第181回コンピュータビジョンとイメージメディア研究発表会　2012年3月15日

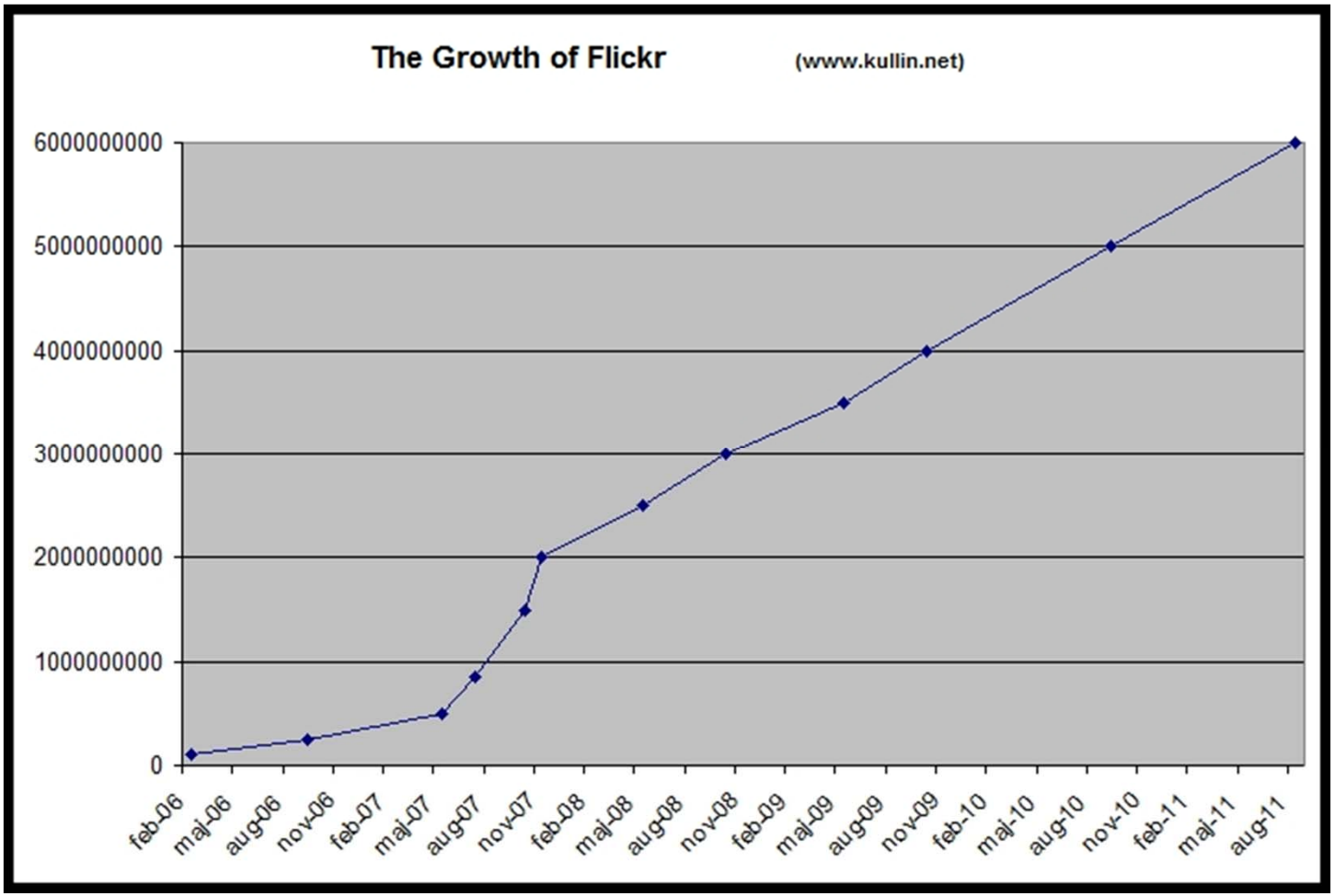
大規模データを用いた
一般物体・シーン認識の潮流と理論

東京大学/JSTさきがけ
原田達也

- Russell A. Kirsch (1929) led a team of colleagues in creating America's first internally programmable computer, the Standards Eastern Automatic Computer (SEAC), capable of scanning digital images in 1957.
- SEAC produced a photograph of Kirsch's three month old son in a mere 176 pixels, measuring 5x5cm. Because of this breakthrough, satellite imaging, CAT scans, bar codes, and desktop publishing were made possible.
- http://en.wikipedia.org/wiki/Russell_A._Kirsch


Flickr reaches 6 billion photos on 1 Aug, 2011.
http://www.flickr.com/photos/eon60/6000000000/

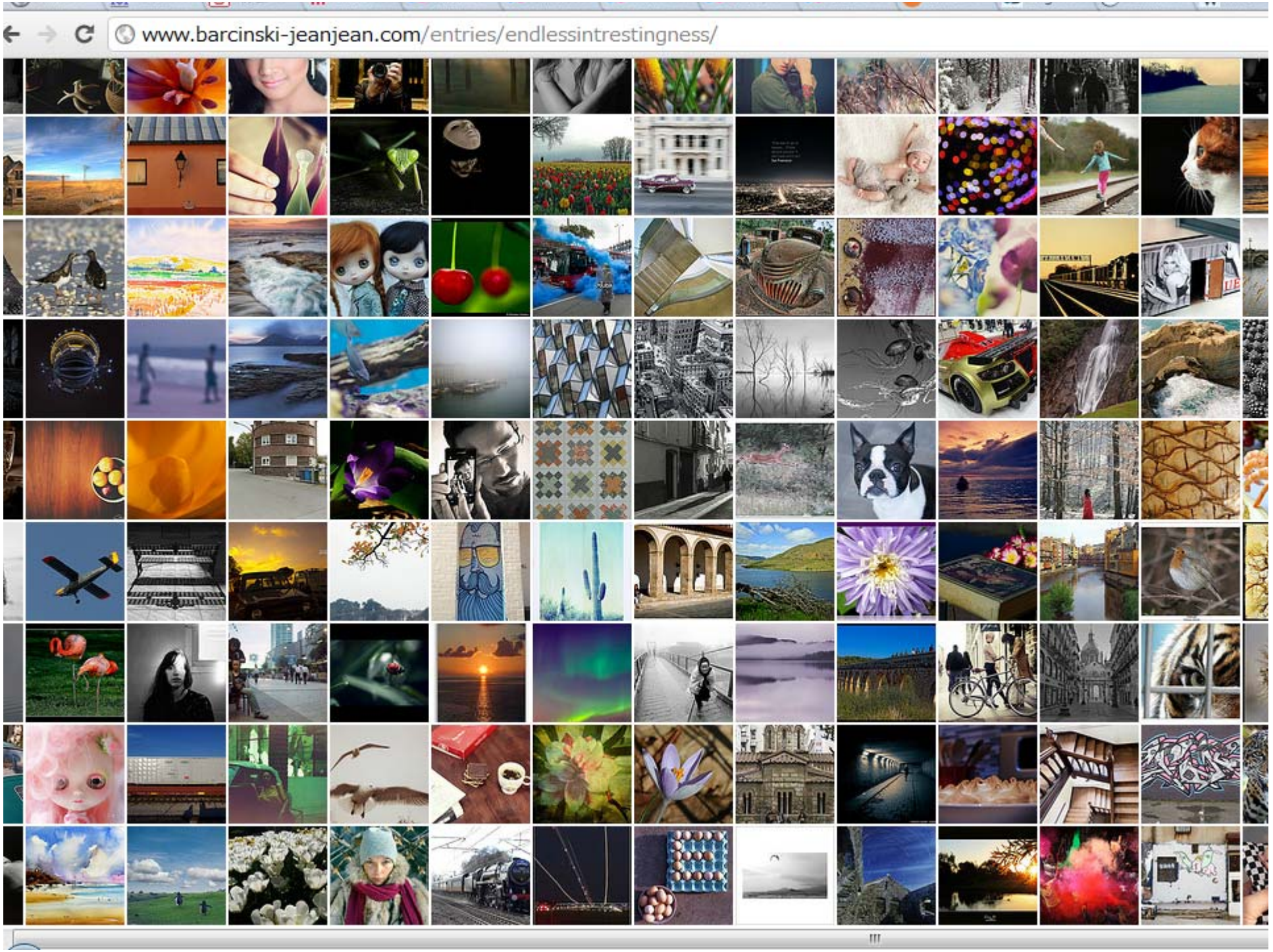
The Growth of Flickr



- http://www.kullin.net/2011/08/flickr-reaches-6-billion-photos/

Statistics	Voting Breakdown (your vote is highlighted in red)
Place: 1 out of 83 **Avg (all users):** 7.3200 **Avg (commenters):** 8.2000 **Avg (participants):** 6.9020 **Avg (non-participants):** 7.5354 **Views since voting:** 2766 **Views during voting:** 330 **Votes:** 150 **Comments:** 64 **Favorites:** 13 (view)	1 1 2 0 3 1 4 2 5 10 6 31 7 39 8 33 9 17 10 16

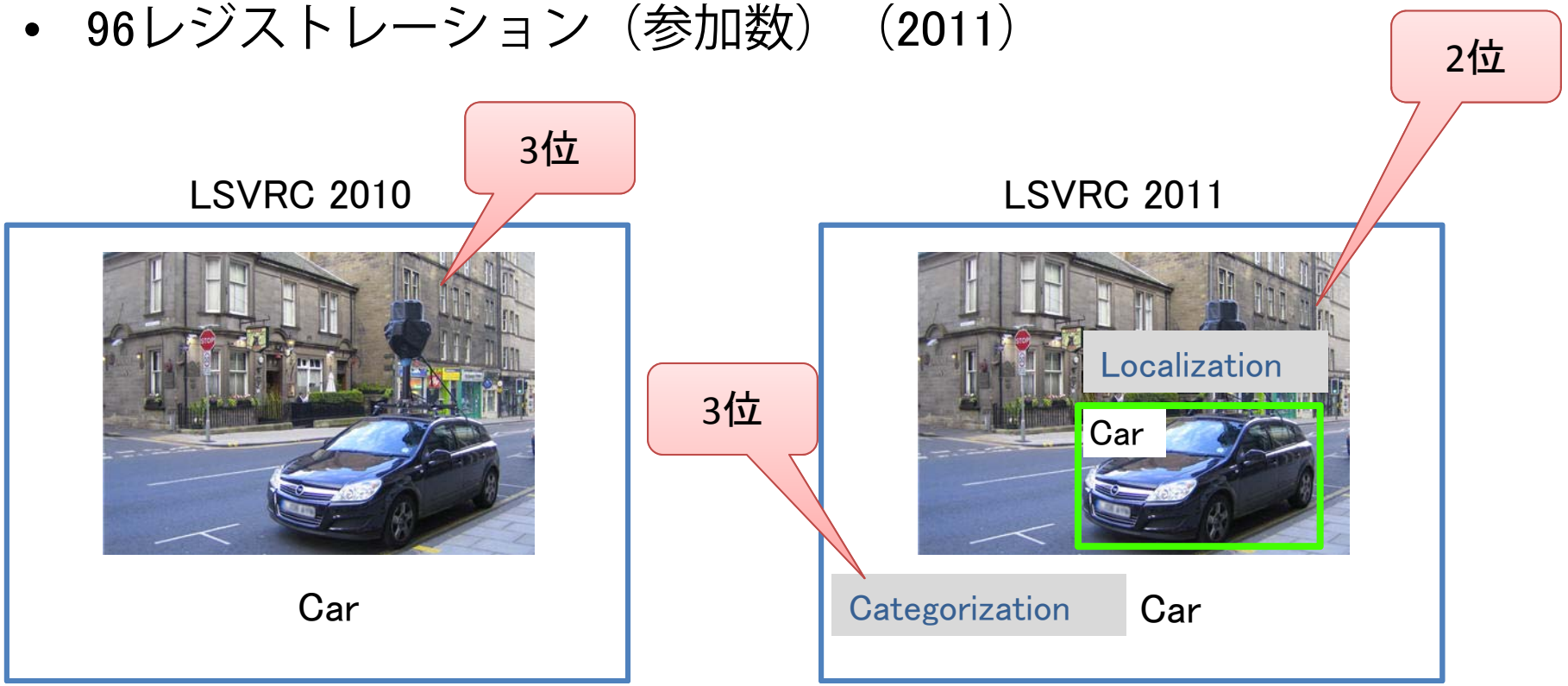
http://www.dpchallenge.com/image.php?IMAGE_ID=997702


www.barcinski-jeanjean.com/entries/endlessintrestingness/

大規模データを用いた画像認識

http://www.image-net.org/challenges/LSVRC/2011/pascal_ilsvrc_2011.pptx

- 120万枚の訓練画像
- 10万枚のテスト画像（2011）
- 1000種類のクラス
- 物体識別タスク：画像に何が写っているのか？
- 物体検出タスク：画像のどこに写っているのか？
- 96レジストレーション（参加数）（2011）



Results

1. neck brace
2. bullet train
3. potter's wheel
4. seat belt
5. barbell

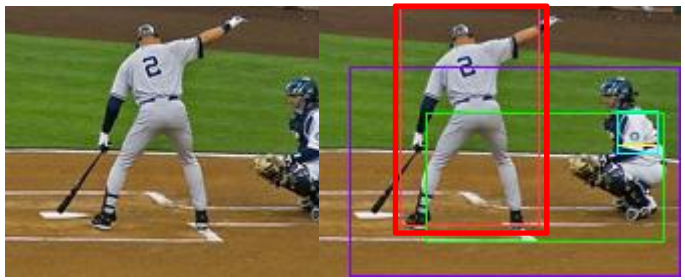
1. brown bear
2. otter
3. hippopotamus
4. raccoon
5. deerhound

1. mountain bike
2. hartebeest
3. yurt
4. bighorn
5. coho

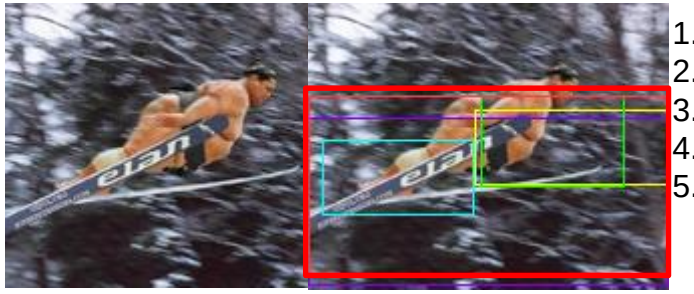
1. volleyball
2. bittern
3. shower curtain
4. crane
5. suspension bridge

1. mask
2. ski mask
3. jack-o'-lantern
4. jellyfish
5. teddy bear

1. toilet seat
2. scanner
3. hard disc
4. scale
5. backpack



1. baseball player
2. racket, racquet
3. solar dish
4. trimaran
5. paddle

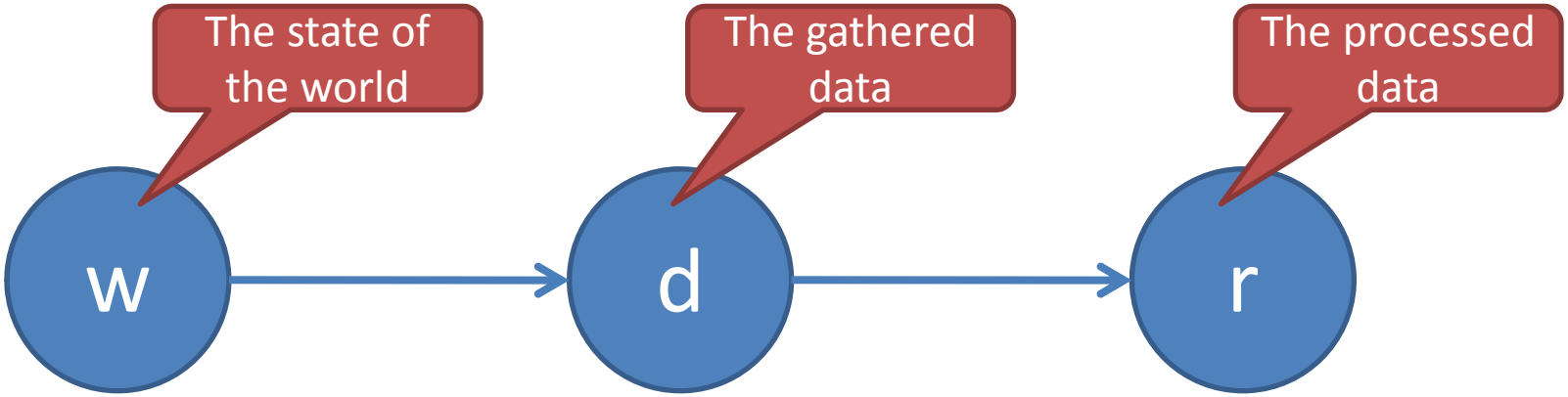


1. aircraft carrier
2. paddle
3. bullfrog
4. water ouzel
5. mantis

画像認識の分類

- 特定物体認識，Specific Object Recognition
  - データベースには認識対象とする物体の画像をすでに持つことを前提として，入力画像に写る物体とデータベース内の画像を照合すること
- 一般物体認識，Generic Object Recognition
  - データベースに存在しない入力画像の物体のカテゴリを予測すること
- 画像アノテーション，Image Annotation
  - 狭義：複数ラベルが付与された画像データセットから，入力画像に複数のラベルを付与すること
  - 広義：特定物体認識，一般物体認識を含む広い概念

The data processing theorem



Markov chain

$$P(w,d,r) = P(w)P(d\,|\,w)P(r\,|\,d)$$

The average information

$$I(W;D) \geq I(W;R)$$

The data processing theorem states that data processing can only destroy information.

David J.C. MacKay. Information Theory, Inference, and Learning Algorithms. Cambridge University Press 2003.

画像認識のプロセス

訓練時

識別時

- 処理を重ねる毎にデータの持つ情報は減少する.
  - データ，特徴抽出，モデルの順に高い質が求められる.
- 従来の画像認識研究の多くはモデル化に重点が置かれていた
  - 小さな実験環境，スモールワールド
- 複雑なモデルは大規模データの前では役に立たない
  - スケーラビリティの重要性
- 高い質のデータ，特徴抽出が適切に行われていればシンプルなモデルで十分な性能が出せる

知識はどこから降ってくる？

- 膨大な情報を利用
  - インターネット上の大規模画像
  - リッチな画像表現，効率な線形識別機の学習

- 人に尋ねる
  - クラウドソーシング
  - 能動学習

- 対象とは別の知識を活用
  - 転移学習，ドメイン適合，マルチタスク学習
  - ゼロショット学習
  - アトリビュート

知識はどこから降ってくる？

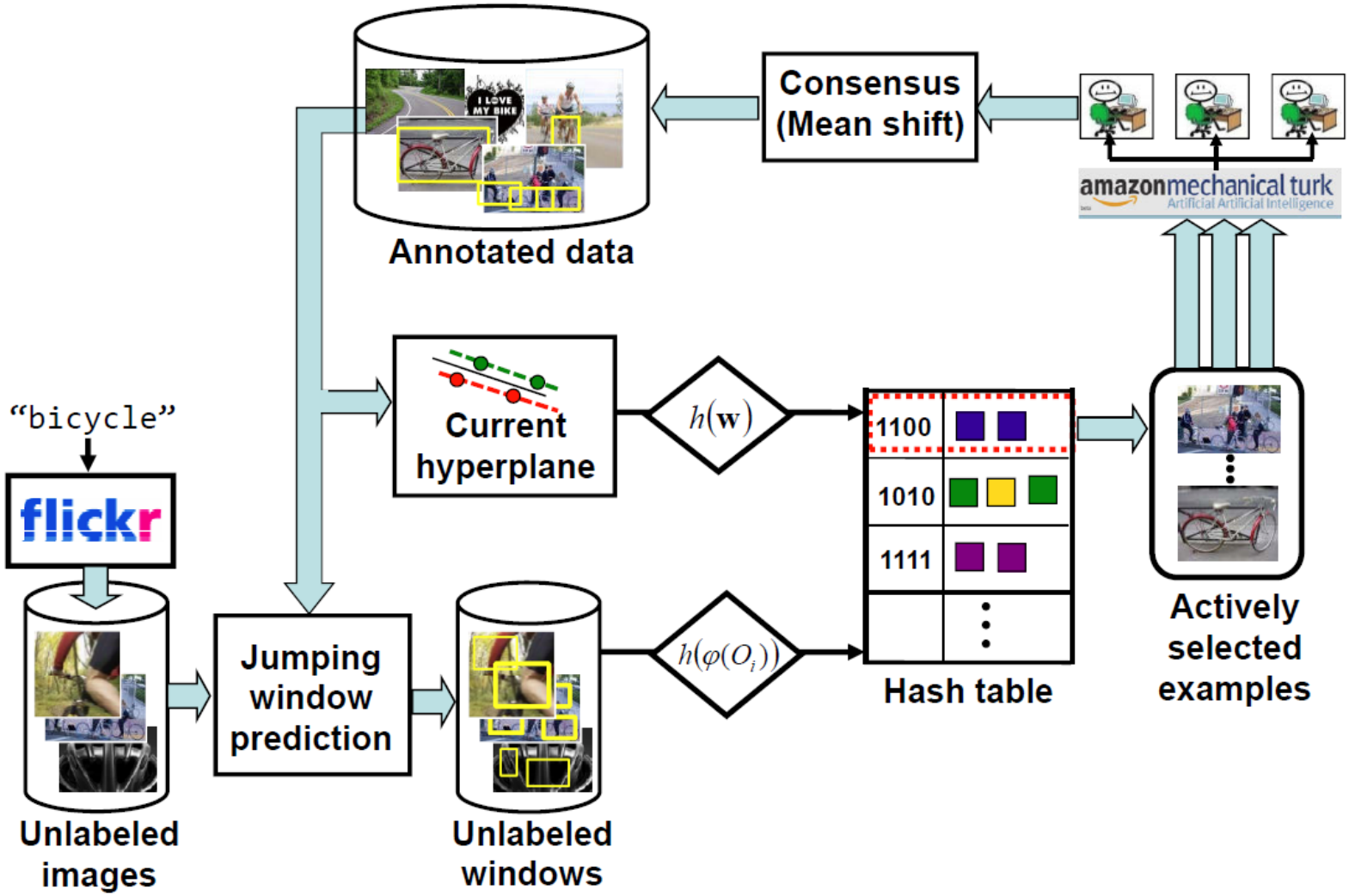
- 膨大な情報を利用
  - インターネット上の大規模画像
  - リッチな画像表現，効率な線形識別機の学習

- 人に尋ねる
  - クラウドソーシング
  - 能動学習

- 対象とは別の知識を活用
  - 転移学習，ドメイン適合，マルチタスク学習
  - ゼロショット学習
  - アトリビュート

人に尋ねる

S. Vijayanarasimhan and K. Grauman. Large-Scale Live Active Learning: Training Object Detectors with Crawled Data and Crowds. In CVPR, 2011.

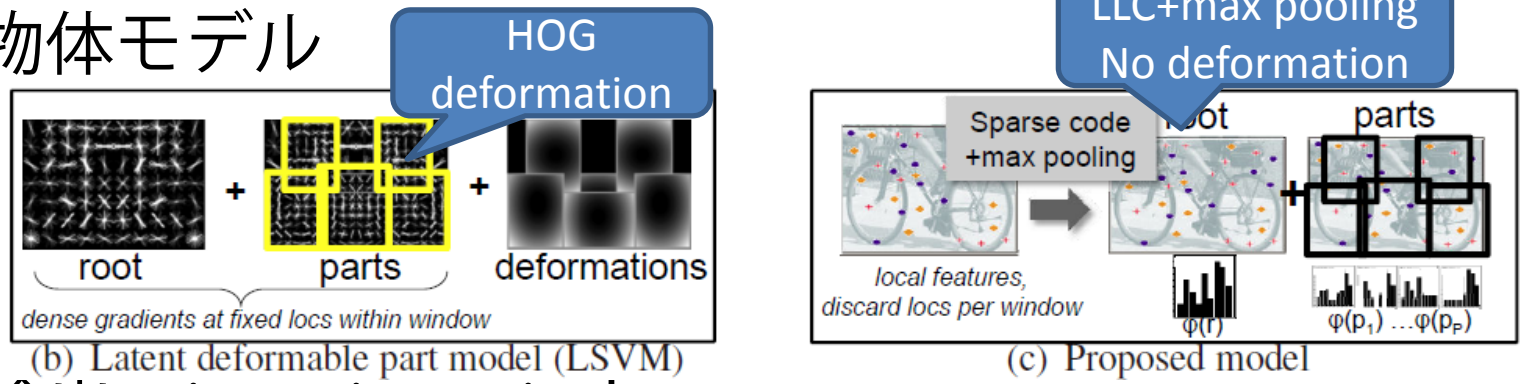


- スケーラブルかつ自動的な物体検出のオンライン学習
- 高速・高性能な物体検出
- 高速かつ適切なクラウドソーシング対象画像選択

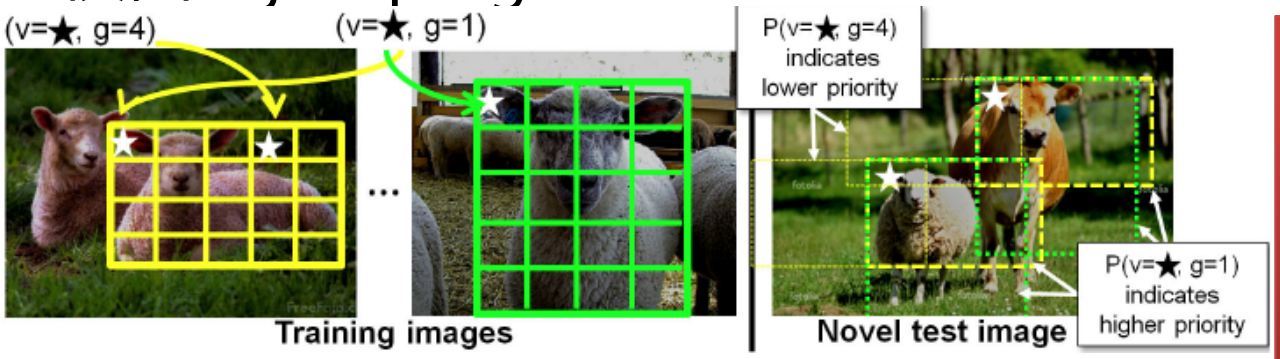
人に尋ねる

S. Vijayanarasimhan and K. Grauman. Large-Scale Live Active Learning: Training Object Detectors with Crawled Data and Crowds. In CVPR, 2011.

- ディテクションの高速化
  - 物体モデル



  - 検出：jumping window



訓練データで，root windowにおけるVisual word (v)と位置(g)の共起確率P(v,g)を求める．

P(v,g)の高い部分のみディテクションを行う．

- クラウドソーシングする画像選択
  - Hyperplane-hashing

NIPS2010

Hyperplaneを入力として，これに近いサンプル群がなるべく同じhash値になるようにする．

$$h_{\mathcal{H}}(\boldsymbol{z}) = \begin{cases} h_{\boldsymbol{u},\boldsymbol{v}}(\phi(O_i), \phi(O_i)), & \text{if } \boldsymbol{z} \text{ is a database vector,} \\ h_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{w}, -\boldsymbol{w}), & \text{if } \boldsymbol{z} \text{ is a query hyperplane,} \end{cases} \qquad h_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{a}, \boldsymbol{b}) = [\text{sign}(\boldsymbol{u}^T \boldsymbol{a}), \text{sign}(\boldsymbol{v}^T \boldsymbol{b})],$$

知識はどこから降ってくる？

- 膨大な情報を利用
  - インターネット上の大規模画像
  - リッチな画像表現，効率な線形識別機の学習

- 人に尋ねる
  - クラウドソーシング
  - 能動学習

- 対象とは別の知識を活用
  - 転移学習，ドメイン適合，マルチタスク学習
  - ゼロショット学習
  - アトリビュート

アトリビュート

アトリビュート（Attribute）

- 物体カテゴリ間で共有される人間が理解可能な属性

- アトリビュートの主な適応先の分類
  1. 一般もしくは見慣れない物体の記述
  2. Zero-shot認識，知識転移，転移学習
  3. 物体識別を補助する中間特徴

S. J. Hwang, F. Sha, and K. Grauman. Sharing Features Between Objects and Their Attributes. CVPR, 2011.

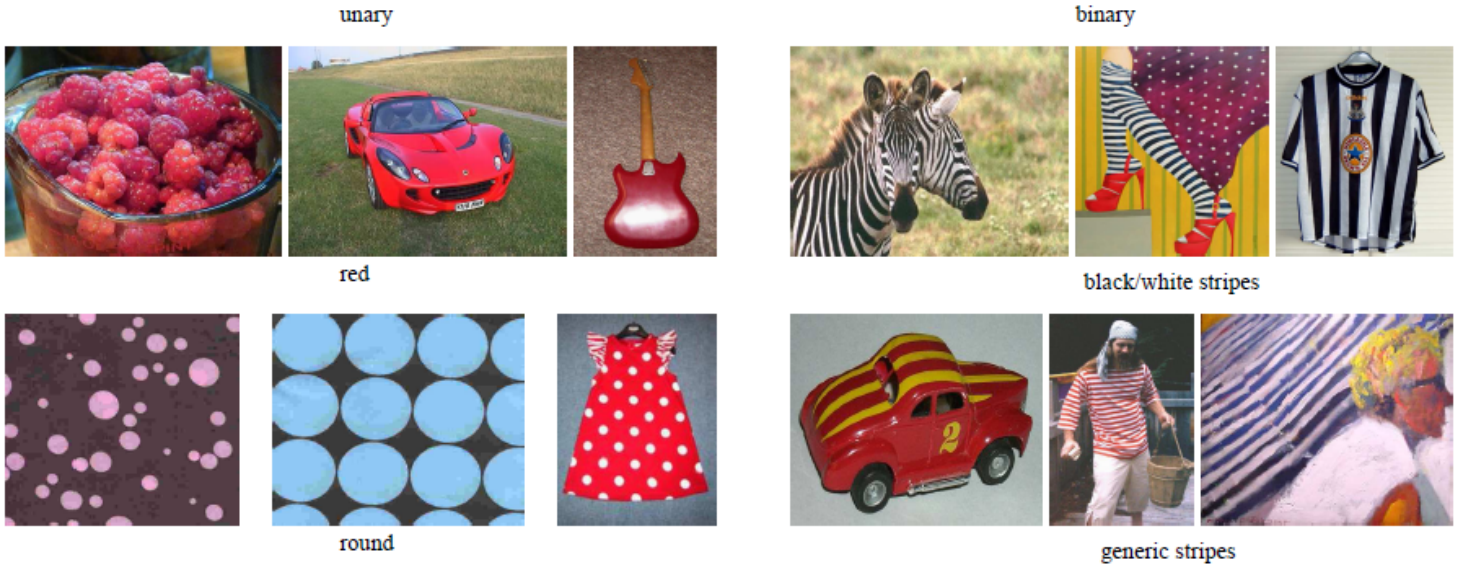


Figure 1: *Examples of different kinds of attributes. On the left we show two simple attributes, whose characteristic properties are captured by individual image segments (appearance for red, shape for round). On the right we show more complex attributes, whose basic element is a pair of segments.*

V. Ferrari and A. Zisserman. Learning visual attributes. In NIPS, 2007.

ICCV2011 program at glance

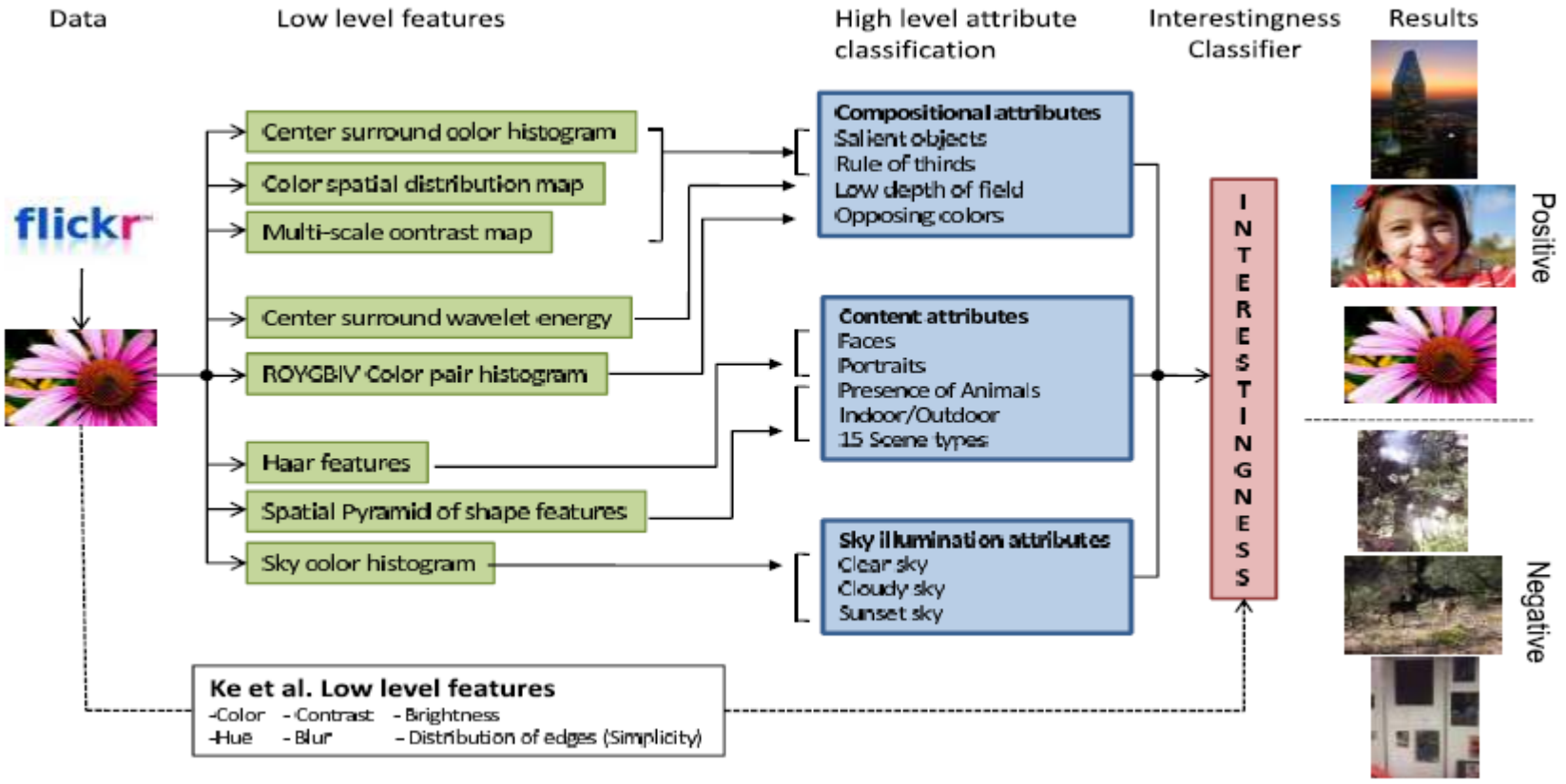
Attributes and Classification

ICCV2011 at a glance

Nov. 6, Sunday	Nov. 7, Monday	Nov. 8, Tuesday	Nov. 9, Wednesday	Nov. 10, Thursday	Nov. 11, Friday	Nov. 12, Saturday	Nov. 13, Sunday
		08:00 **Registration**	08:00 **Registration**	08:00 **Registration**	08:00 **Registration**		
08:30 **Registration**	08:30 **Registration**	08:30-09:20 **Opening Remarks**	08:30-09:30 **Invited Speaker**	08:30-09:30 **Invited Speaker**	08:30-09:30 **Invited Speaker**	08:30 **Registration**	08:30 **Registration**
09:00-19:00 **Tutorials**	09:00-19:00 **Workshops**	09:20-10:20 **Awards Ceremony**	09:30-10:30 **Session 2-1** Illumination and Reflectance	09:30-10:30 **Session 3-1** Optimization Methods	09:30-10:30 **Session 4-1** Matching	09:00-19:00 **Workshops**	09:00-19:00 **Workshops**
		10:20-10:50 **Coffee Break**	10:30-11:00 **Coffee Break**	10:30-11:00 **Coffee Break**	10:30-11:00 **Coffee Break**		
		10:50-12:30 **Session 1-1** Recognition	11:00-12:45 **Session 2-2** Activity Recognition	11:00-12:45 **Session 3-2** Geometric Computer Vision	11:00-12:45 **Session 4-2** Motion and Tracking		
		30-14:00 **Lun**	12:45-14:10 **Lunch**	12:45-14:10 **Lunch**	12:45-14:10 **Lunch**		
		14:00-15:25 **Session 1-2** Statistical Methods and Learning	14:10-15:35 **Session 2-3** Attributes and Classification	14:10-15:35 **Session 3-3** Scene Understanding	14:10-15:35 **Session 4-3** Image Processing		
		15:25-15:55 **Coffee Break**	15:35-15:55 **Coffee Break**	15:35-15:55 **Coffee Break**	15:35-15:55 **Coffee Break**		
		15:55-17:20 **Session 1-3** Detection and Categorization	15:55-17:20 **Session 2-4** Segmentation and Grouping	15:55-17:20 **Session 3-4** Image Restoration and Retargeting	15:55-17:20 **Session 4-4** Faces		
		17:20-20:00 **Poster Session 1**	17:20-20:00 **Poster Session 2**	17:20-20:00 **Poster Session 3**	17:20-20:00 **Poster Session 4**		

アトリビュートの美と魅力予測への応用

- Sagnik Dhar Vicente Ordonez Tamara L Berg. High Level Describable Attributes for Predicting Aesthetics and Interestingness. CVPR, 2011.



データセット

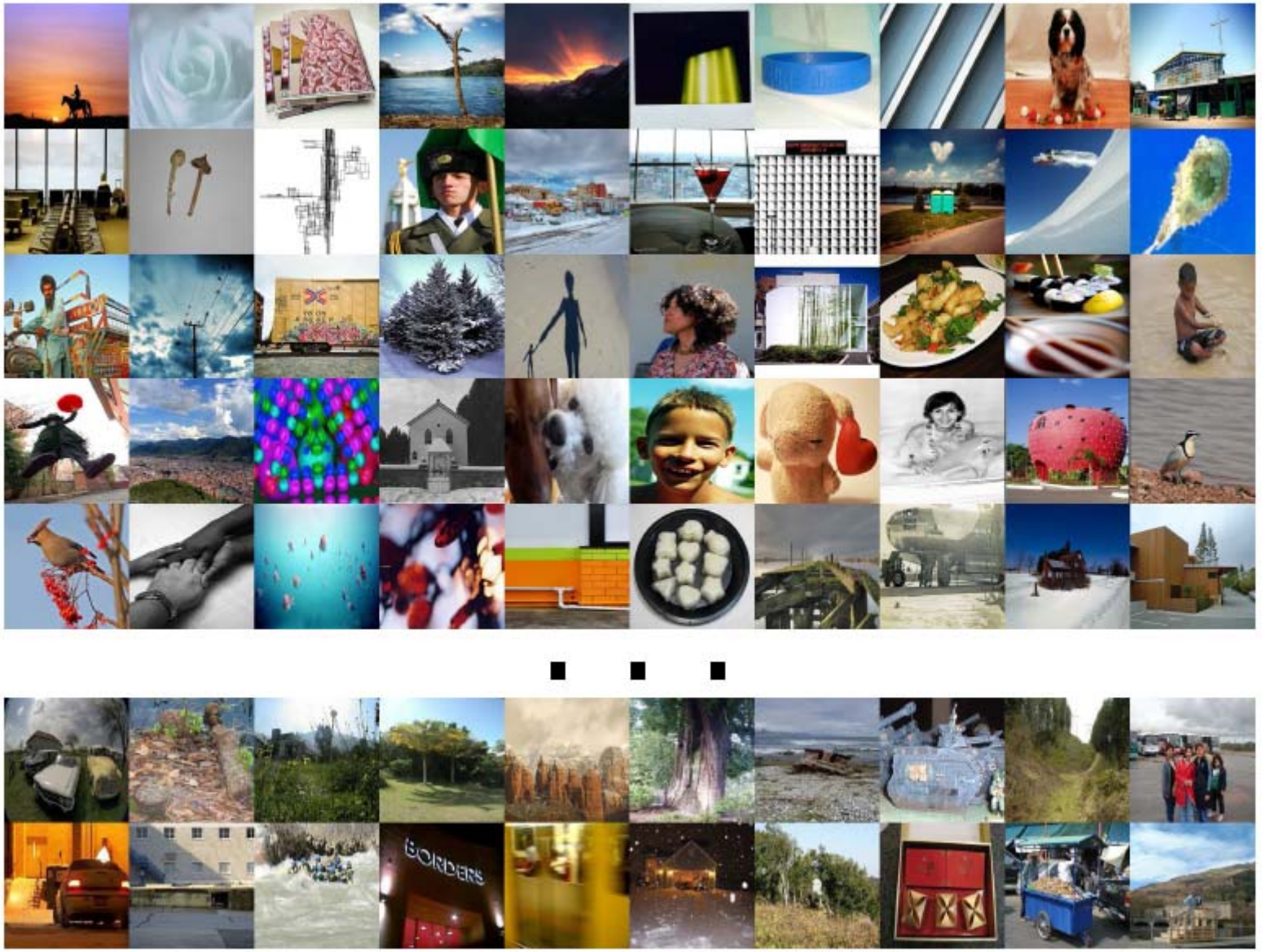
S. Dhar, V. Ordonez, and T. L Berg. High Level Describable Attributes for Predicting Aesthetics and Interestingness. CVPR, 2011.

- 美（Aesthetics）
  - DPChallenge website
  - http://www.dpchallenge.com/
  - 人手によるratingあり

- 魅力（Interestingness）
  - Flickr's "interestingness" measure
  - http://www.flickr.com/explore/interesting/
  - *There are lots of elements that make something 'interesting' (or not) on Flickr. Where the clickthroughs are coming from; who comments on it and when; who marks it as a favorite; its tags and many more things which are constantly changing. Interestingness changes over time, as more and more fantastic content and stories are added to Flickr.*
  - http://www.barcinski-jeanjean.com/entries/endlessintrestingness/

実験結果（interestingness）

S. Dhar, V. Ordonez, and T. L Berg. High Level Describable Attributes for Predicting Aesthetics and Interestingness. CVPR, 2011.



アトリビュートの画像検索への応用

- Matthijs Douze, Arnau Ramisa, and Cordelia Schmid. Combining attributes and Fisher vectors for efficient image retrieval. CVPR, 2011.

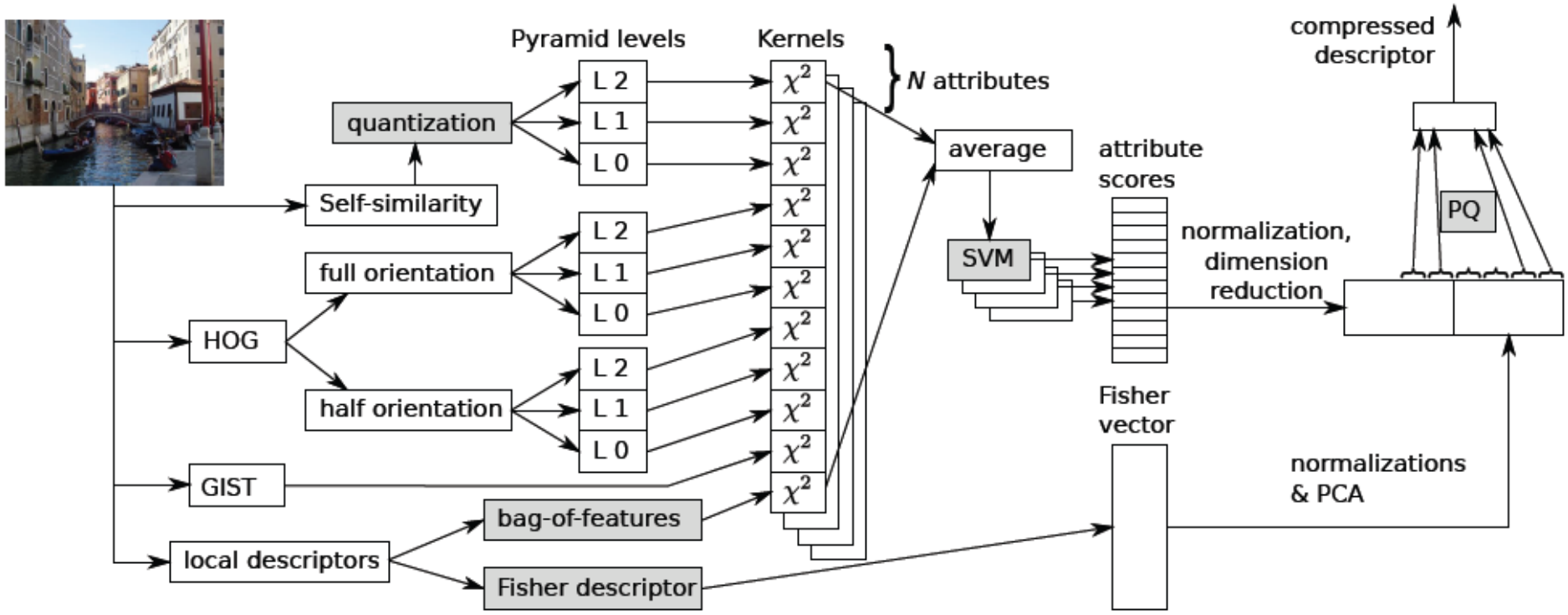


Figure 1. Computation of the attribute + Fisher descriptors for an image. Steps represented in gray require a learning stage.

2659属性
L. Torresani, M. Summer, and A. Fitzgibbon. Efficient object category recognition using classemes. In ECCV, 2010.

実験結果

M. Douze, A. Ramisa, and C. Schmid. Combining attributes and Fisher vectors for efficient image retrieval. CVPR, 2011.

Figure 3. Comparison of the retrieval results obtained with the Fisher vector, the attribute features, and their combination. The top row shows the query image, the remaining rows the first three retrieved images for the different descriptors.

Descriptor	dimension	mAP
BOF k=1000 [6]	1000	41.1
Fisher k=64 [17]	4096	≈ 60
Fisher k=4096 [17]	262144	70.5
VLAD k=64 [8]	8192	52.6
Fisher (F), k=64, L2 dist.	4096	59.5
Attributes (A), L2 dist.	2659	55.0
A + F, F-weight $\times 1$	6755	64.5
A + F, F-weight $\times 2$	6755	69.5
A + F, F-weight $\times 2.3$	6755	69.9

Table 1. Comparison of the different descriptors and their combination on the Holidays dataset.

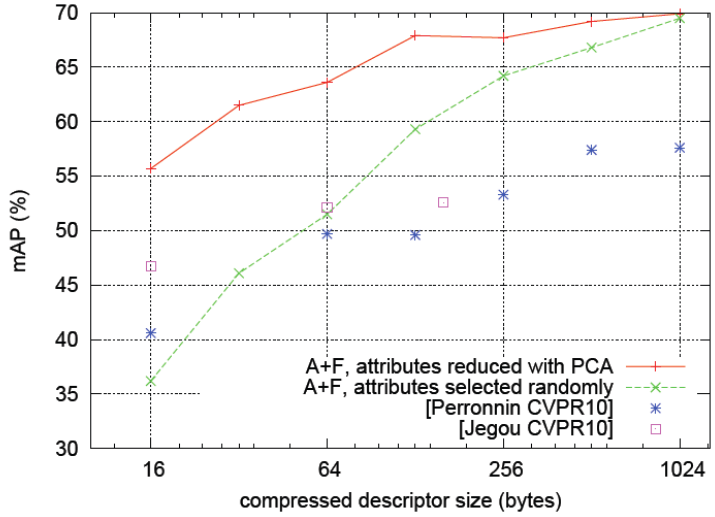
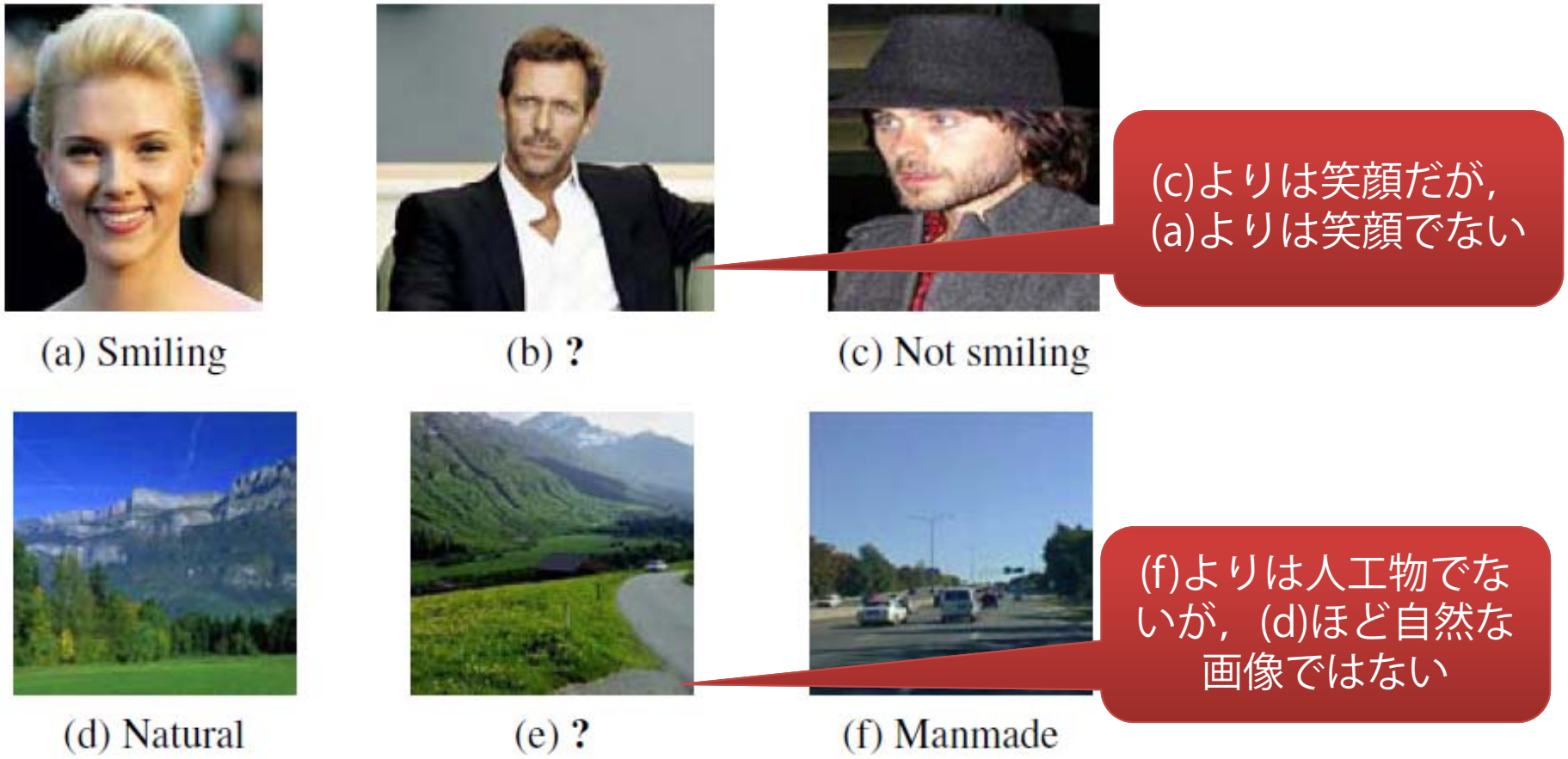


Figure 5. Performance of the A+F descriptor after dimension reduction and descriptor encoding on the Holidays dataset. The Fisher vectors are always reduced with PCA.

ICCV2011 best paper



- D. Parikh and K. Grauman. Relative Attributes. In ICCV, 2011.
- 各アトリビュートに関して，その強さと比較対象となる画像を選択し，新規画像を記述する

Relative Attributes

D. Parikh and K. Grauman. Relative Attributes. In ICCV, 2011.

- 概要
  - ペアサンプルの相対類似度が与えられた時に，各アトリビュートに対しランキング関数を学習する

- 問題の定式化

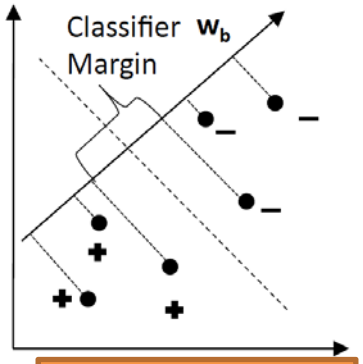
ランキング関数

$$r_m(\boldsymbol{x_i}) = \boldsymbol{w_m^T x_i},$$

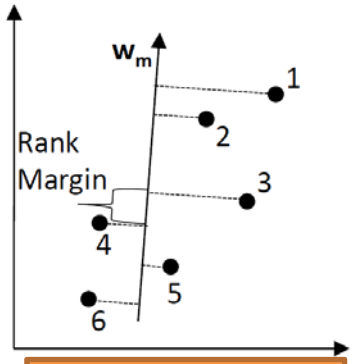
以下の条件を最大限満たす重みwを求める

i, jに順序有り　$\forall (i,j) \in O_m : \boldsymbol{w_m^T x_i} > \boldsymbol{w_m^T x_j}$

i, jがほぼ類似　$\forall (i,j) \in S_m : \boldsymbol{w_m^T x_i} = \boldsymbol{w_m^T x_j}.$



Wide margin binary classifier



Wide margin ranking function

$$\begin{aligned} \text{minimize} \quad & \left(\frac{1}{2}||\boldsymbol{w_m^T}||_2^2 + C\left(\sum \xi_{ij}^2 + \sum \gamma_{ij}^2\right)\right) \\ s.t. \quad & \boldsymbol{w_m^T}(\boldsymbol{x_i} - \boldsymbol{x_j}) \geq 1 - \xi_{ij}; \forall (i,j) \in O_m \\ & |\boldsymbol{w_m^T}(\boldsymbol{x_i} - \boldsymbol{x_j})| \leq \gamma_{ij}; \forall (i,j) \in S_m \\ & \xi_{ij} \geq 0; \gamma_{ij} \geq 0, \end{aligned}$$

主問題をNewton法で解く

知識はどこから降ってくる？

- 膨大な情報を利用
  - インターネット上の大規模画像
  - リッチな画像表現，効率な線形識別機の学習

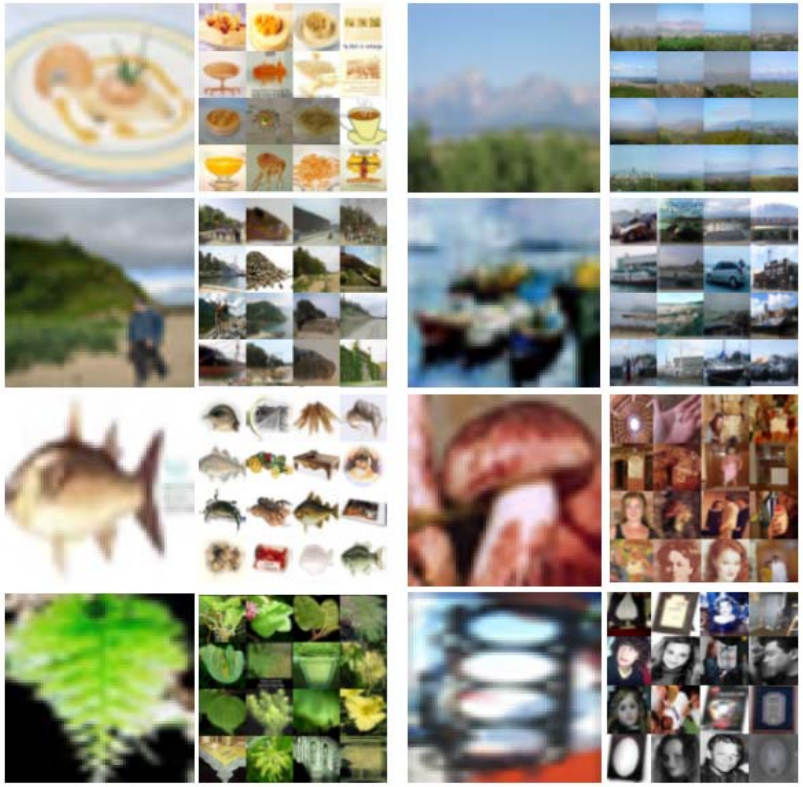
- 人に尋ねる
  - クラウドソーシング
  - 能動学習

- 対象とは別の知識を活用
  - 転移学習，ドメイン適合，マルチタスク学習
  - ゼロショット学習
  - アトリビュート

インターネット上の
大規模画像の活用

TinyImages

- A. Torralba, R. Fergus, W. T. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.30(11), pp. 1958-1970, 2008.
- 8000万枚の画像データセット
- データが大量にあれば最近傍法のみで十分認識可能



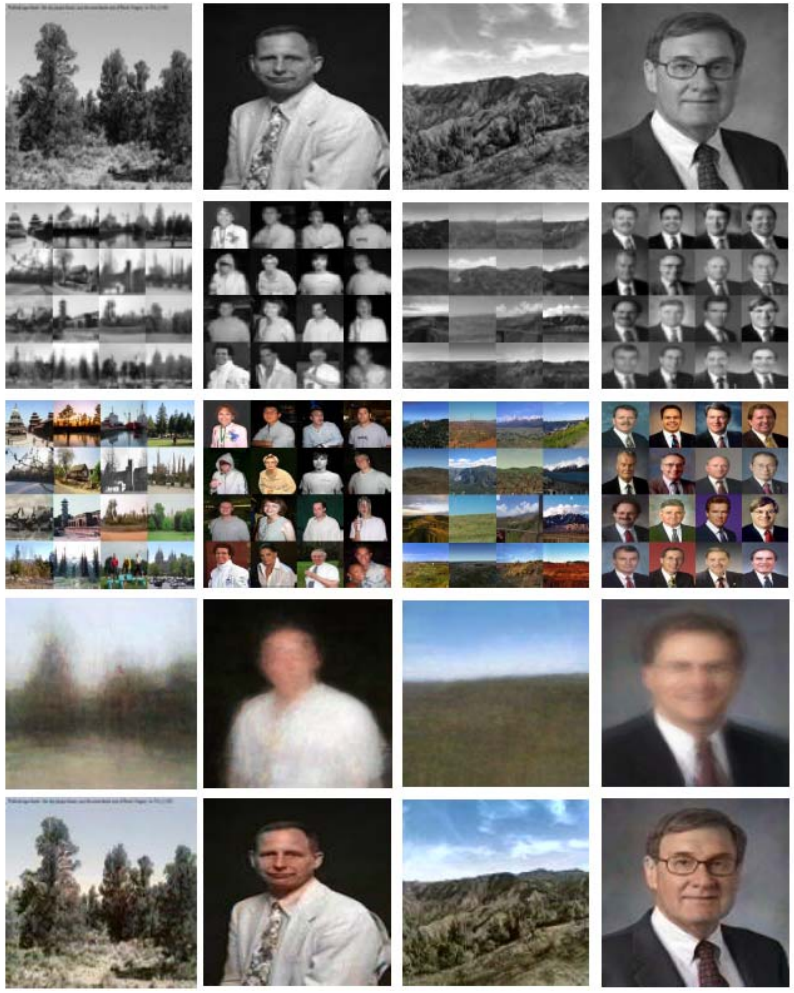


Fig. 1. 1st & 3rd columns: Eight 32 × 32 resolution color images. Despite their low resolution, it is still possible to recognize most of the objects and scenes. These are samples from a large dataset of 10^8 32 × 32 images we collected from the web which spans all visual object classes. 2nd & 4th columns: Collages showing the 16 nearest neighbors within the dataset to each image in the adjacent column. Note the consistency between the neighbors and the query image, having related objects in similar spatial arrangements. The power of the approach comes from the copious amount of data, rather than sophisticated matching methods.

ARISTA

- Xin-Jing Wang, Lei Zhang, Ming Liu, Yi Li, Wei-Ying Ma. ARISTA - Image Search to Annotation on Billions of Web Photos. In CVPR, 2010.
- 20億枚の画像データセットを利用した画像認識
- Near duplicated imageの活用．特定の名称まで認識可能.

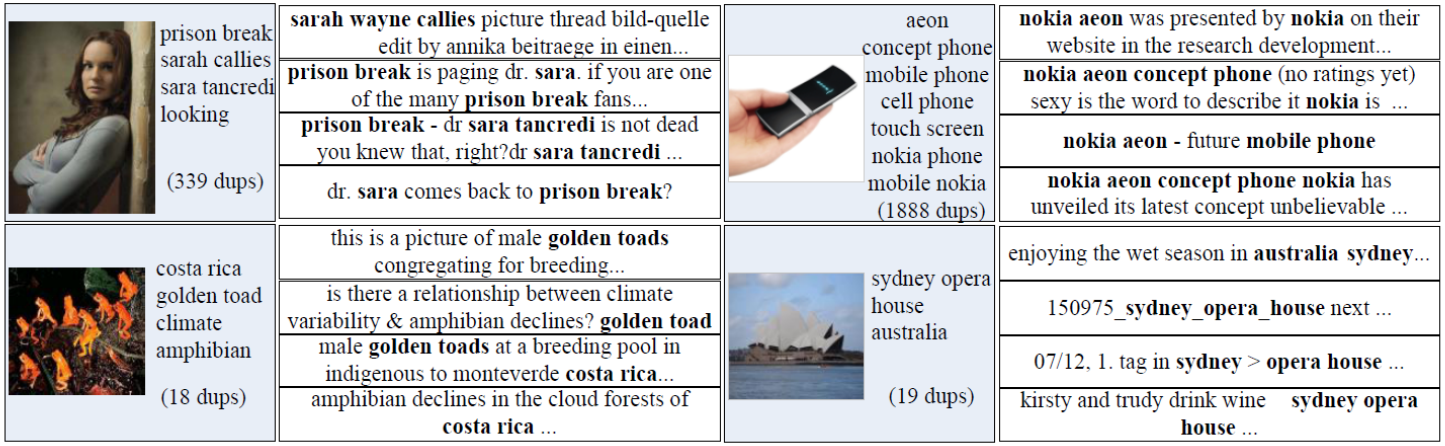


Figure 1. Examples showing that surrounding texts of near-duplicates have common terms which hit the semantics of a query image. The tags inside the image blocks are our annotation outputs. The common terms of each near-duplicate are highlighted in bold. Note that the detected tags are very specific. This is in contrast to most existing works that tend to generate general terms like sky, city, etc.

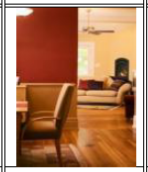
	2.4 M	80M	2B		2.4M	80M	2B
	(no results)	(no results)	*prison break*, **sarah callies**, **sara tancredi**, looking		(no results)	house paint, color	*house*, *paint*, wanta-toos, house painting, hardwood floor, interior design
	**michael jackson**	**michael jackson**, *rock pop*	**michael jackson**, sony music, *cd dvd*, *enter-tainment music*, *pop rock*		linu, *logo*	server, *software*, *logo*, credit card processing, *op-erating system*	**penguin**, *open source*, *virtual server*, *logo*, *operating system*
	**ipod touch**	**apple ipod**, *mp3 player*, **iphone**, wi fi, touch screen	**apple ipod**, *mp3 player*, wi fi, media player, **touch screen**, **mobile phone**		(no results)	(no results)	**bald eagle**, **haliaee-tus leucocephalus**, endangered species, fish wildlife, *eagle flight*

Figure 9. Annotation examples vs. dataset size. Bold-faced tags are perfect terms labeled by human subjects and italic ones are correct terms. Due to space limit, only the top five tags are shown. This figure suggests that larger dataset size ensures more accurate tags.

ImageNet

- ImageNet
  - 12 million images, 15 thousand categories
  - Image found via web searches for WordNet noun synsets
  - Hand verified using Mechanical
  - All new data for validation and testing this year
- WordNet
  - Source of fraction of English nouns
  - Also used tthe labels
  - Semantic hierarchy
  - Contains large o collect other datasets like tiny images (Torralba et al)
  - Note that categorization is not the end goal, but should provide information for other tasks, so idiosyncrasies of WordNet may be less critical

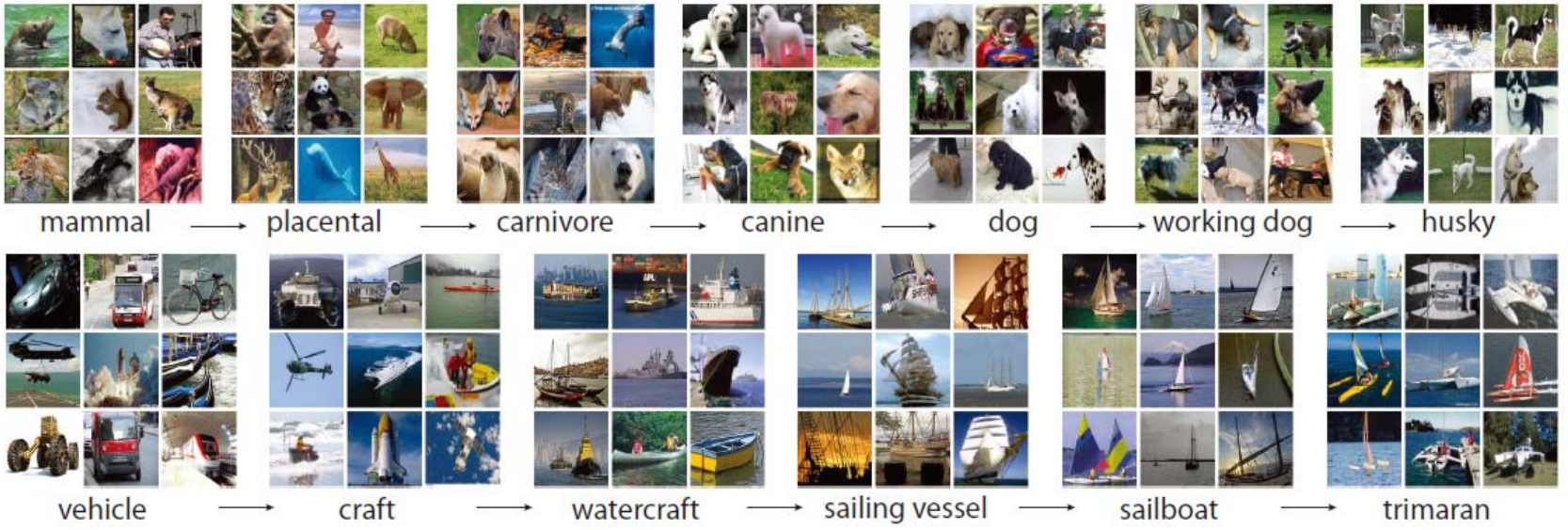
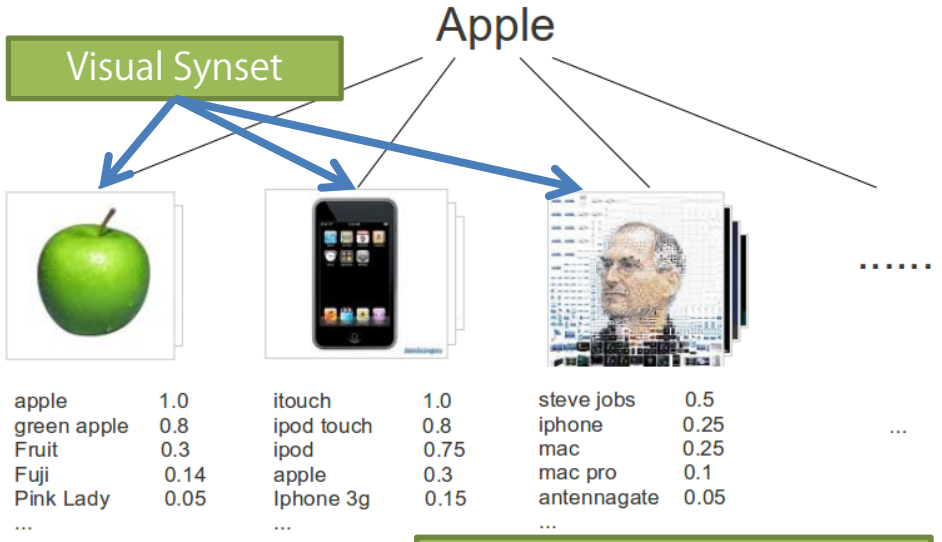


Figure 1: A snapshot of two root-to-leaf branches of ImageNet: the **top** row is from the mammal subtree; the **bottom** row is from the vehicle subtree. For each synset, 9 randomly sampled images are presented.

Deng et al., CVPR2009

Visual Synset

- Webスケールの画像アノテーション
  - D. Tsai, Y. Jing, Y. Liu, H. Rowley, S. Ioffe, and J. M.Rehg. Large-Scale Image Annotation using Visual Synset. In ICCV, 2011.
  - 2億枚の画像，30万のラベル
  - http://cpl.cc.gatech.edu/projects/VisualSynset



- アルゴリズム

1）画像のクラスタリング

プロトタイプへの変換

ciがプロトタイプとしてラベル付けされているか？

$$F(\mathcal{C}) = \sum_{i=1}^{N} S(x_i, L(x_i)) + \sum_{i=1}^{N} \delta_i(\mathcal{C}) \qquad \delta_i(\mathcal{C}) = \begin{cases} -\infty, \text{if} \quad L(c_i) \neq c_i \quad but \; \exists k : L(x_k) = c_i \\ 0 \quad \text{otherwise.} \end{cases}$$

2）Visual Synsetにラベルの付与

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad IDF_i = log\frac{|I|}{1+|\{S:l_i \in S\}|} \qquad S_{i,j} = TF_{i,j} * IDF_i$$

3）Visual Synset識別機を構築

1-vs-all linear SVM

4）投票による画像アノテーション

閾値Tを超えたときだけ1

Synset iのラベルベクトル

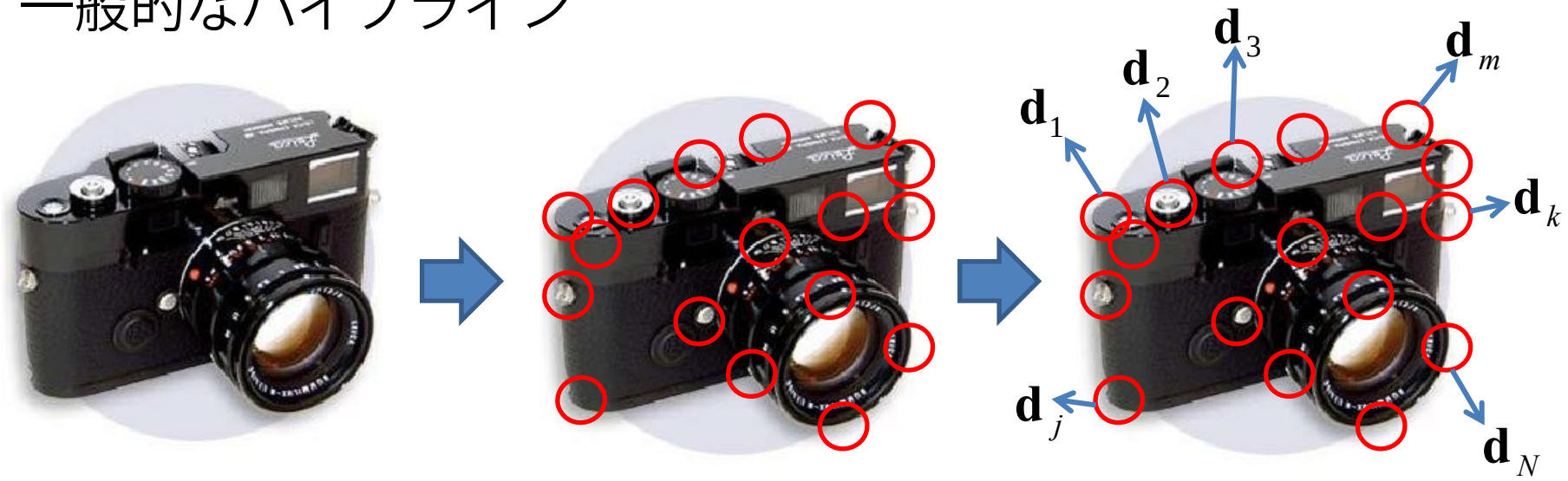
1-vs-all SVM では各識別機出力の大小は比較できない

$$\mathbf{L} = \sum_{i=1}^{n} I(\mathbf{w}_i \cdot \mathbf{x} + b_i > T) \sum_{j=1}^{m_i} \mathbf{K}_{i,j}$$

画像表現

画像特徴とは？

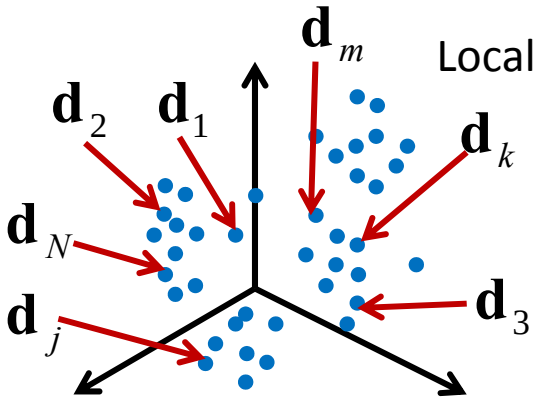
一般的なパイプライン



1) Input Image　2) Detection　3) Description

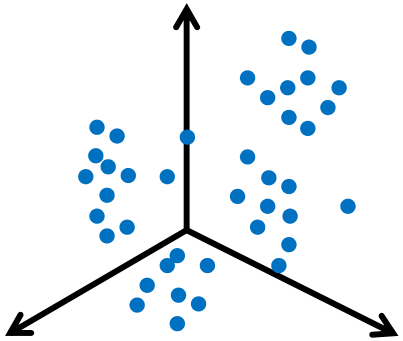
4) Local descriptors in feature space　5) PDF estimation　6) Feature vector

画像表現

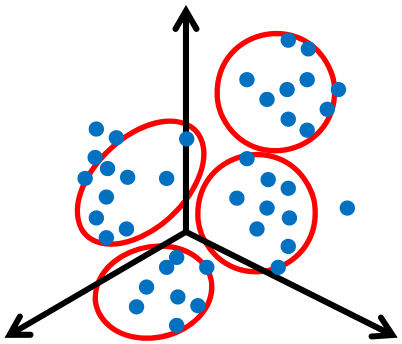


Local descriptors in feature space

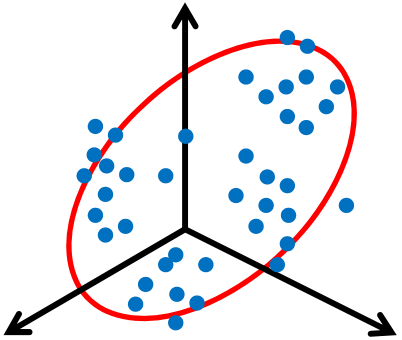
Descriptor matching



Codebook



Global feature



# of anchor points: large
Computational complexity: large

# of anchor points: small
Computational complexity: small

SVM-KNN
Naïve Bayes Nearest Neighbor
Graph Matching Kernel

Bag of Visual Words
Gaussian Mixture Model
ScSPM, Super Vector, LLC
Fisher Vector

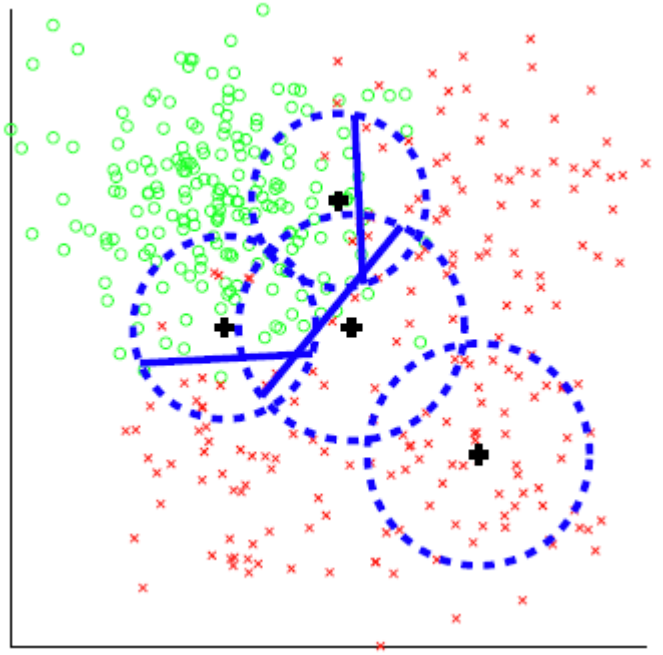
HLAC
GLC
Global Gaussian

画像表現
局所特徴のマッチング

Descriptor matching: SVM-KNN

H. Zhang, A. C. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. In CVPR, 2006.

- Naïve version
  - クエリサンプルと全ての距離を計量し，k-NNを抽出
  - K-NNが全て同じクラスなら，そのクラスをクエリに付与して終了．
  - 全て同じクラスでなければ，k-NNでカーネル行列を作成し，kernel SVMの識別機を構成する．
  - 構成したkernel SVMでクエリを識別する．



画像間距離

Geometric Blur

$$D^A(I_L \to I_R) = \frac{1}{m}\sum_{i=1}^{m} \min_{j=1..n} \|F_i^L - F_j^R\|^2$$

$$D^A(I_L, I_R) = D^A(I_L \to I_R) + D^A(I_R \to I_L) + \lambda \sum_{k=1}^{\text{nfilt}} \|h_k^L - h_k^R\|_{\text{L1}}$$

Texture Histogram

Naïve Bayes Nearest Neighbor

- O. Boiman, E. Shechtman, and M. Irani. In Defense of Nearest-Neighbor Based Image Classification. In CVPR, 2008.
- 非常に単純だが高性能
- 画像-クラス間距離を利用
- アルゴリズム
  - クエリ画像から局所記述子を計算
  - クエリ画像の各局所記述子に関して，クラス内の全局所記述子の中で最近傍のものを探す

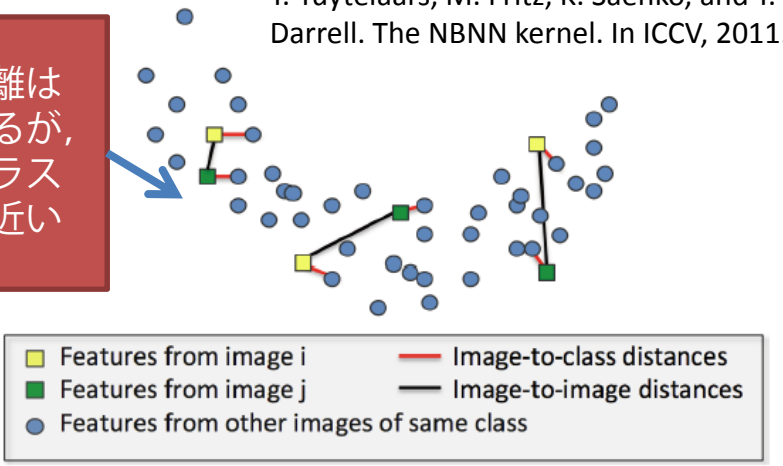
$$\mathrm{NN}_C(d_i)$$

  - クエリ画像の全局所記述子とクラス内の最近傍点とのユークリッド距離の総和を計算し，この距離が最も短いクラスにクエリ画像を割り当てる.

$$\hat{C} = arg\min_C \sum_{i=1}^{n} \| d_i - \mathrm{NN}_C(d_i) \|^2$$

画像間距離は離れているが，画像・クラス間距離は近い

T. Tuytelaars, M. Fritz, K. Saenko, and T. Darrell. The NBNN kernel. In ICCV, 2011.



NBNN Kernel

- T. Tuytelaars, M. Fritz, K. Saenko, and T. Darrell. The NBNN kernel. In ICCV, 2011.

Algorithm 2: the NBNN kernel

1. Compute a set of features $X = \{\mathbf{x}\}$.

2. $\forall \mathbf{x} \forall c$ Compute the NN of $\mathbf{x}$ in c: $NN^c(\mathbf{x})$, and its distance-to-class $d^c_{\mathbf{x}} = ||\mathbf{x} - NN^c(\mathbf{x})||^2$.

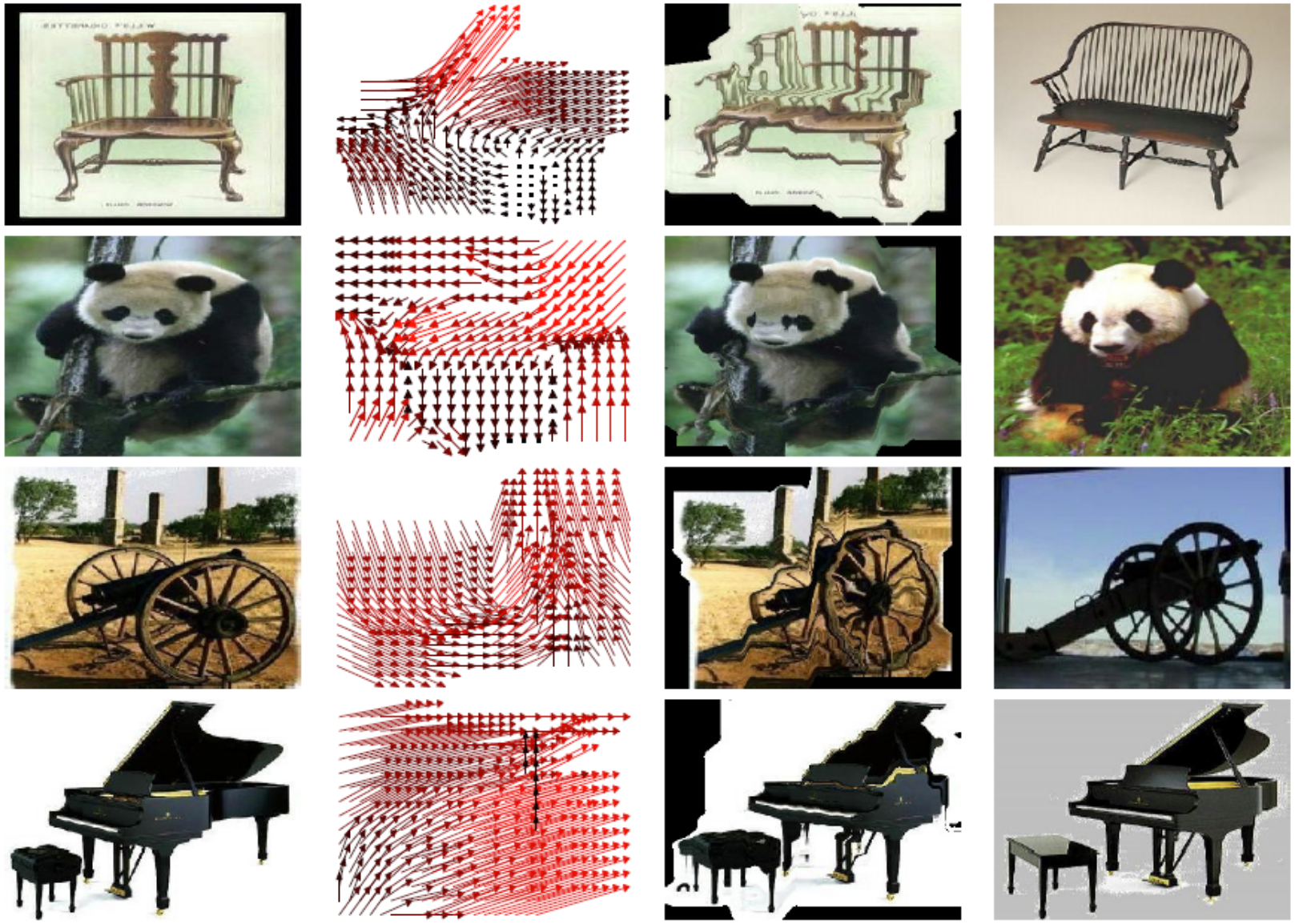
3. $\forall c \;\; \Phi^c(X) = \sum_{\mathbf{x} \in X} f(d^1_{\mathbf{x}}, \ldots, d^{|C|}_{\mathbf{x}})$.

4. $\Phi(X) = [\Phi^1(X) \ldots \Phi^{|C|}(X)]^T$.

5. Repeat steps 1-4 for a second set of features $Y = \{\mathbf{y}\}$.

6. $K(X, Y) = \Phi(X)^T \Phi(Y)$.

Graph Matching Kernel



O. Duchenne , A. Joulin and J. Ponce. A Graph-Matching Kernel for Object Categorization. ICCV, 2011.

画像表現
コードブック

Bag of Visual Words?

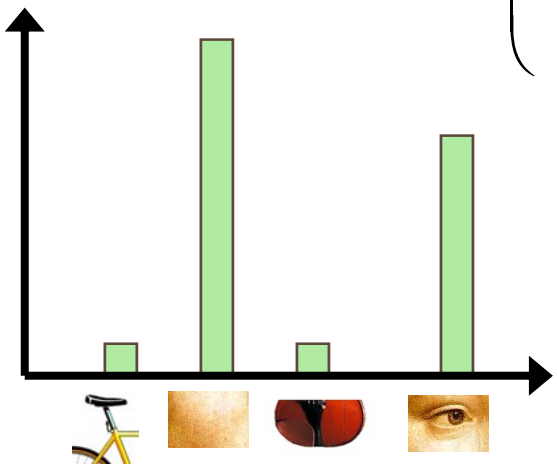
Li Fei Fei, cvpr07 tutorial
より抜粋

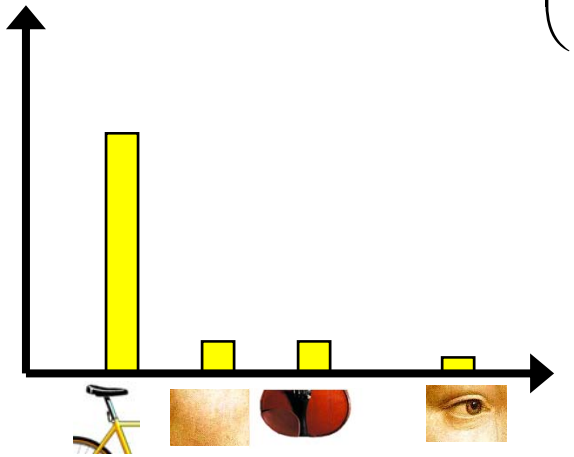
Visual words

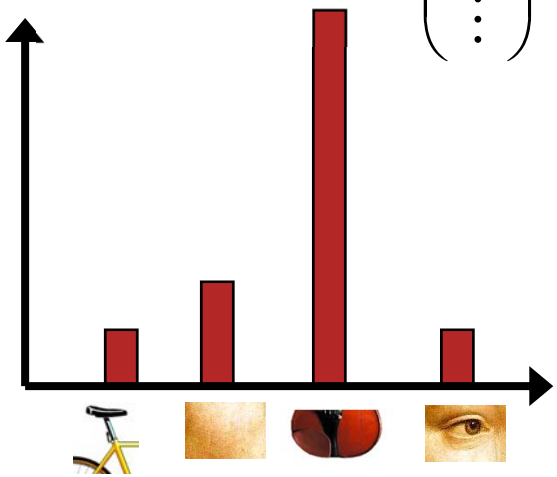
$$\begin{pmatrix} 1 \\ 10 \\ 1 \\ 7 \\ \vdots \end{pmatrix}$$

$$\begin{pmatrix} 7 \\ 2 \\ 2 \\ 1 \\ \vdots \end{pmatrix}$$

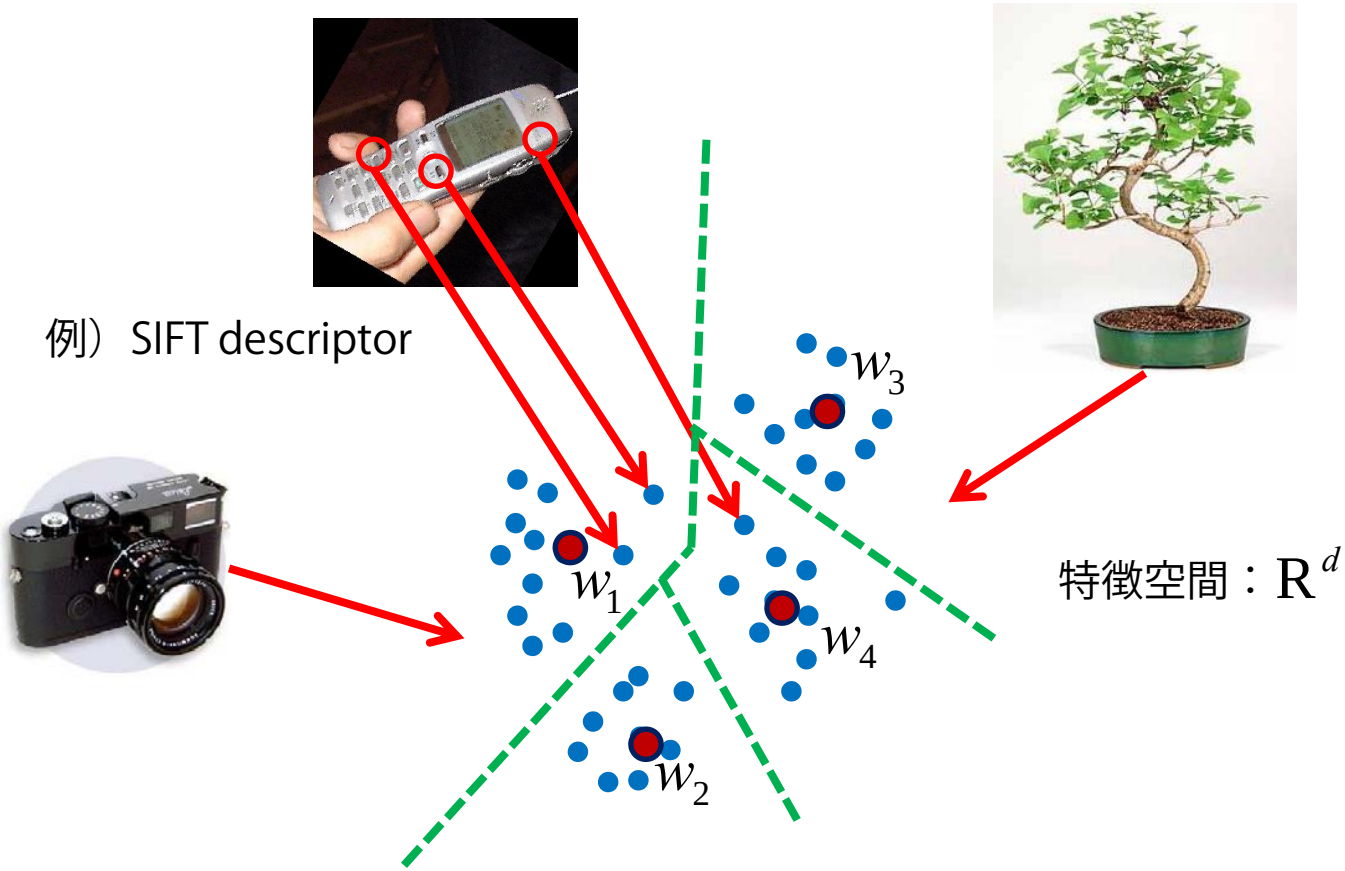
$$\begin{pmatrix} 3 \\ 4 \\ 10 \\ 2 \\ \vdots \end{pmatrix}$$







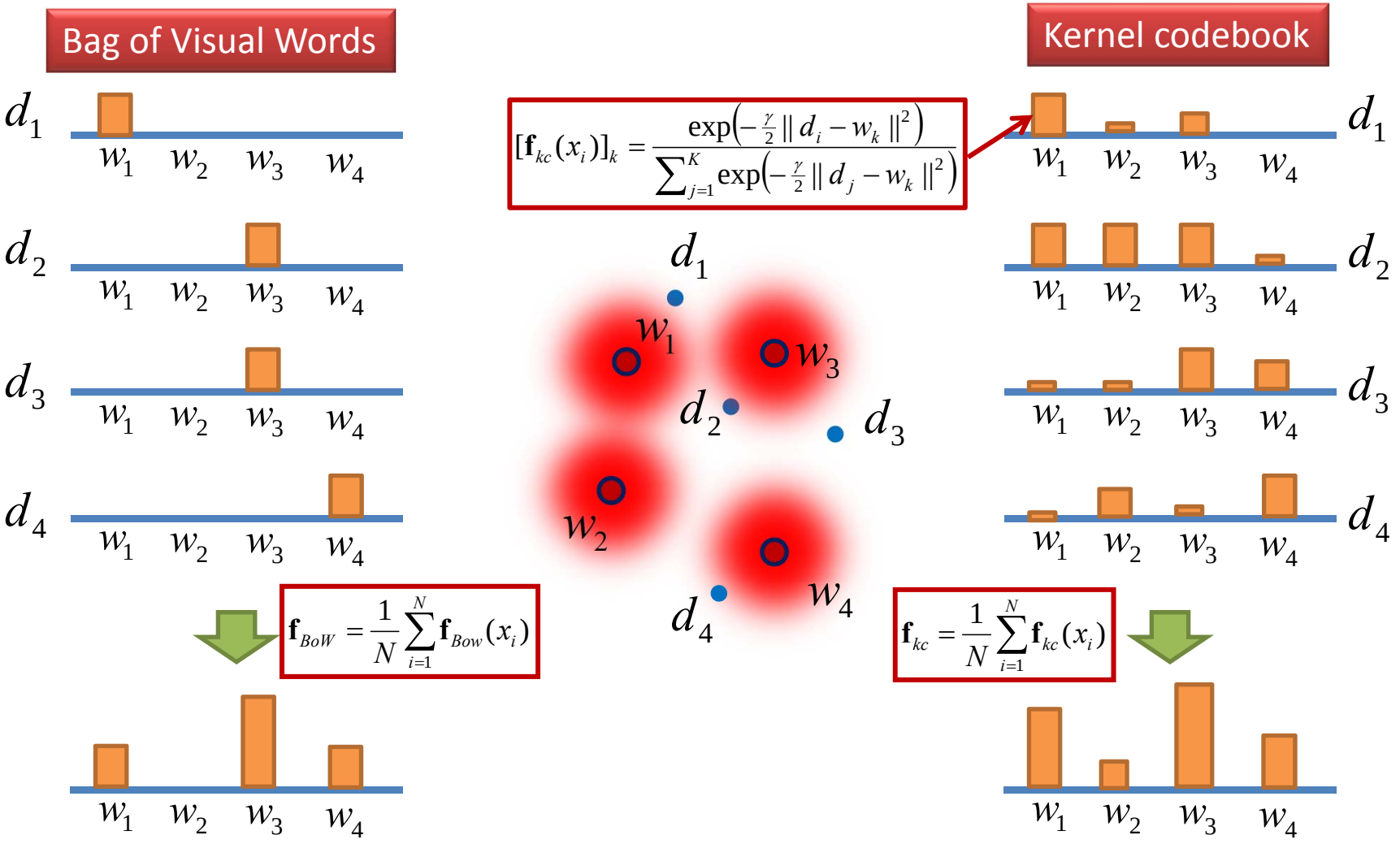
Code wordsの生成：clustering



- ベクトル量子化と呼ばれるプロセス
- 一般的にk-meansによるクラスタリング
  - 階層的クラスタリング：Vocabulary Tree
- 局所記述子にはSIFTがよく用いられる
  - もちろんSURFやRGB，Self Similarityでもよい

Kernel codebook

- 局所記述子を一つのコードワードに割り付けるのではなく，距離に応じた重み付けで全てのコードワードと関連づける．
- Jan C. van Gemert, Jan-Mark Geusebroek, Cor J. Veenman, and Arnold W.M. Smeulders. Kernel Codebooks for Scene Categorization. ECCV, 2008.



BoFのGMM利用による改善

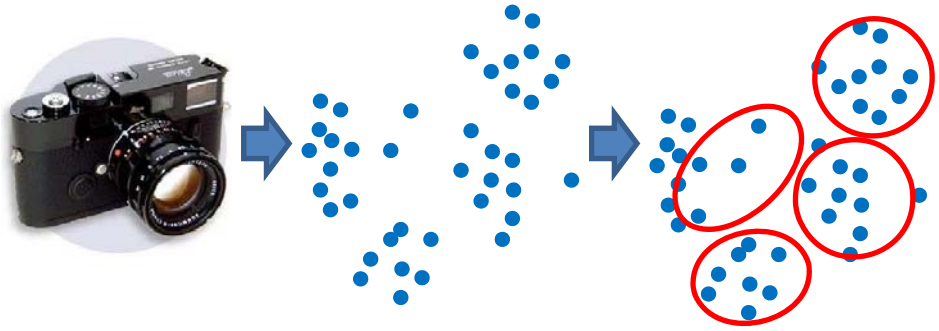


Image　　Local descriptors in feature space　　PDF estimation

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \sum_{k=1}^{K} \pi_k p_k(\boldsymbol{x})$$

$$\gamma_n(k) = p(k|\boldsymbol{x}_n, \boldsymbol{\theta}^{(t)}) = \frac{\pi_k p_k(\boldsymbol{x}_n)}{\sum_{j=1}^{K} \pi_j p_j(\boldsymbol{x}_n)}$$

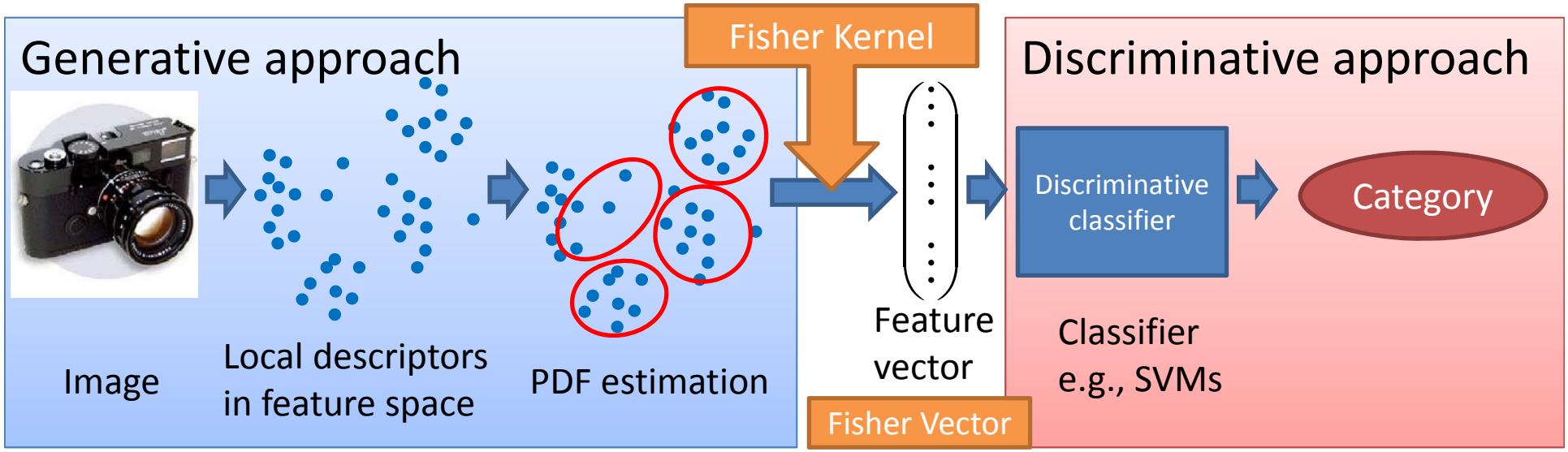
$$\boldsymbol{f} = \frac{1}{N} \sum_{n=1}^{N} [\gamma_n(1), \cdots, \gamma_n(K)]^\top \in R^K$$

- メリット
  - 混合ガウス分布を構成する各ガウス分布がそれぞれ共分散を持つため，共分散を考慮した距離計量を利用できる
  - 混合ガウス分布では局所特徴と多くのコードワードとの関係を表現できるので，特徴空間における局所特徴の位置に関する情報をエンコードできる

- デメリット
  - 混合ガウス分布表現はBoF と比較してパラメータが多い
    - 混合ガウス分布：O(K(D^2/2 + D))，BoF：O(KD)
  - 混合ガウス分布は訓練データに対して過剰適合する可能性があり，学習時に正則化を行う必要

フィッシャーベクトル

F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. CVPR, 2007.



- 混合ガウス分布を用いた確率密度分布推定によるBoF の改良
  - 生成モデル（generative model）
- 生成モデルを識別的なアプローチに適応可能なより洗練された手法があれば識別性能の改善につながる.
- フィッシャーカーネル（Fisher Kernel)
  - 生成的アプローチ（generative approach）と識別的アプローチ（discriminative approach）を結合させる強力な枠組み→確率分布の空間に適切な距離計量を埋め込む
  - 確率分布のなす空間は，Fisher 情報行列を計量とするリーマン空間
  - 手順

1. 局所特徴を生成する確率密度分布から導出される勾配ベクトルの計算
2. 画像を表現する一つの特徴ベクトルの計算
   →フィッシャーベクトル（Fisher Vector）
3. 得られた特徴ベクトルを識別的分類機に入力する.

フィッシャーベクトルのメリット

- 豊かな特徴ベクトル表現
  - BoF と比較してフィッシャーカーネルを利用するメリットは，コードブックサイズが同じであればより要素数の多い特徴ベクトルが得られる.
    - コードブックサイズ：K，局所特徴の次元：d
    - BoFの次元：K
    - フィッシャーベクトル： (2d+1)K-1
  - 特徴ベクトルの表現する情報が多いため計算コストの高いカーネル法を利用して高次元空間へ射影する必要がなく，線形識別機でも十分な識別性能を出すことが可能となる.

大規模データに
最も重要な要素

フィッシャーベクトル詳細

- 局所特徴群

$$\mathcal{X} = \{\boldsymbol{x}_n \in R^D\}_{n=1}^N$$

- あらゆる画像内容を表現する局所特徴の確率密度分布

$$u_\theta$$

- 対数尤度の勾配

$$G_\theta^{\mathcal{X}} = \frac{1}{N}\nabla_\theta \log u_\theta(\mathcal{X}|\boldsymbol{\theta})$$

  - データに最も適合するように確率密度関数のパラメータが修正すべき方向を表現
  - 異なるデータサイズ集合をパラメータ数に依存した特定の長さの特徴ベクトルに変換
  - 内積を利用する識別機には適切な計量が必要！！

- フィッシャー情報行列

$$F_\theta = E_X[\nabla_\theta \log u_\theta(\mathcal{X}|\boldsymbol{\theta})\nabla_\theta \log u_\theta(\mathcal{X}|\boldsymbol{\theta})^\top]$$

- フィッシャーベクトル（Fisher Vector）

$$\mathcal{G}_\theta^X = F_\theta^{-1/2}\nabla_\theta \log u_\theta(\mathcal{X}|\boldsymbol{\theta})$$

フィッシャー情報行列による対数尤度の勾配の正規化

混合ガウス分布におけるフィッシャーベクトル

- 確率密度分布を混合ガウス分布とする
  - 共分散行列は対角行列と仮定

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \sum_{k=1}^{K} \pi_k p_k(\boldsymbol{x})$$

あらゆる画像を生成する確率密度分布

- 対数尤度の微分

$$G_\theta^{\mathcal{X}} = \frac{1}{N} \nabla_\theta \log u_\theta(\mathcal{X}|\boldsymbol{\theta})$$

画像1枚から得られる局所特徴の集合

負担率：局所特徴x_nがGMMのコンポーネントkに属する確率

$$\frac{\partial \mathcal{L}(\mathcal{X}|\boldsymbol{\theta})}{\partial \pi_k} = \sum_{n=1}^{N} \left[\frac{\gamma_n(k)}{\pi_k} - \frac{\gamma_n(1)}{\pi_1} \right]$$

$$\frac{\partial \mathcal{L}(\mathcal{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k^d} = \sum_{n=1}^{N} \gamma_n(k) \left[\frac{\boldsymbol{x}_n^d - \boldsymbol{\mu}_k^d}{(\boldsymbol{\sigma}_k^d)^2} \right]$$

$$\frac{\partial \mathcal{L}(\mathcal{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\sigma}_k^d} = \sum_{n=1}^{N} \gamma_n(k) \left[\frac{(\boldsymbol{x}_n^d - \boldsymbol{\mu}_k^d)^2}{(\boldsymbol{\sigma}_k^d)^3} - \frac{1}{\boldsymbol{\sigma}_k^d} \right]$$

GMMのBoFとほぼ同じ

$$\boldsymbol{f} = \frac{1}{N} \sum_{n=1}^{N} [\gamma_n(1), \cdots, \gamma_n(K)]^\top \in R^K$$

局所特徴x_nとGMMの各コンポーネントkの平均との差分

- 混合比：BoFとほぼ同じ
- 平均，分散：あらゆる画像を表現するpdfの平均との差分
- BoFは0次，Fisher Vectorは1次，2次の統計量を含む
- 分散の表現は平均の表現とあまり差がない？本来は各コンポーネント間の相関が必要

フィッシャー情報行列

- フィッシャー情報行列

$$F_\theta = E_X[\nabla_\theta \log u_\theta(\mathcal{X}|\boldsymbol{\theta}) \nabla_\theta \log u_\theta(\mathcal{X}|\boldsymbol{\theta})^\top]$$

- 混合ガウス分布において近似的に閉じた解が得られる
- 仮定
  - フィッシャー情報行列は対角行列
  - 共分散行列は対角行列
  - 負担率はピーキー
  - 一枚の画像から得られる局所特徴数は一定

$$\frac{\partial \mathcal{L}(\mathcal{X}|\theta)}{\partial \pi_k} \Rightarrow f_{\pi_k} = N\left(\frac{1}{\pi_k} + \frac{1}{\pi_1}\right)$$

$$\frac{\partial \mathcal{L}(\mathcal{X}|\theta)}{\partial \mu_k^d} \Rightarrow f_{\mu_k^d} = \frac{N\pi_k}{(\boldsymbol{\sigma}_k^d)^2}$$

$$\frac{\partial \mathcal{L}(\mathcal{X}|\theta)}{\partial \sigma_k^d} \Rightarrow f_{\sigma_k^d} = \frac{2N\pi_k}{(\boldsymbol{\sigma}_k^d)^2}$$

フィッシャー情報行列の要素

フィッシャーベクトルの性能

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

- Pascal VOC 2007
- 改良されたフィッシャーベクトルを利用
- 識別機：線形SVM

PN	L2	SP	SIFT	Col	S+C
-	-	-	47.9	34.2	45.9
✓	-	-	54.2	45.9	57.6
-	✓	-	51.8	40.6	53.9
-	-	✓	50.3	37.5	49.0
✓	✓	✓	58.3	50.9	60.3

パワー正規化＞L2正規化＞空間ピラミッド，の順で改善の効果が高い

GMM Supervectors

- W. M. Campbell and D. E. Sturim and D. A. Reynolds. Support vector machines using GMM supervectors for speaker verification. IEEE Signal Processing Letters, Vol.13, pp.308-311, 2006.
- もともと音声認識で利用されていたもの.

$$p(x;\theta) = \sum_{k=1}^{K} w_k N(x;\mu_k,\Sigma_k)$$

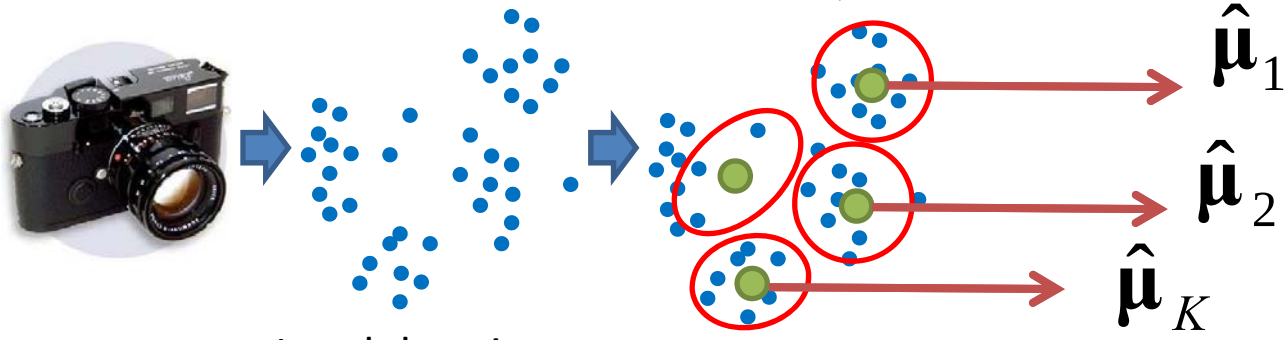


Image　　Local descriptors in feature space　　GMM　　Means of components

•訓練画像全体からユニバーサルな分布を作成
•各画像に適合させる

$$\hat{\mu}_k = \frac{\tau\hat{\mu}_k^{(U)} + \sum_{i=1}^{N}\gamma_i(k)x_i}{\tau + \sum_{i=1}^{N}\gamma_i(k)}$$

MAP適合

$$\tilde{\mu}_k = \sqrt{w_k^{(U)}}\,(\Sigma_k^{(U)})^{-1/2}\,\hat{\mu}_k$$

$$\phi(X) = \begin{pmatrix} \tilde{\mu}_1 \\ \tilde{\mu}_2 \\ \vdots \\ \tilde{\mu}_K \end{pmatrix}$$

コンポーネントkの混合比　　コンポーネントkの共分散行列

GMM supervectors

GMM Supervectors

- W. M. Campbell and D. E. Sturim and D. A. Reynolds. Support vector machines using GMM supervectors for speaker verification. IEEE Signal Processing Letters, Vol.13, pp.308-311, 2006.

GMM supervectors

$$\hat{\boldsymbol{\mu}}_k = \frac{\tau\hat{\mu}_k^{(U)} + \sum_{i=1}^{N}\gamma_i(k)x_i}{\tau + \sum_{i=1}^{N}\gamma_i(k)}$$

$\tau = 0$

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_k &= \sqrt{w_k^{(U)}}(\Sigma_k^{(U)})^{-1/2}\hat{\boldsymbol{\mu}}_k \\ &\approx \frac{\sqrt{w_k^{(U)}}}{\sum_{i=1}^{N}\gamma_k(i)}(\Sigma_k^{(U)})^{-1/2}\sum_{i=1}^{N}\gamma_i(k)x_i \\ &\approx \frac{1}{N\sqrt{w_k^{(U)}}}(\Sigma_k^{(U)})^{-1/2}\sum_{i=1}^{N}\gamma_i(k)x_i\end{aligned}$$

$$Nw_k = \sum_{i=1}^{N}\gamma_i(k)$$

GMM supervectorとFisher Vectorの平均成分はほぼ同一

Fisher Vectorの平均成分

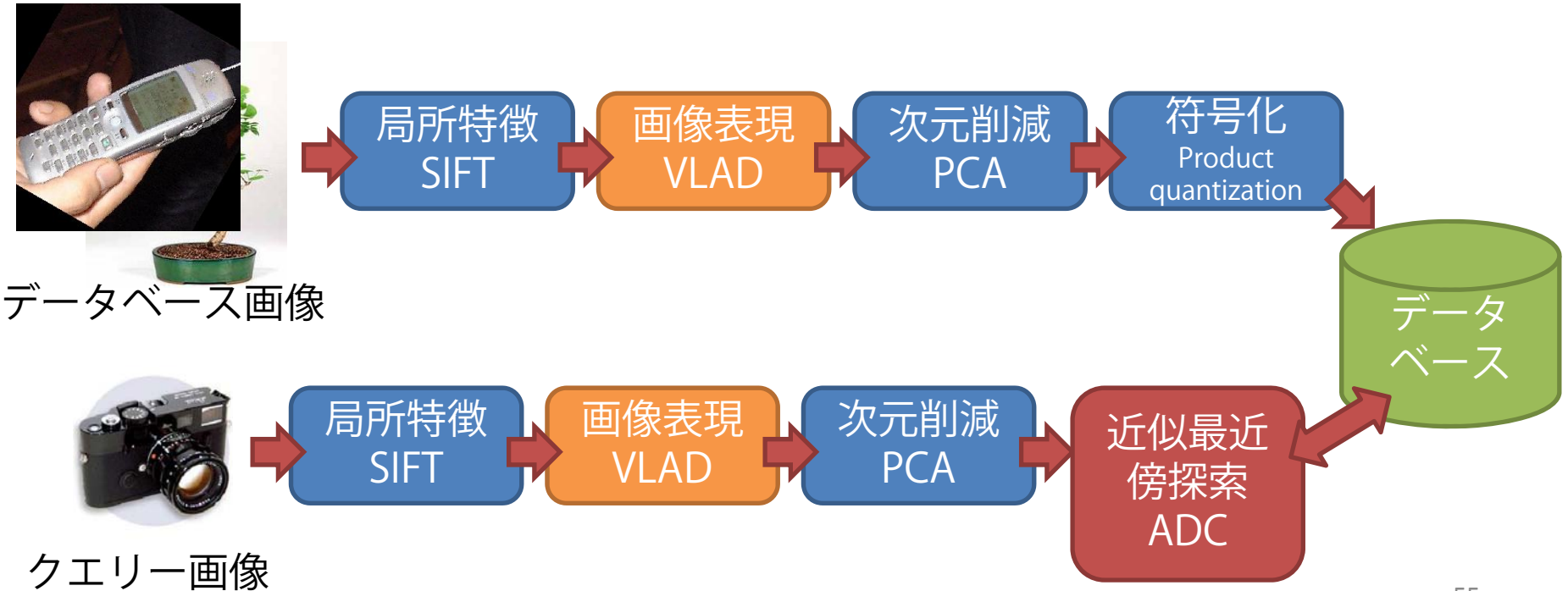
$$g_{\mu,i} = \frac{1}{N\sqrt{w_k}}\sum_{i=1}^{N}\gamma_i(k)(\Sigma_k)^{-1/2}(\mathbf{x}_i - \boldsymbol{\mu}_k)$$

TRECVID 2011ではGMM supervectorの局所特徴の各コンポーネントへの割り付けを高速化させることで第一位の性能を上げている.

N. Inoue and K. Shinoda. A Fast MAP Adaptation Technique for GMMsupervector-based Video Semantic Indexing. ACM Multimedia, 2011.

フィッシャーベクトルの画像検索への応用例

- H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. CVPR, 2010.
- 20bitに画像表現しても，生のBoFを使った検索と同じ検索性能
- パイプライン



VLAD

H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. CVPR, 2010.

- Vector of Locally Aggregated Descriptors

VLADのd番目要素

局所特徴のd番目要素

局所特徴が割り当てられたコードワードiのベクトル

$$\boldsymbol{z}_i^d = \sum_{x \in \mathcal{X}_i} \left(\boldsymbol{x}^d - \boldsymbol{v}_i^d \right)$$

この後 L2正規化

コードワードiに属する局所特徴集合

VLADの例，コードワード数：16

VLADとフィッシャーベクトル

- フィッシャーベクトル

$$\frac{\partial \mathcal{L}(\mathcal{X}|\boldsymbol{\theta})}{\partial \pi_k} = \sum_{n=1}^{N} \left[\frac{\gamma_n(k)}{\pi_k} - \frac{\gamma_n(1)}{\pi_1} \right]$$

$$\frac{\partial \mathcal{L}(\mathcal{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k^d} = \sum_{n=1}^{N} \gamma_n(k) \left[\frac{\boldsymbol{x}_n^d - \boldsymbol{\mu}_k^d}{(\boldsymbol{\sigma}_k^d)^2} \right]$$

$$\frac{\partial \mathcal{L}(\mathcal{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\sigma}_k^d} = \sum_{n=1}^{N} \gamma_n(k) \left[\frac{(\boldsymbol{x}_n^d - \boldsymbol{\mu}_k^d)^2}{(\boldsymbol{\sigma}_k^d)^3} - \frac{1}{\boldsymbol{\sigma}_k^d} \right]$$

GMMのBoFとほぼ同じ

局所特徴x_nとGMMの各コンポーネントkの平均との差分

- VLAD

$$\boldsymbol{z}_i^d = \sum_{x \in \mathcal{X}_i} \left(\boldsymbol{x}^d - \boldsymbol{v}_i^d \right)$$

VLADのd番目要素

局所特徴のd番目要素

局所特徴が割り当てられたコードワードiのベクトル

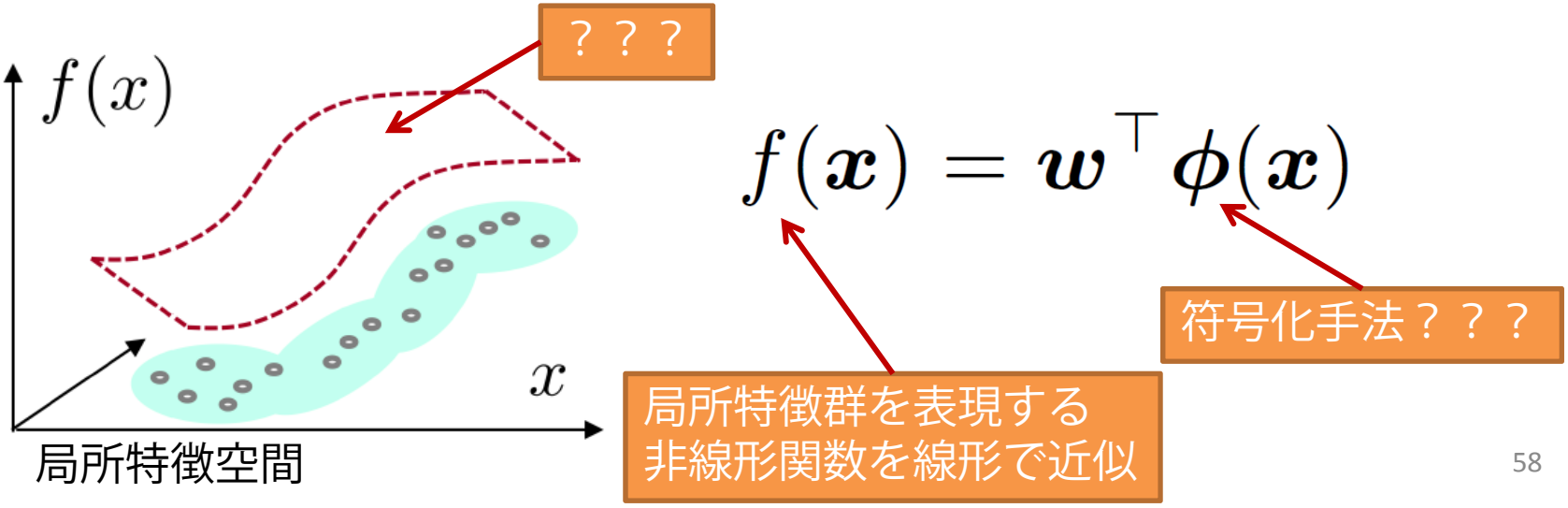
コードワードiに属する局所特徴集合

•負担率：ハードな割り当て
•分散：全てのコンポーネントで同じ

•→VLADはフィッシャーベクトルの平均に関する要素と同じ.
•（注）分散を考えていないのでフィッシャーとは言い難い.

スーパーベクトル符号化
Super-Vector Coding

- X. Zhou, K. Yu, T. Zhang, and T.S. Huang. Image classification using super-vector coding of local image descriptors. ECCV, 2010.
- BoF や混合ガウス分布を用いたBoF の改善手法
  - 特徴空間における局所特徴の分布の表現を得るプロセスと解釈できた.
- ここでも高次元空間における局所特徴分布を表現する，なめらかな非線形関数 *f(x)* の学習について考える.
- 非線形関数 *f(x)* を線形表現可能な符号化手法 *Φ(x)* を求める.



スーパーベクトル符号化の導出

- 局所特徴をコードブックを利用して近似

$$\boldsymbol{x} \approx \sum_{k=1}^{K} \gamma_x(k)\boldsymbol{v}_k \qquad \boldsymbol{\gamma}_x = [\gamma_x(1), \cdots, \gamma_x(K)], \sum_{k=1}^{K} \gamma_x(k) = 1$$

負担率のようなもの

コードワードk

- β Lipschitz derivative smooth

関数f(x)の1次近似のUpper boundに関する式

$$|f(\boldsymbol{x}) - f(\boldsymbol{x}') - \nabla f(\boldsymbol{x}')^\top (\boldsymbol{x} - \boldsymbol{x}')| \leqq \frac{\beta}{2}||\boldsymbol{x} - \boldsymbol{x}'||^2$$

コードワードの代入 $\boldsymbol{x}' = \boldsymbol{v}^x$

$$|f(\boldsymbol{x}) - f(\boldsymbol{v}^x) - \nabla f(\boldsymbol{v}^x)^\top (\boldsymbol{x} - \boldsymbol{v}^x)| \leqq \frac{\beta}{2}||\boldsymbol{x} - \boldsymbol{v}^x||^2$$

$$f(\boldsymbol{x}) = f(\boldsymbol{v}^x) + \nabla f(\boldsymbol{v}^x)^\top (\boldsymbol{x} - \boldsymbol{v}^x) \cdots (\star)$$

||x-v||が小さければ近似精度が向上

- スーパーベクトル符号化

$$f(\boldsymbol{x}) \approx \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x})$$

式(☆)を分解！

Super Vector Coding

$$\boldsymbol{\phi}(\boldsymbol{x}) = \left[s\gamma_x(k), \gamma_x(k)(\boldsymbol{x} - \boldsymbol{v}_k)^\top \right]^\top_{v_k \in \mathcal{V}}$$

$$\boldsymbol{w} = \left[\frac{1}{s} f(\boldsymbol{v}_k), (\nabla f(\boldsymbol{v}_k))^\top \right]^\top_{v_k \in \mathcal{V}}$$

スーパーベクトル符号化の解釈

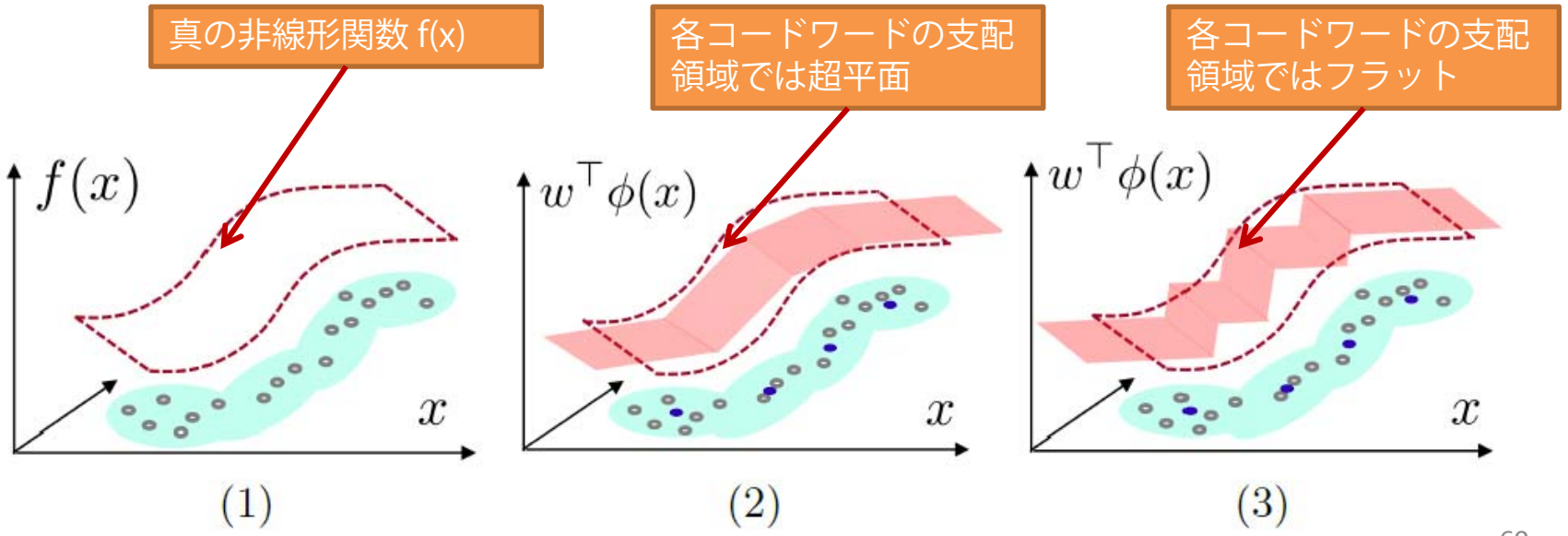
- スーパーベクトル符号化の例
  - コードワード数：3，γ =[0 1 0]'

Super Vector Coding

$$\boldsymbol{\phi}(\boldsymbol{x}) = \left[s\gamma_x(k), \gamma_x(k)(\boldsymbol{x} - \boldsymbol{v}_k)^\top \right]^\top_{v_k \in \mathcal{V}}$$

$$\phi(x) = \left[\underbrace{0, \ldots, 0}_{d+1 \text{ dim.}}, \underbrace{s, (x-v)^\top}_{d+1 \text{ dim.}}, \underbrace{0, \ldots, 0}_{d+1 \text{ dim.}} \right]^\top$$

- スーパーベクトル符号化とBoF



スーパーベクトル符号化とフィッシャーベクトル

- フィッシャーベクトル

$$\frac{\partial \mathcal{L}(\mathcal{X}|\boldsymbol{\theta})}{\partial \pi_k} = \sum_{n=1}^{N} \left[\frac{\gamma_n(k)}{\pi_k} - \frac{\gamma_n(1)}{\pi_1} \right]$$

$$\frac{\partial \mathcal{L}(\mathcal{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k^d} = \sum_{n=1}^{N} \gamma_n(k) \left[\frac{\boldsymbol{x}_n^d - \boldsymbol{\mu}_k^d}{(\boldsymbol{\sigma}_k^d)^2} \right]$$

$$\frac{\partial \mathcal{L}(\mathcal{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\sigma}_k^d} = \sum_{n=1}^{N} \gamma_n(k) \left[\frac{(\boldsymbol{x}_n^d - \boldsymbol{\mu}_k^d)^2}{(\boldsymbol{\sigma}_k^d)^3} - \frac{1}{\boldsymbol{\sigma}_k^d} \right]$$

GMMのBoFとほぼ同じ

局所特徴x_nとGMMの各コンポーネントkの平均との差分

- スーパーベクトル符号化

$$\boldsymbol{\phi}(\boldsymbol{x}) = \left[s\gamma_x(k), \gamma_x(k)(\boldsymbol{x} - \boldsymbol{v}_k)^\top \right]^\top_{v_k \in \mathcal{V}}$$

負担率

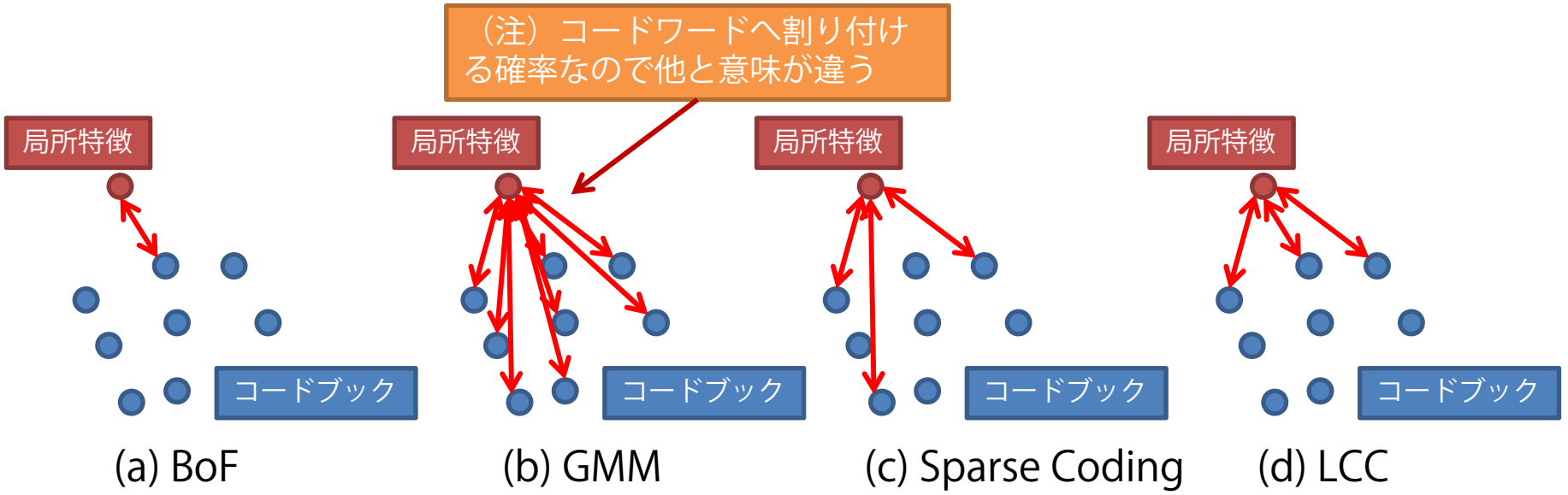
局所特徴x_nとコードワードとの差分

•混合比：一定
•分散：一定

•→スーパーベクトル符号化はフィッシャーベクトルの混合比と平均に関する要素と同じ
•（注）分散を考えていないのでフィッシャーとは言い難い.

スパース符号化の比較

- BoF
  - 局所特徴が一つのコードワードに割り当てられる
- BoFのGMMによる表現
  - 局所特徴が全てのコードワードと関係を持つ
- スパース符号化
  - 局所特徴が少数のコードワードと関係を持つ
- 局所線形制約符号化
  - 局所特徴が局所の少数コードワードと関係を持つ



スパース符号化の定式化

- Bag of Visual Words
  - ベクトル量子化（VQ）

$$\min_{U,V} \sum_{n=1}^{N} ||\boldsymbol{x}_n - \boldsymbol{V}\boldsymbol{u}_n||^2$$

$$\text{s.t. } \text{Card}(\boldsymbol{u}_n) = 1, |\boldsymbol{u}_n| = 1, \boldsymbol{u}_n \succeq 0, \forall n$$

局所特徴

コードブック

局所特徴がどのコードワードに所属するかを示す指標

一つのコードワードに属する制約→厳しすぎる！！！

- スパース符号化（Sparse Coding）

$$\min_{U,V} \sum_{n=1}^{N} ||\boldsymbol{x}_n - \boldsymbol{V}\boldsymbol{u}_n||^2 + \lambda|\boldsymbol{u}_n|$$

$$\text{s.t. } ||\boldsymbol{v}_k|| \leqq 1, \forall k$$

L1ノルム正則化項
→少数のコードワードへの所属を許容

L1正則化の役割

- コードブックは局所特徴の次元数よりも多く，過剰（K > D）なため，under determinedな系である．つまり情報が不足して解を定められない状況にある．そのためL1 正則化により解を定めることが可能となる．

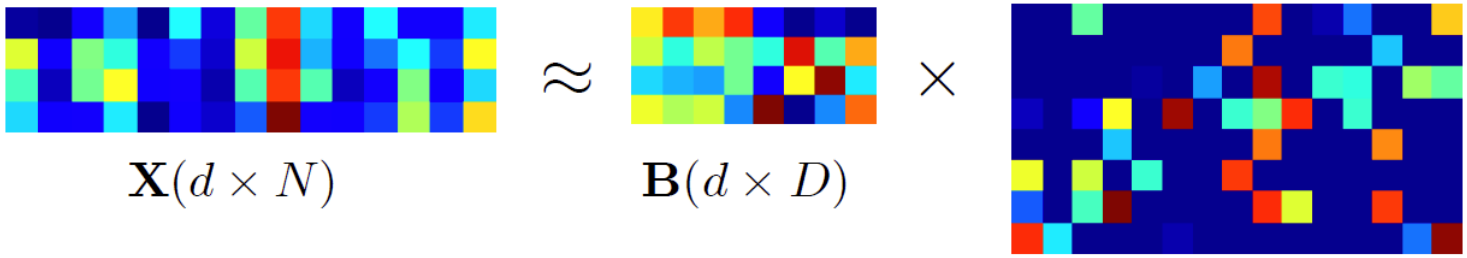
- *スパース性の事前知識を用いることによって*局所特徴の顕著なパターンを捉えることができる．

- *ベクトル量子化よりもスパース符号化の方が*量子化誤差を低減させられる．

局所座標符号化
Local Coordinate Coding (LCC)

- K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. NIPS, 2009.

http://www.image-net.org/challenges/LSVRC/2010/ILSVRC2010_NEC-UIUC.pdf



$\mathbf{X}(d \times N)$ $\approx$ $\mathbf{B}(d \times D)$ $\times$ $\mathbf{Z}(D \times N)$

Assume $\mathbf{B}$ is given.

Sparse coding:

$$\mathbf{z}^* = \arg\min_{\mathbf{z}} \tfrac{1}{2}\|\mathbf{x} - \mathbf{B}\mathbf{z}\|^2 + \lambda \sum_{i=1}^{D} |z_i|$$

局所性がスパースネスよりも本質！！

LCC: K. Yu et. al, NIPS 2009

$$\mathbf{z}^* = \arg\min_{\mathbf{z}} \tfrac{1}{2}\|\mathbf{x} - \mathbf{B}\mathbf{z}\|^2 + \lambda \sum_{i=1}^{D} \|\mathbf{x} - \mathbf{b}_i\|^2 |z_i|$$

Explicitly enforcing locality constraint

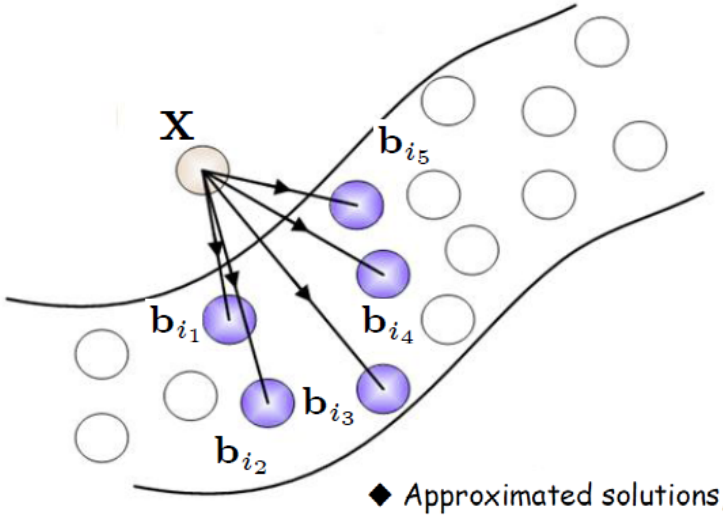
局所座標符号化の高速な実装

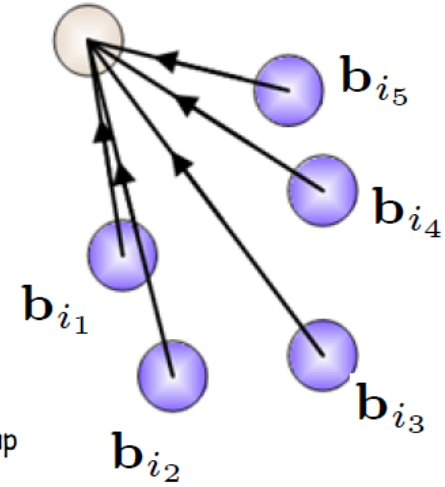
- 局所制約線形符号化
  - Locality-constrained Linear Coding (LLC)
  - J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. CVPR, 2010.

Step 1: be local to the test point $\mathbf{x}$

-- given $\mathbf{x}$, find its KNNs.

Step 2: small reconstruction error -- solve LMS fitting using only the KNNs





◆ Approximated solutions, but significant speedup

For a regular image (7k patches), with D=8192:
sparse coding needs ~10mins, (approximate) LCC needs only ~2s

http://www.image-net.org/challenges/LSVRC/2010/ILSVRC2010_NEC-UIUC.pdf

局所線形埋込み（Local Linear Embedding, LLE）と比較して，
局所制約線形符号化はコードブックの学習が入る点で異なる．

スパース符号化空間ピラミッド

- 空間ピラミッド
  - 符号化された局所特徴群 $\boldsymbol{U}$ から一つの特徴ベクトル $\boldsymbol{f}$ を得る手段

- プーリング（pooling）

$$\boldsymbol{f} = \mathcal{F}(\boldsymbol{U})$$

局所特徴集合

プーリング関数

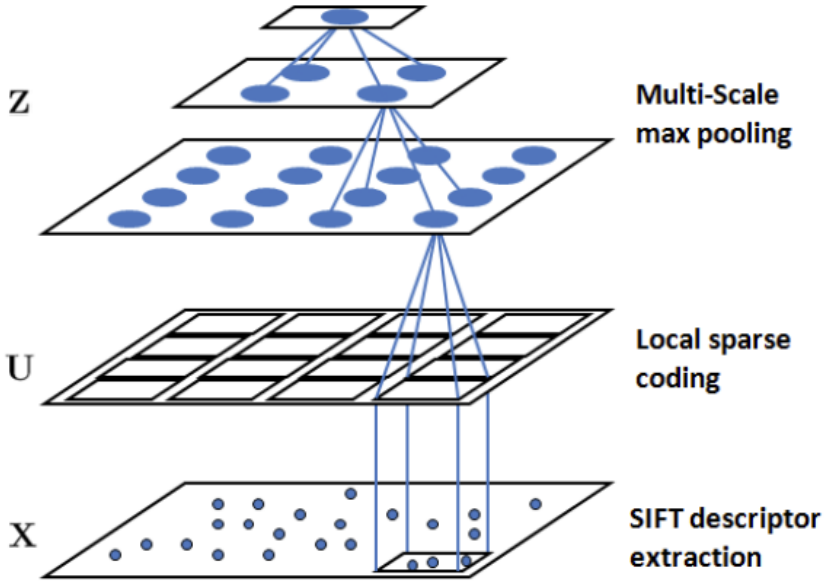
  - 平均プーリング
    average pooling

$$\boldsymbol{f} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{u}_n$$

BoFはこれを利用

  - 最大値プーリング
    max pooling

$$\boldsymbol{f}^d = \max\{|\boldsymbol{u}_1^d|, |\boldsymbol{u}_2^d|, \cdots, |\boldsymbol{u}_N^d|\}$$



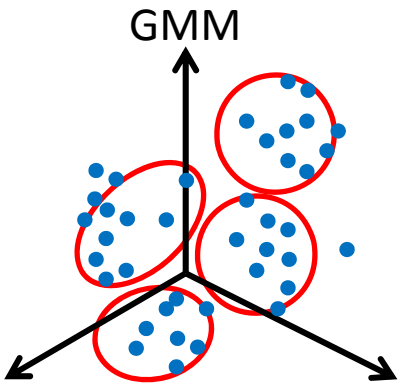
J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. CVPR, 2009.

画像表現
大域特徴

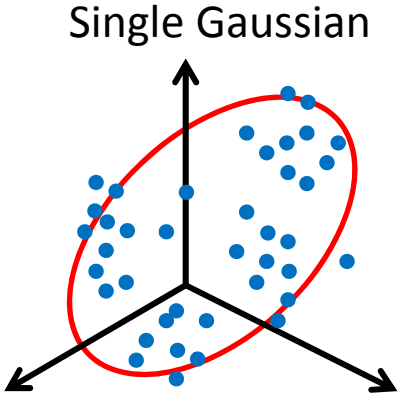
Generalized Local Correlation (GLC)

H. Nakayama, T. Harada, and Y. Kuniyoshi. Dense Sampling Low-Level Statistics of Local Features. In CIVR, 2009.

- GMM
  - 表現能力が高い
  - GMMはパラメータが多いので，共分散行列の非対角成分を0とする場合が多い．
  - 計算コストが高い



- Single Gaussian
  - 表現能力に限界有り
  - パラメータが少ないので，共分散行列推定可能
  - 共分散行列の非対角成分を有効活用
  - 計算コストが低い



局所記述子

平均 $\boldsymbol{\mu}^{(j)} = \frac{1}{p^{(j)}} \sum_k^{p^{(j)}} \boldsymbol{v}_k^{(j)}$

自己相関行列 $R^{(j)} = \frac{1}{p^{(j)}} \sum_k^{p^{(j)}} \boldsymbol{v}_k^{(j)} \boldsymbol{v}_k^{(j)T}$

GLC

$$\boldsymbol{x}^{(j)} = \begin{pmatrix} \boldsymbol{\mu}^{(j)} \\ upper(R^{(j)}) \end{pmatrix}$$

Generalized Local Correlation (GLC)

H. Nakayama, T. Harada, and Y. Kuniyoshi. Dense Sampling Low-Level Statistics of Local Features. In CIVR, 2009.

- GLCは単純であるが結構いける

Table 2: Comparison of the performance in two scene datasets and Caltech-101 (%). (*)approximate value read from the graph.

Dataset	GLC + PLDA			Previous	
	L1	L2	L3	no SI	with SI
OT8	88.8	90.5	91.1	82.3 [19]	90.2 [19]
				82.5 [3]	87.8 [3]
LSP15	80.0	83.2	84.1	72.7 [3]	83.7 [3]
				74.8 [11]	81.4 [11]
Caltech-101	55.0	63.3	64.8		72.0* [1]
					67.7 [3]
					66.2 [20]
				41.2 [11]	64.6 [11]
				58.2 [7]	
				39.6 [8]	

SPM+SVM

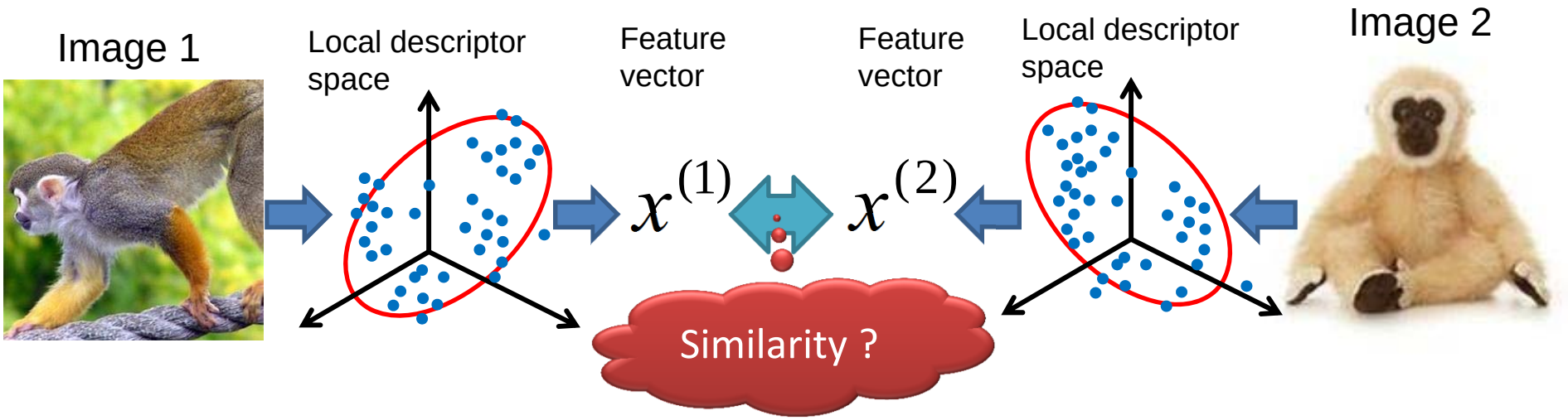
Figure 1: Sample images from the OT8 dataset.

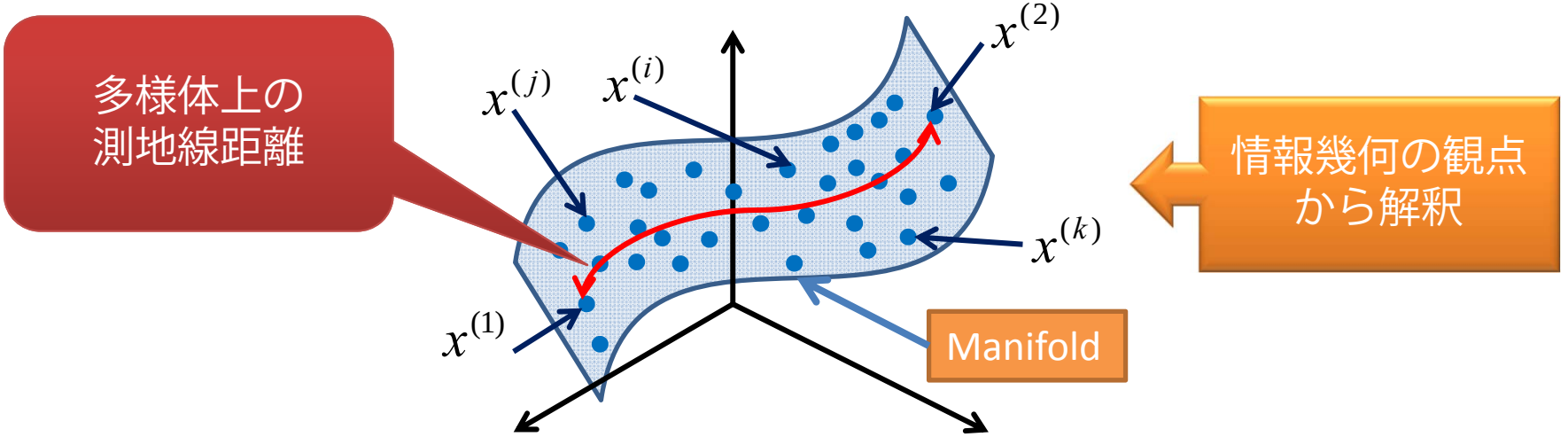
Figure 2: Additional seven classes in the LSP15 dataset.

Global Gaussian (GG)

H. Nakayama, T. Harada, and Y. Kuniyoshi. Global Gaussian Approach for Scene Categorization Using Information Geometry. In CVPR, 2010.

- 平均と分散を並べたGLCの表現は適切か？
- GLC間の距離計量は適切か？





Global Gaussian

H. Nakayama, T. Harada, and Y. Kuniyoshi. Global Gaussian Approach for Scene Categorization Using Information Geometry. In CVPR, 2010.

- 確率密度分布間の距離計量を正しく設定
- 情報幾何の考え方からGLCが自然に出てくる

η座標系におけるsingle Gaussianの表現

$$\begin{aligned}\boldsymbol{\eta} &= \sum_{1\le i\le d}\eta_i \boldsymbol{e}_i + \sum_{1\le i\le j\le d}\eta_{ij}\boldsymbol{e}_{ij} \\ &= (\eta_1,...,\eta_d,\eta_{11},...,\eta_{1d},\eta_{22},...\eta_{2d},...,\eta_{dd})^T \\ &= \left(\hat{\mu}_1,...,\hat{\mu}_d,\hat{\Sigma}_{11}+\hat{\mu}_1^2,...,\hat{\Sigma}_{1d}+\hat{\mu}_1\hat{\mu}_d,\right. \\ &\qquad \left.\hat{\Sigma}_{22}+\hat{\mu}_2^2,...,\hat{\Sigma}_{dd}+\hat{\mu}_d^2\right)^T.\end{aligned}$$

$=$

GLC

$$\boldsymbol{x}^{(j)} = \begin{pmatrix} \boldsymbol{\mu}^{(j)} \\ upper(R^{(j)}) \end{pmatrix}$$

GLCの厳密な距離

$$\begin{aligned}dist\big(\boldsymbol{\eta}(P),\boldsymbol{\eta}(Q)\big) &= tr(\Sigma_P\Sigma_Q^{-1}) + tr(\Sigma_Q\Sigma_P^{-1}) - 2d + \\ &\quad tr\left((\Sigma_P^{-1}+\Sigma_Q^{-1})(\boldsymbol{\mu}_P-\boldsymbol{\mu}_Q)(\boldsymbol{\mu}_P-\boldsymbol{\mu}_Q)^T\right)\end{aligned}$$

$$K_{kl}(P,Q) = \exp\big(-a\, dist\big(\boldsymbol{\eta}(P),\boldsymbol{\eta}(Q)\big)\big)$$

Gauss分布間のsymmetric KL-divergence

GLCの近似的な距離

Fisher Information Matrix

$$K_{ct}(P,Q) = \boldsymbol{\eta}(P)^T G^{\eta}(\boldsymbol{\eta}_c)\boldsymbol{\eta}(Q) \quad \Rightarrow \quad \boldsymbol{\zeta} = \big(G^{\eta}(\boldsymbol{\eta}_c)\big)^{1/2}\boldsymbol{\eta}$$

Linear-SVMにそのまま利用可

Global Gaussian (GG)

H. Nakayama, T. Harada, and Y. Kuniyoshi. Global Gaussian Approach for Scene Categorization Using Information Geometry. In CVPR, 2010.

- 性能評価

Table 5. Performances of global Gaussian, BoK, and combined approach (%). $L = 2$ spatial pyramid is implemented. Kernel PDA is used for classification. SURF descriptor is used for LSP15 and SIFT descriptor is used for 8-sports.

	LSP15	8-sports
GG (KL)	86.1±0.5	84.4±1.4
GG (ct-linear)	82.3±0.4	82.9±1.0
BoK200	81.1±0.7	79.6±1.1
BoK1000	82.5±0.7	81.5±1.7
GG (ct-linear) + BoK200	85.0±0.5	83.2±0.9
GG (ct-linear) + BoK1000	85.3±0.5	83.4±0.7

提案手法のスコア

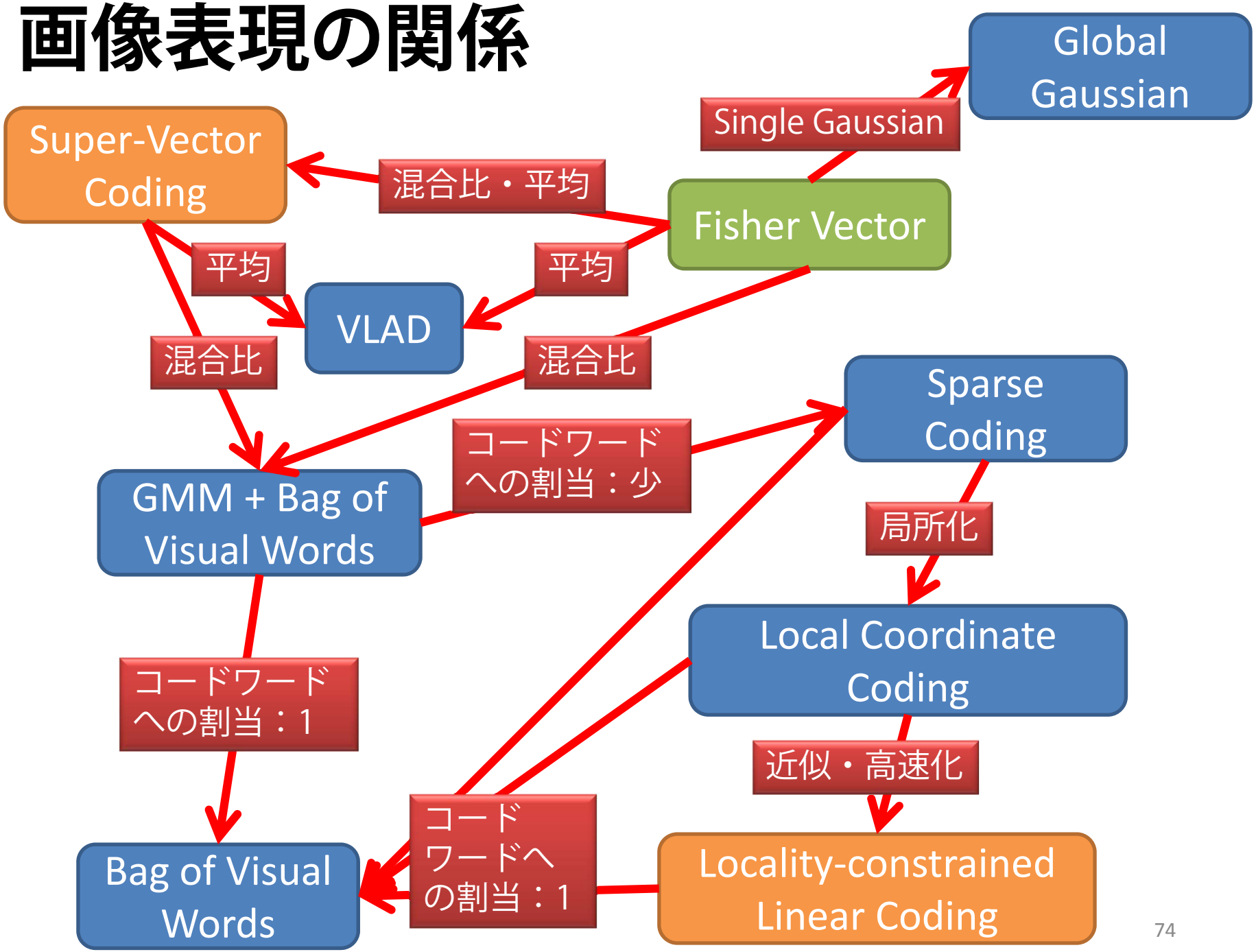
Table 6. Performance comparison with previous work (%). For our method, $L = 2$ spatial pyramid is implemented, and kernel PDA is used for classification. We use the SURF descriptor for LSP15 and Indoor67, and the SIFT descriptor for 8-sports.

Method	LSP15	8-sports	Indoor67
GG (KL-div.)	86.1±0.5	84.4±1.4	45.5±1.1
GG (ct-linear) + BoK1000	85.3±0.5	83.4±0.7	44.9±1.3
Previous	85.2 [30]	84.2 [29]	25.0 [23]
	84.1 [29]	73.4 [14]	
	83.7 [6]		

従来手法のスコア

いずれのデータセットにおいても従来手法を上回る性能
平均，共分散といったパラメータのみで計算可能

画像表現の関係


Global Gaussian
Single Gaussian
Super-Vector Coding
混合比・平均
Fisher Vector
平均
平均
VLAD
混合比
混合比
Sparse Coding
コードワードへの割当：少
GMM + Bag of Visual Words
局所化
Local Coordinate Coding
コードワードへの割当：1
近似・高速化
コードワードへの割当：1
Bag of Visual Words
Locality-constrained Linear Coding

画像表現の性能比較

- K. Chatfield, V. Lempitsky, A. Vedaldi and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. BMVC, 2011.

Method				mAP								
(a) FK	Lin	ss3	256	**61.69**	78.97	67.43	51.94	70.92	30.79	72.18	79.94	61.35
(b) SV	Lin	ss3	1024	**58.16**	74.32	63.79	47.02	69.44	29.06	66.46	77.31	60.18
(c) LLC	Lin	ss2	25k	**57.60**	71.05	62.85	47.40	67.67	25.21	62.70	77.02	59.59
(d) LLC-F	Lin	ss2	25k	**59.32**	74.10	64.92	51.48	68.33	27.18	62.89	78.44	61.39
(e) VQ	Chi	ss2	25k	**56.07**	70.00	58.90	42.86	66.75	26.59	62.27	75.67	57.09
(f) LLC	Lin	ss3	25k	**57.27**	71.35	62.65	46.12	68.98	26.04	63.92	76.98	59.71
(g) LLC	Sqr	ss3	25k	**56.71**	71.24	61.75	42.73	68.21	25.85	62.33	76.40	59.31
(h) LLC	Chi	ss3	25k	**57.66**	72.41	62.19	47.30	68.91	25.78	63.95	77.27	59.83
(i) LLC-F	Lin	ss3	25k	**59.74**	74.17	65.39	51.15	69.69	28.67	64.40	78.48	63.00
(j) VQ	Chi	ss3	25k	**55.30**	70.10	59.24	44.14	66.34	26.79	60.88	75.62	55.42
(k) KCB	Chi	ss3	25k	**56.26**	70.83	60.60	44.50	66.52	27.02	62.07	76.29	57.61
(l) LLC	Lin	ss5	25k	**56.96**	69.82	61.63	46.71	68.27	25.66	63.78	76.32	59.83
(m) LLC-F	Lin	ss5	25k	**58.70**	73.44	62.90	50.22	67.90	27.85	64.35	77.91	62.44
(n) VQ	Chi	ss5	25k	**53.87**	68.74	57.14	41.24	64.54	25.20	61.12	74.06	53.22
(o) LLC	Lin	ss3	14k	**56.18**	70.71	59.67	44.81	67.20	26.03	60.99	76.25	58.54
(p) VQ	Chi	ss3	14k	**54.82**	69.09	58.61	41.27	66.30	26.49	61.46	75.42	55.77
(q) LLC	Lin	ss3	10k	**56.01**	69.66	60.44	44.21	67.78	24.66	61.84	75.42	57.70
(r) VQ	Chi	ss3	10k	**54.98**	69.56	57.97	42.86	65.84	23.52	61.06	75.89	55.55
(s) LLC	Lin	ss3	4k	**53.79**	69.83	57.63	42.04	66.46	22.44	55.62	72.77	56.98
(t) LLC	Sqr	ss3	4k	**52.07**	68.52	54.62	40.14	65.34	21.53	51.89	71.54	55.19
(u) LLC	Chi	ss3	4k	**53.47**	70.17	56.20	42.73	65.27	22.23	55.18	72.78	56.95
(v) LLC-1	Lin	ss3	4k	**36.06**	53.39	43.20	22.47	46.32	11.40	29.48	64.66	45.41
(w) LLC-F	Lin	ss3	4k	**55.87**	72.27	61.41	44.08	67.85	24.97	57.92	75.40	59.44
(x) VQ	Sqr	ss3	4k	**51.97**	67.29	55.22	36.58	64.42	21.89	56.31	72.90	52.11
(y) VQ	Chi	ss3	4k	**53.42**	68.65	57.04	39.86	64.59	21.96	58.79	73.89	53.77
(z) VQ	Lin	ss3	4k	**46.54**	60.63	48.80	32.76	58.54	16.26	50.44	68.42	45.97
(α) KCB	Chi	ss3	4k	**54.60**	69.82	59.20	41.97	64.85	23.90	59.02	74.98	54.63

Fisher Vectorは結構いい.

【データセット】
- Pascal VOC 2007

【比較した画像表現】
- VQ: Bag of Words
- FK: Fisher Vector
- SV: Super Vector
- LLC: Locality-constrained Linear Coding
- KCB: Kernel Codebook

効率的な識別機

膨大なクラス識別における識別機

- 特徴ベクトルの次元を高次元に保つ

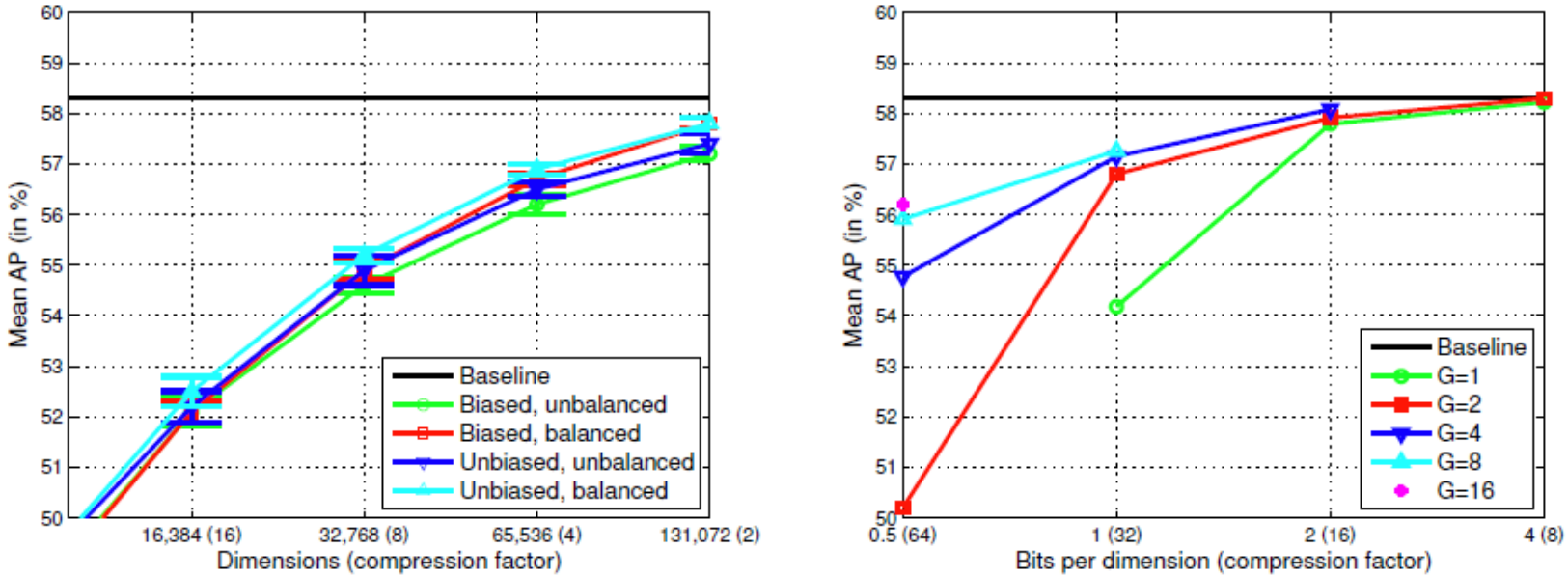


Figure 2. Compression results on VOC 2007. Left: HK results as a function of the number of dimensions. Right: PQ results as a function of the number b of bits per dimension and the group size G (without sparsity encoding). The baseline corresponds to the uncompressed signature (262,144 dimensions). For a given compression factor, PQ performs much better than HK.

J. Sanchez, and F. Perronnin. High-Dimensional Signature Compression for Large-Scale Image Classification. In CVPR, 2011.

高次元の特徴であっても破綻しない識別機が必要→SVMの利用が多い

SVMの逐次学習

- 多クラスSVM
  - One-vs-all SVMがほとんど
  - 各クラスのSVM出力値の大小関係が適切である保証はない．しかしながら十分高い識別率が得られることが知られている．
  - 各クラス毎に独立に学習，識別ができる→容易に並列化でき，大規模データに有効

- SVMのオンライン学習
  - 訓練サンプルを1個メモリにロードする．
  - 現状の識別機で識別し，誤っていたら識別機を更新する．
  - →訓練サンプルを全てメモリに展開する必要がないの，大規模データに適する．

- 確率的勾配降下法（Stochastic Gradient Decent）によるSVM

評価関数

$$L = \sum_{t=1}^{T} L(\boldsymbol{w}, b, \boldsymbol{x}_t, y_t) = \sum_{t=1}^{T} \frac{\lambda}{2} ||\boldsymbol{w}||^2 + \max\left[0, 1 - y_t(\boldsymbol{w}^\top \boldsymbol{x}_t + b)\right],$$

パラメータの更新

$$\boldsymbol{w}^t = \boldsymbol{w}^{t-1} - \eta \nabla_w L(\boldsymbol{w}, b, \boldsymbol{x}_t, y_t), \quad b^t = b^{t-1} - \eta \nabla_b L(\boldsymbol{w}, b, \boldsymbol{x}_t, y_t).$$

平均化：高速化

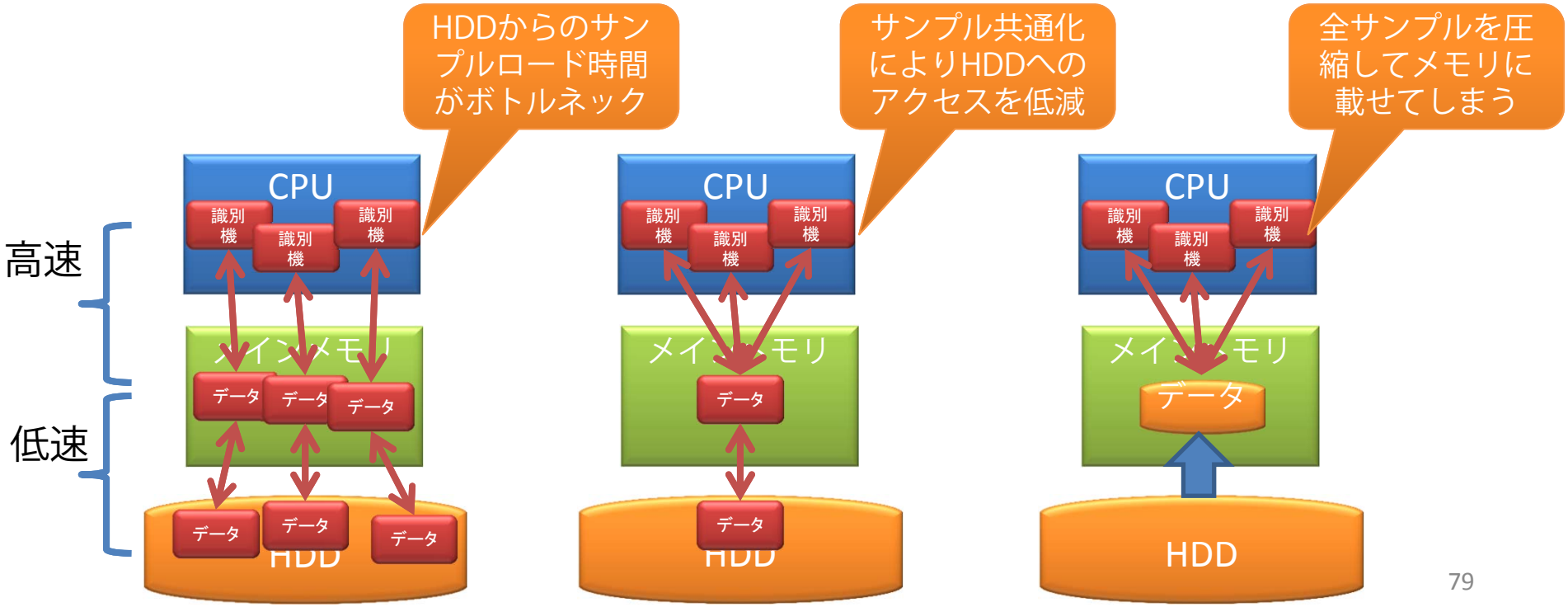
$$\bar{\boldsymbol{w}}^t = (1 - 1/t)\bar{\boldsymbol{w}}^{t-1} + \boldsymbol{w}^t/t, \quad \bar{b}^t = (1 - 1/t)\bar{b}^{t-1} + b^t/t.$$

I/Oのボトルネック対策

- 計算速度 （高速）>> HDDのアクセス速度（低速）
  - ボトルネックは補助記憶装置への頻繁なアクセス
  - データが膨大なので全てのデータをロードするだけで一苦労

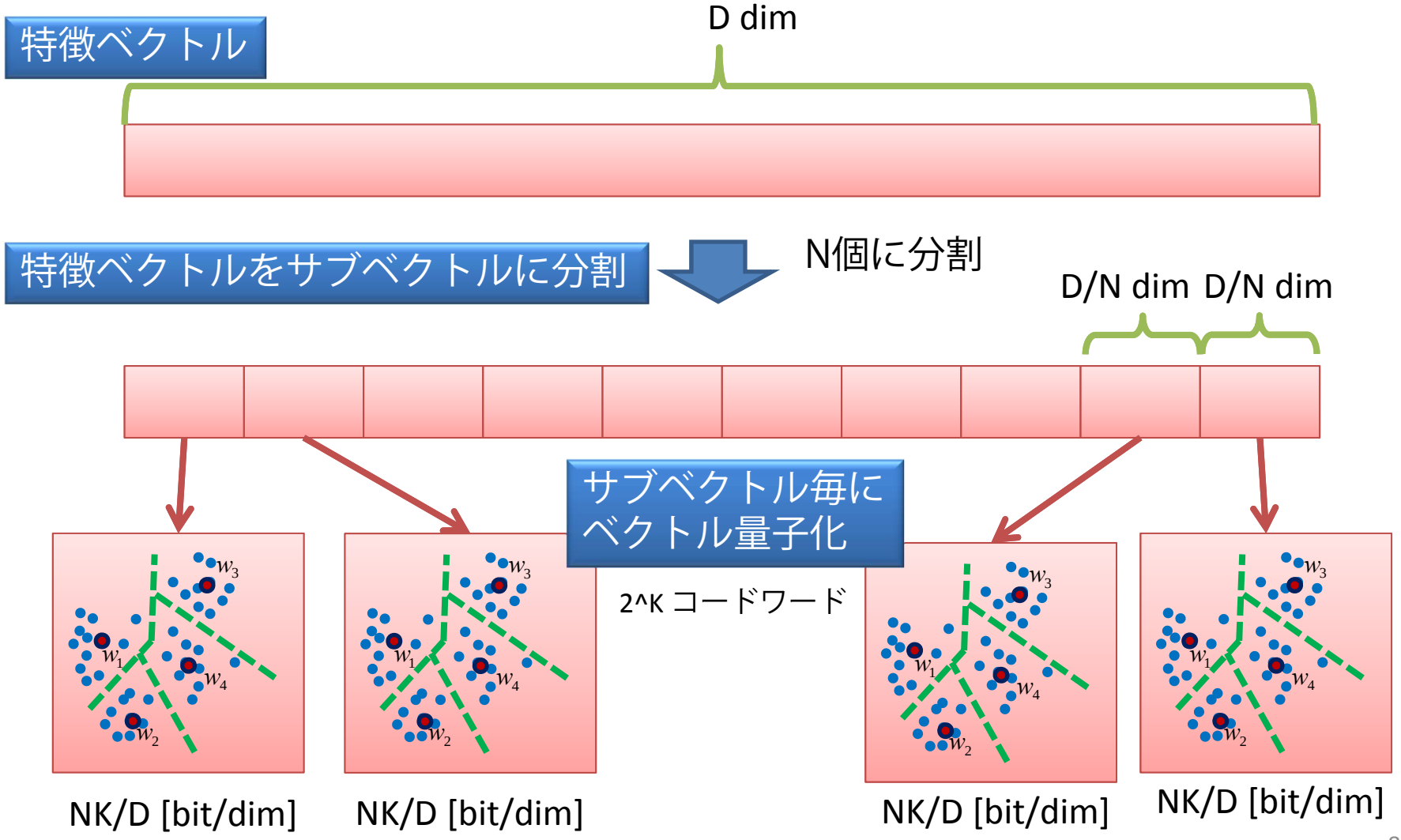
- NEC system
  - Hadoopによる並列化
  - One-vs-all SVMの識別機を並列に学習する際に，学習サンプルを共通化.

- XRCE system
  - 学習データを圧縮し，全ての学習データをメインメモリに押し込む.

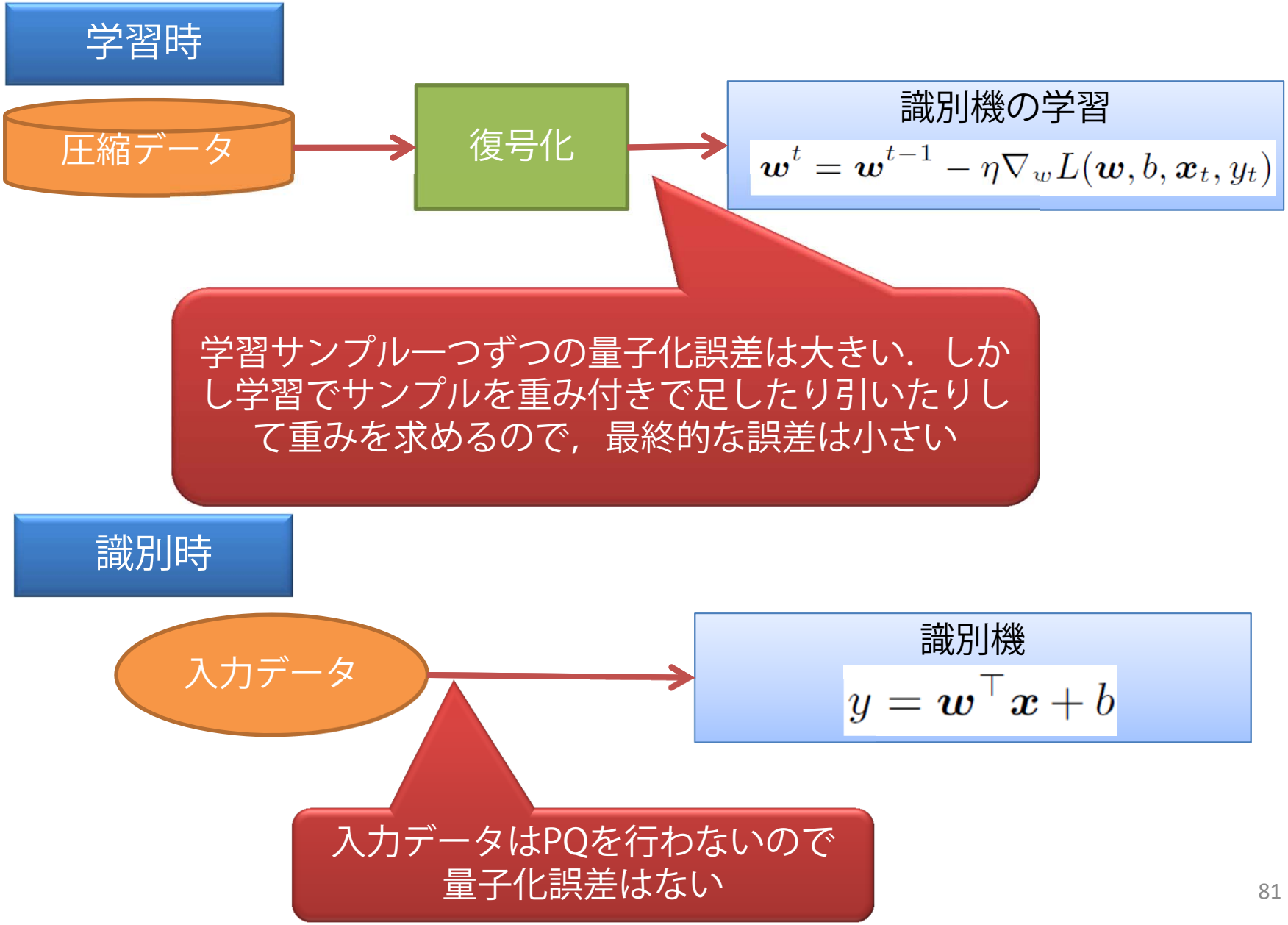


プロダクト量子化

- H. Jegou, M. Douze, and C. Schmid. Product Quantization for Nearest Neighbor Search. IEEE Trans. on PAMI, Vol.33, pp.117-128, 2011.

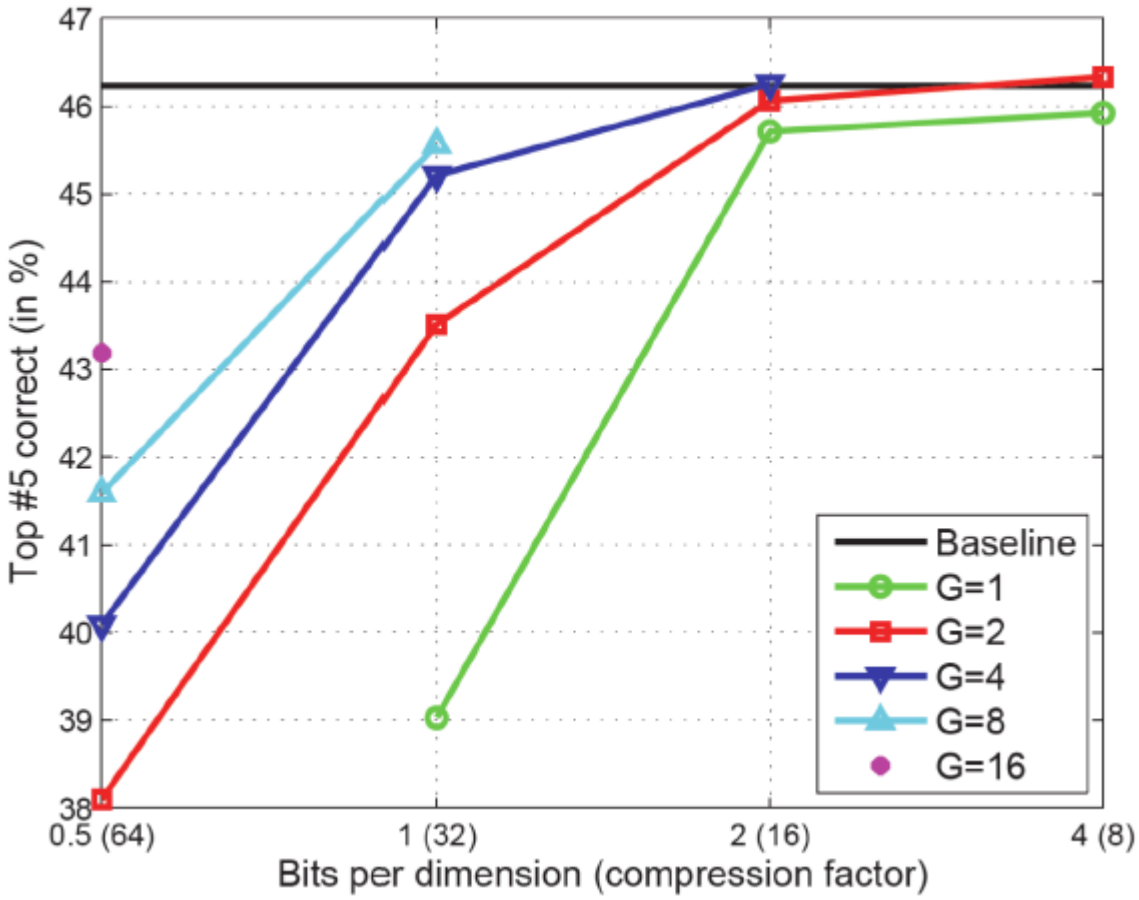


SGD-SVM + PQ


学習時
圧縮データ
復号化
識別機の学習
$\boldsymbol{w}^t = \boldsymbol{w}^{t-1} - \eta \nabla_w L(\boldsymbol{w}, b, \boldsymbol{x}_t, y_t)$
学習サンプル一つずつの量子化誤差は大きい．しかし学習でサンプルを重み付きで足したり引いたりして重みを求めるので，最終的な誤差は小さい
識別時
入力データ
識別機
$y = \boldsymbol{w}^\top \boldsymbol{x} + b$
入力データはPQを行わないので
量子化誤差はない

PQ + SGD-SVMの実験結果

- 1次元1bitにしても性能低下は少ない



http://www.image-net.org/challenges/LSVRC/2011/ilsvrc11.pdf

まとめ

- 大規模画像データセットを用いた画像認識のトレンドについて紹介した.

- 近年，大規模画像識別に用いられている画像表現を紹介し，それらの体系化の試みを解説した.