Nlab-UTokyo at ImageCLEF 2013 Plant Identification Challenge

Augmenting descriptors for fine-grained categorization

Hideki Nakayama

Graduate School of Information Science and Technology
The University of Tokyo

Contents

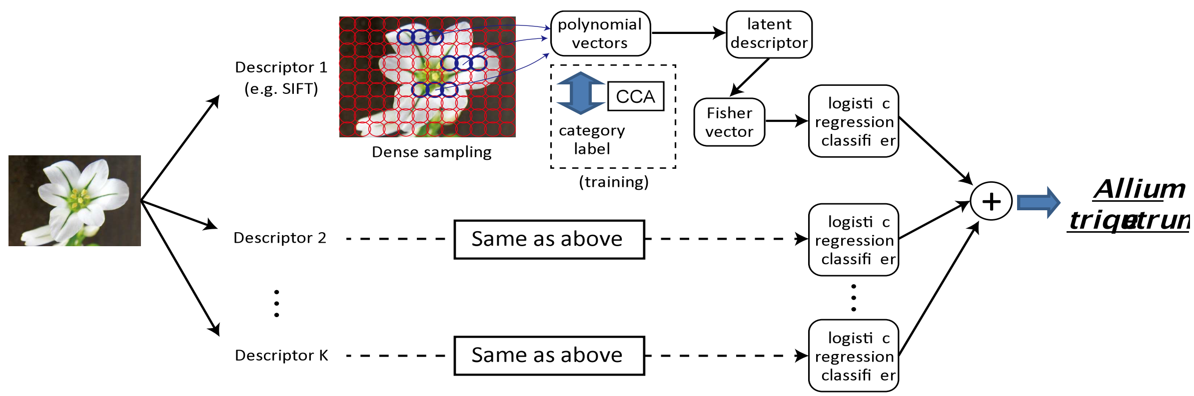
- Overview
- Motivation & Problems
- Our solution
- Challenge results
- Conclusion & Discussion

Overview of our participation

- Basically follows a standard object recognition pipeline based on bag-of-features
  - We implemented our recently proposed method for general-purpose **fine-grained visual categorization**

    Hideki Nakayama, "Augmenting descriptors for fine-grained visual categorization using polynomial embedding", *Proc. IEEE ICME*, 2013.

- We focused on extracting powerful image signatures, rather than segmentation and classification algorithms.
  - Obtain strong local descriptors by embedding local spatial information in a supervised learning framework



Fine-grained visual categorization (FGVC)

- **Distinguish hundreds of very similar object categories** under a specific domain (e.g., species of plants, dogs, birds, etc.)
  - Complementary to traditional object recognition problems
- We need highly discriminative image features

Generic Object Recognition



Airplane Monitor Dog

Caltech-256
[Griffin *et al.*, 2007]

V.S.

Plant Identification



Tilia cordata Tilia dasystyla Tilia platyphyllos

Two basic ideas

- 1. Co-occurrence (correlation) of neighboring local descriptors
  - Shaplet [Sabzmeydani et al., 2007] Covariance feature [Tuzel et al., 2006] CoHOG [Ito et al., 2010] GGV [Harada et al., 2012]

☺ Expected to capture middle-level local information

☹ Results in high-dimensional local features

How to relax these problems?

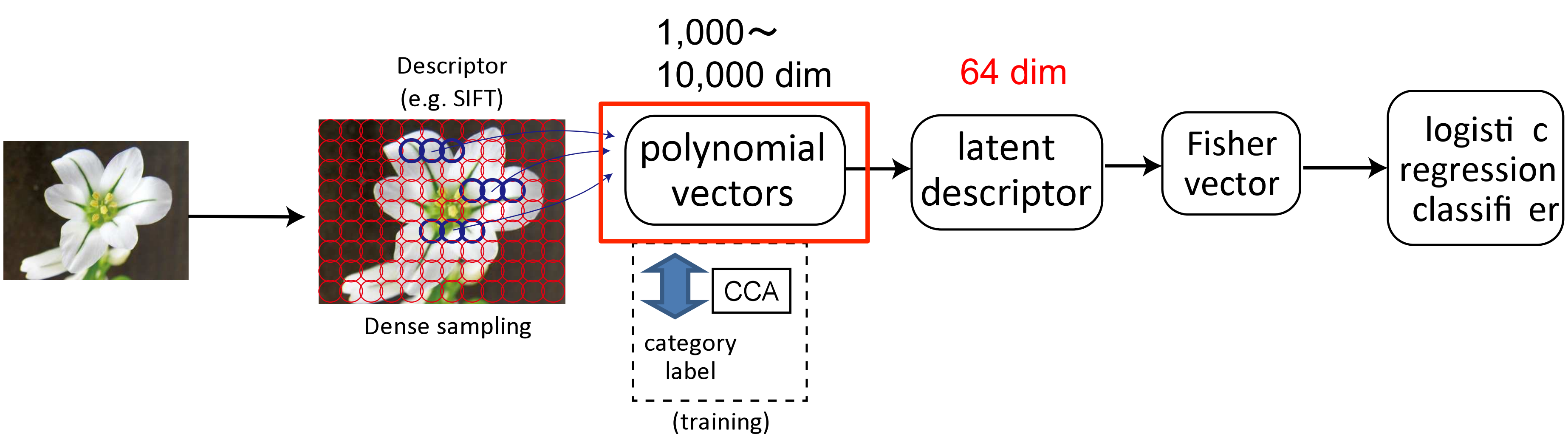
- 2. Fisher Vector encoding [Perronnin et al., 2010]
  - State-of-the-art bag-of-words representation based on higher-order statistics of local features

☺ Remarkably high-performance, enables linear classification

☹ Dimensionality increases in linear to the size of local features

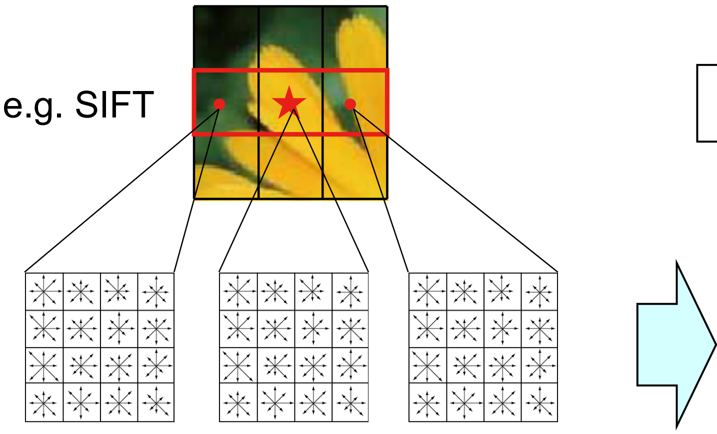
Our approach

- Densely sample local decriptors
- Compress co-occurrence patterns (polynomials) of neighboring local descriptors
  - ⇒ Discriminative latent descriptor
- Encode by means of bag-of-words (Fisher vector)
- Logistic regression classifier



Exploit co-occurrence information

e.g. SIFT



$\mathbf{v}_{(x-\sigma,y)}$ $\mathbf{v}_{(x,y)}$ $\mathbf{v}_{(x+\sigma,y)}$

Neighbor (Left)

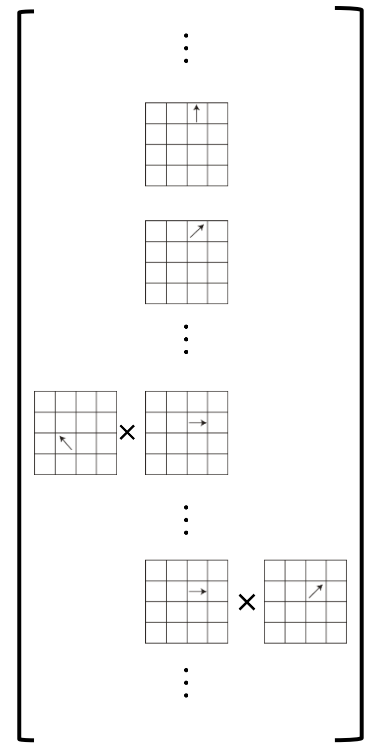
Descriptor at target position

Neighbor (Right)

Polynomial Vector

$$\mathbf{p}^2_{(x,y)} = \begin{pmatrix} \mathbf{v}_{(x,y)} \\ upperVec\left(\mathbf{v}_{(x,y)}\mathbf{v}^T_{(x,y)}\right) \\ Vec\left(\mathbf{v}_{(x,y)}\mathbf{v}^T_{(x-\delta,y)}\right) \\ Vec\left(\mathbf{v}_{(x,y)}\mathbf{v}^T_{(x+\delta,y)}\right) \end{pmatrix}$$

$Vec(\,)$: flattened vector of a Matrix



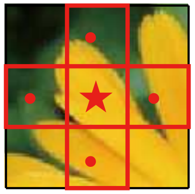
Exploit co-occurrence information

- More spatial information can be integrated with more neighbors (but become high-dimensional)

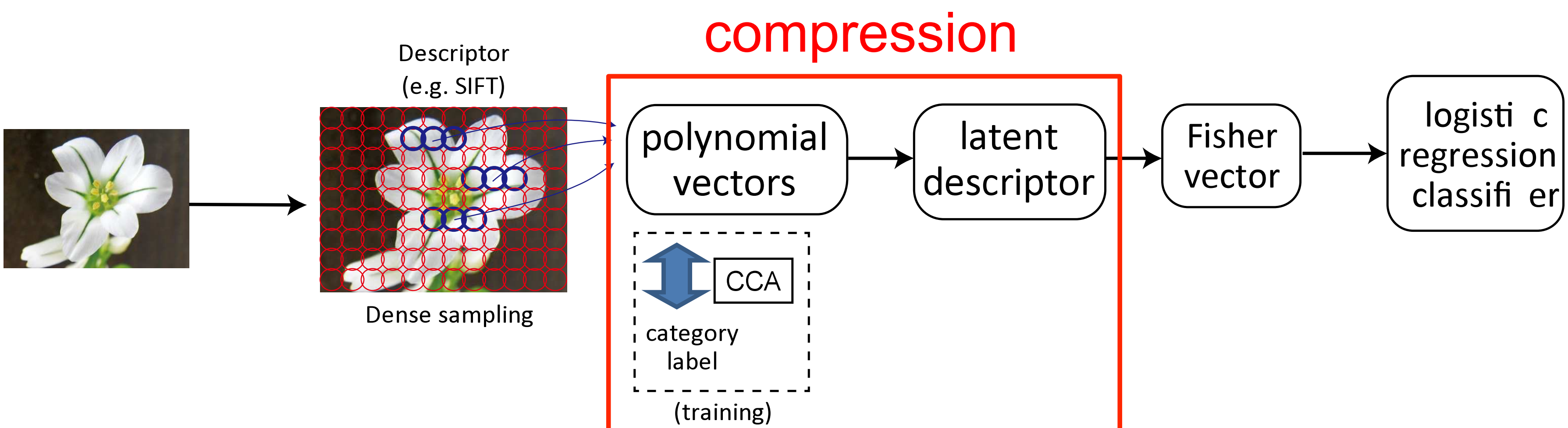
$$\mathbf{p}^2_{(x,y)} = \begin{pmatrix} \mathbf{v}_{(x,y)} \\ upperVec\left(\mathbf{v}_{(x,y)}\mathbf{v}^T_{(x,y)}\right) \\ Vec\left(\mathbf{v}_{(x,y)}\mathbf{v}^T_{(x-\delta,y)}\right) \\ Vec\left(\mathbf{v}_{(x,y)}\mathbf{v}^T_{(x+\delta,y)}\right) \end{pmatrix}$$

2-neighbors 10,336dim

$$\mathbf{p}^4_{(x,y)} = \begin{pmatrix} \mathbf{v}_{(x,y)} \\ upperVec\left(\mathbf{v}_{(x,y)}\mathbf{v}^T_{(x,y)}\right) \\ Vec\left(\mathbf{v}_{(x,y)}\mathbf{v}^T_{(x,y-\delta)}\right) \\ Vec\left(\mathbf{v}_{(x,y)}\mathbf{v}^T_{(x-\delta,y)}\right) \\ Vec\left(\mathbf{v}_{(x,y)}\mathbf{v}^T_{(x+\delta,y)}\right) \\ Vec\left(\mathbf{v}_{(x,y)}\mathbf{v}^T_{(x,y+\delta)}\right) \end{pmatrix}$$



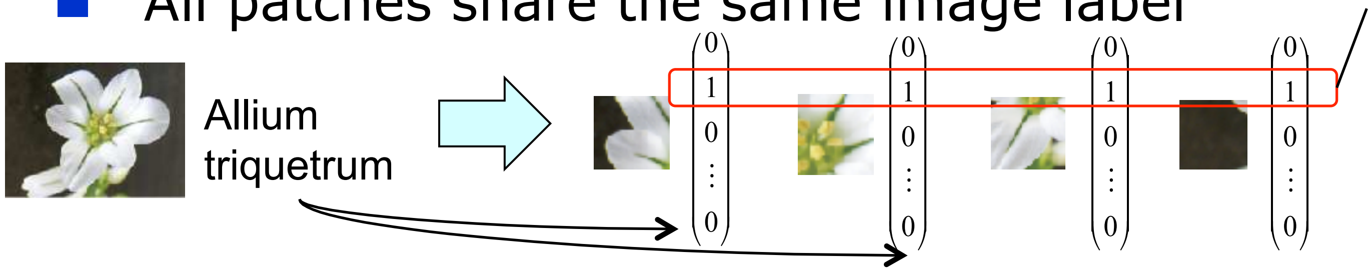
4-neighbors 18,528dim



compression
Descriptor
(e.g. SIFT)
Dense sampling
polynomial vectors
latent descriptor
CCA
category label
(training)
Fisher vector
logisti c regression classifi er

Supervised dimensionality reduction to compress polynomial vector

- Training set: patch features (polynomial vectors) and category labels
  - All patches share the same image label



- Strong supervision assumption
  - Most patches should be related to the category
  - (Somewhat) justified for FGVC considering the applications
  - Users will more or less target the object

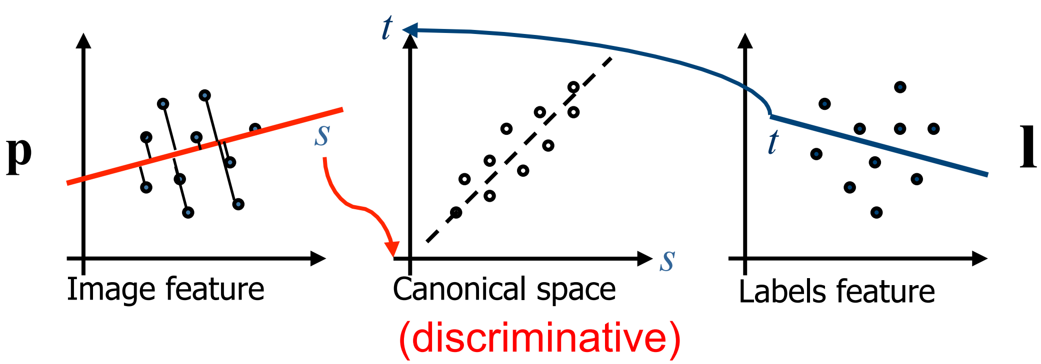
Supervised dimensionality reduction

- Canonical Correlation Analysis (CCA) [Hotelling, 1936]

  $\mathbf{p}$: patch feature (polynomials), $\mathbf{l}$: label feature

  CCA finds linear transformations

  $\mathbf{s} = A^T(\mathbf{p} - \bar{\mathbf{p}})$, $\mathbf{t} = B^T(\mathbf{l} - \bar{\mathbf{l}})$ that maximize the correlation between $\mathbf{s}$ and $\mathbf{t}$



$$C_{pl}C_{ll}^{-1}C_{lp}A = C_{pp}A\Lambda^2 \quad \left(A^T C_{pp} A = I\right)$$

C : covariance matrices

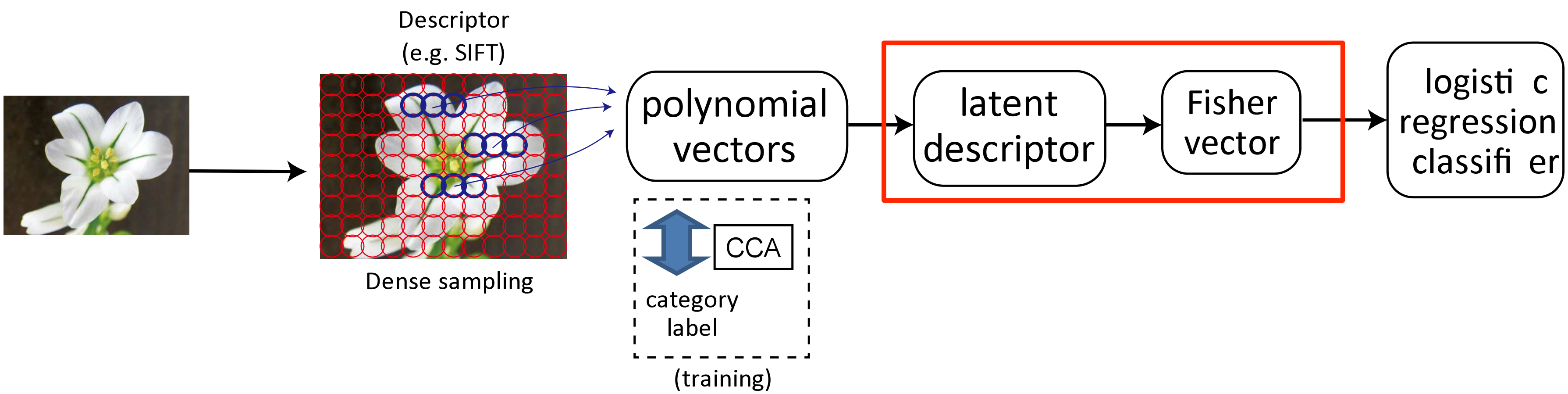
Λ : canonical correlations

Latent descriptor

$$\mathbf{s} = A^T\left(\mathbf{p} - \bar{\mathbf{p}}\right)$$

64 dim

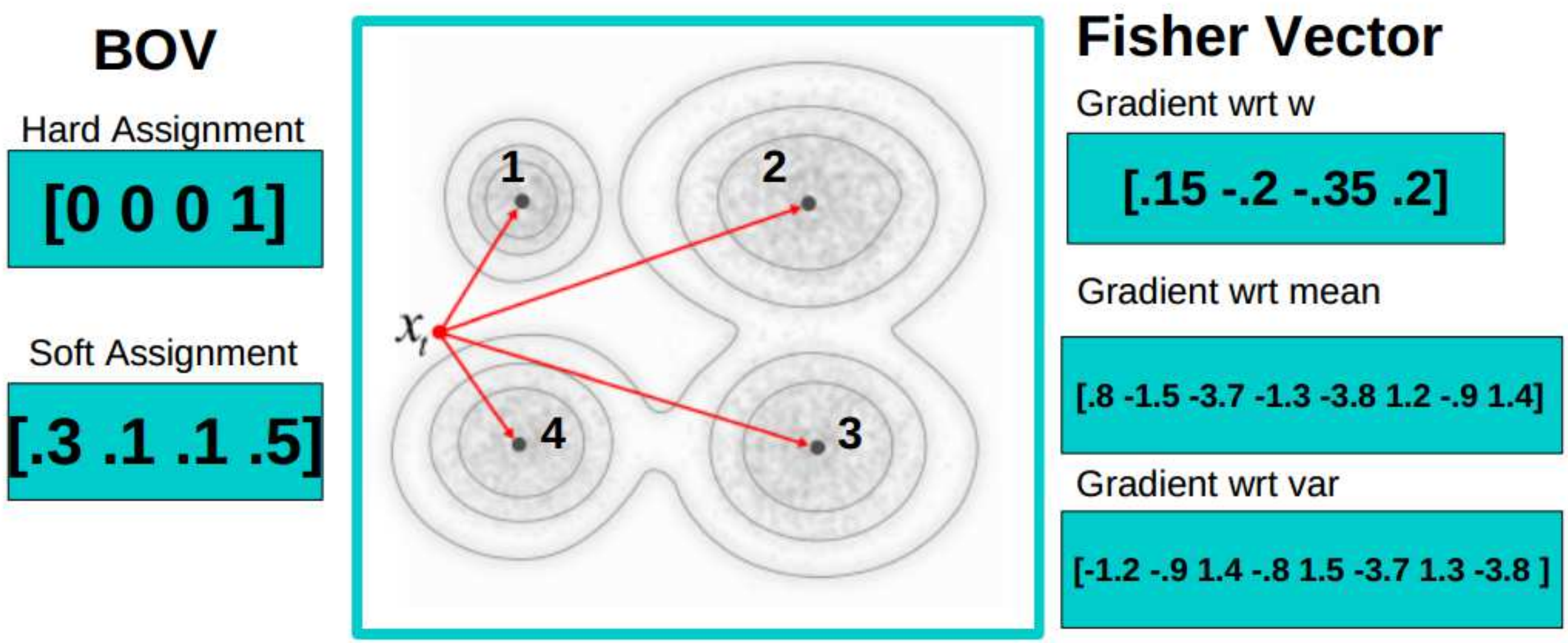
1,000～
10,000 dim


Descriptor
(e.g. SIFT)
Dense sampling
polynomial vectors
CCA
category label
(training)
latent descriptor
Fisher vector
logistic regression classifier

Fisher Vector [Perronnin *et al.*, 2010]

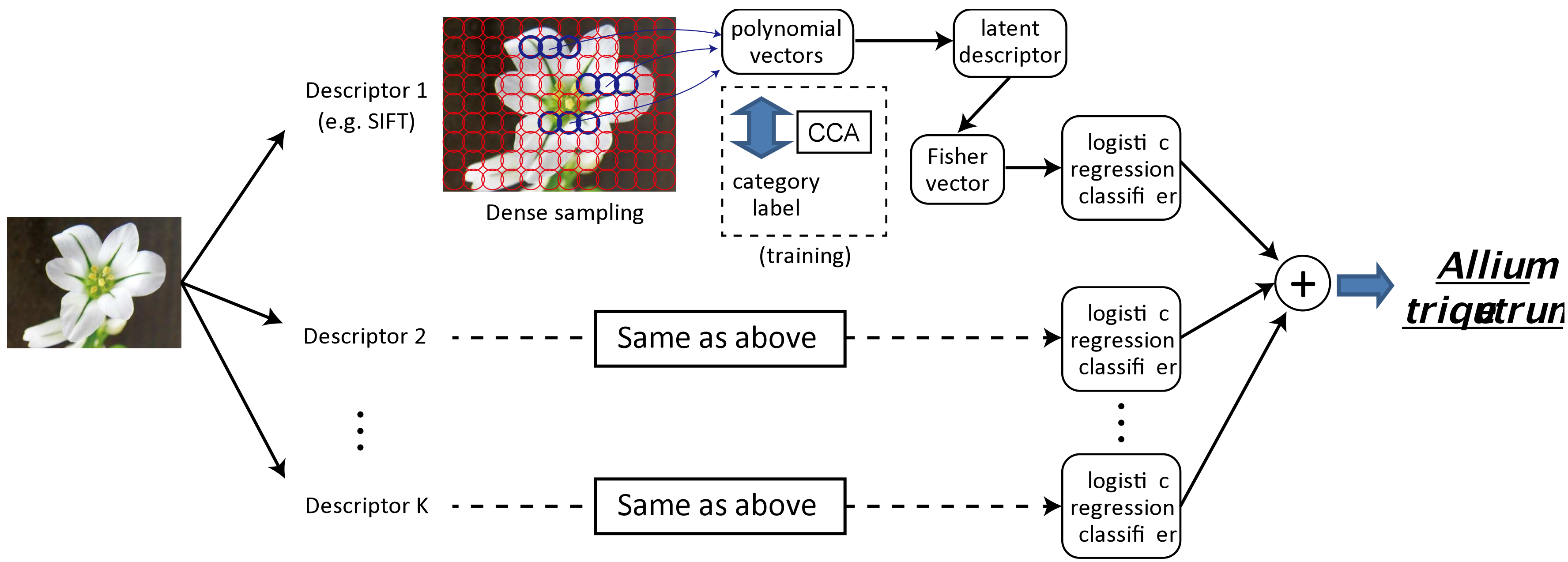
- State-of-the-art bag-of-words encoding method using higher-level statistics of descriptors (mean and var)

http://www.image-net.org/challenges/LSVRC/2010/ILSVRC2010_XRCE.pdf



Our final system

- Combine multiple descriptors in late-fusion approach (SIFT, C-SIFT, Opp.-SIFT, HSV-SIFT, Self similarity)
- Sum of log-likelihoods output by each classifier

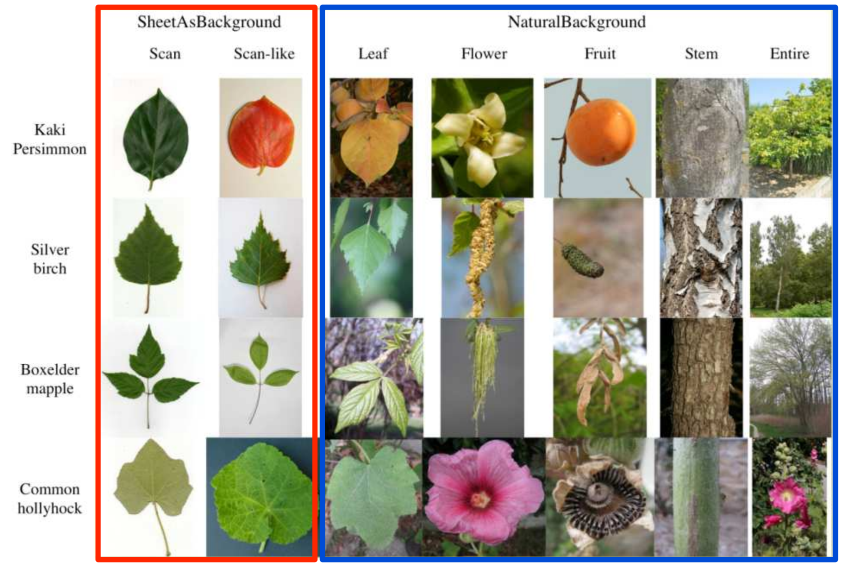


Plant Identification Challenge

Challenge Overview

- Identify 250 plant species from images of different organs (Leaf, Flower, Fruit, etc.)

- Two main categories:
  - Sheet As Background
  - Natural Background



http://www.imageclef.org/2013/plant

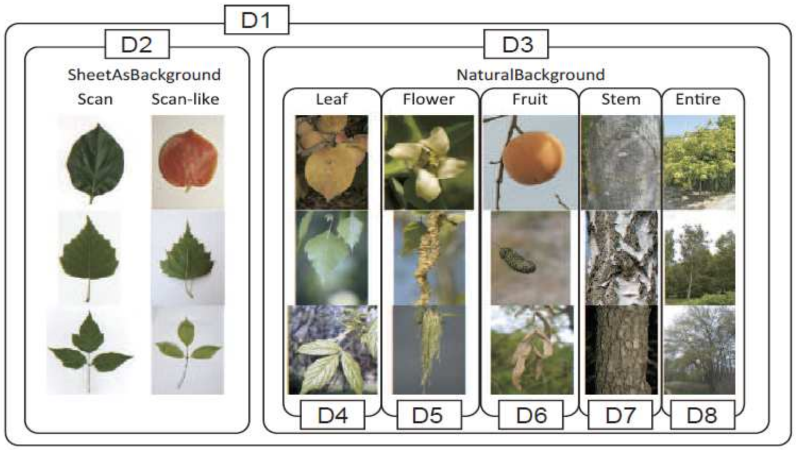
“Natural Background” has more generic nature (e.g, cluttered background, view, etc.) and is the primary interest in our participation

Setup

Our submitted runs

We trained classifiers independently for each (sub)category

- Run 1: All
- Run 2: SheetAsBackground + NaturalBackground
- Run 3: SheetAsBackground + Leaf, Flower, Fruit, Stem, Entire



Validation

- We used roughly 10% of training samples **(in terms of individual plants)** for validation set
- Parameter tuning and selection of local descriptors

Results on the validation set

- Our method consistently improves the performance from the baseline for all descriptors & domains.
- **Particularly effective for Natural Background task.**

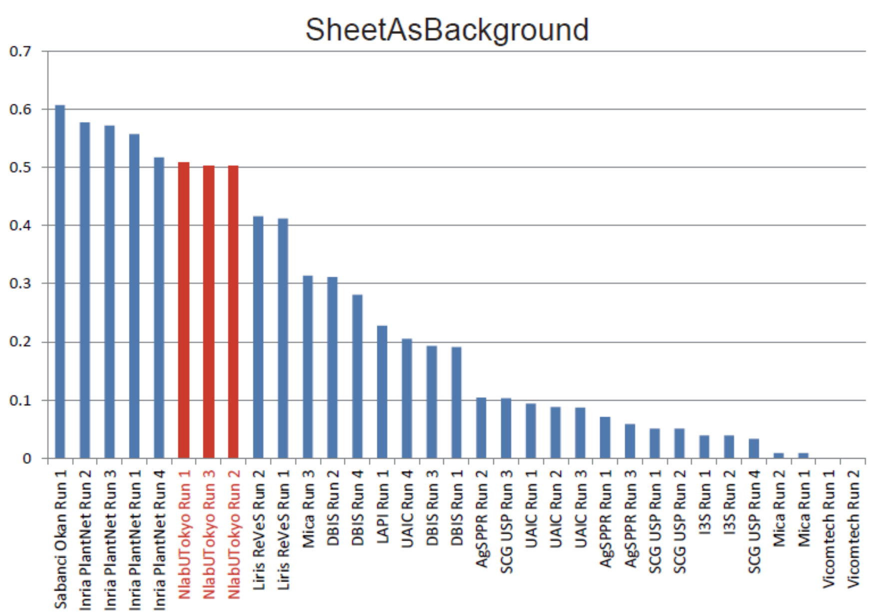
Standard implementation of Fisher Vector.

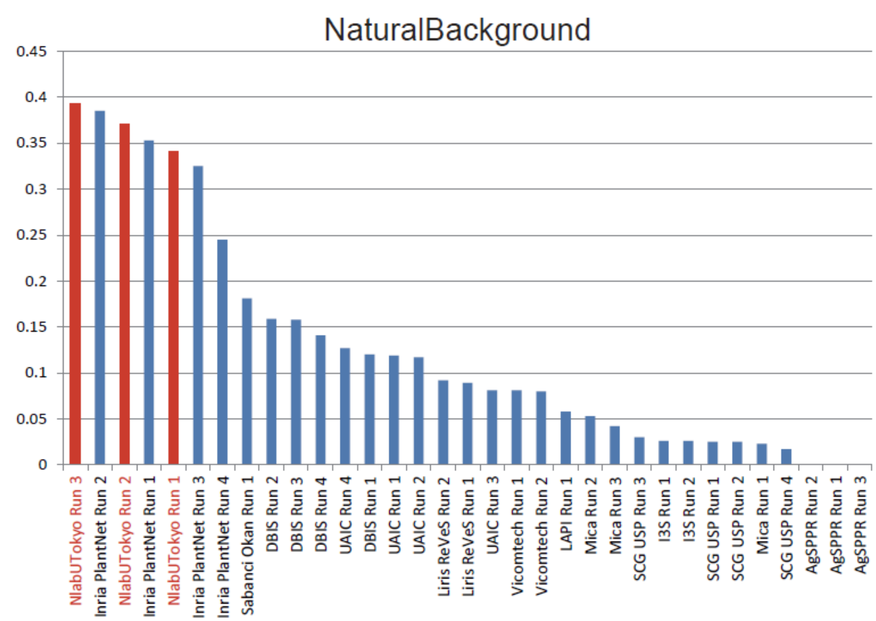
Used descriptor types and classification rates (%) on the validation set.

		SIFT	C-SIFT	Opp.-SIFT	HSV-SIFT	SSIM	Baseline	Ours	Rel. Imp. (%)
Run 1	All	✔	✔	✔	✔		38.2	38.8	1.6
Run 2,3	SB	✔					50.8	52.5	3.3
Run 2	NB		✔	✔	✔	✔	15.9	17.8	11.9
Run 3	Leaf		✔	✔	✔	✔	15.2	17.3	13.8
Run 3	Flower		✔	✔	✔	✔	21.2	24.7	16.5
Run 3	Fruit		✔	✔	✔	✔	7.4	11.1	50.0
Run 3	Stem		✔	✔	✔	✔	13.8	16.5	19.6
Run 3	Entire		✔	✔	✔	✔	8.2	8.5	3.7

Final results

- **We achieved:**
  - The 1st place in NaturalBackground category (and 4/5 subcategories). Run 3
  - The 3rd place in SheetAsBackground category. Run 1





Conclusion

- A simple but effective method for FGVC
  - Embedding co-occurrence patterns of neighboring descriptors.
  - Obtain discriminative and small-dimensional latent descriptor to make Fisher vector encoding feasible.
  - Particularly effective for Natural Background task.
- Patch-level strong supervision approximation
  - Not always perfect but reasonable for FGVC problems.
- Discussion
  - Standard object recognition approach is not bad, as the task becomes more general.
  - Features are the most important key to success, of course better segmentation & classification algorithms should be implemented as well.

Implementation Details

- Low-level descriptors
  - SIFT, C-SIFT, Opponent-SIFT, HSV-SIFT, Self Similarity
  - Dense sampling (5 pixels apart)

  http://koen.me/research/colordescriptors/
  http://www.robots.ox.ac.uk/~vgg/software/SelfSimilarity/

- Fisher Vector
  - 64 Gaussians (visual words)
  - Entire image + 3 horizontal spatial regions

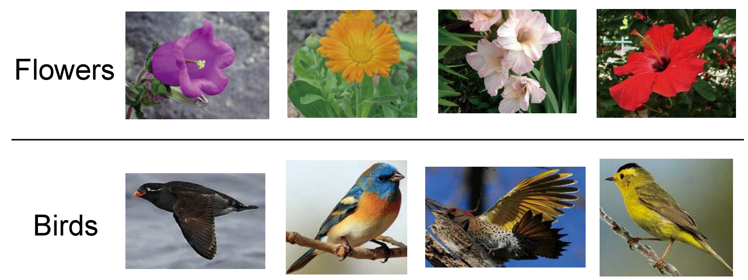
  http://lear.inrialpes.fr/src/inria_fisher/

- Classifier
  - Logistic regression (LIBLINEAR)
  - Average scores of multiple classifiers

  http://www.csie.ntu.edu.tw/~cjlin/liblinear/

Experimental setup

- FGVC Datasets
  - Oxford-Flowers-102
  - Caltech-Bird-200

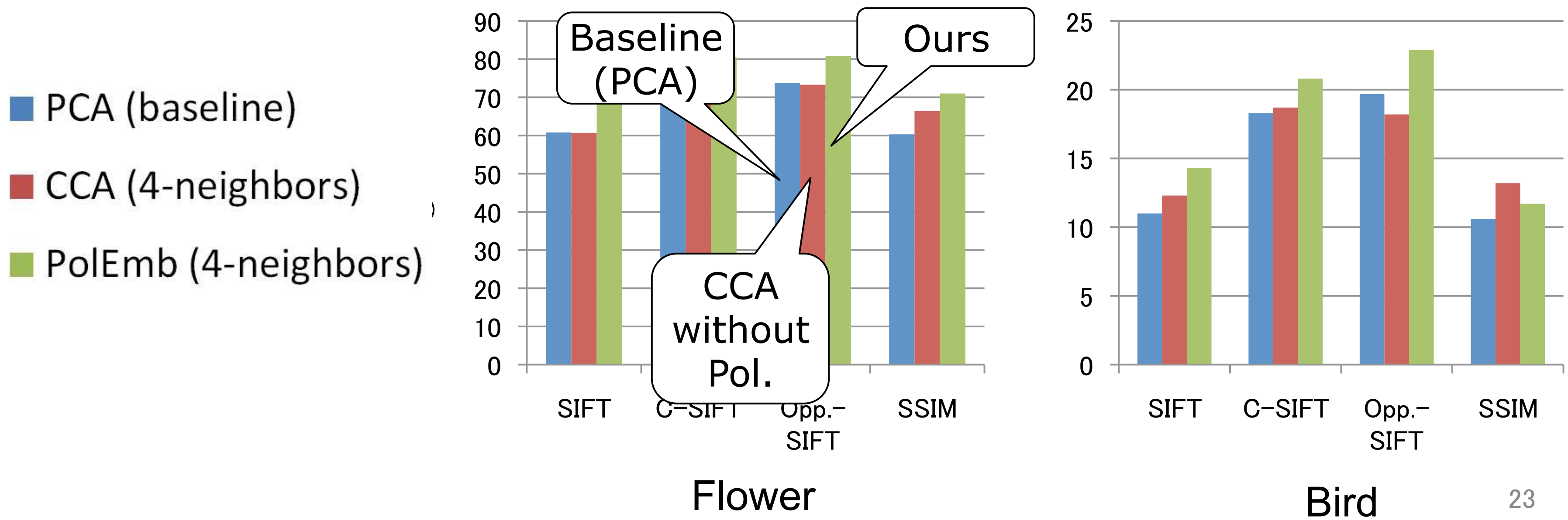


- Descriptors
  - SIFT, C-SIFT, Opponent-SIFT, Self Similarity
  - Compressed into 64dim using several methods
- Fisher Vector
  - 64 Gaussians (visual wods)
  - Global + 3 horizontal spatial regions
- Classifier
  - Logistic regression
- Evaluation
  - Mean classification accuracy

Results: comparison with PCA and CCA

- Our method substantially improves performance for all descriptors
- Just applying CCA to concatenated neighbors does not improve performance
  - Polynomial embedding makes sense (non-linear convolution)

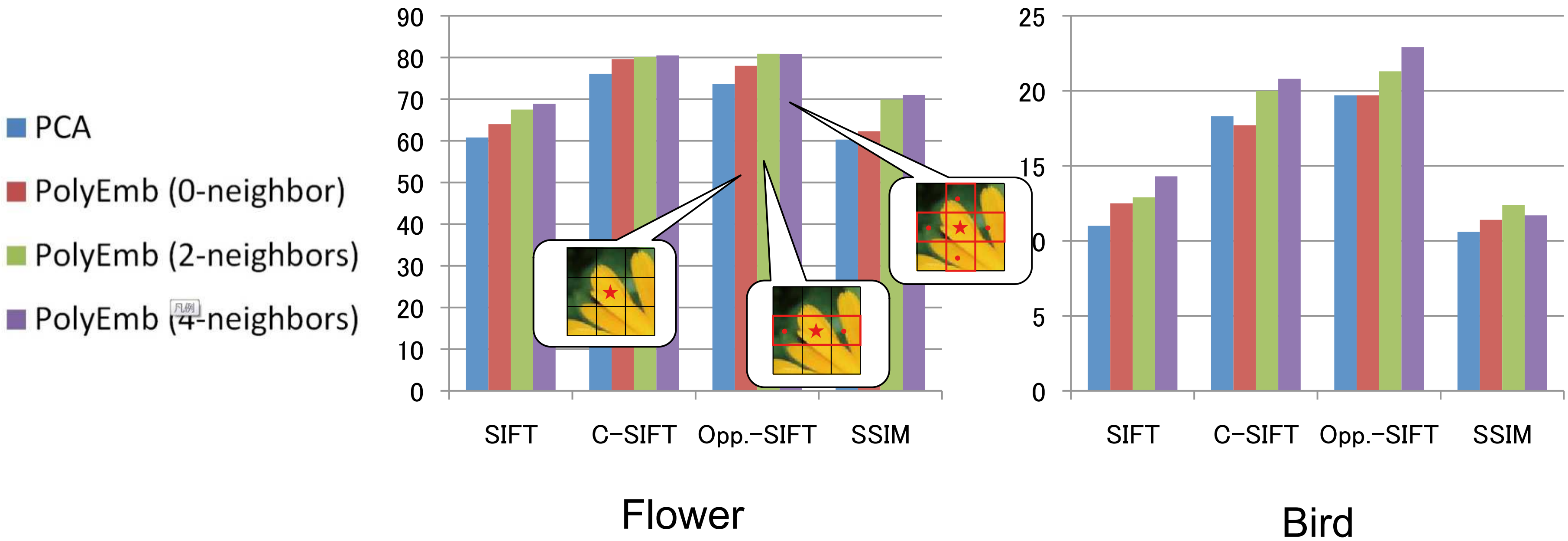
Classification performance (%) with different embedding methods (all 64dim)



Results: number of neighbors

- Including more neighbors improves performance

Classification performance (%) of our method with different number of neighbors



Comparison on FGVC datasets

- Our method outperforms previous work on bird and flower datasets

Mean classification accuracy (%)

	Flowers	Birds
4 desc. (PCA)	81.6	23.9
4 desc. (PolEmb)	**87.2**	28.1
8 desc. (PCA+PolEmb)	85.7	**28.8**
Previous Work	85.6 [32]	28.2 [33]
	80.0 [34]	26.7 [32]
	76.3 [35]	26.4 [36]
	73.3 [37]	22.4 [37]
		19.2 [31]
		19.0 [38]
		18.0 [7]

← baseline

For the bird dataset, [32] uses the bounding box only for training images, therefore the result is not directly comparable to ours.