

第14回情報論的学習理論ワークショップ (IBIS 2011) 2011年11月11日

画像認識検索における特徴量表現

東京大学/JSTさきがけ
原田達也

これは何でしょう？



beach, water, people, kauai, tree



ocean, coral, fish, angelfish, reefs

Retrieving by “flight”



Image Annotation Results on Corel5K



birds, booby,
flight, rocks,
water



buildings, ships,
bridge, flag, sky



church, stone,
buildings, chapel,
people



sky, people, close-
up, statue, clouds



buildings,
water, city,
light, night



people, woman,
indian, pots,
baby



cat, tiger, water,
rocks, forest

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

一般画像認識:

上記の分類を包含したセマンティックスレベルの画像認識

一般的な視覚認識機能の困難さ

- 人の認識の曖昧性: weak labeling



Jet plane sky



cat tiger forest tree



beach people water oahu

- 文脈の考慮



皿,
鍋,
やかん,
包丁



コップ,
急須

- Data drivenの特徴と意味との不一致: semantic gap
- 大量の学習データへのスケーラビリティ
- 多様な環境への対応：高速かつ安定な追加学習

一般画像認識 (image annotation) のモデル

Blob to Label

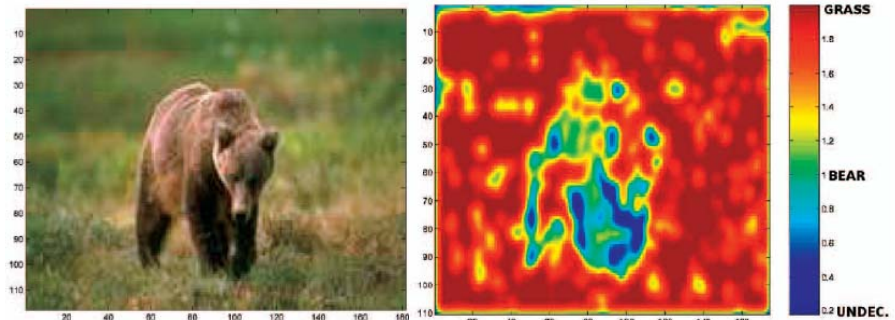
Translation Model. Duygulu et al., 2002.
CMRM. Jeon et al., 2003.



From Duygulu et al., ECCV2002.

Pixel to Label

SML. Carneiro et al., 2007.



From Carneiro et al., PAMI2007.

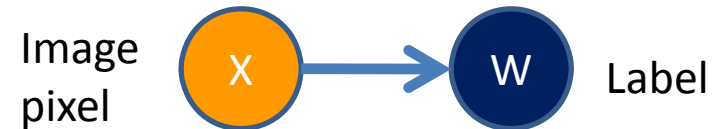
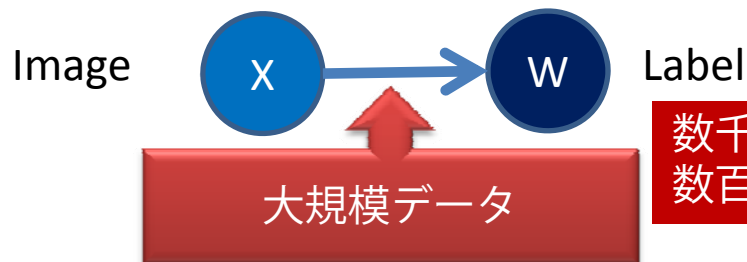


Image Instance to Label

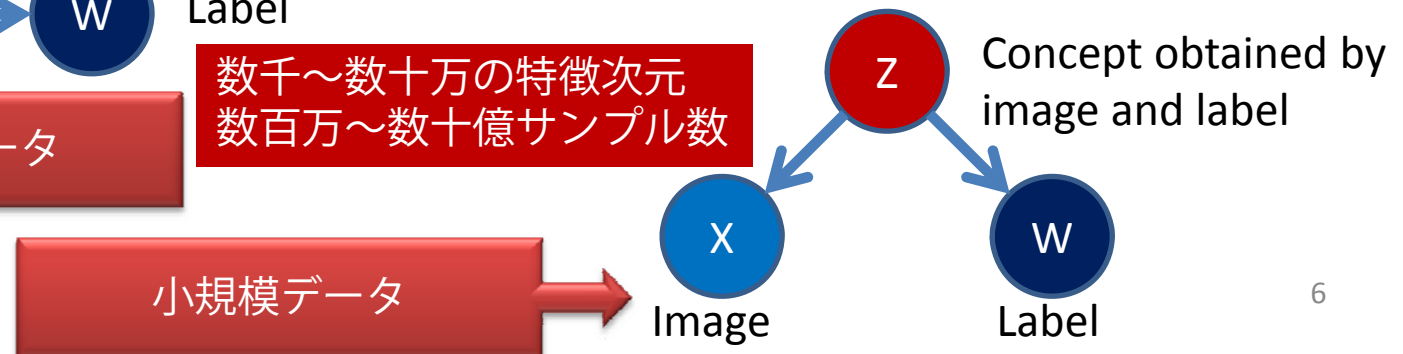
MBRM. Feng et al., 2004.
JEC. Makadia et al., 2008.



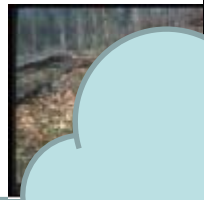
数千~数十万の特徴次元
数百万~数十億サンプル数

Concept to Label

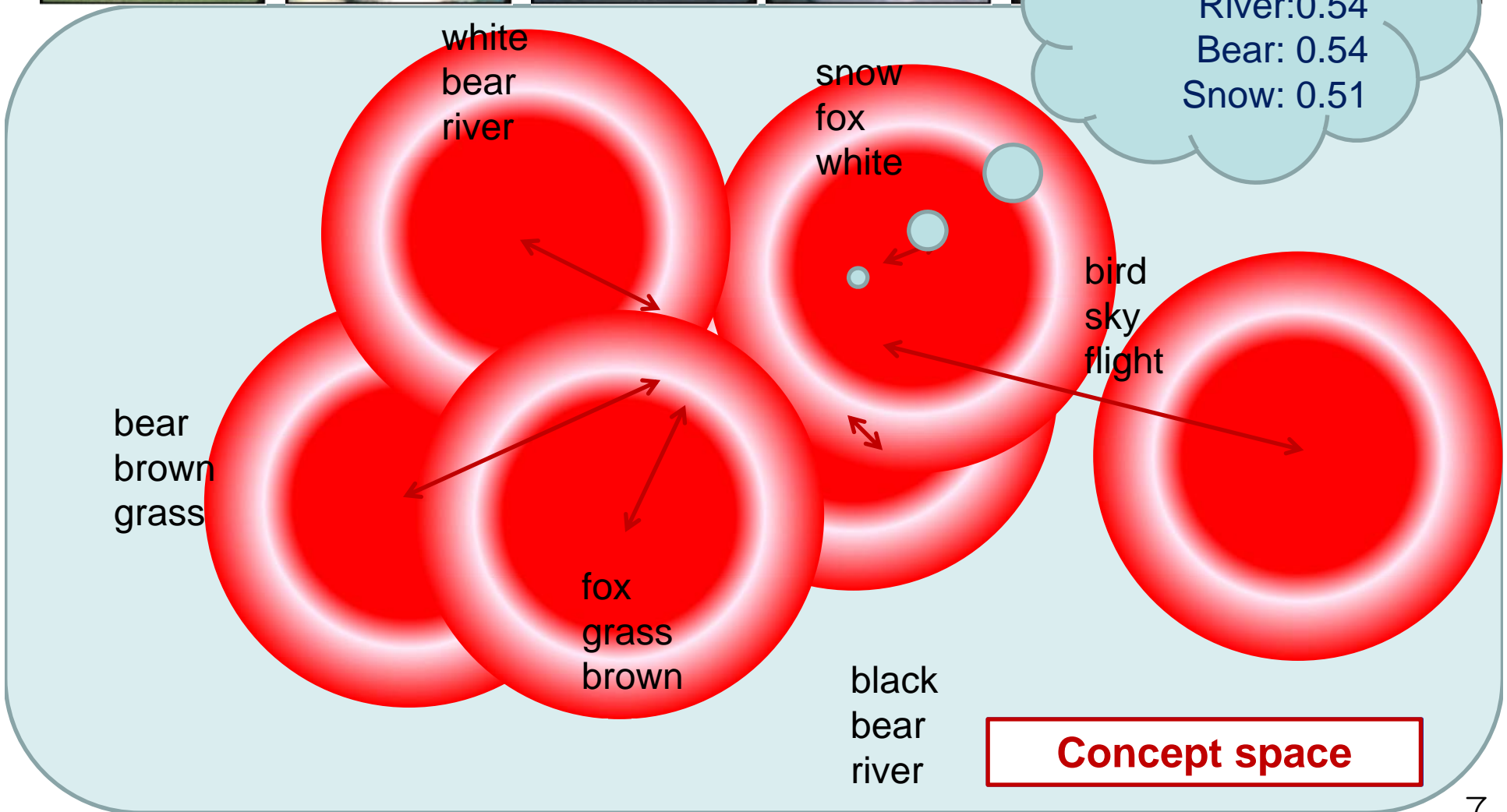
CorrLDA. Blei and Jordan, 2003.
KCCA. Hardoon et al., 2006.
CCA. Kurita et al., 1992.



Concept Learning and Image Annotation



Fox: 0.90
White: 0.83
River: 0.54
Bear: 0.54
Snow: 0.51

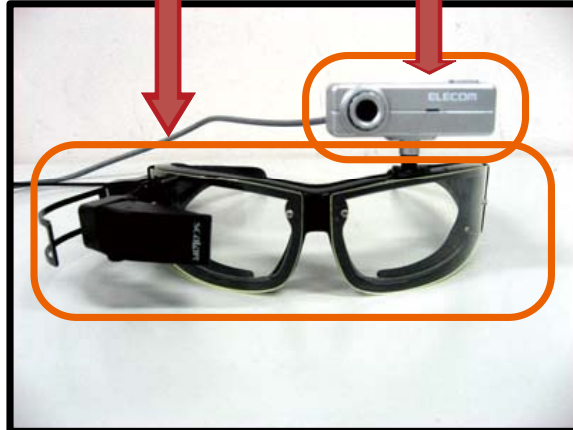


興味：実世界認識知能の構築

- AI Goggles
 - 実世界のリアルタイム記述と言葉による視覚記憶の検索
 - 高速かつ安定なオンライン学習

Head Mount Display
displays labels of objects and scene with the image.

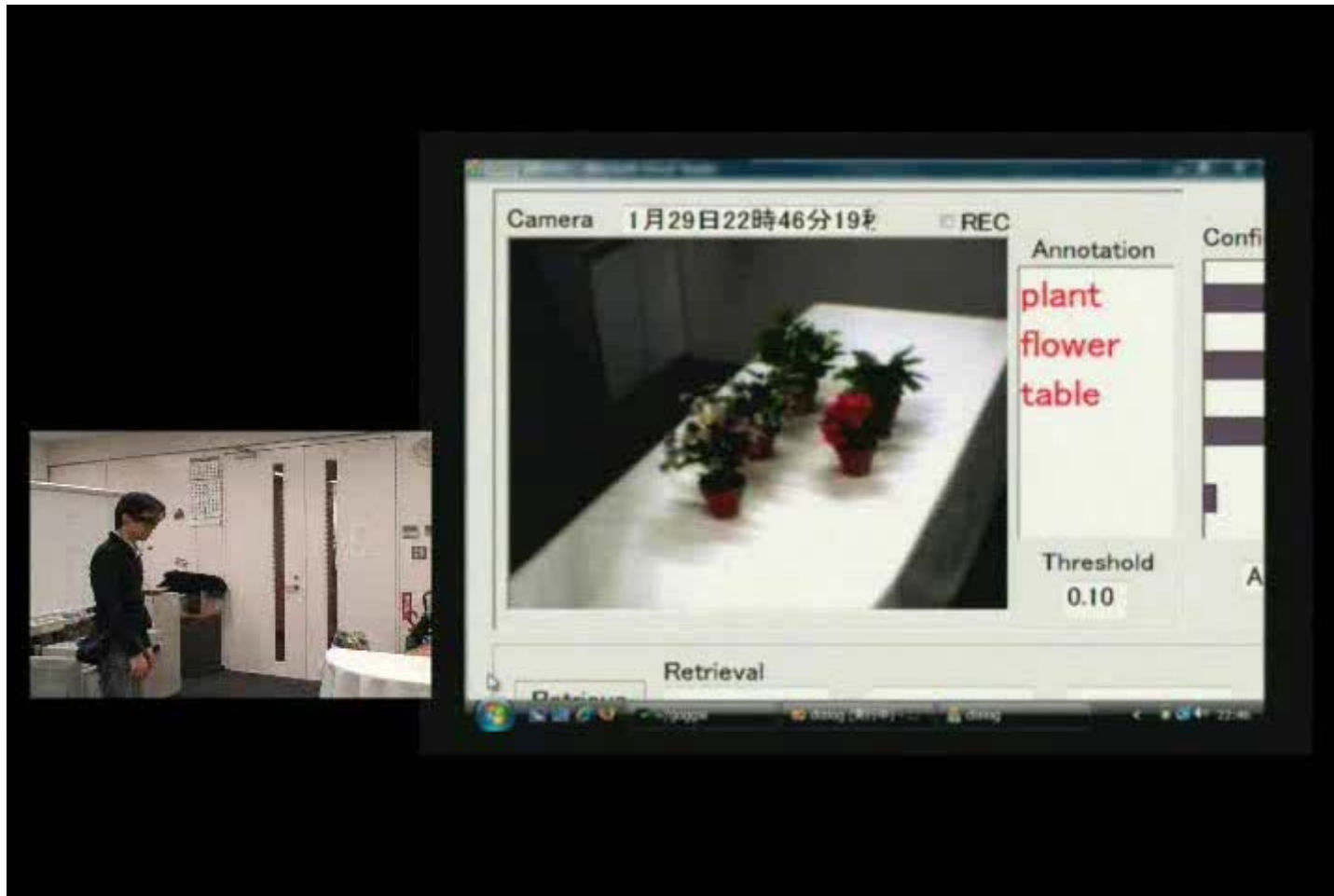
Camera
captures images, which the wearer is seeing now.



Portable Computer
recognizes the images quickly that the camera is capturing, and shows the results on HMD. Moreover, it accumulates the images and labels, and enables the wearer to search those images by labels.



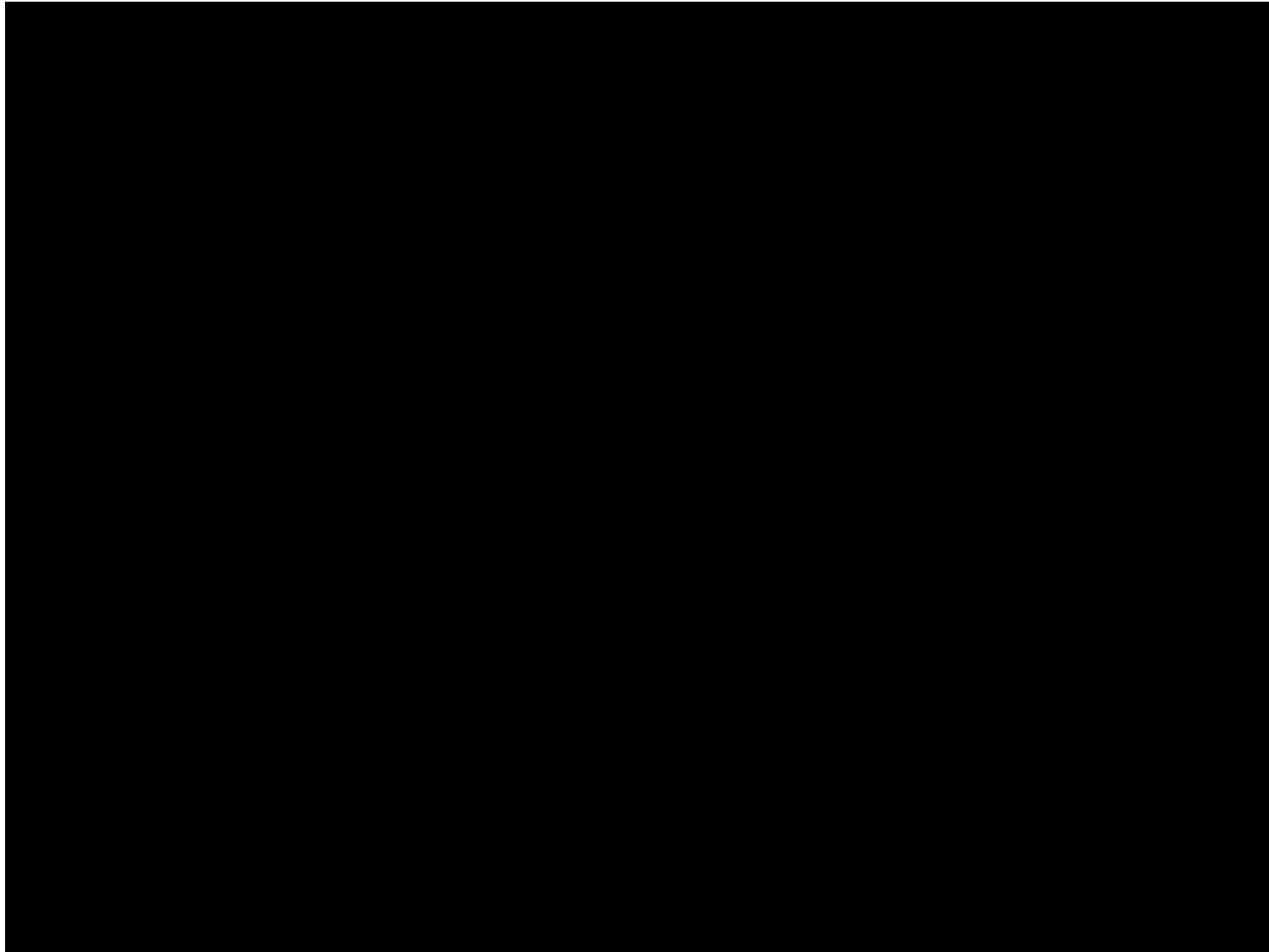
AI Goggles 実世界におけるリアルタイムアノテーション



AI Goggles on iPhone Project



AI Goggles on iPhone prototype



Déjà Vu demo

Demo déjà-vu system

- 1 sec
 - 16 byte/frame
- 1 day
 - 16 byte x 1 fps x 60 sec x 60 min x 16 hour = 1 MB
- 1 year
 - 1 MB x 365 = 365 MB
- 1 life
 - 365 MB x 80 = 30 GB



www.apple.com

= 1 year

iPhone4S: 512MB



www.apple.com

= 1 life

Mac Pro: 64GB

[PR] 今なら!光ファイバー申込みで最大52,300円キャッシュバック中!

[ログイン] [新規ID登録] [閲覧履歴] [ご利用ガイド]

USBフラッシュメモリ 32GB 価格比較

ホーム > パソコン > USBメモリ > メーカー問わず > USBフラッシュメモリ 32GB

メーカー問わず
USBフラッシュメモリ 32GB

USB memory 32GB

最安価格(税込): **¥2,138** | 登録価格一覧(13店舗) | 価格変動履歴

価格帯: ¥2,138~¥11,779(13店舗) | メーカー希望小売価格: ¥- | 登録日: 2008/04/10

スペック
メモリ容量: 32GB (スペック詳細)

さらに賢くお買い物! 光ファイバー加入で
現金最大52,300円 キャッシュバック!!

✖ 下取り価格を調べる
➡ メーカー問わずのUSBメモリー

売れ筋ランキング: **1位** (4069製品中)
満足度: **5.00** (採点者2人)
クチコミ: **28件** (最終更新: 2011年6月24日)

ご利用の前にお読みください | ツイートする | メッセージ共有 | チェック | 気になる | いいね!

価格比較 | スペック | 特長 | レビュー | クチコミ | 画像・動画 |ピックアップリスト | 関連製品 | オークション

USBフラッシュメモリ 32GB の価格比較 - 送料別価格 (取り扱い店舗数: 13店舗)

支払方法アイコンについて: 代:代金引換 | 振:銀行振込 | ク:クレジットカード | 価格.com 安心支払い | 振:銀行振込 | ク:クレジットカード | コ:コンビニ支払い

順位	価格(差額) 条件指定	送料	在庫・発送 の目安 詳細	地域	ショップ/評価/支払方法 詳細	コメント	ショップサイト
1位	¥2,138 (最安)	¥210~	有	千葉	上海問屋 - 本店 評価: [5.0] ショップ詳細 代: [] 振: [] ク: [] コ: []	カード支払OK! 平日17時(土日祝日16時)までのご注文は当日発送!	ショップの売り場へ行く
1位	¥2,138 (最安)	¥200~	有	東京	デジタル達人 評価: [5.0] ショップ詳細 代: [] 振: [] ク: [] コ: []	★即納! 楽天ポイント	ショップの売り場へ行く

Ads by Google

社内のUSB管理できてる?
USBで持ち出した情報の流出を防ぐ ALSIのUSB暗号化ソフトを体験
www.alsico.jp

USBメモリ修復&データ復元
認識しないUSBメモリ、データが消失したUSBメモリからデータ復元可能!
www.rescue-center.jp/

情報漏洩をしたくないなら
米政府機関も使用している 暗号化USBメモリ「鉄」

Your life is 2000 yen!

知識はどこから降ってくる？

- 人による教示
 - クラウドソーシング 
- インターネットにある膨大な情報を利用
 - 大規模画像処理 
- 対象とは別の知識を活用
 - 転移学習

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.

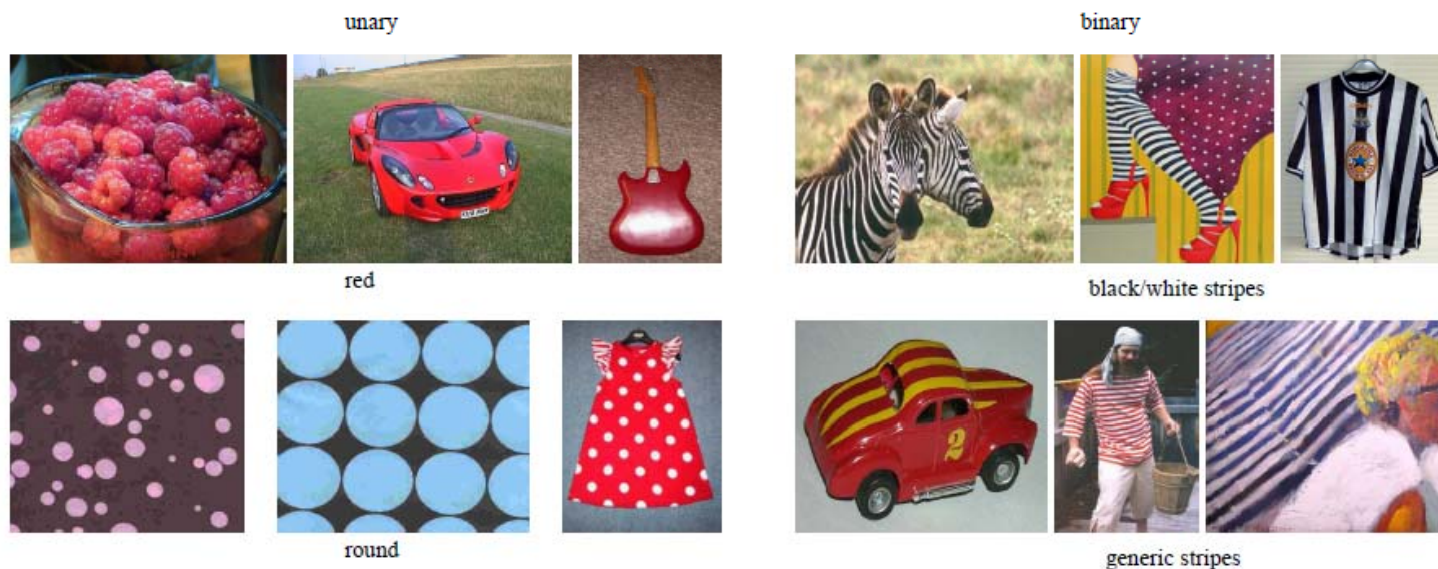


Figure 1: *Examples of different kinds of attributes. On the left we show two simple attributes, whose characteristic properties are captured by individual image segments (appearance for red, shape for round). On the right we show more complex attributes, whose basic element is a pair of segments.*

V. Ferrari and A. Zisserman. Learning visual attributes. In NIPS, 2008.

知識転移のアプローチ

M. Rohrbach, M. Stark, and B. Schiele. Evaluating Knowledge Transfer and Zero-Shot Learning in a Large-Scale Setting. CVPR, 2011.

- 階層構造の利用
 - Inner WordNet nodes model
 - All WordNet nodes model
 - Leaf nodes, cost sensitive

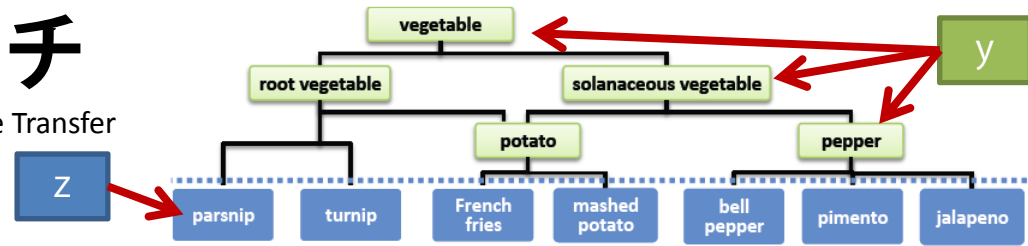


Figure 1: ISVLR10 subgraph. Leaf (blue), inner nodes (green).

$$s^{inn}(z_l|x) = \frac{\sum_{y_i \in H_{z_l}} s(y_i|x)}{|H_{z_l}|}$$

内ノードの識別機 (inner node classifier)
親ノードの集合 (parent node set)

$$s^{all}(z_l|x) = \frac{s(z_l|x) + \sum_{y_i \in H_{z_l}} s(y_i|x)}{1 + |H_{z_l}|}$$

葉ノードの識別機 (leaf node classifier)

$$s^{cost}(z_l|x) = - \sum_{z_i \in Z_l} c_{z_i}^{z_l} p(z_i|x)$$

属性識別機 (attribute classifier)
階層コスト (hierarchical cost)
肩にかかっているけど積では??? (should be multiplied but it's not?)

$$s^{attr}(z_l|x) = \frac{\sum_{m=1}^M s(a_m|x)^{a_m^{z_l}}}{\sum_{m=1}^M a_m^{z_l}}$$

属性と葉ノードの関係を示す指標 (indicator showing relationship between attribute and leaf node)

前提：物体とアトリビュートの関係はgiven (premise: relationship between object and attribute is given)

アトリビュートの利用

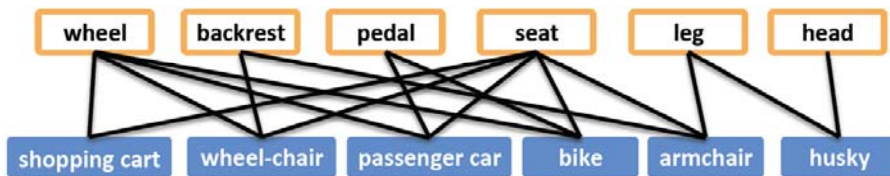


Figure 2: Example part attributes (orange), object classes (blue).

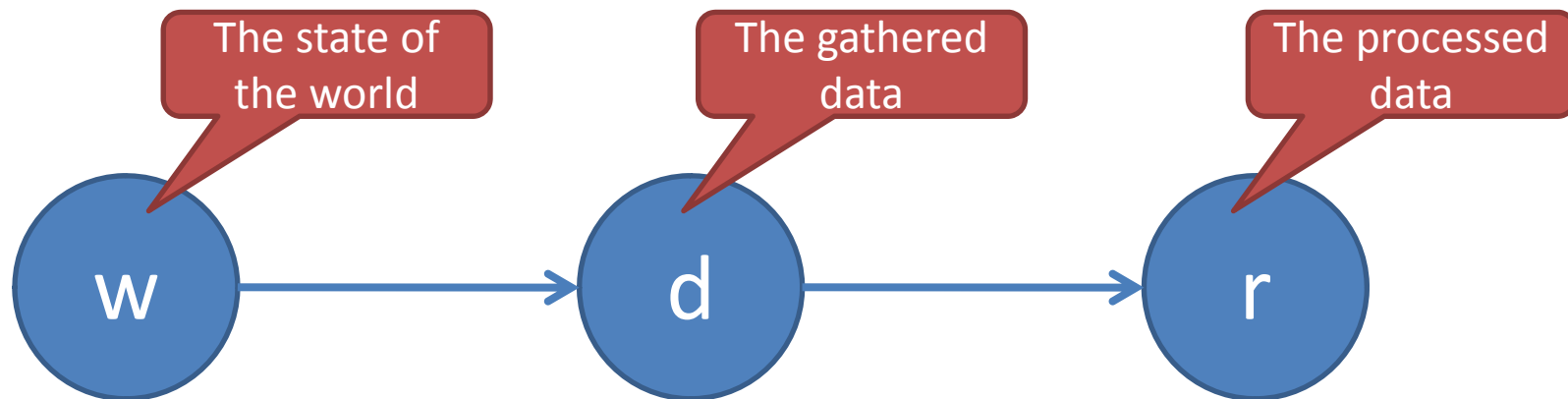
直接的類似度の利用

前提：物体間の関係はgiven (premise: relationship between objects is given)

$$s^{dir}(z_l|x) = \frac{\sum_{k=1}^K s(z_k|x)}{K}$$

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		
		12:30-14:00 Lunch	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		



Markov chain

The data processing theorem:

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

The average information

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

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

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

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

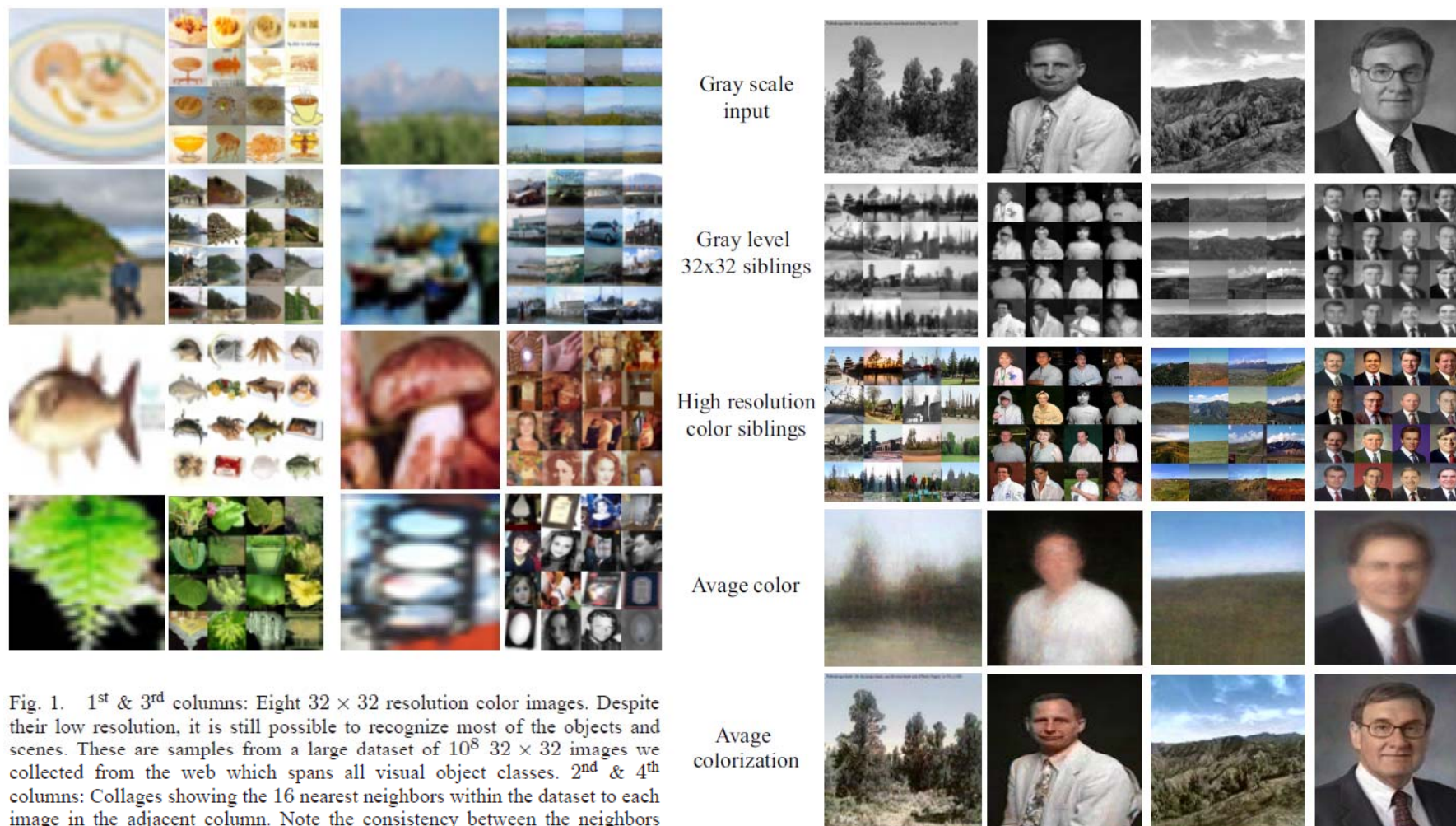


Fig. 1. 1st & 3rd columns: Eight 32 × 32 resolution color images. Despite their low resolution, it is still possible to recognize most of the objects and scenes. These are samples from a large dataset of 10⁸ 32 × 32 images we collected from the web which spans all visual object classes. 2nd & 4th columns: 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.





	prison break sarah callies sara tancredi looking (339 dups)	sarah wayne callies picture thread bild-quelle edit by annika beitraege in einen... prison break is paging dr. sara. if you are one of the many prison break fans... prison break - dr sara tancredi is not dead you knew that, right?dr sara tancredi ... dr. sara comes back to prison break ?		aeon concept phone mobile phone cell phone touch screen nokia phone mobile nokia (1888 dups)	nokia aeon was presented by nokia on their website in the research development... nokia aeon concept phone (no ratings yet) sexy is the word to describe it nokia is ... nokia aeon - future mobile phone nokia aeon concept phone nokia has unveiled its latest concept unbelievable ...
	costa rica golden toad climate amphibian (18 dups)	this is a picture of male golden toads congregating for breeding... is there a relationship between climate variability & amphibian declines? golden toad male golden toads at a breeding pool in indigenous to monteverde costa rica ... amphibian declines in the cloud forests of costa rica ...		sydney opera house australia (19 dups)	enjoying the wet season in australia sydney ... 150975_ sydney_opera_house next ... 07/12. 1. tag in sydney > opera house ... kirsty and trudy drink wine sydney opera house ...

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.


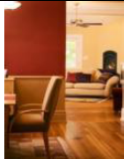




	2.4 M	80M	2B		2.4M	80M	2B
	(no results)	(no results)	<i>prison break,</i> sarah callies, sara tancredi, looking		(no results)	house paint, color	<i>house, paint,</i> wanta- toos, house painting, hardwood floor, interior design
	michael jackson	michael jackson, <i>rock pop</i>	michael jackson, <i>sony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

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 the 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

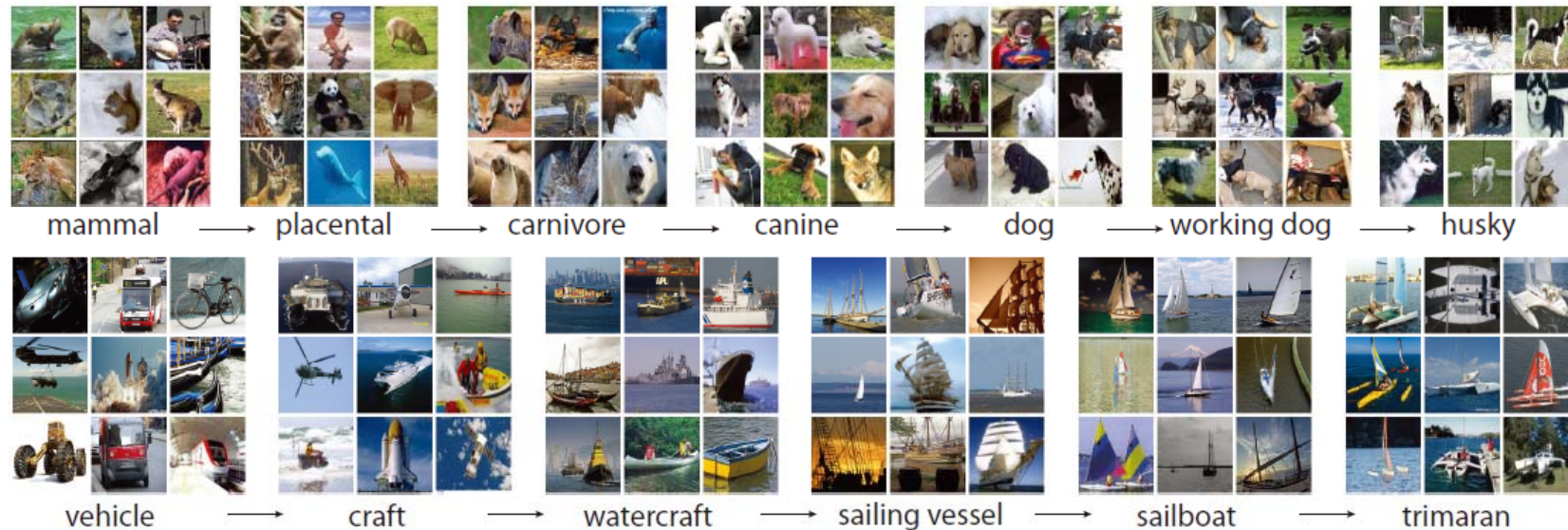


Figure 1: A snapshot of two root-to-leaf branches of ImageNet: the **top** row is from the mammal subtree; the **bottom** row is from the vehicle subtree. For each synset, 9 randomly sampled images are presented.

大規模画像認識コンテスト

- ImageNet
 - <http://www.image-net.org/challenges/LSVRC/2010/index>
- 今年度の挑戦課題
 - 画像識別
 - 1000 カテゴリ
 - 120万枚の訓練画像
 - 5万枚の検証画像
 - 15万枚のテスト画像

テスト画像



カテゴリ

[french fries](#) [mashed potato](#) [black olive](#) [face powder](#) [crab apple](#) [Granny Smith](#) [strawberry](#) [blueberry](#) [cranberry](#) [currant](#) [blackberry](#) [raspberry](#) [persimmon](#) [mulberry](#) [orange](#) [kumquat](#) [lemon](#) [grapefruit](#) [plum](#) [fig](#) [pineapple](#) [banana](#) [jackfruit](#) [cherry](#) [grape](#) [custard](#)
[apple](#) [durian](#) [mango](#) [elderberry](#) [guava](#) [litchi](#) [pomegranate](#) [quince](#) [kidney bean](#) [soy](#) [green pea](#) [chickpea](#) [chard](#) [lettuce](#) [cress](#) [spinach](#) [bell pepper](#) [pimento](#) [jalapeno](#) [cherry tomato](#) [parsnip](#) [turnip](#) [mustard](#) [bok choy](#) [head cabbage](#) [broccoli](#) [cauliflower](#) [brussels sprouts](#) [zucchini](#) [spaghetti squash](#) [acorn squash](#) [butternut squash](#) [cucumber](#) [artichoke](#) [asparagus](#) [green onion](#) [shallot](#) [leek](#) [cardoon](#) [celery](#) [mushroom](#) [pumpkin](#) [cliff](#) [lunar crater](#) [valley](#) [alp](#) [volcano](#) [promontory](#) [sandbar](#) [dune](#) [coral reef](#) [lakeside](#) [seashore](#) [geyser](#) [bakery](#) [juniper berry](#) [gourd](#) [acorn](#) [olive](#) [hip](#) [ear](#) [pumpkin seed](#) [sunflower seed](#) [coffee](#) [bean](#) [rapeseed](#) [corn](#) [buckeye](#) [bean](#) [peanut](#) [walnut](#) [cashew](#) [chestnut](#) [hazelnut](#) [coco nut](#) [pecan](#) [pistachio](#) [lentil](#) [pea](#) [peanut](#) [okra](#) [sunflower](#) [lesser celandine](#) [wood anemone](#) [blue columbine](#) [delphinium](#) [nigella](#) [calla lily](#) [sandwort](#) [pink](#) [baby's breath](#) [ice plant](#) [globe amaranth](#) [four o'clock](#) [Virginia spring beauty](#) [wallflower](#) [damask violet](#) [candytuft](#) [Iceland poppy](#) [prickly poppy](#) [oriental poppy](#) [celandine](#) [blue poppy](#) [Welsh poppy](#) [celandine poppy](#) [corydalis](#) [pearly everlasting](#) [strawflower](#) [yellow chamomile](#) [dusty miller](#) [tansy](#) [daisy](#) [common marigold](#) [China aster](#) [cornflower](#) [chrysanthemum](#) [mistflower](#) など

カテゴリ (Google翻訳後, , ,)

フライドポテトは ジャガイモのマッシュアップ ブラックオリーブ 白粉の カニリンゴ グラニースミスの イチゴ ブルーベリー クランベリー カラント ブラックベリー ラズベリー の 柿 桑 オレンジ キンカン レモン グレープフルーツ 梅 図 パイナップル バナナ ジャックフルーツ 桜の ブドウ カスタードアップルを ドリアン マンゴー ニフトコ グアバ ライチ ザクロ カリン インゲン 大豆 グリーンピースの ひよこ豆の フダンソウ の レタス クレソンの ほうれん草 ピーマン ピーマン ハラペーニョ ミニトマト パースニップ カブ マスタード ボクは菜 ヘッドキャベツ ブロッコリー カリフラワー 芽キャベツ の ズッキーニの スパゲティは、スカッシュ スカッシュドングリ バタースカッシュ キュウリは アーティチョーク アスパラガス ねぎ エシャロットを ネギ カルドンの セロリ マッシュルーム カボチャの 崖の 月面のクレーター の 谷 アルプスの 火山 岬 砂州 の 砂丘に サンゴ礁を 湖畔 海岸 間欠泉の パン屋さん ジュニパーベリーは、 ヒョウタン ドングリ オリーブ ヒップ 耳 カボチャの種 ヒマワリの種 コーヒー豆の 菜種 トウモロコシ バックアイ 豆 ピーナッツ クルミ カシューナッツ 栗 ヘーゼルナッツ ココナッツ ピーカンナッツ ピスタチオ 豆 豆 ピーナッツ オクラ ヒマワリ 低いクサノオウの 木のクマノミ ブルーコロンバイン デルフィニウム ニゲラの カーラリ リー sandwort ピンク 赤ちゃんの呼吸 アイスプラントの 世界をアマランサス 四〇の'ク ロック バージニア春の美しさの 壁の花の ダマスクバイオレット キャンディータフト アイランドポピー 厄介ポピー オリエンタルポピー クサノオウ 青いケシ ウェルシュポピー クサノオウケシ キケマン 真珠のような永遠の ストローフィールド 黄色のカモミール ダスティミラーの ヨモギギクに デイジーチェーン 共通マリーゴールド エゾギク コーンフラワー キク キク科ヒヨドリバナ属の多年草の など

結果

- 参加チーム
 - 150以上参加, 最終的には11チームの結果報告
 - データ規模が膨大! ダウンロードだけで1週間以上!
 - ベースラインの結果が決まっている
- 順位
 1. NEC-UIUC, USA
 2. XRCE, France
 3. ISIL, University of Tokyo, Japan
 4. UC Irvine, USA
 5. MIT, USA
 6. Nanyang Technological University, Singapore
 7. LIG Grenoble, France
 8. IBM-ensemble, USA
 9. SRI International, USA
 10. National Institute of Informatics, Tokyo, Japan
 11. Harbin Institute of Technology, China



demo

- 大規模画像を用いた一般物体認識

単語による理解で十分か？

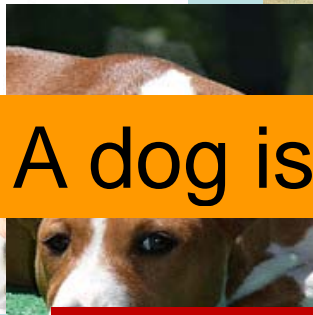


<http://www.totallifecounseling.com/services-therapy-orlando-florida-clermont-east-southwest-therapy-therapists-counselors/ptsd/>

labels	dog, person, bit	=	dog, person, bit
a sentence	A dog bit a person.	≠	A person bit a dog.
		≠	

画像からの文章生成手法

Input



A dog is lying in the garden.

summarize

A dog wearing a suit. A motorbike parked in a garden.

A bear is riding on a horse.

A fox is lying in a garden.

A dog is wearing a suit.

A fox is lying in a garden.

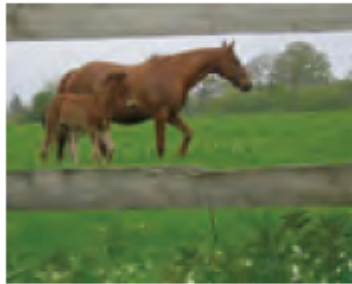
A motorbike parked in a garden.

A bear is riding on a horse.

Concept space

Automatic Sentence Generation from Images

GOOD EXAMPLES



A brown horse standing in a lush green field.



A jet flies high in the blue sky.



A silver car parked in a residential street.



A man stands in front of a train



A sheep with a tree in the foreground.



A city bus driving past a building.

Automatic Sentence Generation from Images

BAD EXAMPLES



A white jeep parked on a large white horse.



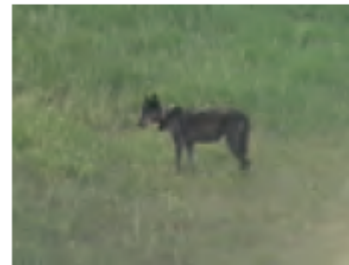
A man standing in front of spiderman poster.



Yellow and white plane flying through the air.



Three cows are grazing in a grassy field.

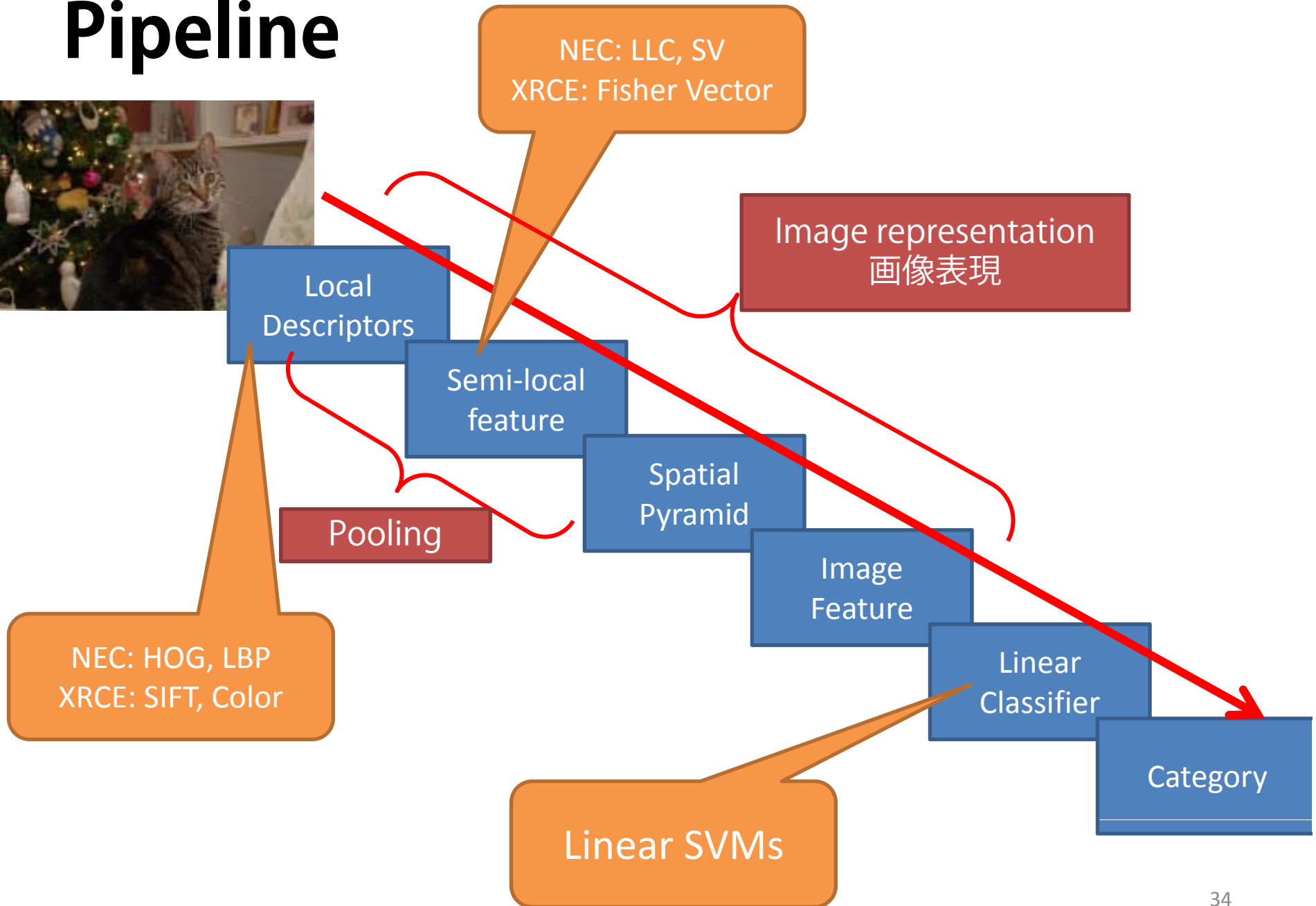


A black dog looking at a black sheep.



A man and woman is riding a brown horse.

Pipeline



Spatial Pyramid Representation

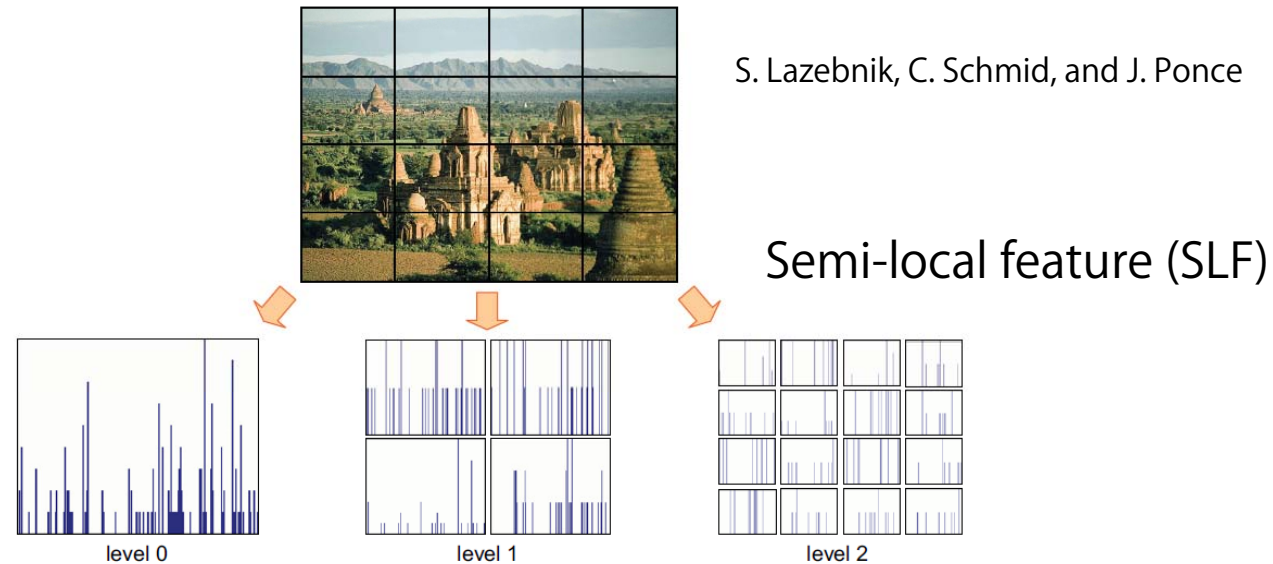


Fig. 1.1. A schematic illustration of the spatial pyramid representation. A spatial pyramid is a collection of orderless feature histograms computed over cells defined by a multi-level recursive image decomposition. At level 0, the decomposition consists of just a single cell, and the representation is equivalent to a standard bag of features. At level 1, the image is subdivided into four quadrants, yielding four feature histograms, and so on. Spatial pyramids can be matched using the *pyramid kernel*, which weights features at higher levels more highly, reflecting the fact that higher levels localize the features more precisely (see Section 1.2).

- Level0: Global featureと同じ
- Level1: 2x2のcellに分割し各cellでSLFを計算
- Level2: 4x4のcellに分割し各cellでSLFを計算

Pooling

- Average pooling
- Max pooling

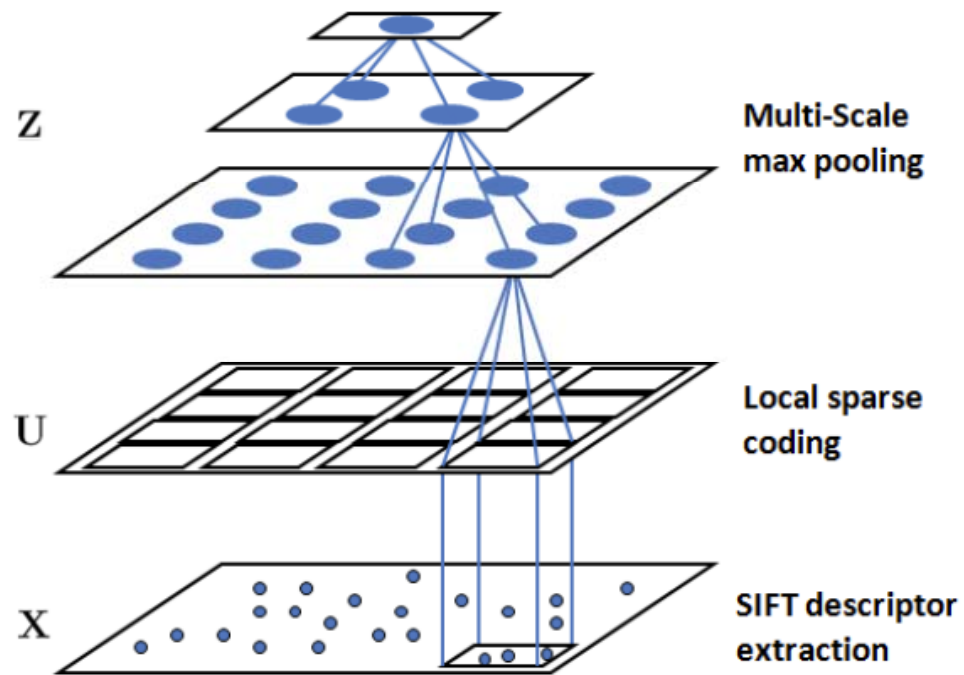


Figure 2. The illustration architecture of our algorithm based on sparse coding. Sparse coding measures the responses of each local descriptor to the dictionary's "visual elements". These responses are pooled across different spatial locations over different spatial scales.

J. Yang, K. Yu, Y. Gong and T. Huang. Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. CVPR, 2009.

Discriminative Spatial Pyramid

- Tatsuya Harada, Yoshitaka Ushiku, Yuya Yamashita, and Yasuo Kuniyoshi. Discriminative Spatial Pyramid. In CVPR, 2011.

Purpose

Proposal of the discriminative, compact and efficient Spatial Pyramid Representations

Background

Spatial Pyramid Representation (SPR)

- a widely used method for embedding spatial information
- good performance in terms of generic image recognition

Problems of SPR

- a huge feature vector
- hand-crafted pyramid structure

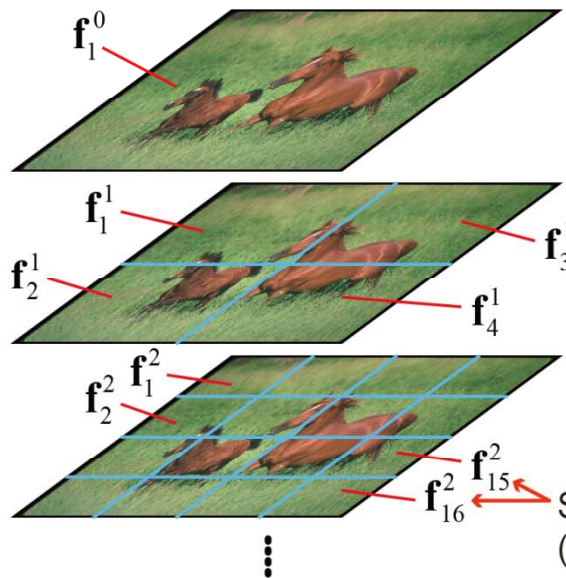


Discriminative Spatial Pyramid

Weighted Spatial Pyramid Representation

Spatial Pyramid Representation

Weighted Spatial Pyramid Representation

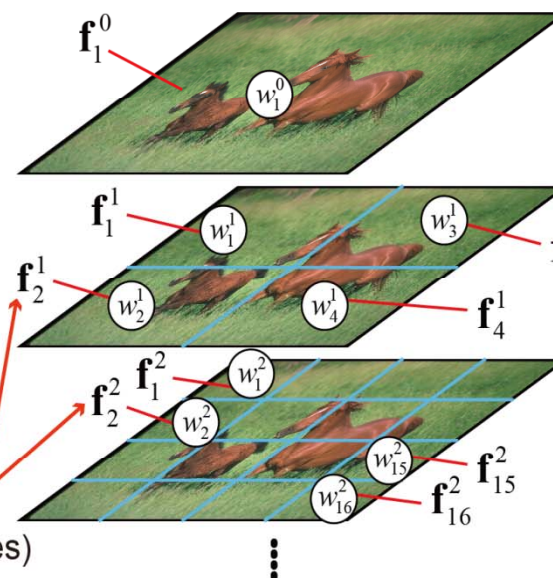


Level 0

Level 1

Level 2

Semi-local features
(e.g. Bag of Features)



Proposed

Pooling

Concatenation of semi-local features

Weighted sum of semi-local features

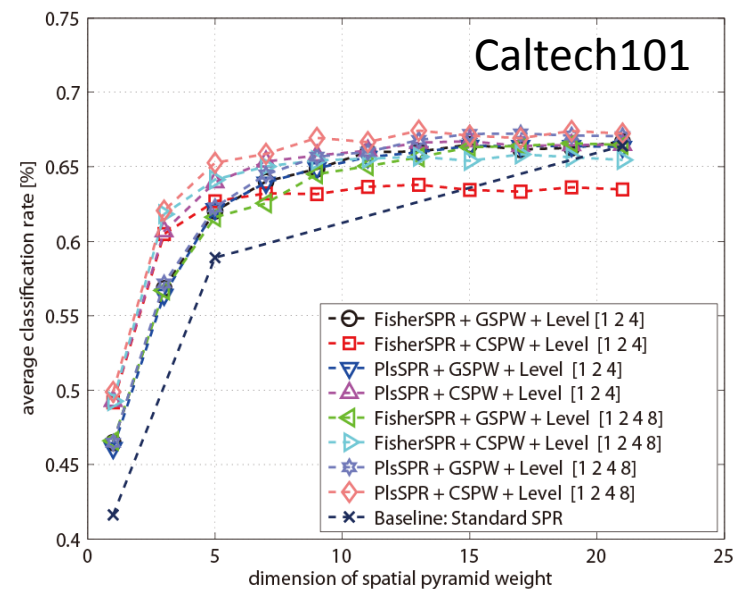
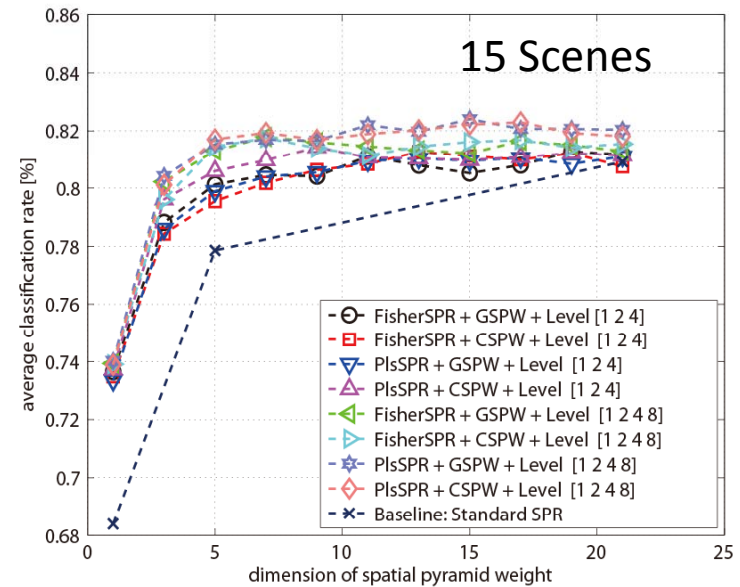
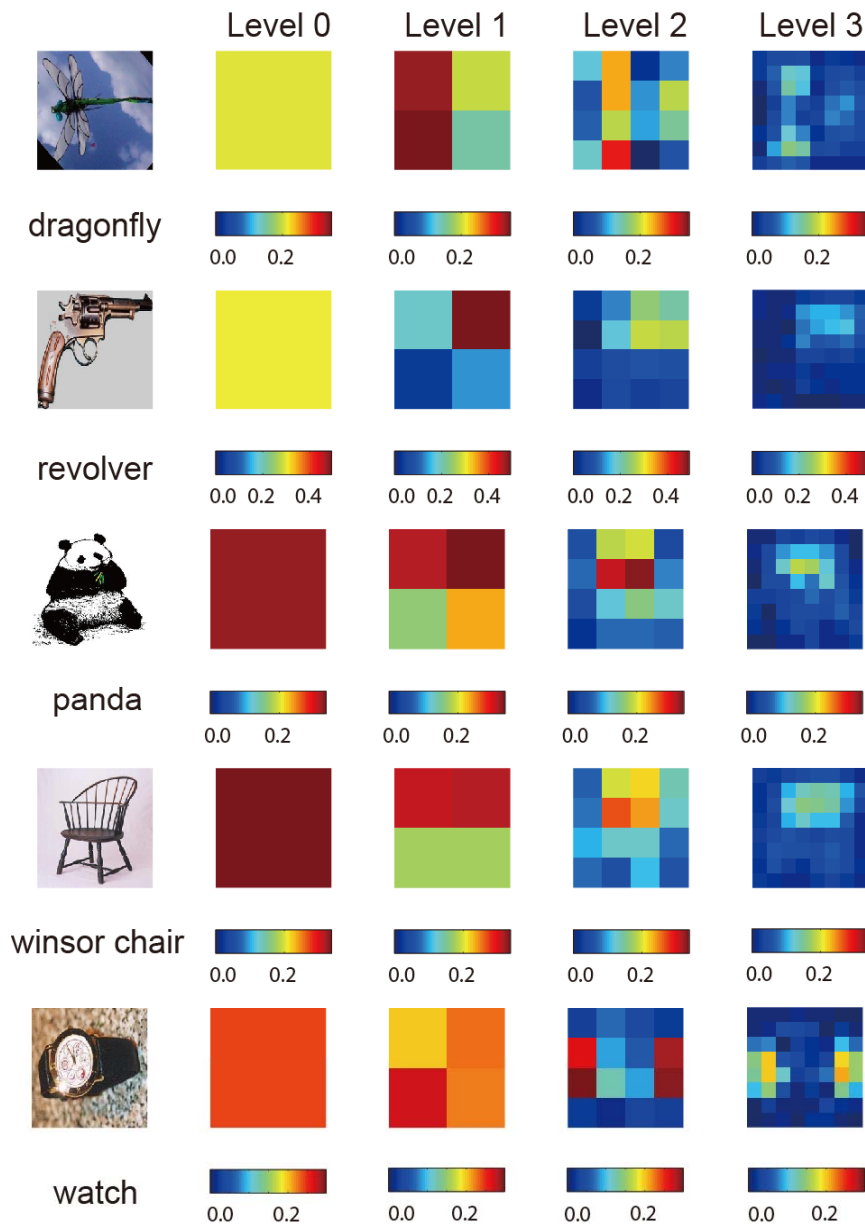
$$\mathbf{f}_S = \begin{pmatrix} \mathbf{f}_1^0 \\ \mathbf{f}_1^1 \\ \vdots \\ \mathbf{f}_{c(L-1)}^{L-1} \end{pmatrix} \Rightarrow \text{Huge vector :-}$$

$$\mathbf{f}_W = w_1^0 \mathbf{f}_1^0 + w_1^1 \mathbf{f}_1^1 + \dots + w_{c(L-1)}^{L-1} \mathbf{f}_{c(L-1)}^{L-1}$$

Compact vector :-

But, how to obtain weights?

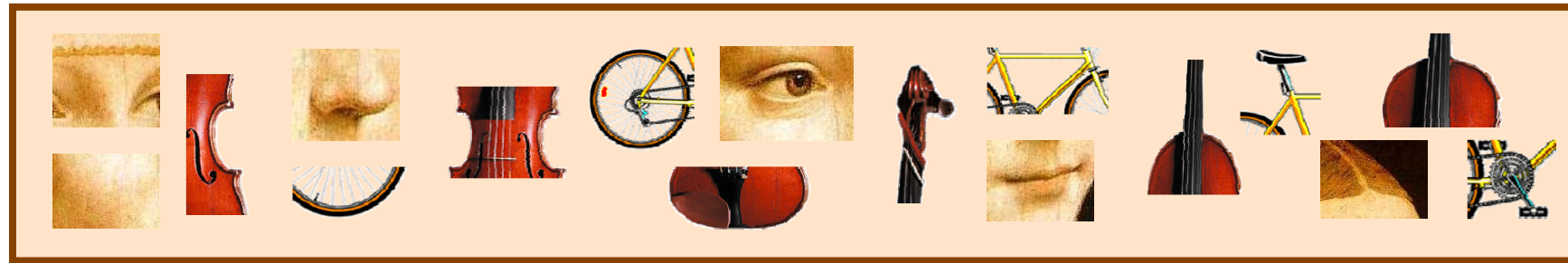
Discriminative Spatial Pyramid



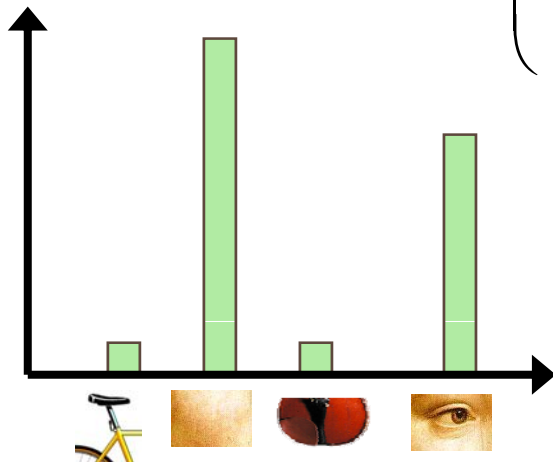
Bag of Visual Words?

Visual words

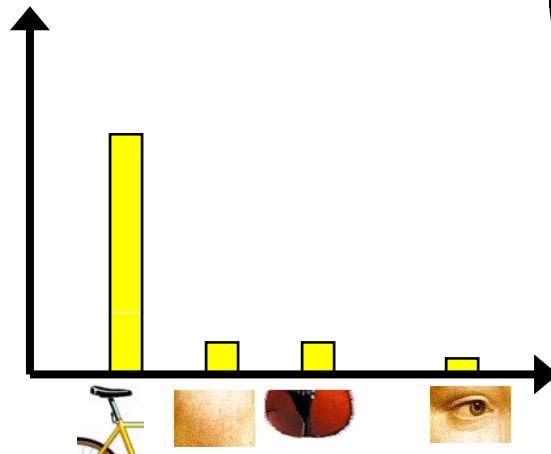
Li Fei Fei, cvpr07 tutorial
より抜粋



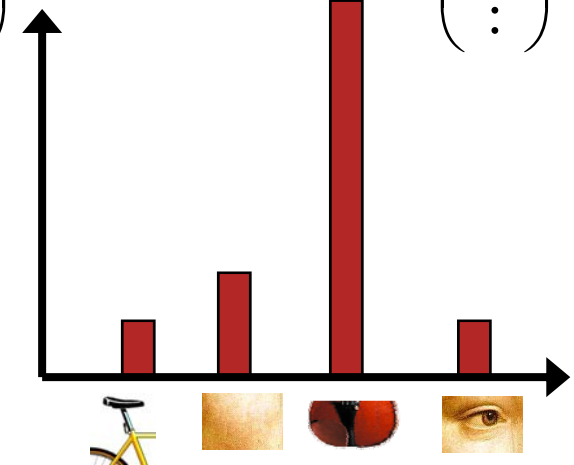
- $$\begin{pmatrix} 1 \\ 10 \\ 1 \\ 7 \\ \vdots \end{pmatrix}$$



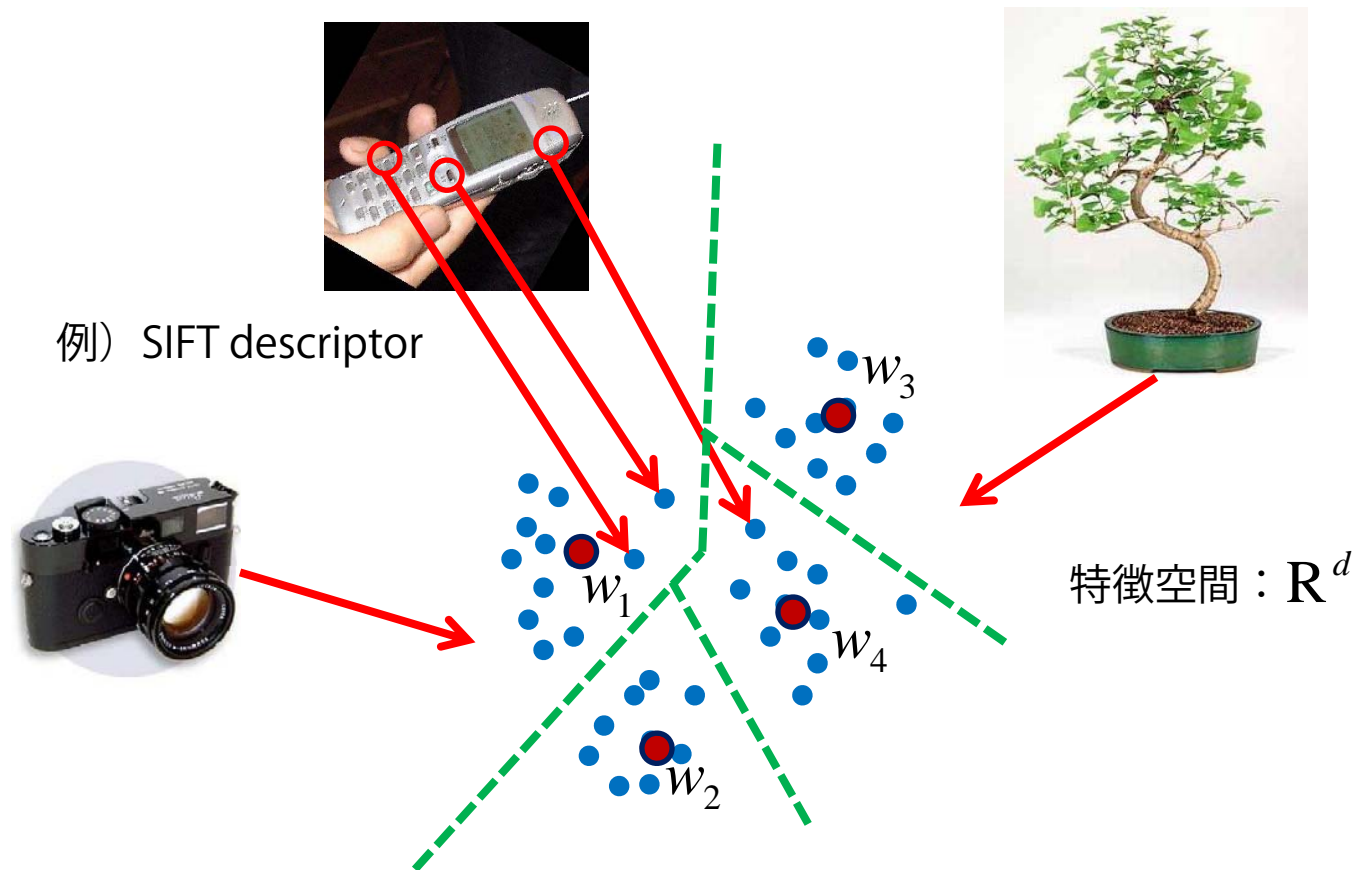
- $$\begin{pmatrix} 7 \\ 2 \\ 2 \\ 1 \\ \vdots \end{pmatrix}$$



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

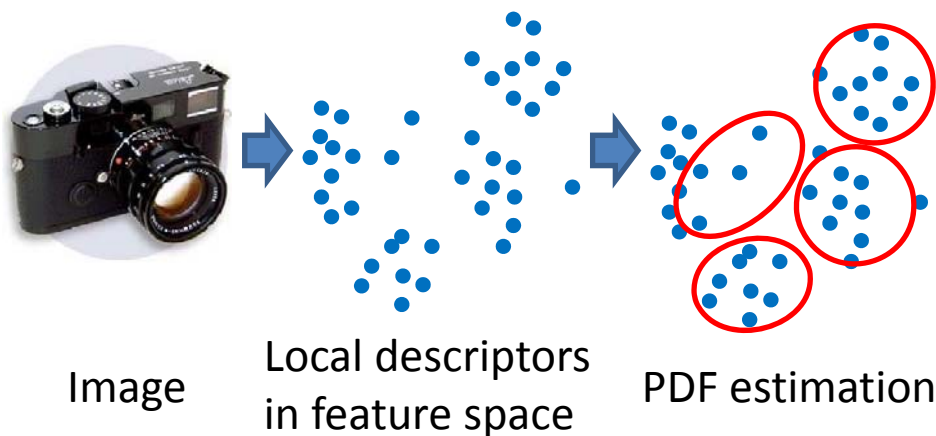


Code wordsの生成：clustering



- 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

BoFのGMM利用による改善



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

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

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

- メリット

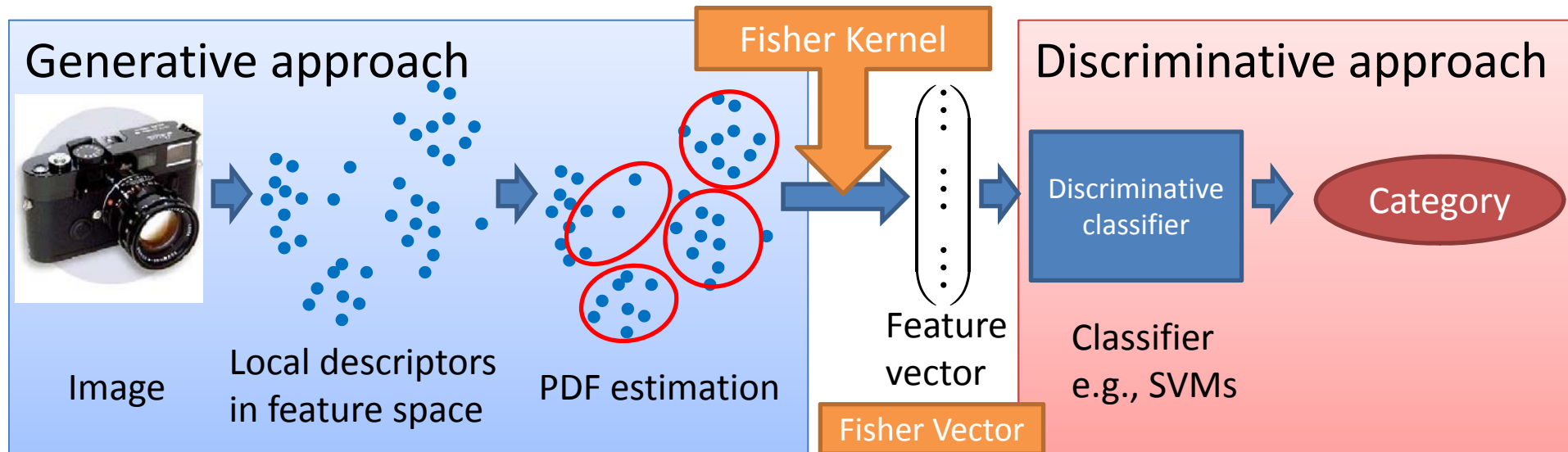
- 混合ガウス分布を構成する各ガウス分布がそれぞれ共分散を持つため、共分散を考慮した距離計量を利用できる
- 混合ガウス分布では局所特徴と多くのコードワードとの関係を表現できるので、特徴空間における局所特徴の位置に関する情報をエンコードできる

- 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。

フィッシャーベクトル詳細

- 局所特徴群

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

- 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

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

- 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正規化が必要！！

- フィッシャー情報行列

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

- フィッシャーベクトル (Fisher Vector)

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

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

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

- 対数尤度の微分

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

画像1枚から得られる局所特徴の集合

あらゆる画像を生成する確率密度分布

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

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

GMMのBoFとほぼ同じ

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

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

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

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

- 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

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

- 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

$$\frac{\partial \mathcal{L}(\mathcal{X}|\theta)}{\partial \pi_k}$$



$$f_{\pi_k} = N \left(\frac{1}{\pi_k} + \frac{1}{\pi_1} \right)$$

$$\frac{\partial \mathcal{L}(\mathcal{X}|\theta)}{\partial \mu_k^d}$$



$$f_{\mu_k^d} = \frac{N \pi_k}{(\sigma_k^d)^2}$$

$$\frac{\partial \mathcal{L}(\mathcal{X}|\theta)}{\partial \sigma_k^d}$$



$$f_{\sigma_k^d} = \frac{2N \pi_k}{(\sigma_k^d)^2}$$

フィッシャー情報行列の要素

フィッシャーベクトルの直感的解釈

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

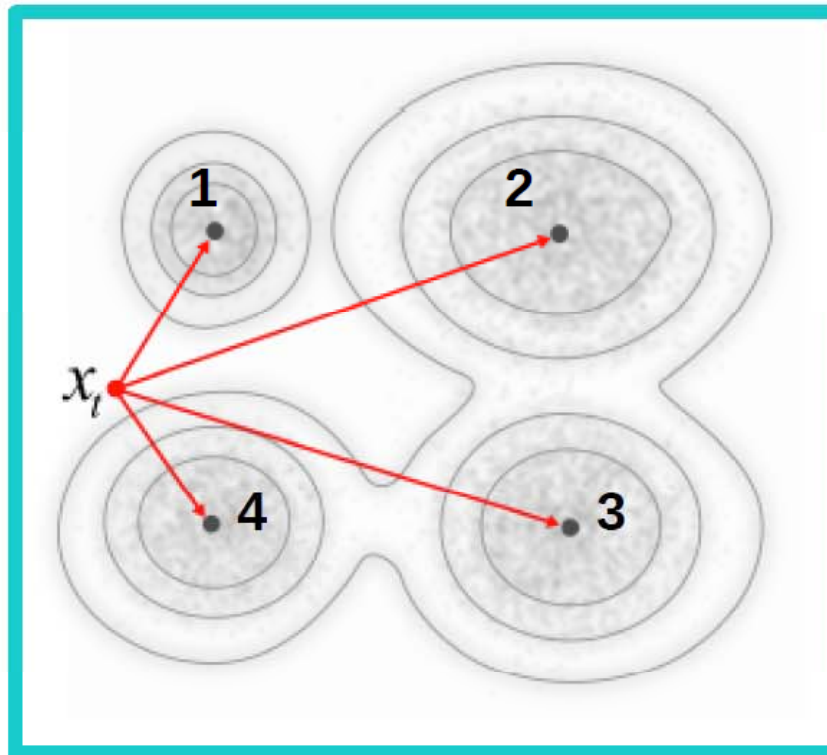
BOV

Hard Assignment

[0 0 0 1]

Soft Assignment

[.3 .1 .1 .5]



Fisher Vector

Gradient wrt w

[.15 -.2 -.35 .2]

Gradient wrt mean

[.8 -1.5 -3.7 -1.3 -3.8 1.2 -.9 1.4]

Gradient wrt var

[-1.2 -.9 1.4 -.8 1.5 -3.7 1.3 -3.8]

Bag of Visual Words (GMM)

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

Fisher Vector

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

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

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

フィッシャーベクトルの改善

- フィッシャーベクトルはBoFと比較して豊かな表現
 - しかしながら、そのまま画像識別に利用してもBoFとさほど性能に差がない。

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

GMMのBoFとほぼ同じ

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

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

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

- 改善方法
 - L2正規化
 - パワー正規化
 - 空間ピラミッドの導入
- F. Perronnin, J. Sanchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. ECCV, 2010.

L2正規化によるフィッシャーベクトルの改善

- 対数尤度の勾配

$$G_{\theta}^{\mathcal{X}} = \frac{1}{N} \nabla_{\theta} \log u_{\theta}(\mathcal{X}|\theta) \longrightarrow G_{\theta}^{\mathcal{X}} \approx \nabla_{\theta} \int_{\mathbf{x}} p(\mathbf{x}) \log u_{\theta}(\mathbf{x}) d\mathbf{x}$$

1枚の画像から得られた局所特徴群 \mathcal{X} は $p(\mathbf{x})$ に従うとする

- 確率密度分布の分解



画像に特定の分布：前景

あらゆる画像を表現する分布：背景

$q(\mathbf{x})$

$u(\mathbf{x})$

$$p(\mathbf{x}) = \omega q(\mathbf{x}) + (1 - \omega) u_{\theta}(\mathbf{x})$$

最尤法によりパラメータを求めた場合ゼロとなる！

前景・背景の混合比

$$G_{\theta}^{\mathcal{X}} \approx \omega \nabla_{\theta} \int_{\mathbf{x}} q(\mathbf{x}) \log u_{\theta}(\mathbf{x}) d\mathbf{x} + (1 - \omega) \nabla_{\theta} \int_{\mathbf{x}} u_{\theta}(\mathbf{x}) \log u_{\theta}(\mathbf{x}) d\mathbf{x}$$

画像に特定の分布のみが残る！！
ただし前景・背景の混合比の影響が残るのでL2正規化を行う。

$$G_{\theta}^{\mathcal{X}} \approx \omega \nabla_{\theta} \int_{\mathbf{x}} q(\mathbf{x}) \log u_{\theta}(\mathbf{x}) d\mathbf{x}$$

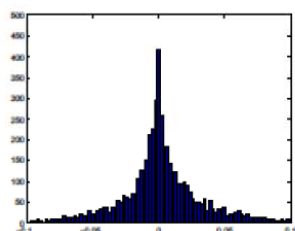
パワー正規化, 空間ピラミッドによる フィッシャーベクトルの改善

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

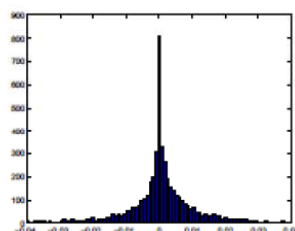
- パワー正規化

- 混合数の増加に伴いフィッシャーベクトルがスパースになる
- スパースベクトルにおけるL2距離は性能が悪い
- 方針1: カーネル法は計算コストが高い
- 方針2: スパースにしない

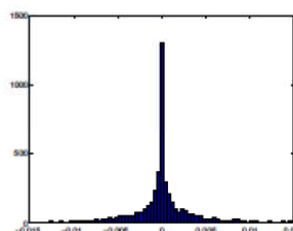
$$f(z) = \text{sign}(z)|z|^\alpha$$



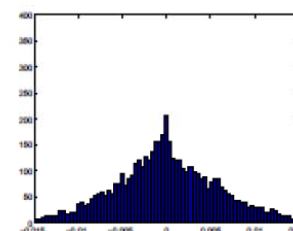
K = 16



K = 64

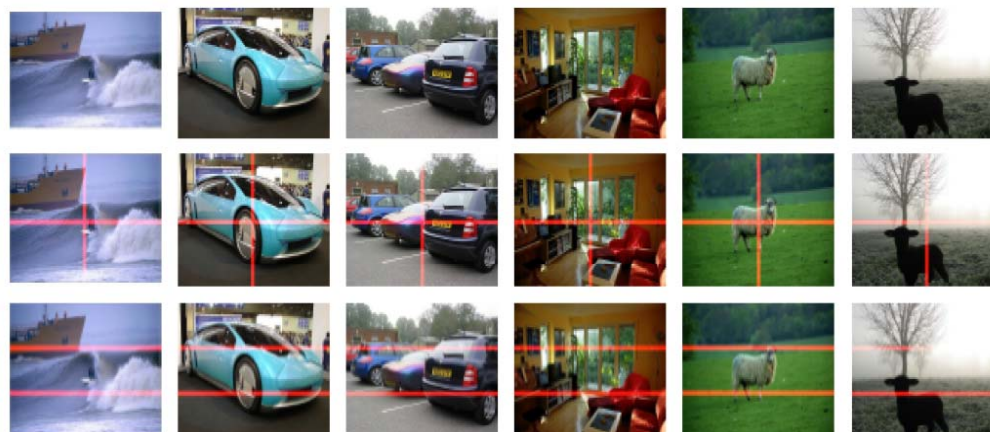


K = 256



K = 256
Power Normalized

- 空間ピラミッド



画像1枚あたり8個
のフィッシャーベ
クトルを抽出

フィッシャーベクトルの性能

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

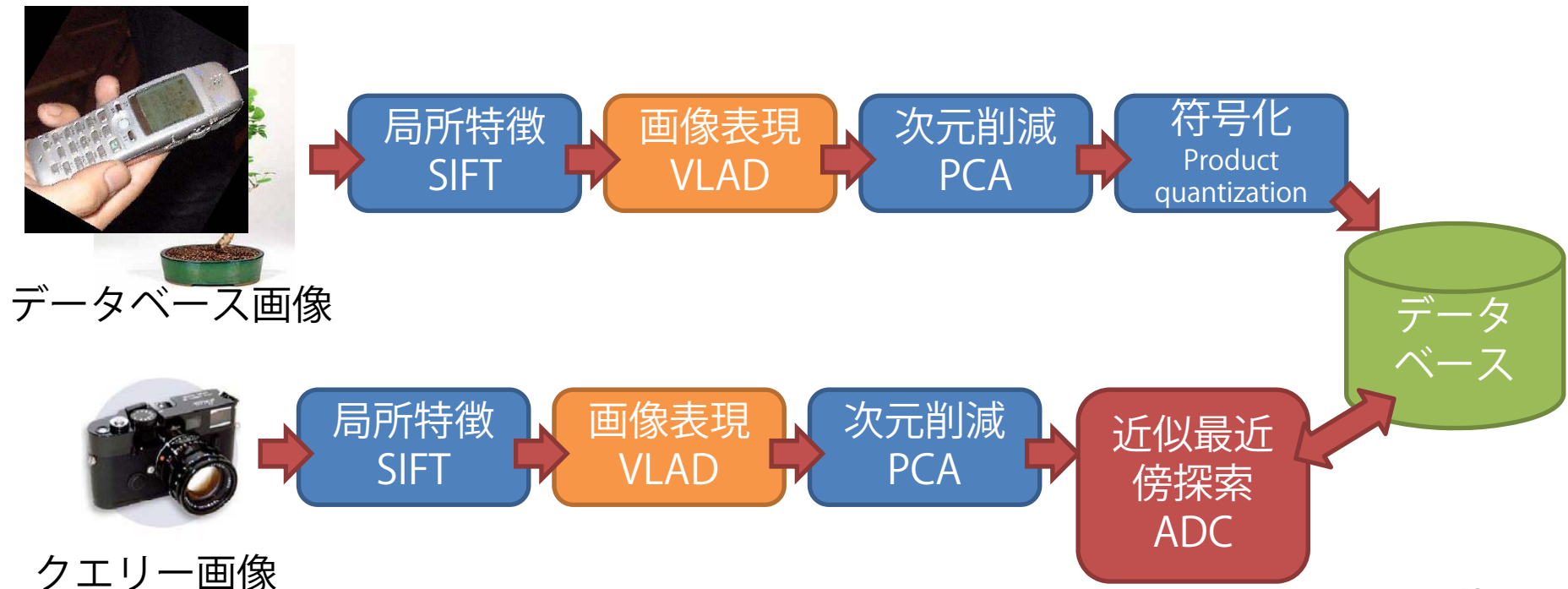
- Pascal VOC 2007
- 改良されたフィッシャーベクトルを利用
- 識別機：線形SVM

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

フィッシャーベクトルの 画像検索への応用例

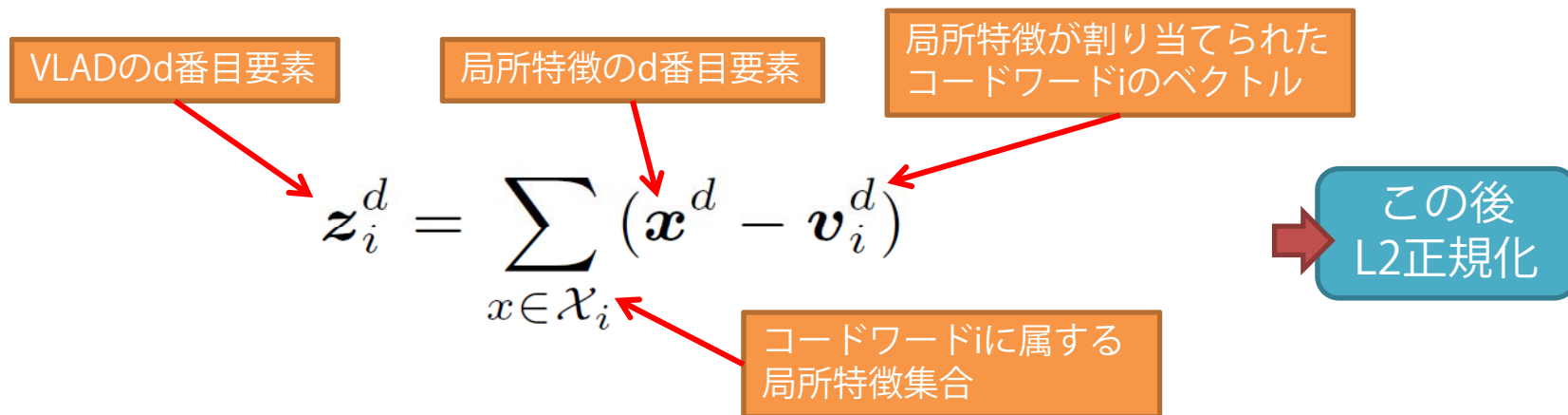
- H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. CVPR, 2010.
- 20bitに画像表現しても，生のBoFを使った検索と同じ検索性能
- パイプライン



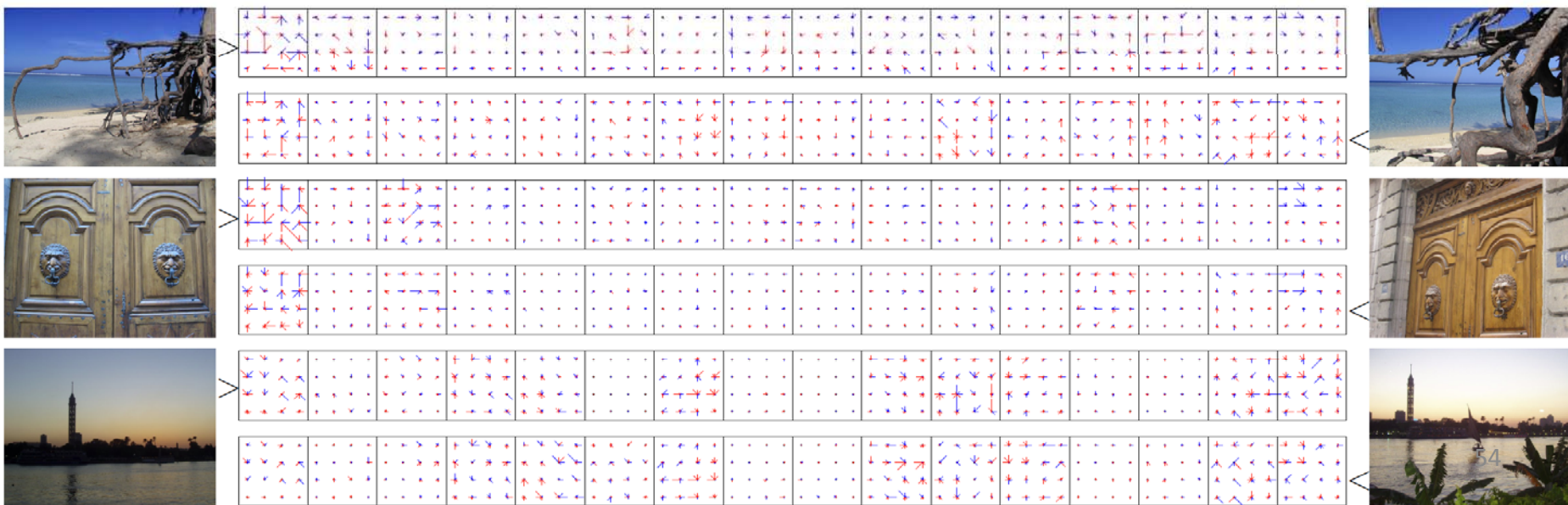
VLAD

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

$$\frac{\partial \mathcal{L}(\mathcal{X}|\theta)}{\partial \pi_k} = \sum_{n=1}^N \left[\frac{\gamma_n(k)}{\pi_k} - \frac{\gamma_n(1)}{\pi_1} \right]$$
$$\frac{\partial \mathcal{L}(\mathcal{X}|\theta)}{\partial \mu_k^d} = \sum_{n=1}^N \gamma_n(k) \left[\frac{\mathbf{x}_n^d - \mu_k^d}{(\sigma_k^d)^2} \right]$$
$$\frac{\partial \mathcal{L}(\mathcal{X}|\theta)}{\partial \sigma_k^d} = \sum_{n=1}^N \gamma_n(k) \left[\frac{(\mathbf{x}_n^d - \mu_k^d)^2}{(\sigma_k^d)^3} - \frac{1}{\sigma_k^d} \right]$$

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

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

$$z_i^d = \sum_{x \in \mathcal{X}_i} (\mathbf{x}^d - \mathbf{v}_i^d)$$

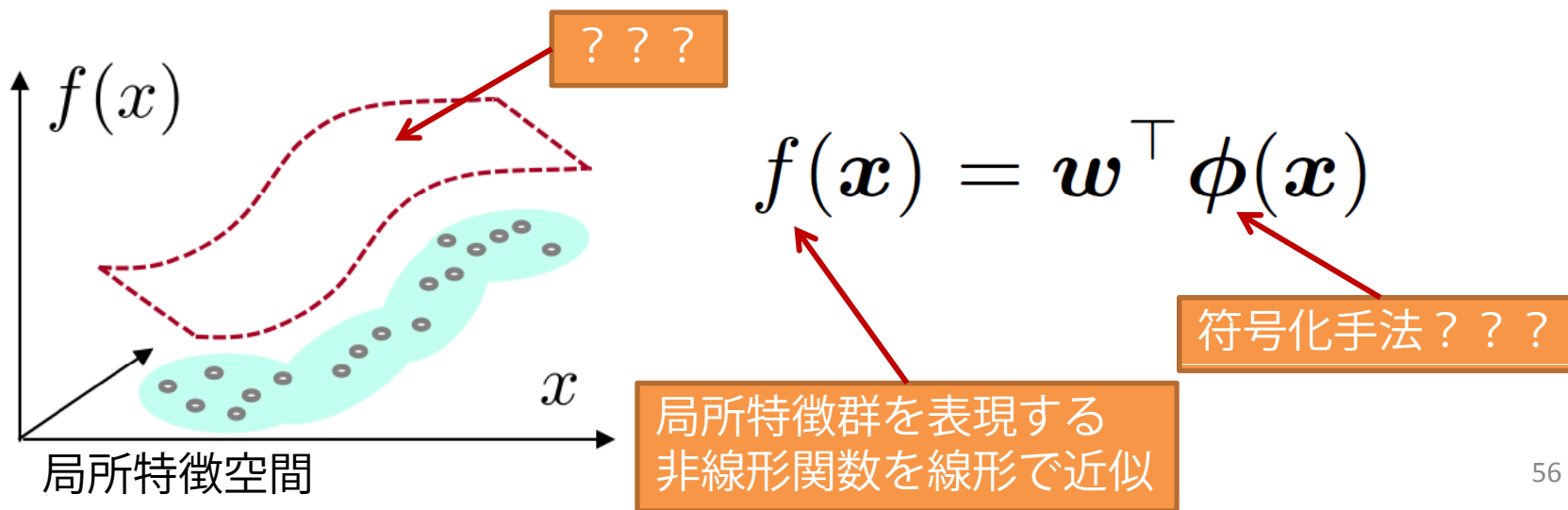
コードワード i に属する局所特徴集合

•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

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 $\phi(x)$ 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

$$\mathbf{x} \approx \sum_{k=1}^K \gamma_x(k) \mathbf{v}_k \quad \gamma_x = [\gamma_x(1), \dots, \gamma_x(K)], \quad \sum_{k=1}^K \gamma_x(k) = 1$$

負担率のようなもの (points to $\gamma_x(k)$)
コードワードk (points to \mathbf{v}_k)

- β Lipschitz derivative smooth

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

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

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

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

関数f(x)の1次近似のUpper boundに関する式

$\|\mathbf{x} - \mathbf{v}\|$ が小さければ
近似精度が向上

- スーパーベクトル符号化

$$f(\mathbf{x}) \approx \mathbf{w}^\top \phi(\mathbf{x}) \quad \rightarrow$$

式(☆)を分解!

Super Vector Coding

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

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

スーパーベクトル符号化の解釈

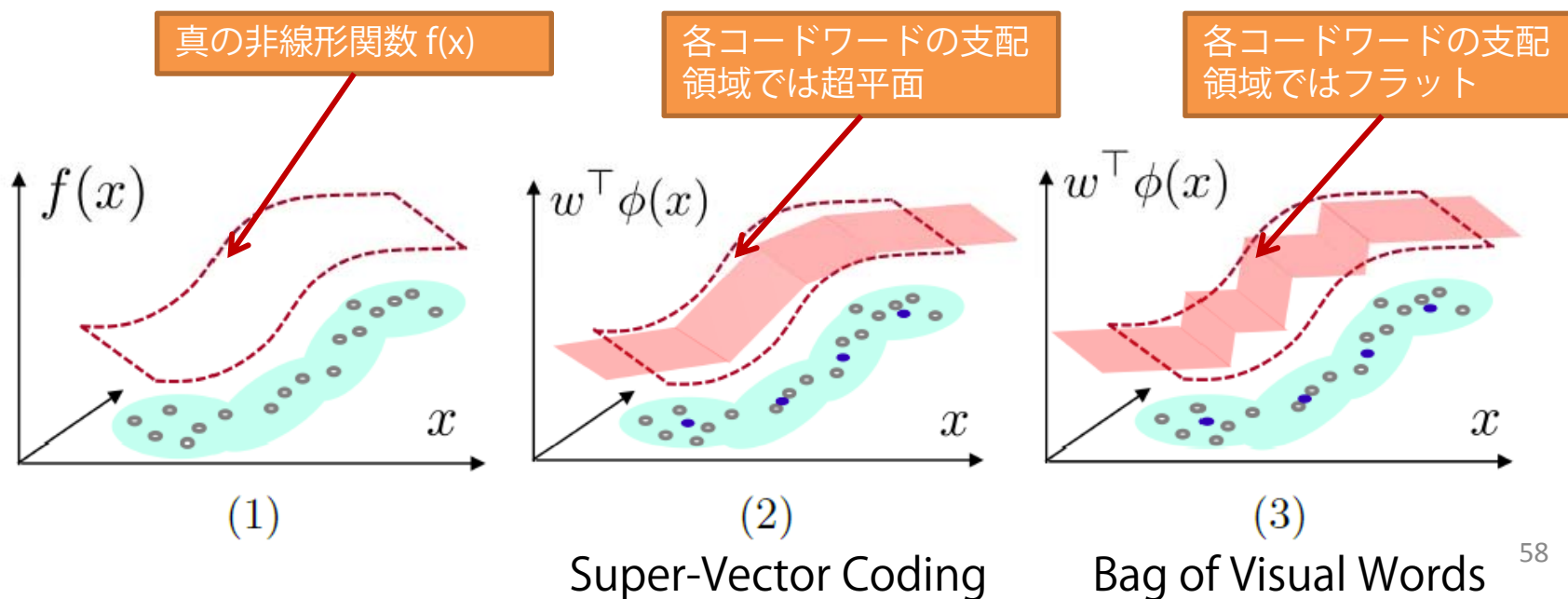
- スーパーベクトル符号化の例
 - コードワード数：3, $\gamma = [0 \ 1 \ 0]^T$

Super Vector Coding

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

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

- 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

$$\frac{\partial \mathcal{L}(\mathcal{X}|\theta)}{\partial \pi_k} = \sum_{n=1}^N \left[\frac{\gamma_n(k)}{\pi_k} - \frac{\gamma_n(1)}{\pi_1} \right]$$
$$\frac{\partial \mathcal{L}(\mathcal{X}|\theta)}{\partial \mu_k^d} = \sum_{n=1}^N \gamma_n(k) \left[\frac{\mathbf{x}_n^d - \mu_k^d}{(\sigma_k^d)^2} \right]$$
$$\frac{\partial \mathcal{L}(\mathcal{X}|\theta)}{\partial \sigma_k^d} = \sum_{n=1}^N \gamma_n(k) \left[\frac{(\mathbf{x}_n^d - \mu_k^d)^2}{(\sigma_k^d)^3} - \frac{1}{\sigma_k^d} \right]$$

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

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

負担率

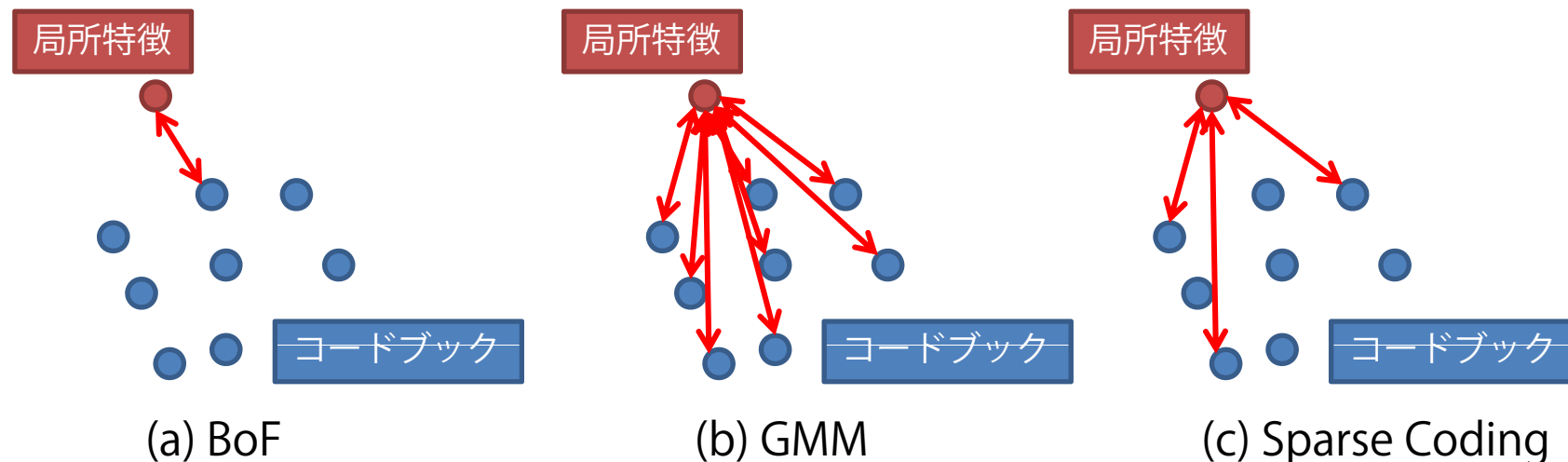
局所特徴 x_n とコードワードとの差分

- 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

スパース符号化 (Sparse Coding)

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



スパース符号化の定式化

- Bag of Visual Words
 - ベクトル量子化 (VQ)

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

Codebook (コードブック) → \mathbf{V}

Local Feature (局所特徴) → \mathbf{x}_n

Indicator (局所特徴がどのコードワードに所属するかを示す指標) → \mathbf{u}_n

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

Constraint (一つのコードワードに属する制約 → 厳しすぎる!!!) → $\text{Card}(\mathbf{u}_n) = 1$

- スパース符号化 (Sparse Coding)

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

L1 Norm Regularization Term (L1ノルム正則化項) → $\lambda \|\mathbf{u}_n\|$
→ 少数のコードワードへの所属を許容

$$\text{s.t. } \|\mathbf{v}_k\| \leq 1, \forall k$$

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。

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 U から一つの特徴ベクトル f を得る手段

- プーリング (pooling)

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

局所特徴集合

プーリング関数

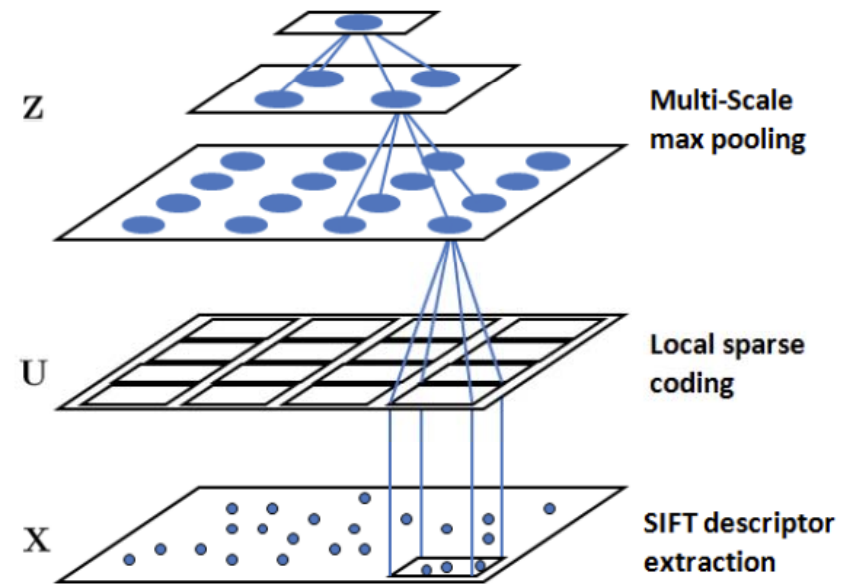
- 平均プーリング
average pooling

$$f = \frac{1}{N} \sum_{n=1}^N u_n$$

BoFはこれを利用

- 最大値プーリング
max pooling

$$f^d = \max\{|u_1^d|, |u_2^d|, \dots, |u_N^d|\}$$



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

最大値プーリングの効果

- Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. CVPR, 2010.

Method	Caltech-101, 30 training examples		15 Scenes, 100 training examples	
	Average Pool	Max Pool	Average Pool	Max Pool
Results with basic features, SIFT extracted each 8 pixels				
Hard quantization, linear kernel	51.4 ± 0.9 [256]	64.3 ± 0.9 [256]	73.9 ± 0.9 [1024]	80.1 ± 0.6 [1024]
Hard quantization, intersection kernel	64.2 ± 1.0 [256] (1)	64.3 ± 0.9 [256]	80.8 ± 0.4 [256] (1)	80.1 ± 0.6 [1024]
Soft quantization, linear kernel	57.9 ± 1.5 [1024]	69.0 ± 0.8 [256]	75.6 ± 0.5 [1024]	81.4 ± 0.6 [1024]
Soft quantization, intersection kernel	66.1 ± 1.2 [512] (2)	70.6 ± 1.0 [1024]	81.2 ± 0.4 [1024] (2)	83.0 ± 0.7 [1024]
Sparse codes, linear kernel	61.3 ± 1.3 [1024]	71.5 ± 1.1 [1024] (3)	76.9 ± 0.6 [1024]	83.1 ± 0.6 [1024] (3)
Sparse codes, intersection kernel	70.3 ± 1.3 [1024]	71.8 ± 1.0 [1024] (4)	83.2 ± 0.4 [1024]	84.1 ± 0.5 [1024] (4)
Results with macrofeatures and denser SIFT sampling				
Hard quantization, linear kernel	55.6 ± 1.6 [256]	70.9 ± 1.0 [1024]	74.0 ± 0.5 [1024]	80.1 ± 0.5 [1024]
Hard quantization, intersection kernel	68.8 ± 1.4 [512]	70.9 ± 1.0 [1024]	81.0 ± 0.5 [1024]	80.1 ± 0.5 [1024]
Soft quantization, linear kernel	61.6 ± 1.6 [1024]	71.5 ± 1.0 [1024]	76.4 ± 0.7 [1024]	81.5 ± 0.4 [1024]
Soft quantization, intersection kernel	70.1 ± 1.3 [1024]	73.2 ± 1.0 [1024]	81.8 ± 0.4 [1024]	83.0 ± 0.4 [1024]
Sparse codes, linear kernel	65.7 ± 1.4 [1024]	75.1 ± 0.9 [1024]	78.2 ± 0.7 [1024]	83.6 ± 0.4 [1024]
Sparse codes, intersection kernel	73.7 ± 1.3 [1024]	75.7 ± 1.1 [1024]	83.5 ± 0.4 [1024]	84.3 ± 0.5 [1024]

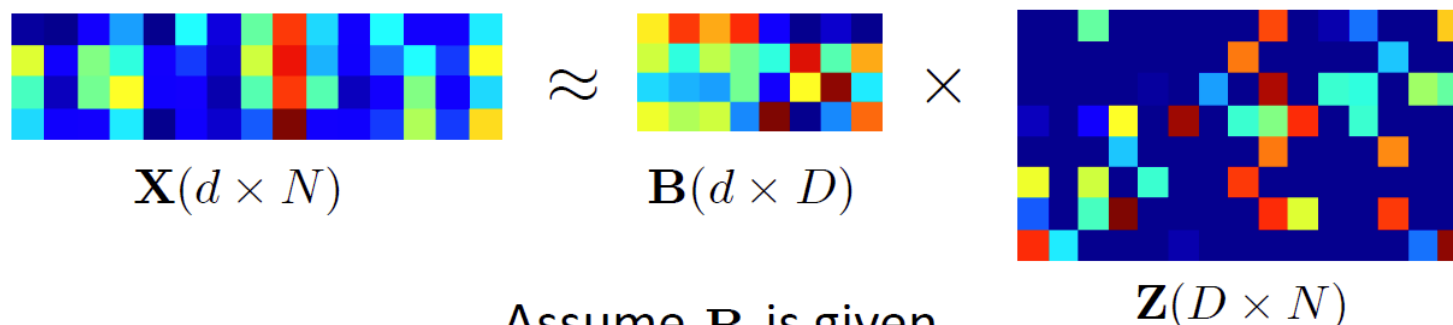
Table 1. Average recognition rate on Caltech-101 and 15-Scenes benchmarks, for various combinations of coding, pooling, and classifier types. The codebook size shown inside brackets is the one that gives the best results among 256, 512 and 1024. Linear and histogram intersection kernels are identical when using hard quantization with max pooling (since taking the minimum or the product is the same for binary vectors), but results have been included for both to preserve the symmetry of the table. Top: Results with the baseline SIFT sampling density of 8 pixels and standard features. Bottom: Results with the set of parameters for SIFT sampling density and macrofeatures giving the best performance for sparse coding.

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



Sparse coding:

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

局所性がスパースネスよりも本質！！

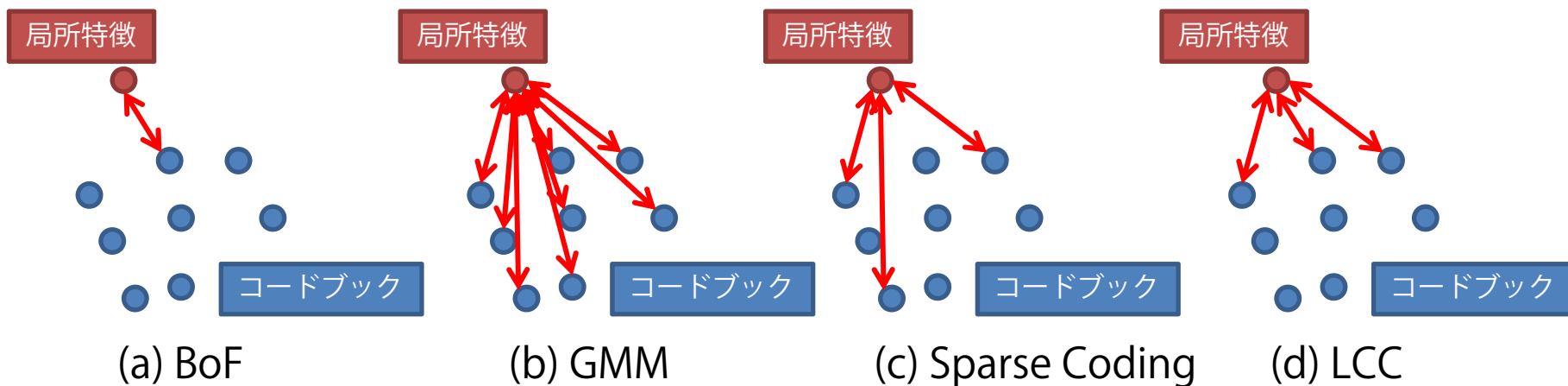
LCC: K. Yu et. al, NIPS 2009

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

Explicitly enforcing locality constraint

局所線形制約符号化と他符号化の比較

- 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



なぜ局所座標符号化が良いのか？

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

$$f(\mathbf{x}) \approx \sum_{i=1}^D z_i(\mathbf{x}) w_i$$

e.g. nonlinear separating hyperplane

$$|f(\mathbf{x}) - \sum_{i=1}^D z_i(\mathbf{x}) f(\mathbf{b}_i)| \leftarrow \text{Functional approximation error}$$

$$\leq \alpha \underbrace{\|\mathbf{x} - \mathbf{Bz}(\mathbf{x})\|}_{\text{Coding error}} + \beta \underbrace{\sum_{i=1}^D \|\mathbf{x} - \mathbf{b}_i\|^2 |z_i(\mathbf{x})|}_{\text{Locality term}}$$

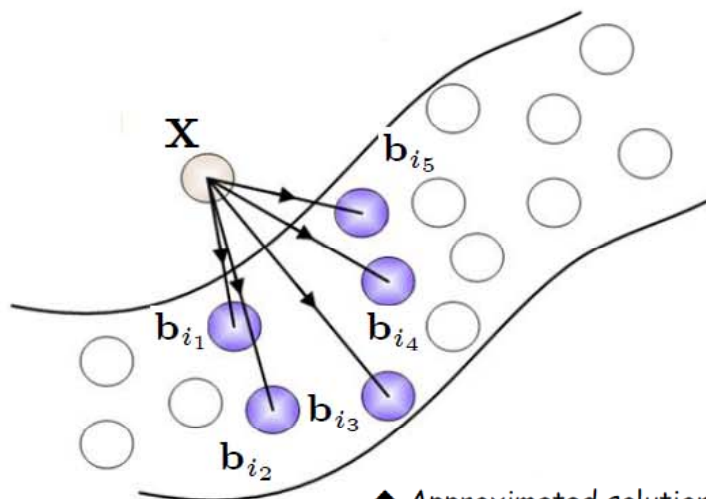
- よりよく近似するためには
 - 局所特徴に対して局所性を有すること
 - 局所特徴の再構築誤差を減らすこと

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.

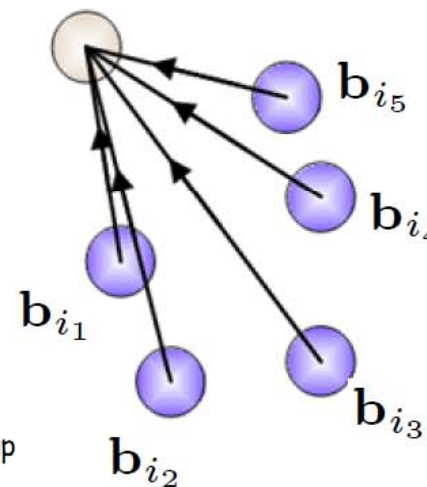
Step 1: be local to the test point \mathbf{x}

-- 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*

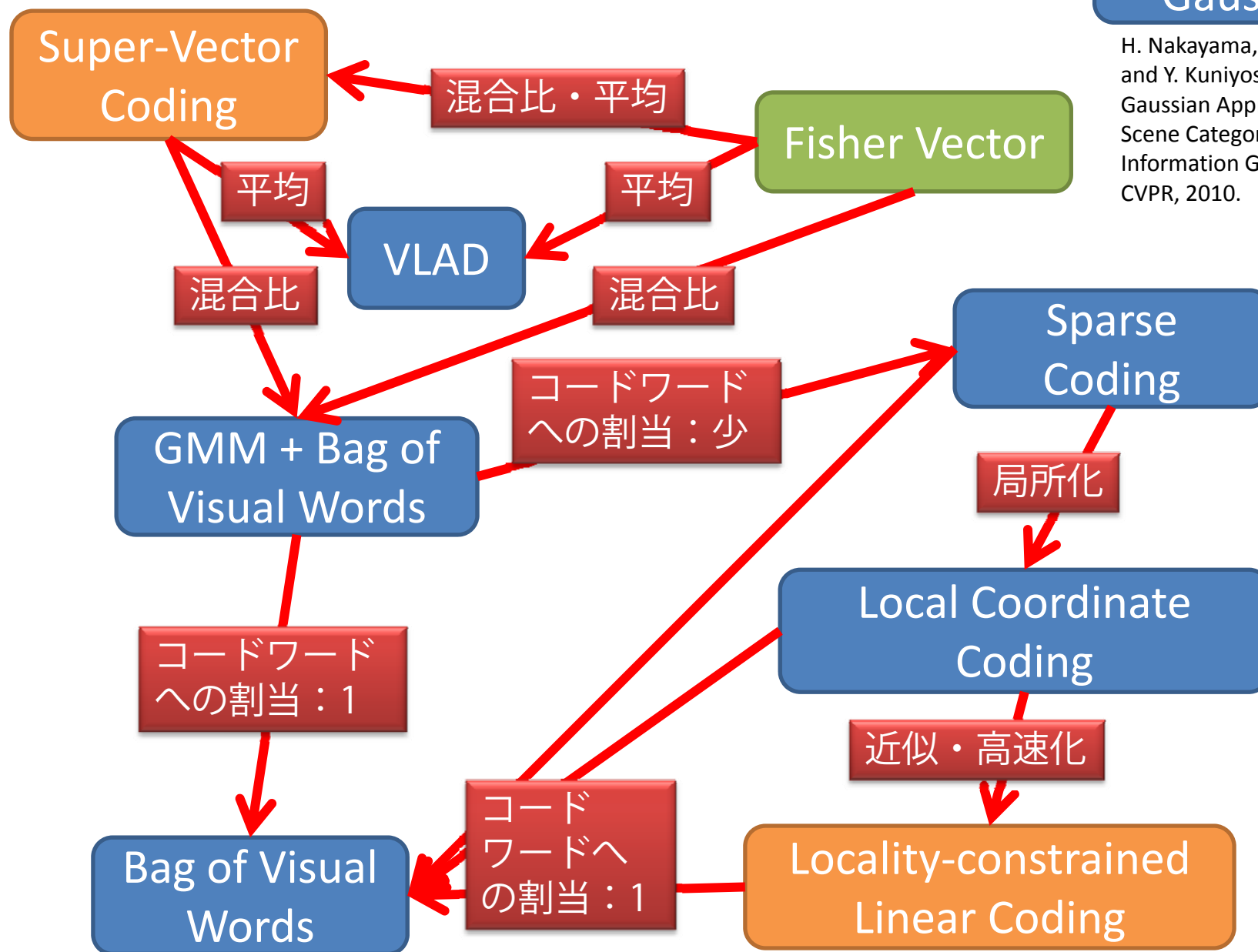
http://www.image-net.org/challenges/LSVRC/2010/ILSVRC2010_NEC-UIUC.pdf

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。

画像表現の関係

Global Gaussian

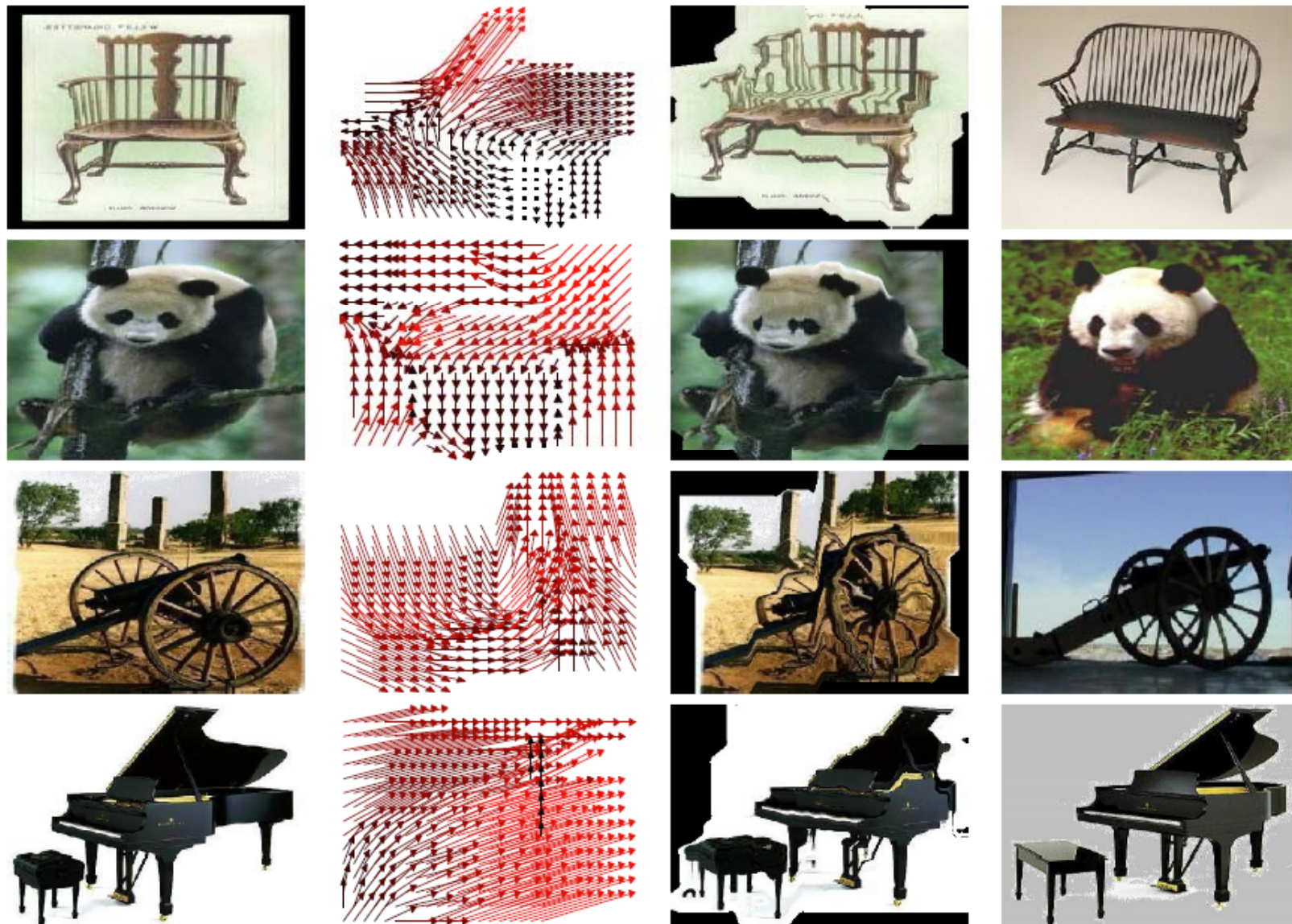
H. Nakayama, T. Harada, and Y. Kuniyoshi. Global Gaussian Approach for Scene Categorization Using Information Geometry. In CVPR, 2010.



物体認識の二大巨頭

- コードブック
 - 局所記述子→画像表現（局所記述子の空間における多様体の表現）→識別機
- 局所記述子のマッチング
 - 画像間の局所記述子間の類似度
 - 局所記述子間の画像空間での関係をグラフ構造で表現

Graph Matching Kernel



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

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.