Image Classification Using Gaussian Mixture and Local Coordinate Coding

Kai Yu
NEC Laboratories America, Cupertino, California, USA

Contributors:

Jinjun Wang, Fengjun Lv, Wei Xu, Yihong Gong	NEC Laboratories America
Xi Zhou, Jianchao Yang, Thomas Huang,	Univ. of Illinois at Urbana-Champaign
Tong Zhang	Rutgers University
Chen Wu	Stanford University

PASCAL VOC Challenge, ICCV, at Kyoto, Japan, October 3rd, 2009

Where We Are in This Competition

	Our 4 submissions				Our Best	Other's Best	Our Improvement
Aeroplane	88.1	88.0	87.1	87.7	88.1	86.6	1.5
Bicycle	68.0	68.6	67.4	67.8	68.6	63.9	**4.7**
Bird	68.0	67.9	65.8	68.1	68.1	66.7	1.4
Boat	72.5	72.9	72.3	71.1	72.9	67.3	**5.6**
Bottle	41.0	44.2	40.9	39.1	44.2	43.7	0.5
Bus	78.9	79.5	78.3	78.5	79.5	74.1	**5.4**
Car	70.4	72.5	69.7	70.6	72.5	64.7	**7.8**
Cat	70.4	70.8	69.7	70.7	70.8	64.2	**6.6**
Chair	58.1	59.5	58.5	57.4	59.5	57.4	2.1
Cow	53.4	53.6	50.1	51.7	53.6	46.2	**7.4**
Diningtable	55.7	57.5	55.1	53.3	57.5	54.7	2.8
Dog	59.3	59.0	56.3	59.2	59.3	53.5	**5.8**
Horse	73.1	72.6	71.8	71.6	73.1	68.1	**5.0**
Motorbike	71.3	72.3	70.8	70.6	72.3	70.6	1.7
Person	84.5	85.3	84.1	84.0	85.3	85.2	0.1
Pottedplant	32.3	36.6	31.4	30.9	36.6	39.1	-2.5
Sheep	53.3	56.9	51.5	51.7	56.9	48.2	**8.7**
Sofa	56.7	57.9	55.1	55.9	57.9	50.0	**7.9**
Train	86.0	85.9	84.7	85.9	86.0	83.4	2.6
Tvmonitor	66.8	68.0	65.2	66.7	68.0	68.6	-0.6
Average	65.4	66.5	64.3	64.6			

Accuracy measured by average precision (AP)

Comparative Overview

Paradigm	State of the Art	Ours
Feature Detection	multiple detectors	dense sampling
Feature Extraction	multiple descriptors	SIFT (gray)
Coding Scheme	VQ	GMM, LCC
Spatial Pooling	SPM	SPM
Classifier	nonlinear classifiers	linear classifiers

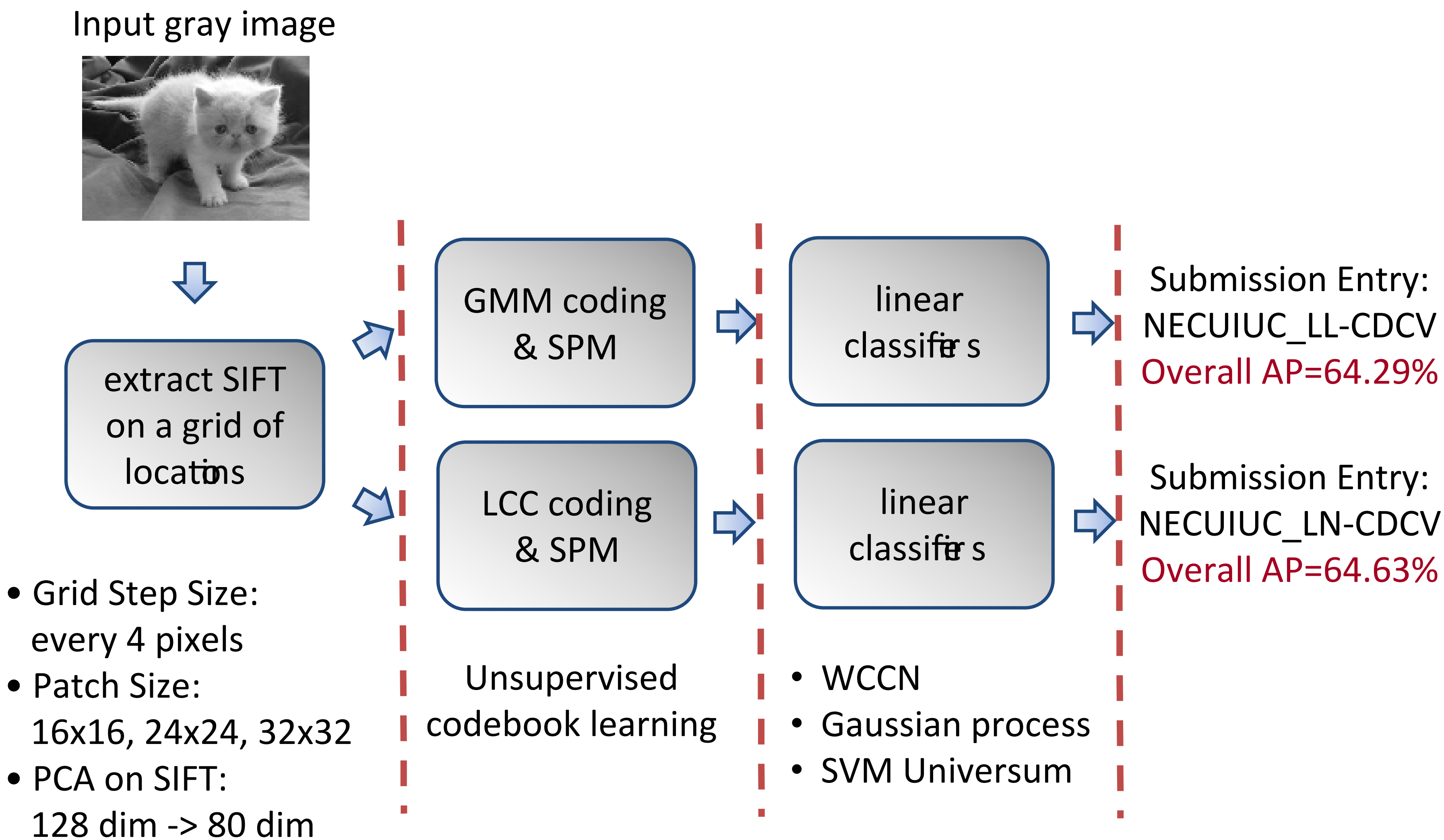
Our Strategy

Minimum feature engineering

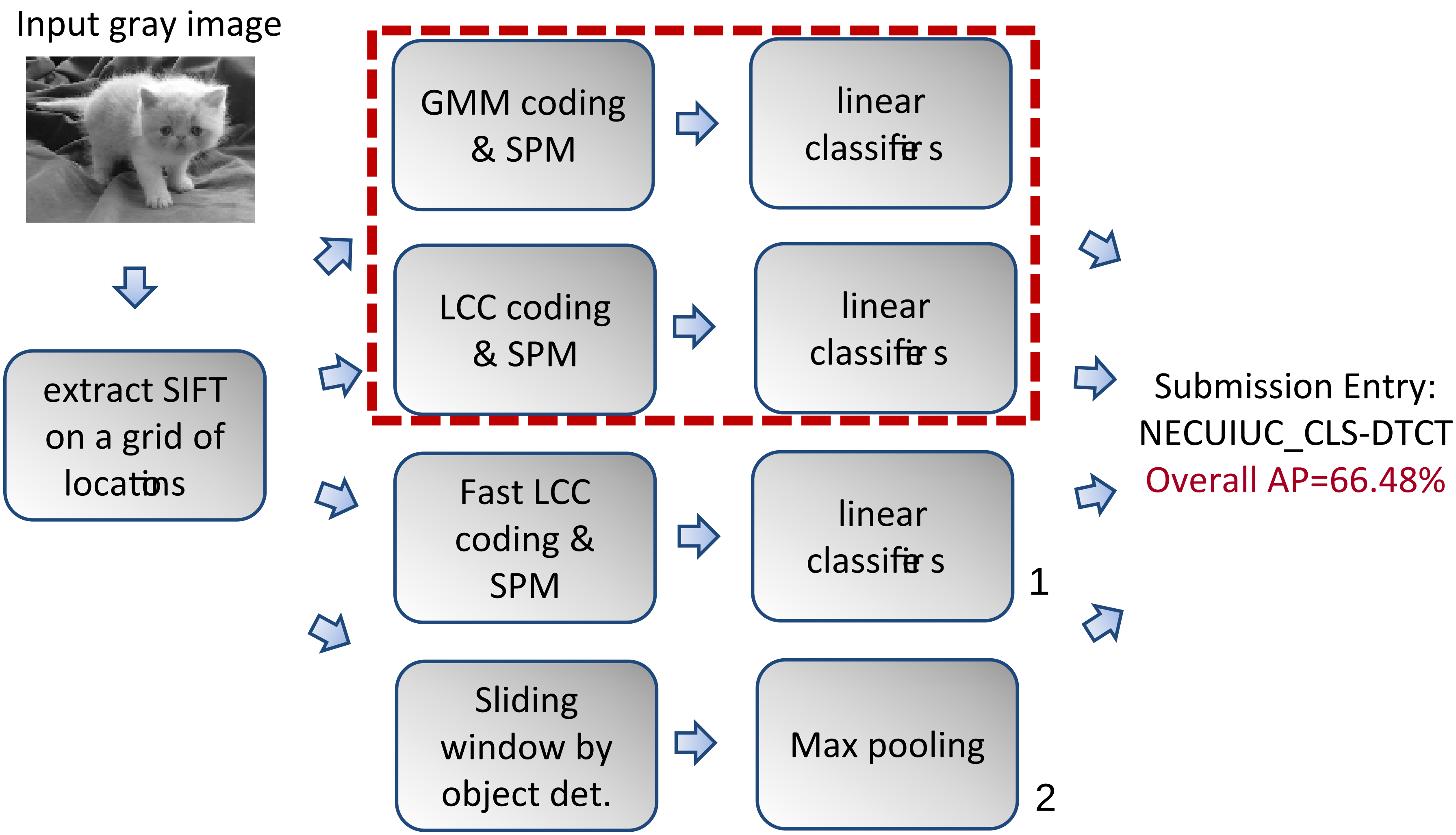
Paradigm	State of the Art	Ours
Feature Detection	multiple detectors	dense sampling
Feature Extraction	multiple descriptors	SIFT (gray)
Coding Scheme	VQ	GMM, LCC
Spatial Pooling	SPM	SPM
Classifier	nonlinear classifiers	linear classifiers

We bet on machine learning techniques.

Pipeline Overview - I



Pipeline Overview - II



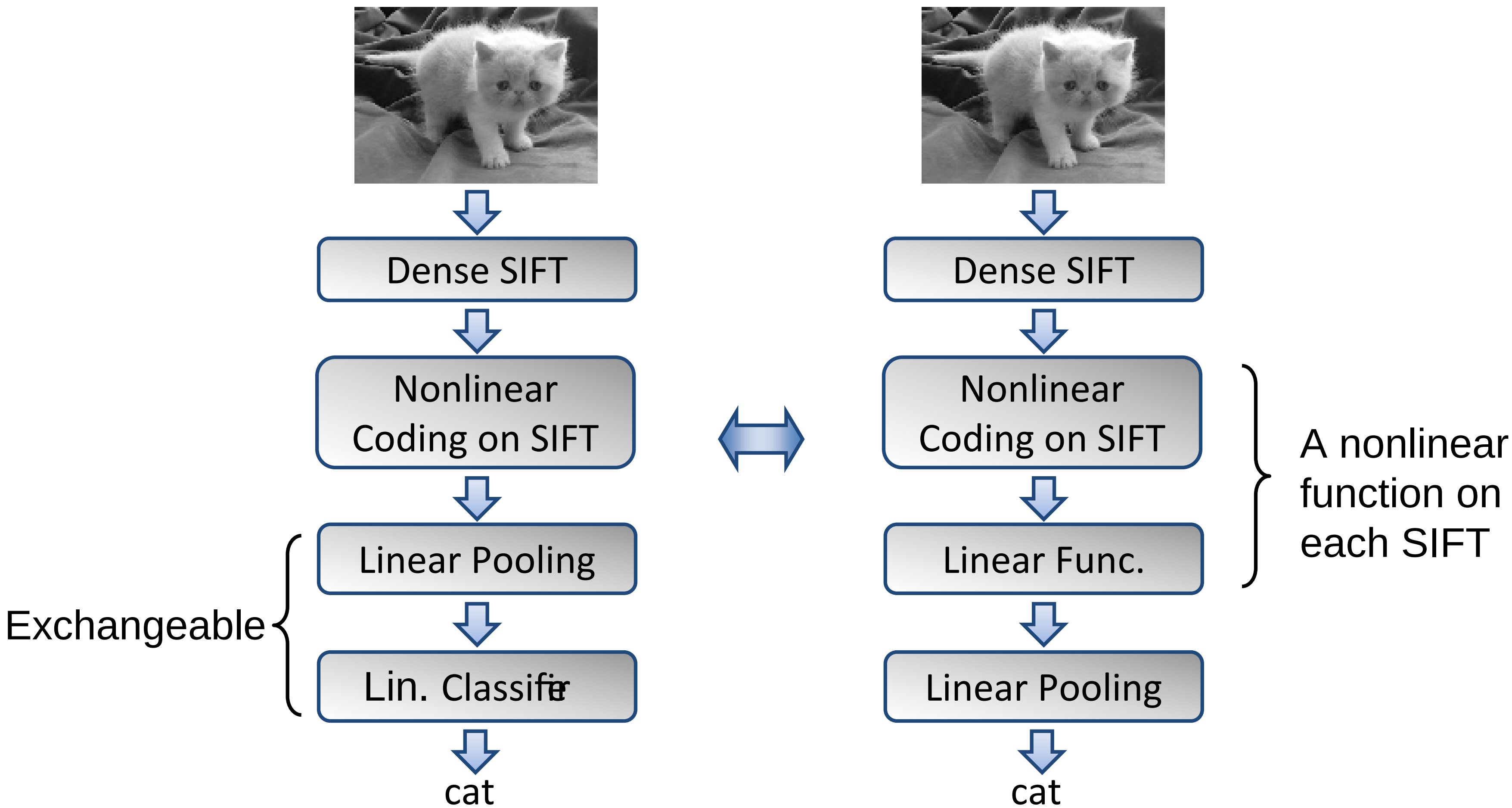
Note: 1. Overall AP is around 58.0%; 2. Overall AP is around 46% (estimation based on 5-fold cross validation)

Prior Publications

- Local Coordinate Coding
  - Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification
    Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang, CVPR 2009
  - Nonlinear Learning using Local Coordinate Coding
    Kai Yu, Tong Zhang, and Yihong Gong, NIPS 2009, to appear
- GMM
  - Hierarchical Gaussianization for Image Classification
    Xi Zhou, Na Cui, Zhen Li, Feng Liang, and Thomas S. Huang, ICCV 2009
  - SIFT-Bag Kernel for Video Event Analysis
    Xi Zhou, Xiaodan Zhuang, Shuicheng Yan, Shih-Fu Chang, Mark Hasegawa-Johnson, Thomas S. Huang, ACM Multimedia 2008

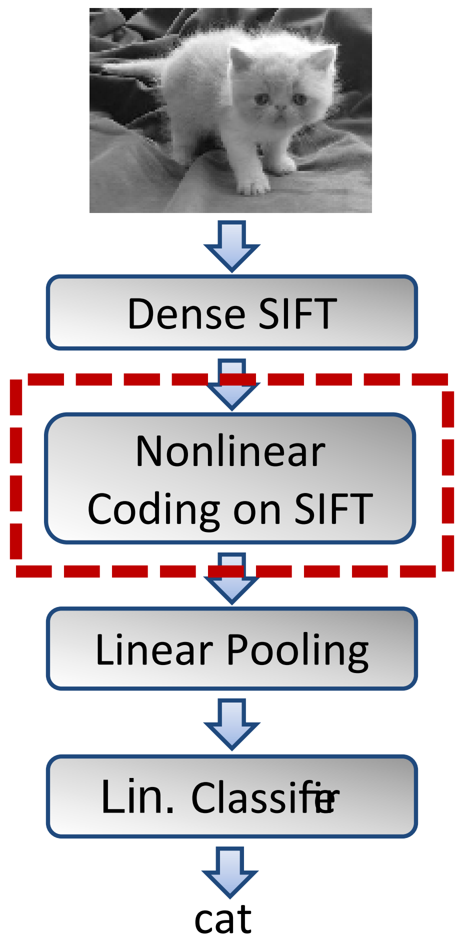
In our work on PASCAL challenge, we made further extensions of the above work in both engineering and theory.

A Unified Framework



- What matters is to learn nonlinear function on SIFT vectors.
- This boils down to learning a good coding scheme of SIFT.

Coding of SIFT



Some Notation

$X \in \mathbb{R}^D$ a SIFT feature vector

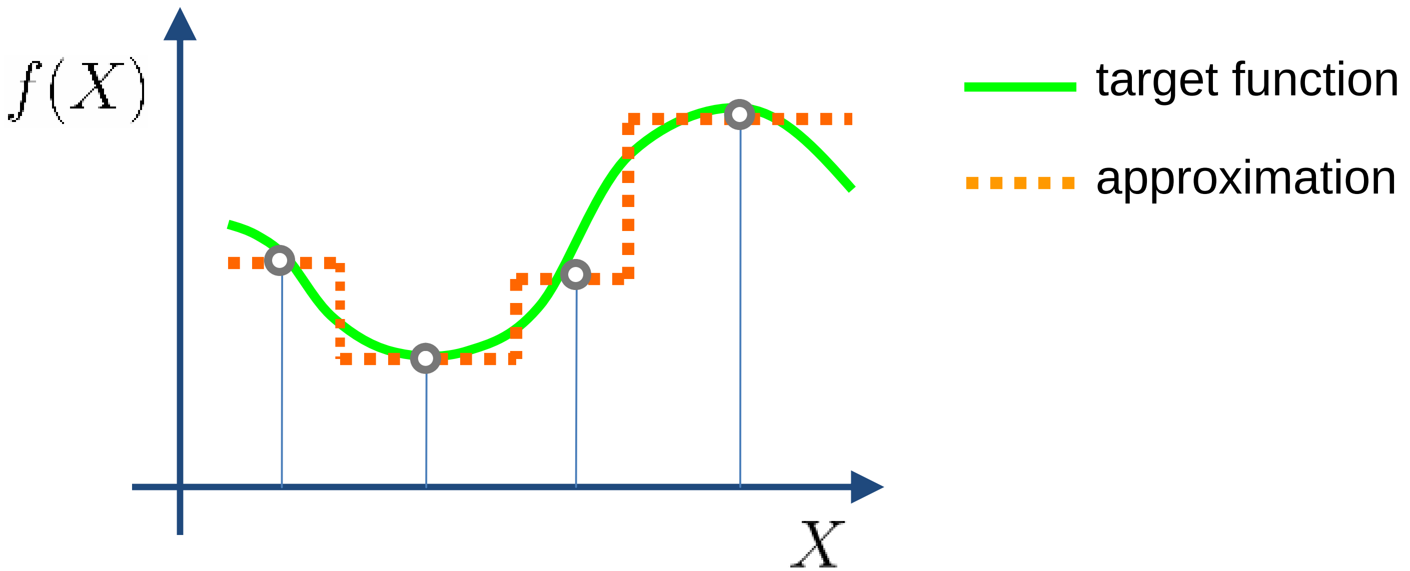
$\Phi(X) : \mathbb{R}^D \rightarrow \mathbb{R}^L$ encoding function

$f(X) : \mathbb{R}^D \longrightarrow \mathbb{R}$ unknown function on local features

$\hat{f}(X) = W^\top \Phi(X)$ approximating function

Supervised Learning (↑ W)   Unsupervised Learning (↑ Φ)

Example 1: Vector Quantization Coding (VQ)



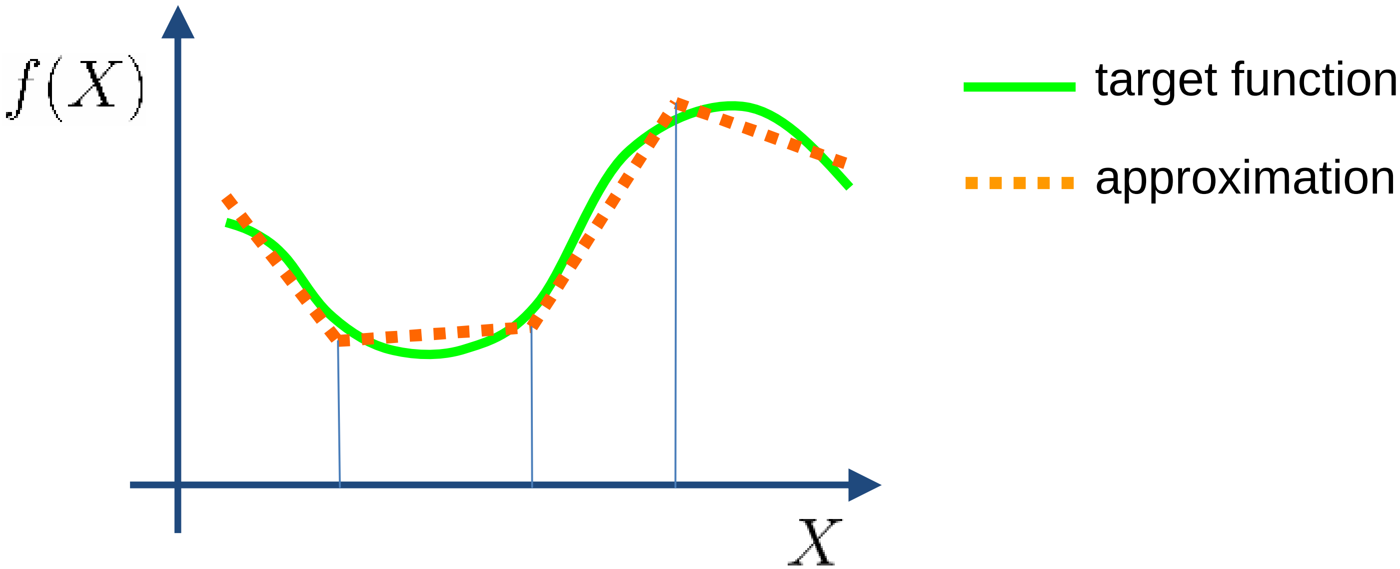
- The approximating function is

$$\hat{f}(X) = W^\top \Phi(X),$$

where $W = [W_1, W_2, \ldots, W_K]^\top$, $\Phi(X)$ is the code of X.

- If X belongs to class 2, $\Phi(X) = [0, 1, 0, \ldots, 0]^\top$, then $\hat{f}(X) = W^\top \Phi(X) = W_2$.

Example 2: "Supervector" Coding



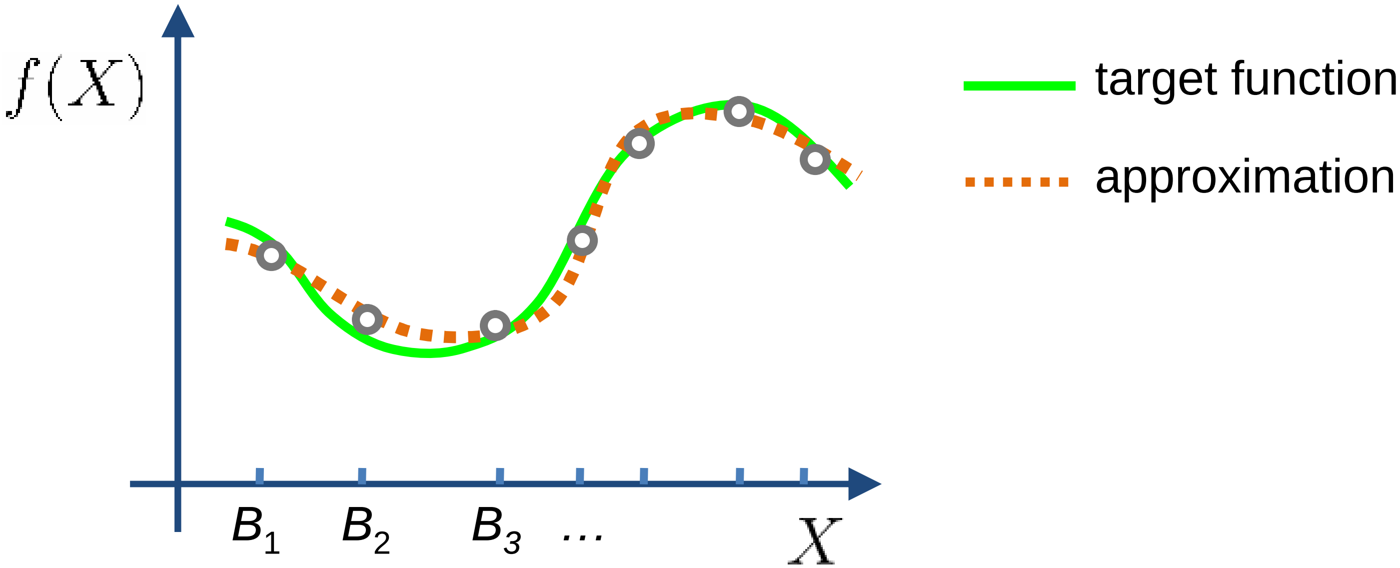
- Given K clusters in X space, let $W = [W_1^\top, W_2^\top, \ldots, W_K^\top]^\top$, where $W_k \in \mathbb{R}^D$, and

$$\Phi(X) = [C_1(X) * X^\top, C_2(X) * X^\top, \ldots, C_K(X) * X^\top]^\top,$$

with $C_k(X) = 1$ if X belongs to cluster k, otherwise $C_k(X) = 0$.

- Then $\hat{f}(X) = W^\top \Phi(X) = \sum_k C_k(X) * W_k^\top X$. — a **locally piecewise linear function**
- $C_k(X)$ can be soft probability given by GMM, then $\Phi(X)$ is **GMM supervector**.

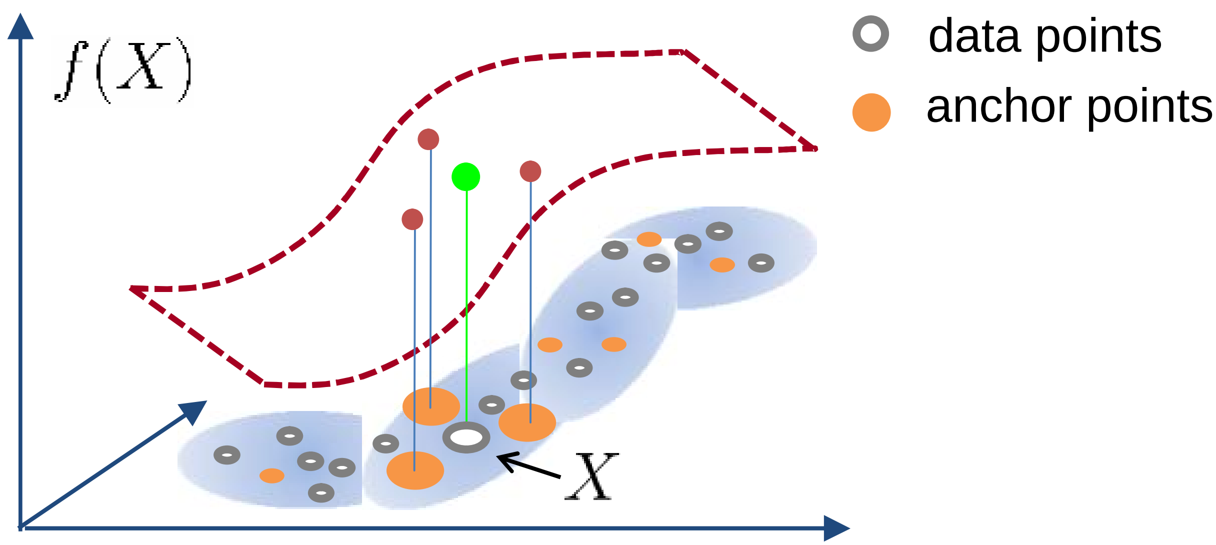
Example 3: Local Coordinate Coding



- Given **anchor points** $[B_1, \ldots, B_K]$, if the coding scheme $\Phi(X) = [\phi_1, \ldots, \phi_K]$ satisfies
  1. **low reconstruction error**: $X \approx \sum_{k=1}^{K} \phi_k B_k$;
  2. **good locality**: ϕ_k tends to be nonzero if B_k is in X's neighborhood, otherwise 0.
- Then $\hat{f}(X) = W^\top \Phi(X)$ provides a close approximation to $f(X)$.

LCC: How It Works

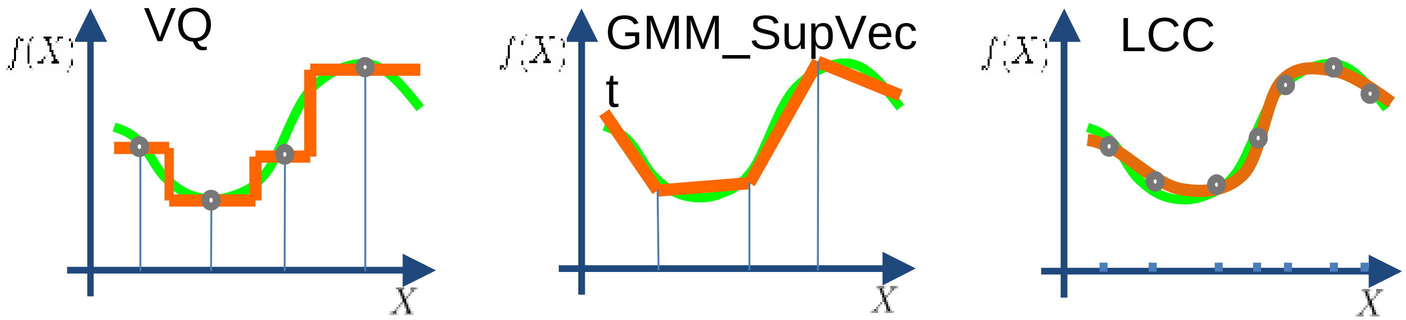
$$\hat{f}(X) = \sum_{k=1}^{K} \Phi_k W_k = \sum_{k=1}^{K} \Phi_k \hat{f}(B_k)$$ forms a **local interpolation**



$$\Phi(X) = \arg\max_{\Phi} \left\| X - \sum_{k=1}^{K} \Phi_k B_k \right\|^2 + \lambda \sum_k \alpha_k(X) |\Phi_k|$$

where $\alpha_k(X)$ is a distance from X to B_k

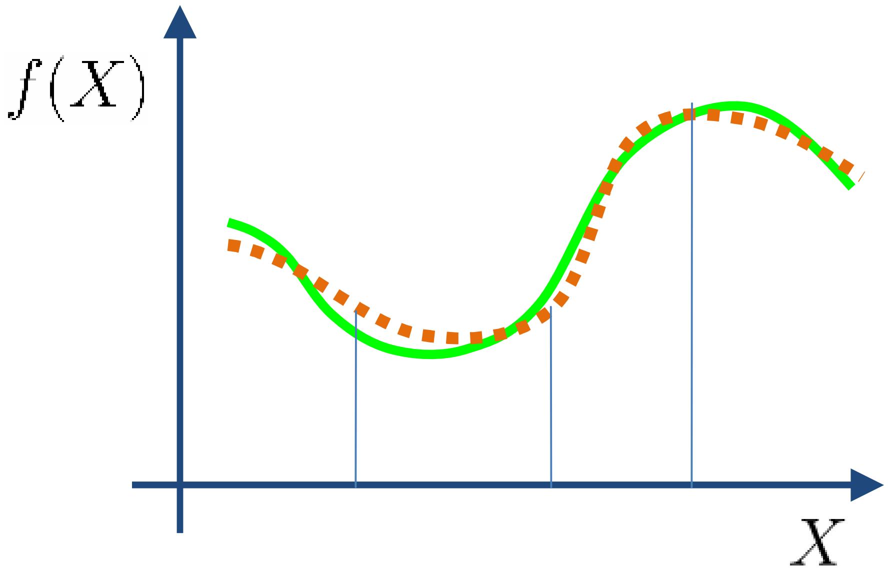
Comparison of Coding Methods



Function Approximation	Poor	Good	Excellent
Computation	Low	Medium	High
Locality	Yes	Yes	Yes
Caltech-101	~65% [1]	~73% [2]	~73% [3]
		⇩ Improve its fitting power	⇩ Reduce its Computation

1. Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, CVPR, 2006
2. Xi Zhou, Na Cui, Zhen Li, Feng Liang, and Thomas S. Huang, ICCV, 2009
3. Jianchao Yang, Kai Yu, Yihong Gong, and Thomas S. Huang, CVPR, 2009

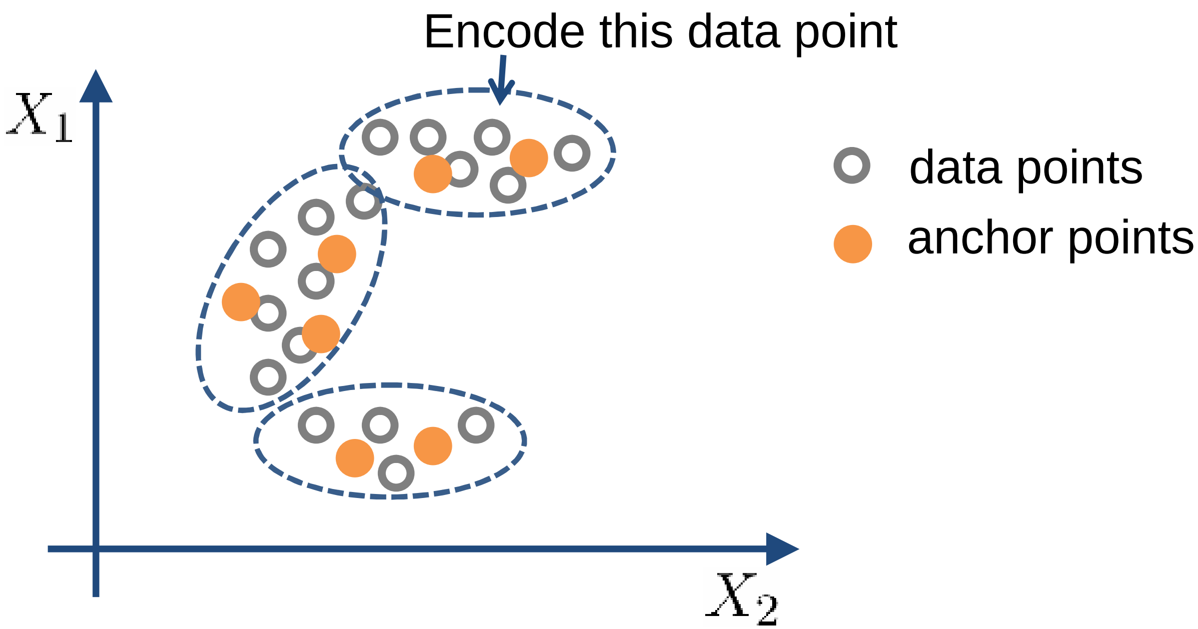
Improve GMM Supervector Coding



- "local linear" $\rightarrow$ "local nonlinear"
- the code of X is

$$\Phi(X) = \left[C_1(X) * (X, X^2)^\top, \ldots, C_K(X) * (X, X^2)^\top\right]$$

Improve LCC's Efficiency



- Pre-computation: partition data and anchor points
- Eliminate those anchor points in different partitions

Equivalent to "Mixture of Coding Experts"

- Use a **soft-max gating function** $G_k(X)$ indicating if X is in local partition k.
- Optimize the following cost

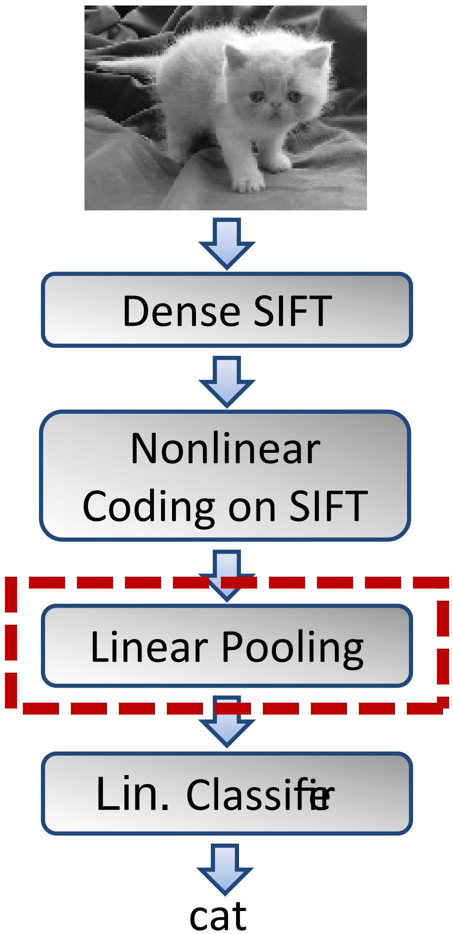
$$\Phi(X) = \arg\min_{\Phi} \sum_{k=1}^{K} G_k(X) \left(\left\| X - \sum_{m=1}^{M} \Phi_m^{(k)} B_m^{(k)} \right\|^2 + \lambda \sum_{m} \left| \Phi_m^{(k)} \right| \right)$$

- This is equivalent to

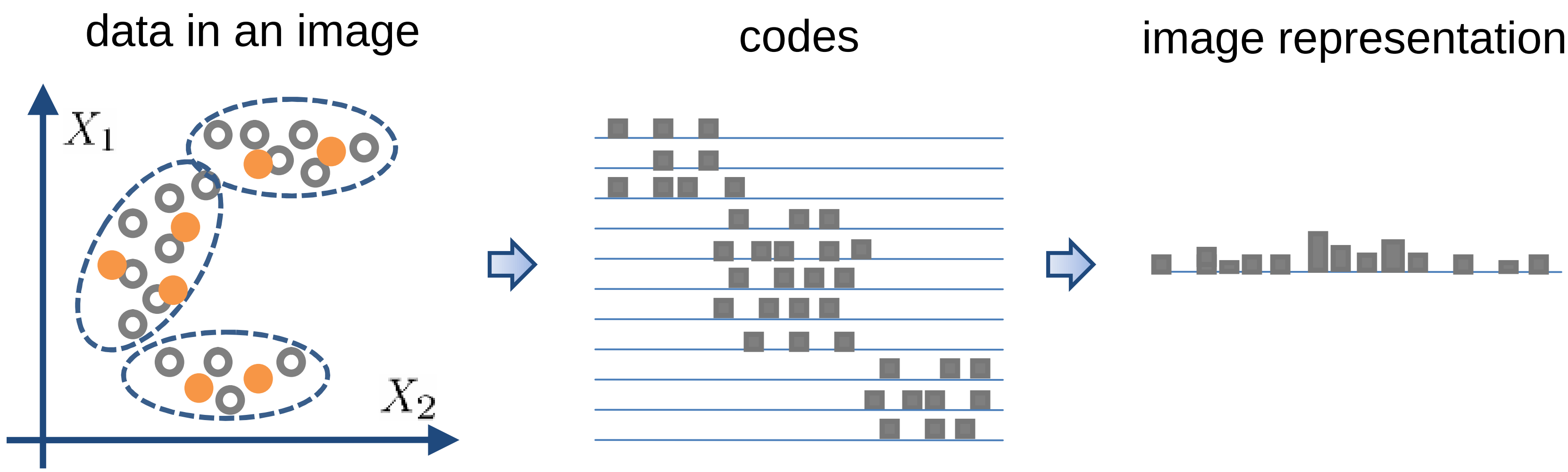
$$\Phi(X) = \arg\max_{\Phi} \left\| X - \sum_{k=1}^{M*K} \Phi_k B_k \right\|^2 + \lambda \sum_{k=1}^{M*K} \alpha_k(X) |\Phi_k|$$

where $\alpha_k(X)$ is 1 if X and B_k belong to the same partition, overwise $+\infty$.

Linear Pooling



(Local) Linear Pooling



$$Z_I^{(k)} = \frac{\sum_{i \in I} G_k(X_i) \Phi_{\text{nmlz}}^{(k)}(X_i)}{\sqrt{\sum_{j \in I} G_k(X_j)}}$$

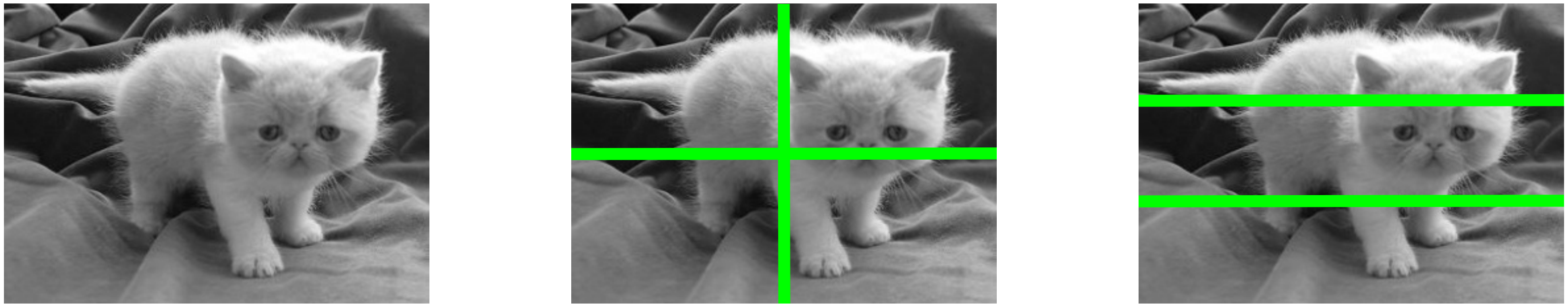
where $\Phi_{\text{nmlz}}^{(k)}(X)$ is the normalized version of $\Phi^{(k)}(X)$, obtained by subtracting mean and then dividing by variance.

- The classification function on image I is

Nonlinear function on local features

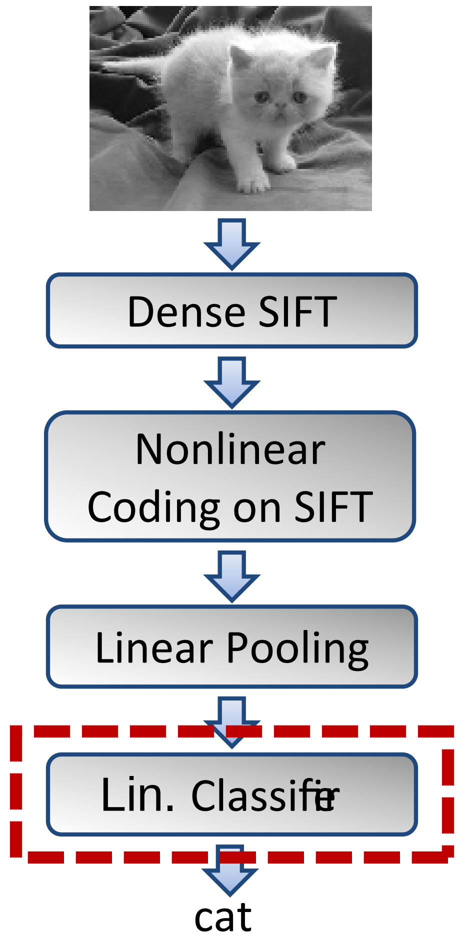
$$c(I) = \sum_{k=1}^{K} {W^{(k)}}^\top Z_I^{(k)} = \sum_{k=1}^{K} \frac{\sum_{i \in I} G_k(X_i) {W^{(k)}}^\top \Phi_{\text{nmlz}}^{(k)}(X_i)}{\sqrt{\sum_{j \in I} G_k(X_j)}} = \sum_{k=1}^{K} \frac{\sum_{i \in I} G_k(X_i) f^{(k)}(X_i)}{\sqrt{\sum_{j \in I} G_k(X_j)}}$$

SPM representation



See also in “SurreyUVA_SRKDA method”, presentatin at PASCAL VOC workshop 08.

Linear Classifier



Support Vector Machines

- Use our own implementation, training using gradient based method LBFGS.

$$\min_{W} \left\{ J(W) = ||W||^2 + C \sum_{i=1}^{n} \ell\left(W; Y_i, Z_i\right) \right\}$$

.

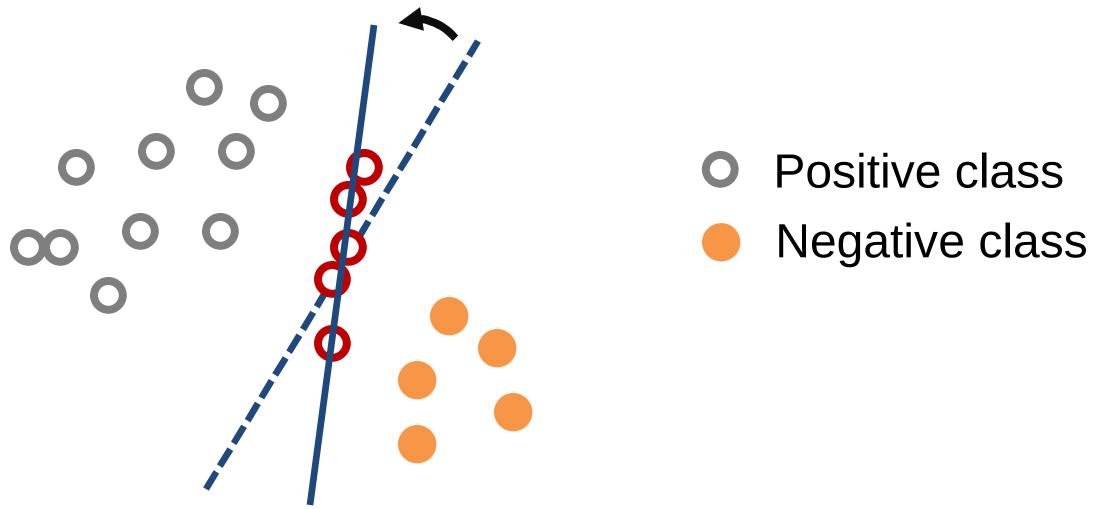
- Use a differentiable hinge loss

$$\ell\left(W; Y_i, Z_i\right) = \left[\max\left(0, W^\top Z_i \cdot Y_i - 1\right)\right]^2$$

Universum SVMs

- Use the **Universum approach**: if image i is a difficult case, let the loss be

$$\ell(W; Y_i, Z_i) = (W^\top Z_i)^2$$



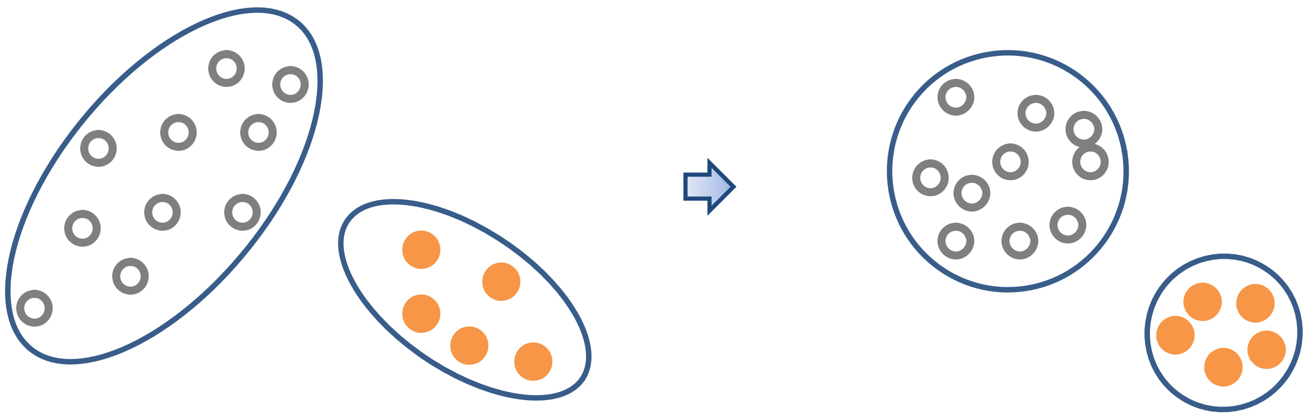
Vapnik, V. N. (1998). Statistical learning theory. New York: Wiley.
Weston, Inference with Universum, ICML 2006

Within-class Covariance Normalization

- Within-class normalization

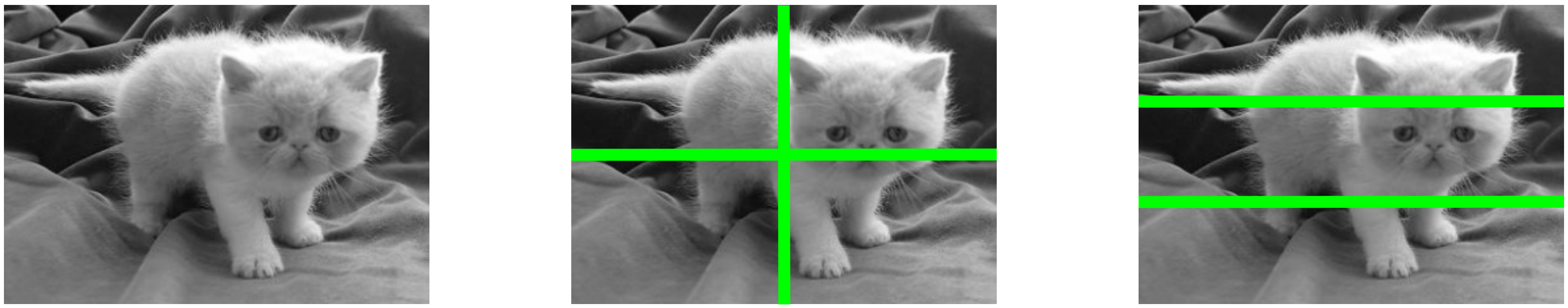
$$K_{i,j} = Z_i^\top (\gamma S + (1-\gamma) I)^{-1} Z_j$$

where S is the average within-class covariance matrix.



Vapnik, V. N. (1998). Statistical learning theory. New York: Wiley.
Weston, Inference with Universum, ICML 2006

Improve SPM using Gaussian Process



- The SPM approach uses 8 linear kernels.
- We can learn the kernel weights.

$$\min_{\{\alpha_s \geq 0\}} -\log P\left(Y \,\middle|\, \sum_{s=1}^{8} \alpha_s K_s\right) + \lambda \sum_{s=1}^{8} (\alpha_s - \alpha_0)^2$$

.

- We learn a set of global weights for all classes.

Some Details

- Number of partitions or components
  - GMM: 1024 and 2048
  - LCC: 1024 and 2048

- Dimensionality of feature vector for each image (e.g. in case of 1024 partitions)
  - GMM: 1024x80x8 (1024 components, 80 PCA-SIFT, 8 SPM sub kernels)
  - LCC: 1024x256x8 (1024 partitions, 256 codebook size, 8 SPM sub kernels)

Conclusion Remarks

- Highly nonlinear, highly local encoding of image local features make difference!

- Still a long way to go
  - No high-level (semantic) features used so far
  - how to get compact image representations?
  - Supervised training of coding schemes
  - Better methods to use the bounding box information

- More details will be provided in forthcoming TR and papers.