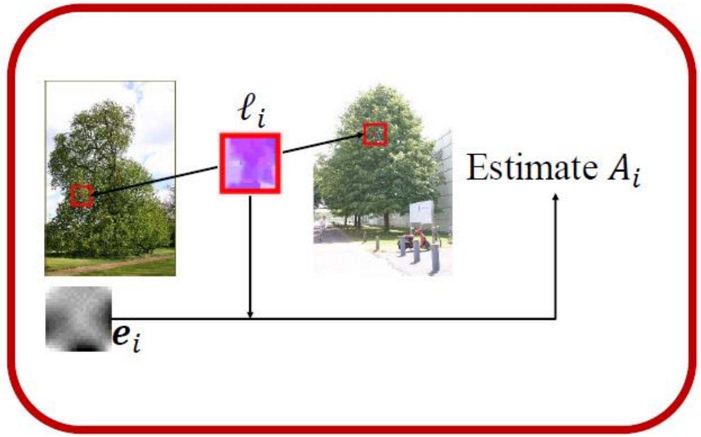
where $\mathbf{1}$ is a column vector whose elements are one, and

$$J(\ell_i, A_i, \boldsymbol{e}_i) = \gamma \|\boldsymbol{p}_i' - A_i\, \boldsymbol{p}_i\|^2 + \sum_{j \in \mathcal{N}_i} e_{ij} \left\|\boldsymbol{p}_j' - A_i\, \boldsymbol{p}_j\right\|^2$$
$$+\alpha \sum_{j \in \mathcal{N}_i} \left(e_{ij} - s_{ij}\right)^2 + \beta \sum_{j \in \mathcal{N}_i} e_{ij} [\ell_i \neq \ell_j].$$

- We adopt an alternate optimization procedure

The optimization problem: On optimizing A





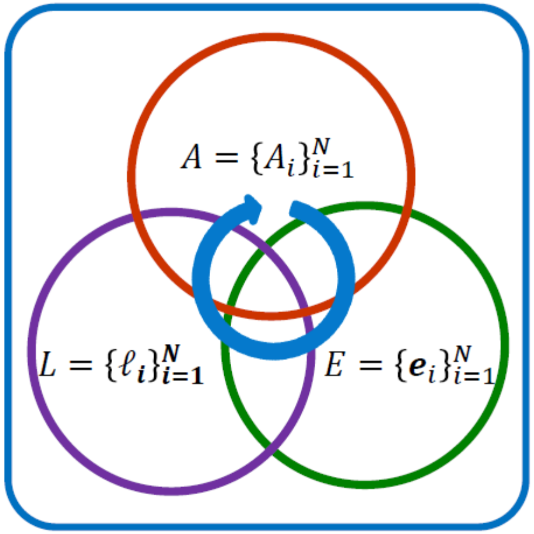
- By fixing L and E, it leads to a weighted least square problem:

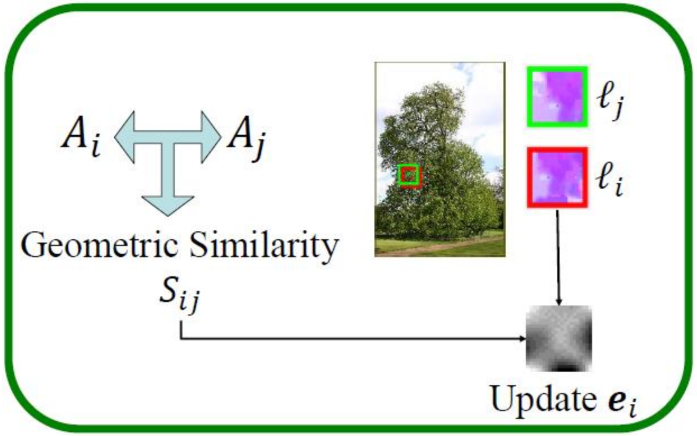
$$\min_{A} \sum_{i=1}^{N} V(A_i)\,, \text{where}$$

$$V(A_i) = \gamma \|\boldsymbol{p}_i' - A_i\, \boldsymbol{p}_i\|^2 + \sum_{j \in \mathcal{N}_i} e_{ij} \left\|\boldsymbol{p}_j' - A_i\, \boldsymbol{p}_j\right\|^2 .$$

- A closed-form solution exists.

The optimization problem: On optimizing E





- By fixing L and A, it leads to a convex quadratic programming problem:

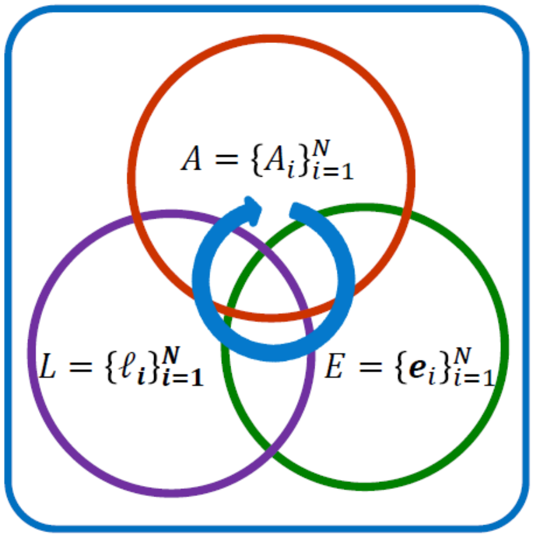
$$\min_{\boldsymbol{e}_i} \ \alpha \|\boldsymbol{e}_i - \boldsymbol{s}_i\|^2 + \boldsymbol{e}_i^T \boldsymbol{r}_i$$

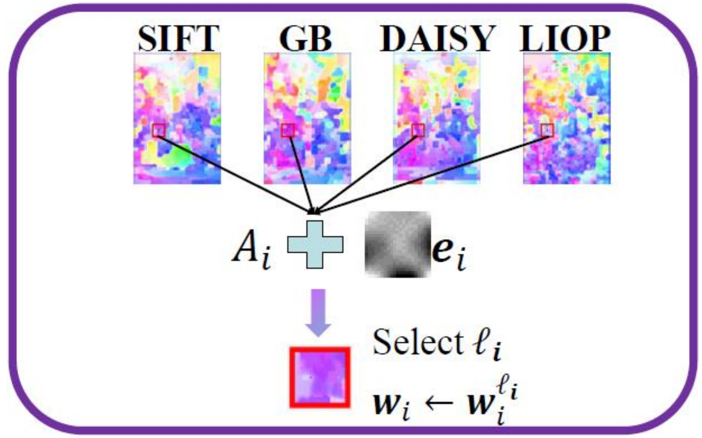
$$\text{s.t.} \ \ \boldsymbol{e}_i \succcurlyeq 0, \boldsymbol{e}_i^T \mathbf{1} = 1,$$

$$\text{where} \ \ \boldsymbol{s}_i = \left[s_{ij} \mid j \in \mathcal{N}_i\right]^T,$$

$$\boldsymbol{r}_i = \left[\left\|\boldsymbol{p}_j' - A_i \, \boldsymbol{p}_j\right\|^2 + \beta[\ell_i \neq \ell_j] | j \in \mathcal{N}_i\right]^T.$$

The optimization problem: On optimizing L

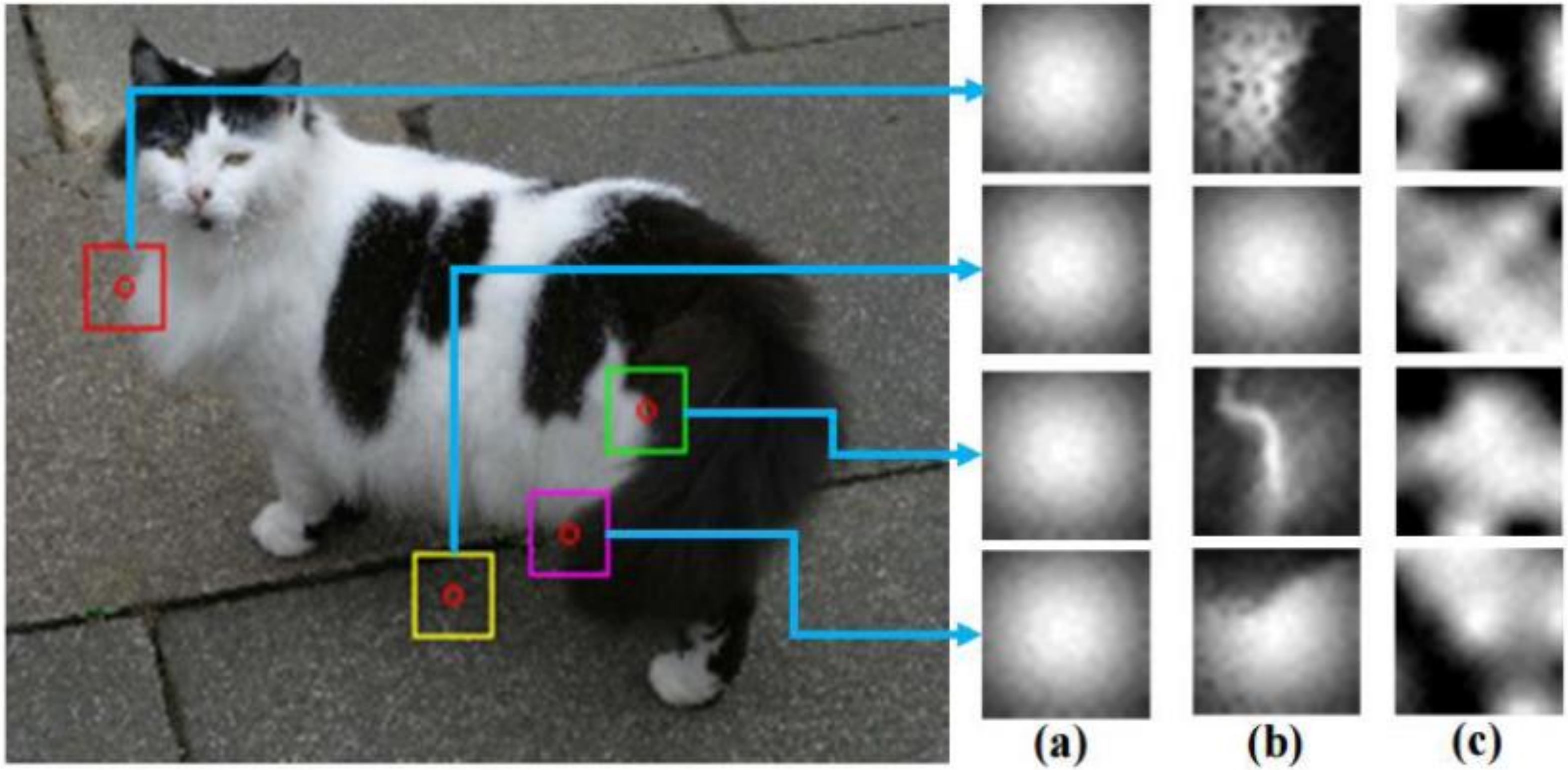




- By fixing E and A, it leads to a labeling problem, and is solved by graph cuts:

$$\min_L \sum_{i=1}^{N} \left(\gamma \left\| \mathbf{p}_i + \mathbf{w}_i^{\ell_i} - A_i\, \mathbf{p}_i \right\|^2 + \sum_{j \in \mathcal{N}_i} e_{ij} \left\| \mathbf{p}_i + \mathbf{w}_i^{\ell_i} - A_i\, \mathbf{p}_j \right\|^2 + \beta \sum_{j \in \mathcal{N}_i} e_{ij} [\ell_i \neq \ell_j] \right).$$

Matching-guided neighborhood



(a) Gaussian weight. (b) Bilateral weight. (c) Ours.

Quantitative results

Method	Descriptor	Dataset			
		VGG	MSRC	Caltech	LMO
SF [24]	SIFT	45.02	41.99	46.08	68.86
	GB	42.71	38.33	50.66	67.80
	DAISY	44.97	41.22	46.68	68.78
	LIOP	47.76	34.87	41.92	66.24
	Con.	49.47	42.20	48.77	63.29
	Avg.	41.18	41.90	49.48	70.27
	FF [21]	47.76	40.40	47.79	69.03
GPM [3]	SIFT	43.99	32.58	37.41	57.44
	GB	41.95	31.41	38.37	57.49
	DAISY	48.44	37.92	39.56	61.05
	LIOP	42.22	25.15	29.44	52.41
	Con.	50.42	38.30	41.47	61.96
	Avg.	39.86	32.04	34.34	56.77
	FF [21]	42.22	34.02	37.15	59.73
DFF [41]	DAISY (R + S)	51.09	40.37	50.95	69.93
Ours + G.	All	53.70	44.67	52.72	71.07
Ours + B.	All	53.64	44.55	53.13	71.02
Ours	All	**54.31**	**46.22**	**54.87**	**72.42**

Datasets and evaluation metrics:

- VGG: ratio of correct matches
- MSRC: intersection / union
- Caltech: intersection / union
- LMO: accuracy

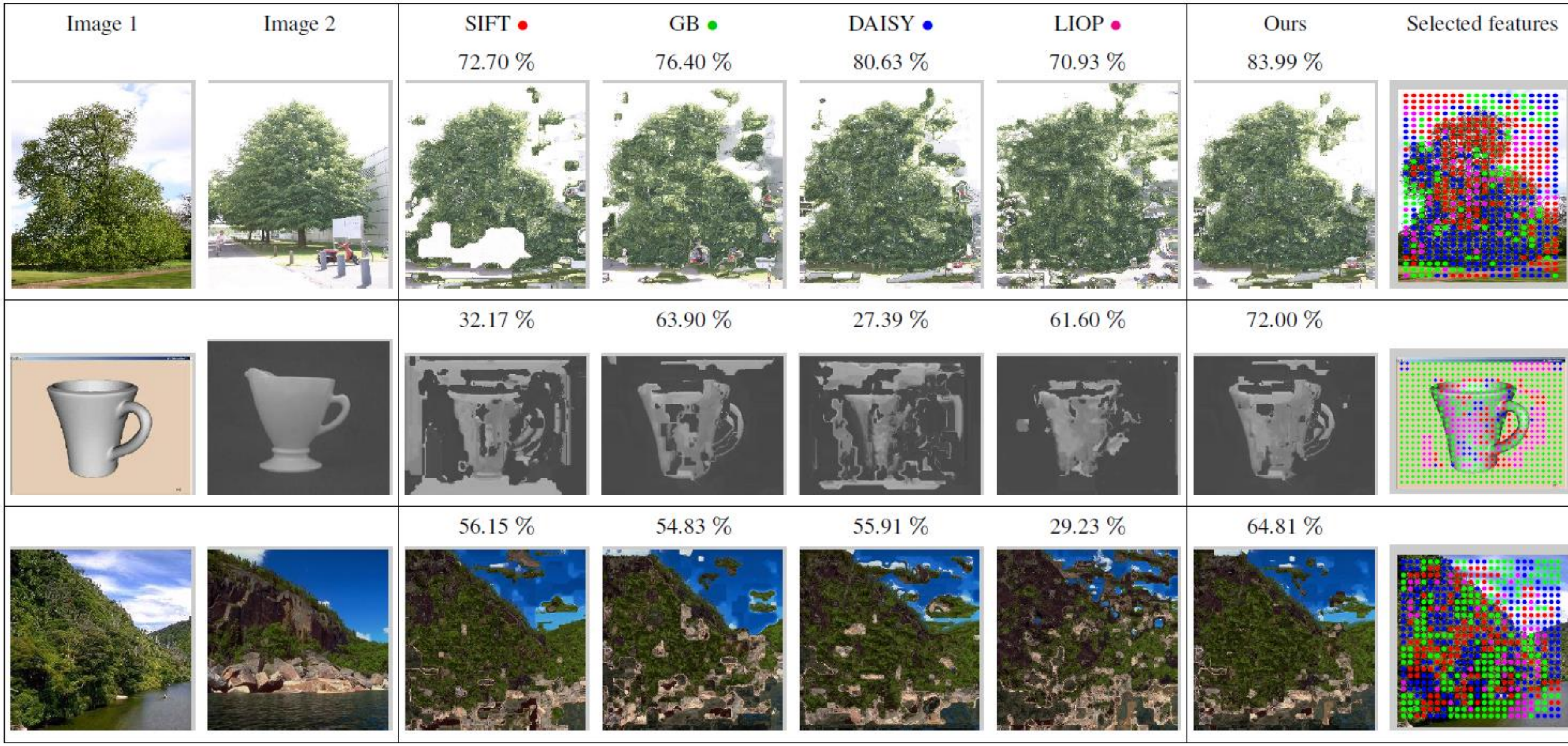
Descriptors:

- [SIFT] Lowe, *IJCV*, 2004
- [GB] Berg and Malik, *CVPR*, 2001
- [DAISY] Tola et al., *TPAMI*, 2010
- [LIOP] Wang, et al., *ICCV*, 2011

Competing approaches:

- [3] Barnes et al., *ECCV*, 2010
- [24] Liu et al., *TPAMI*, 2011
- [41] Yang et al., *CVPR*, 2014

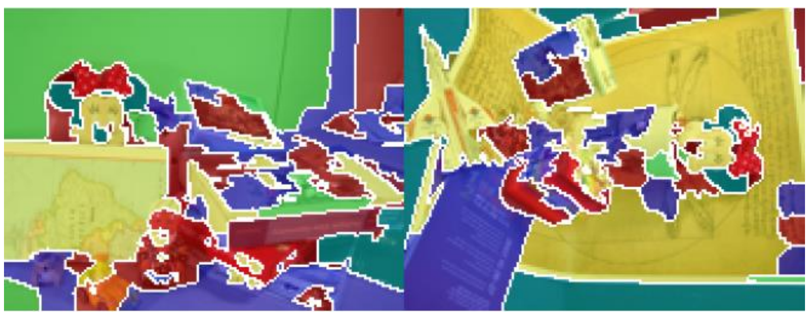
Visualization results

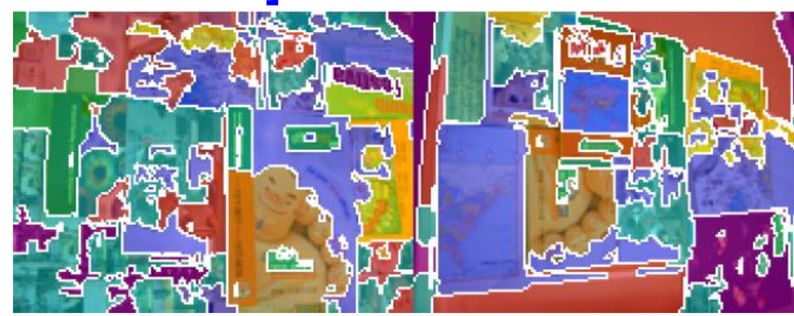


Matching-guided co-segmentation

Transformation clustering

Image co-segmentation by DALCIM [Joulin et al. CVPR'12]





Ours: matching guided co-segmentation

Summary

- ***DescriptorEensemble***: An unsupervised approach to descriptor fusion in the domain of affine transformations
- Ongoing research
  - Seek a unified formulation of DescriptorEnsmble
  - Explore its computer vision applications
  - Combine sparse matching and dense matching (alignment)
- Related papers:
  - H.-Y. Chen, Y.-Y. Lin, and B.-Y. Chen, "Robust Feature Matching with Alternate Hough and Inverted Hough Transforms," *IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013
  - K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "Robust Image Alignment with Multiple Feature Descriptors and Matching-Guided Neighborhoods," *IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015
  - H.-Y. Chen, Y.-Y. Lin, and B.-Y. Chen, "Co-Segmentation Guided Hough Transform for Robust Feature Matching," to appear in *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI).*
  - Y.-T. Hu, Y.-Y. Lin, H.-Y. Chen, K.-J. Hsu and B.-Y. Chen, "Matching Images with Multiple Descriptors: An Unsupervised Approach for Locally Adaptive Descriptor Selection," ," to appear in *IEEE Trans. on Image Processing (TIP).*

Thank You for Your Attention!

Yen-Yu Lin (林彥宇)
yylin@citi.sinica.edu.tw