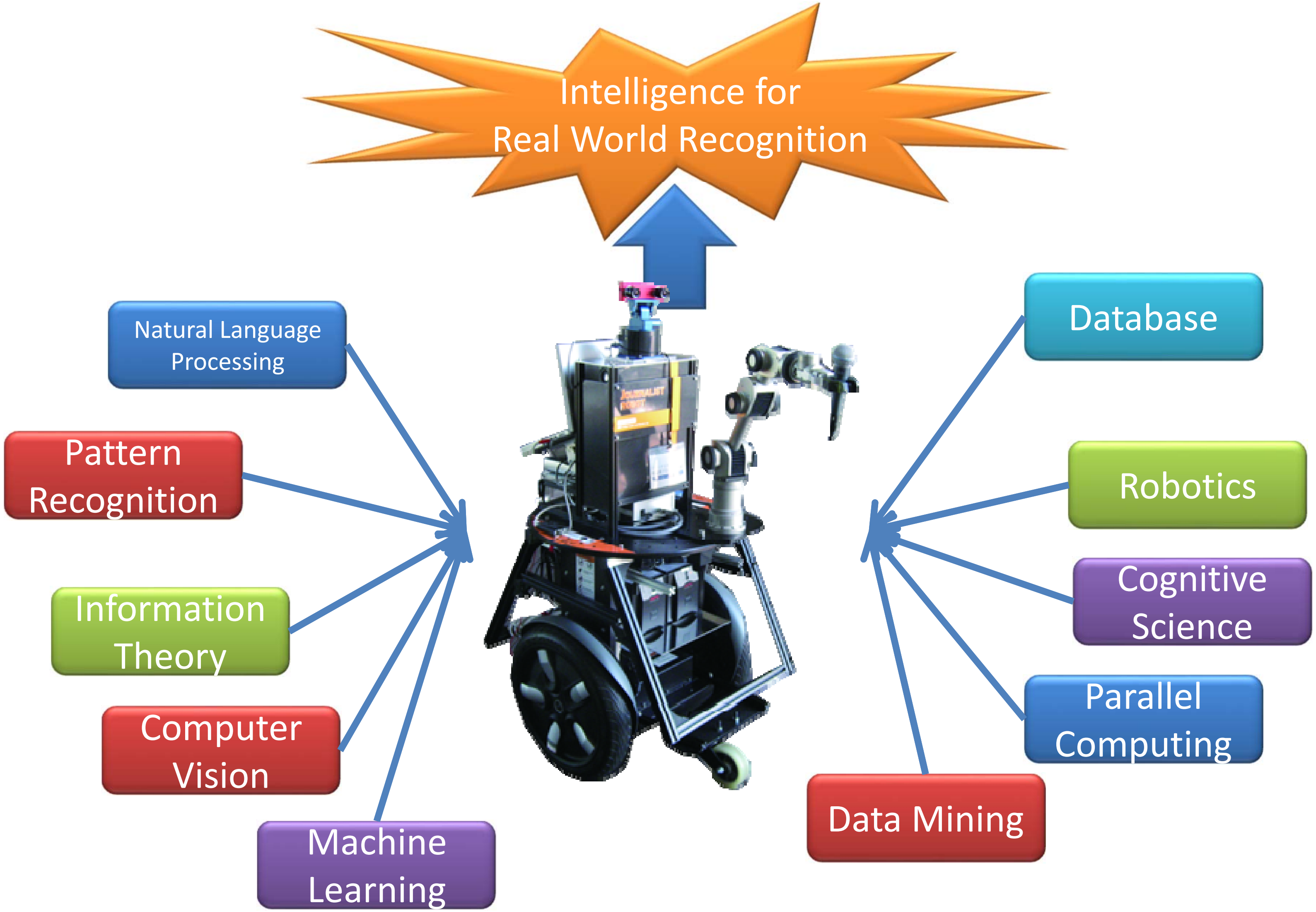
IBIS 11/13/2013

大規模データを用いた
画像の識別と言語記述

The Univ. of Tokyo
Tatsuya Harada

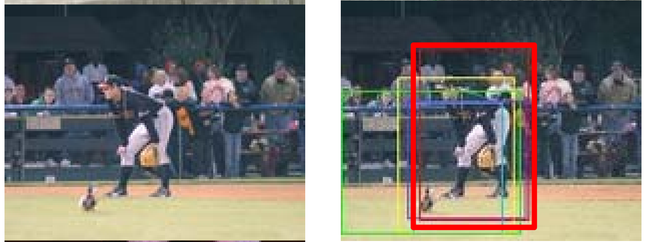
Intelligent Robot System



Intelligent Robot is “Mixed Martial Arts” in computer science!

Results (2012)

1. brown bear
2. Tibetan mastiff
3. sloth bear
4. American black bear
5. bison



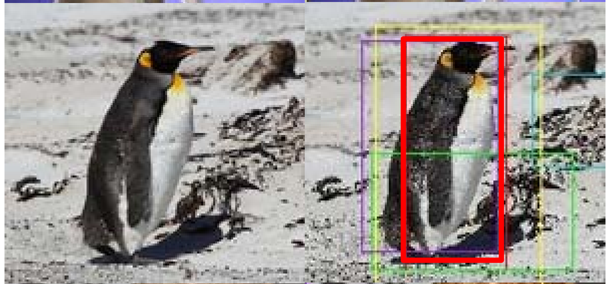
1. baseball player
2. unicycle
3. racket
4. rugby ball
5. basketball

1. digital watch
2. Band Aid
3. syringe
4. slide rule
5. rubber eraser

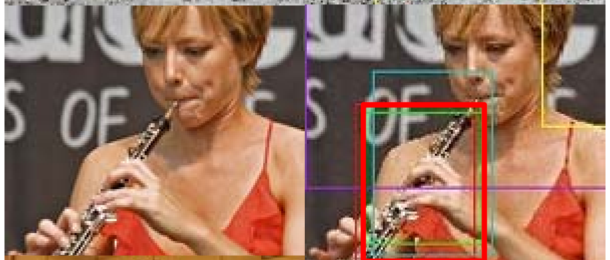
1. shower cap
2. bonnet
3. bath towel
4. bathing cap
5. ping-pong ball

1. diaper
2. swimming trunks
3. bikini
4. miniskirt
5. cello

1. Siamese cat
2. Egyptian cat
3. Ibizan hound
4. balance beam
5. basenji



1. king penguin
2. sea lion
3. drake
4. magpie
5. oystercatcher

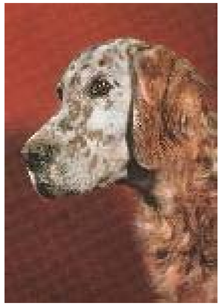


1. oboe
2. flute
3. ice lolly
4. bassoon
5. cello

1. beer bottle
2. pop bottle
3. wine bottle
4. Polaroid camera
5. microwave

1. butcher shop
2. swimming trunks
3. miniskirt
4. barbell
5. feather boa

Fine-grained object recognition results (2012)


English setter


Siberian husky


Australian terrier


English springer


malamute


Great Dane

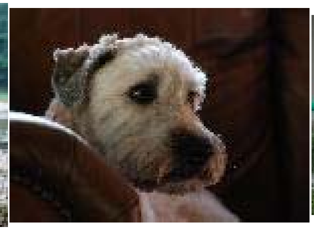

Walker hound

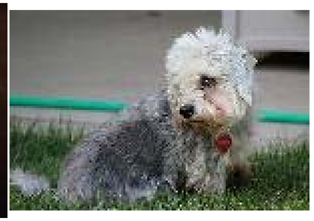

Welsh springer spaniel


whippet


Scottish deerhound


Weimaraner


soft-coated wheaten terrier


Dandie Dinmont


Old English sheepdog


otterhound


bloodhound


Airedale


giant schnauzer


black-and-tan coonhound


papillon


Staffordshire bullterrier


Mexican hairless

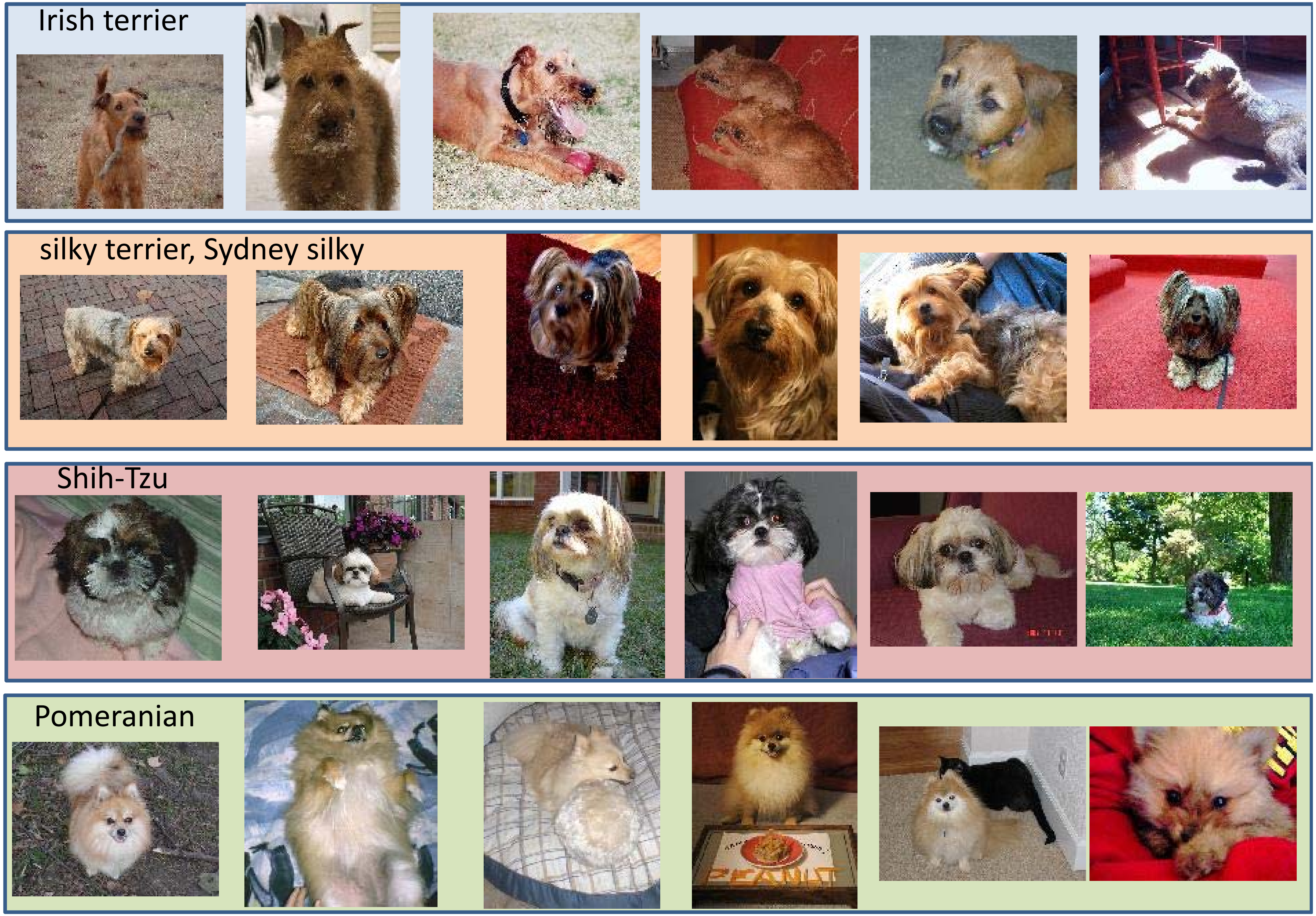

Bouvier des Flandres


miniature poodle


Cardigan


malinois

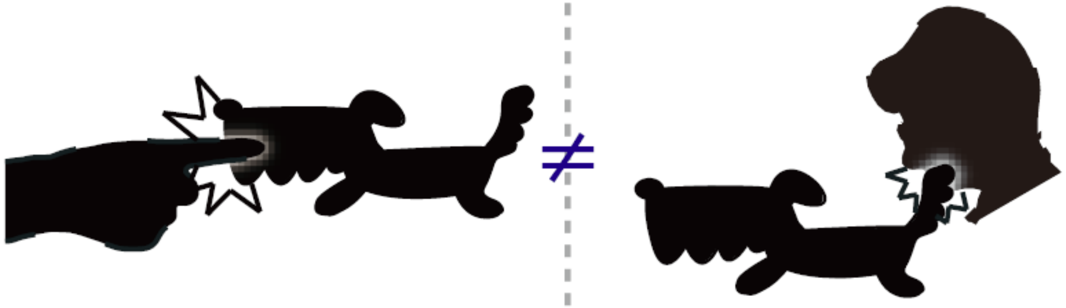
Fine-grained object recognition results (2012)



Is it enough to label images with words?

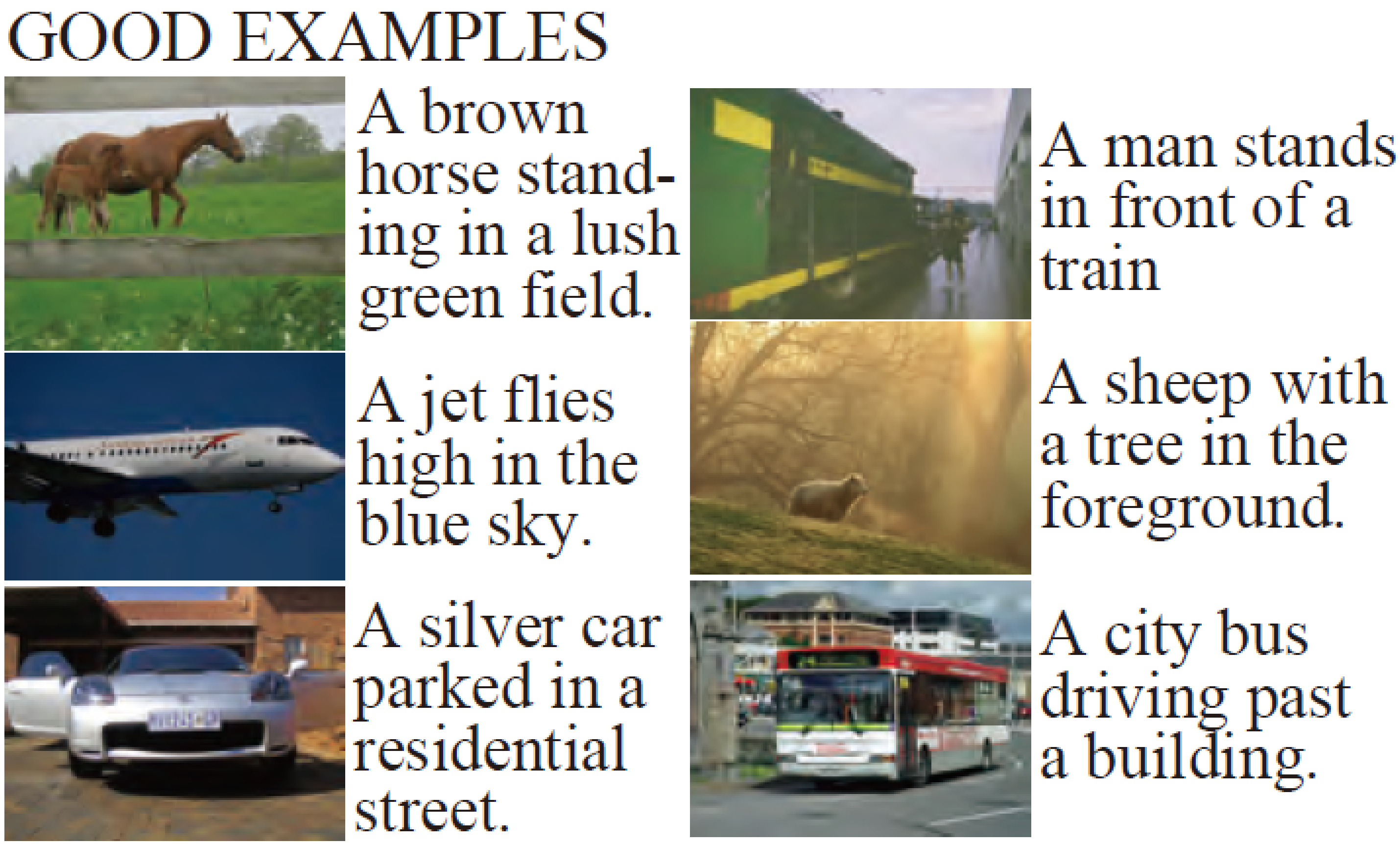
http://www.totallifecounseling.com/services-therapy-orlando-florida-clermont-east-southwest-therapy-therapists-counselors/ptsd/

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



Automatic Sentence Generation from Images

Best Application of a Theory Framework Special Prize at ACM Multimedia 2011 Grand Challenge.



Yoshitaka Ushiku, Tatsuya Harada, and Yasuo Kuniyoshi. Automatic Sentence Generation from Images. ACM Multimedia, 2011.

Automatic Sentence Generation from Images

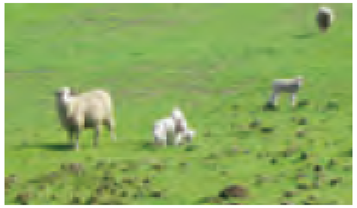
Best Application of a Theory Framework Special Prize at ACM Multimedia 2011 Grand Challenge.

BAD EXAMPLES

A white jeep parked on a large white horse.

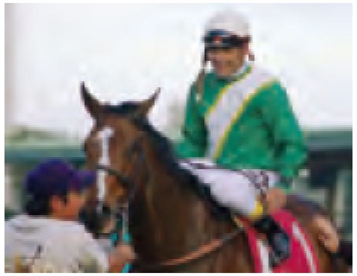
A man standing in front of spiderman poster.

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.

Yoshitaka Ushiku, Tatsuya Harada, and Yasuo Kuniyoshi. Automatic Sentence Generation from Images. ACM Multimedia, 2011.

Learning Images and Texts

Assumption: pairs of images and texts are given.

A Golden Retriever puppy jumps to catch a treat.

http://en.wikipedia.org/wiki/Golden_Retriever

画像と文章との学習方法：

１）画像と文章全体との学習

→文章表現の自由度が高すぎるため学習困難☹

２）画像とシングルラベル間の学習

→教師データとして与えられる単語間の順序を無視☹

Multi-Keyphrase Problem

Y. Ushiku, T. Harada, and Y. Kuniyoshi. Efficient Image Annotation for Automatic Sentence Generation. ACM MM, 2012.

- Proposal of multi-keyphrase problem
  - 画像とbi-gramなど順序を持つラベル間の関係性を学習し，画像に対して適切な複数の順序付きラベルを推定する.

- A man bites a white dog in his arms.

- man bites
- white dog
- his arms

- Assumption
  - 画像のコンテンツはいくつかのキーフレーズと文法で表現可能.

Keyphrase		Grammer		Sentence
his arms *man bites* *white dog*	+	in→*his arms* a→*man* a→*white dog*		*A man bites a white dog in his arms.*

Solution of Multi-Keyphrase Problem

Y. Ushiku, T. Harada, and Y. Kuniyoshi. Efficient Image Annotation for Automatic Sentence Generation. ACM MM, 2012.

man bites → Label_1

white dog → Label_2

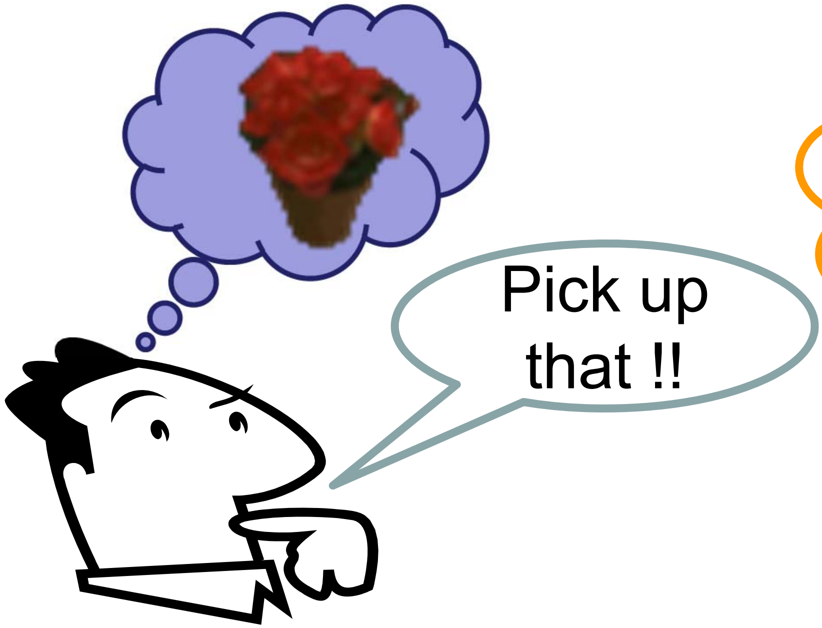
his arms → Label_3

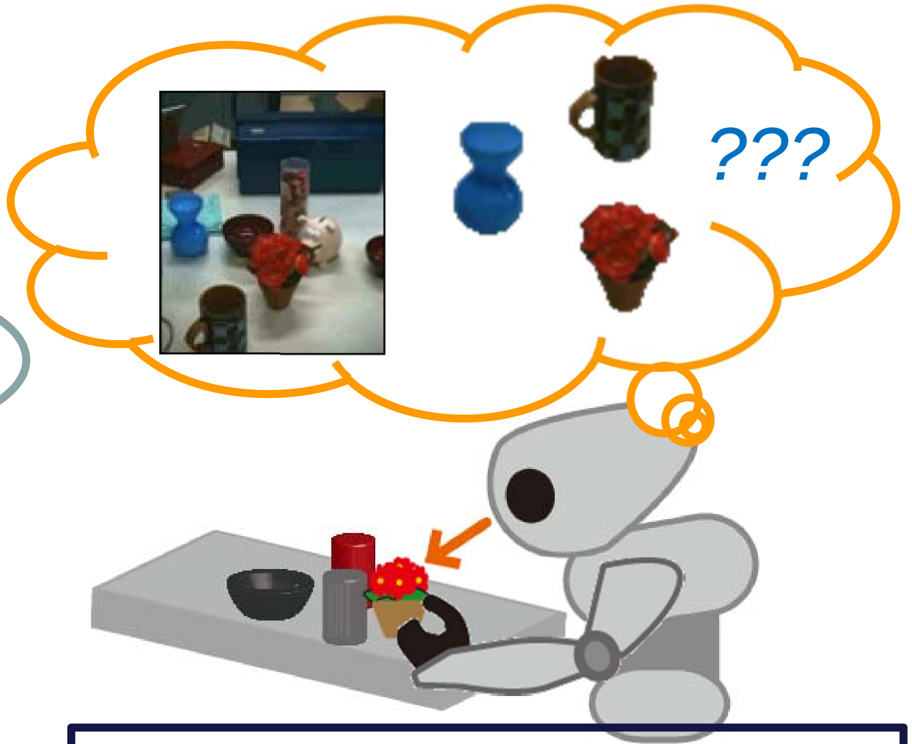
⇒ **Multi-class classification problem**

- Problem
  - キーフレーズは単語の組み合わせであるため，識別するクラス数が膨大になる．
  - 単語数nとするとbi-gramならn^2，tri-gramならn^3のオーダー．

- 従来のマルチクラス識別問題よりも，膨大なクラスを効率的に学習可能なアルゴリズムが求められる．
  - Passive Aggressive with Averaged Pairwise Lossの提案

Problems of Assistant Robot





How to **convey user's intention**

difficult and stressful to solve

How to **extract the target** from environment

complicated, variable shape and texture

Other Interesting Results

N. Shibuya, Y. Shimohata, T. Harada and Y. Kuniyoshi. Smart Extraction of Desired Object from Color-Distance Image with User's Tiny Scribble. IROS, 2008.

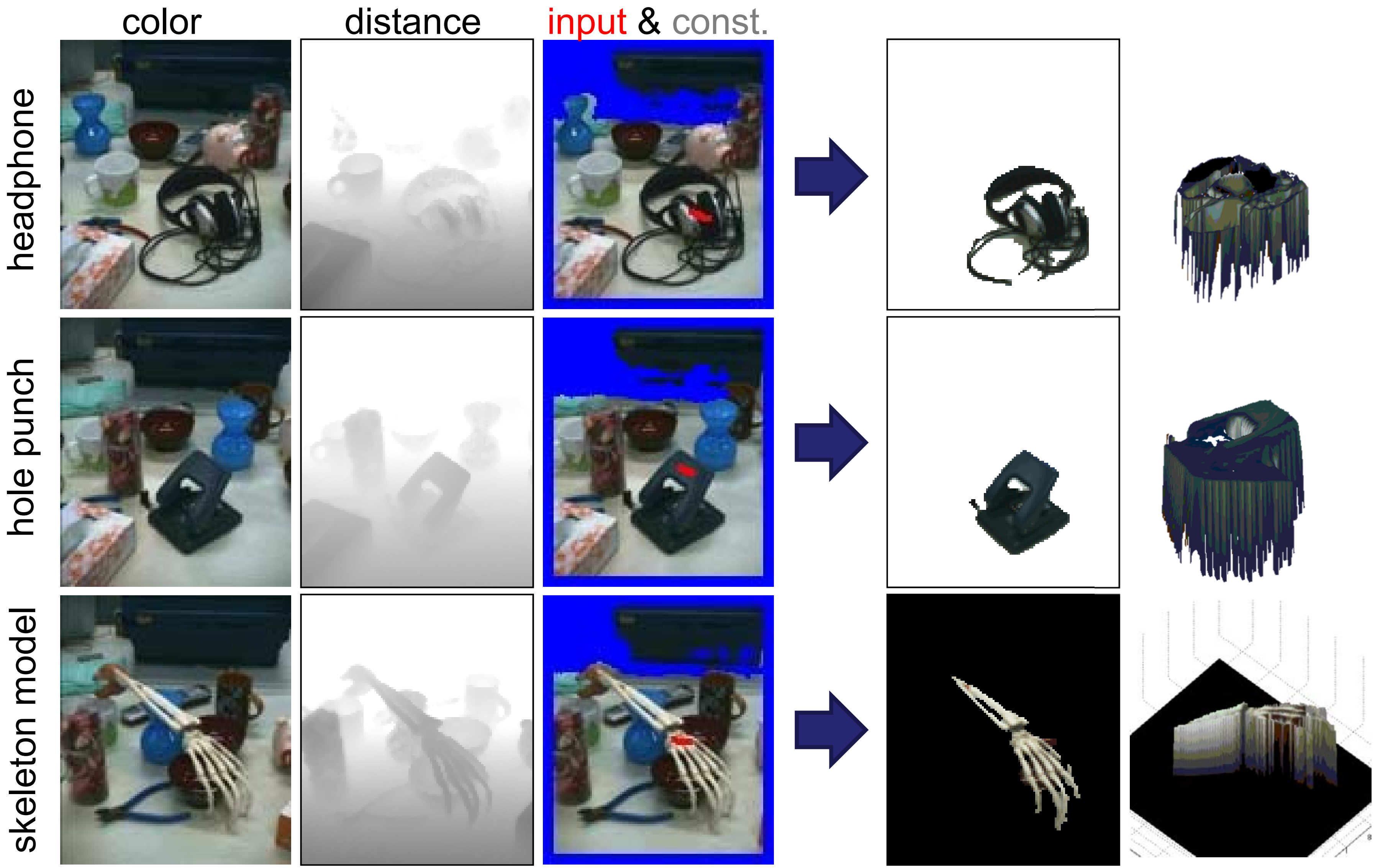
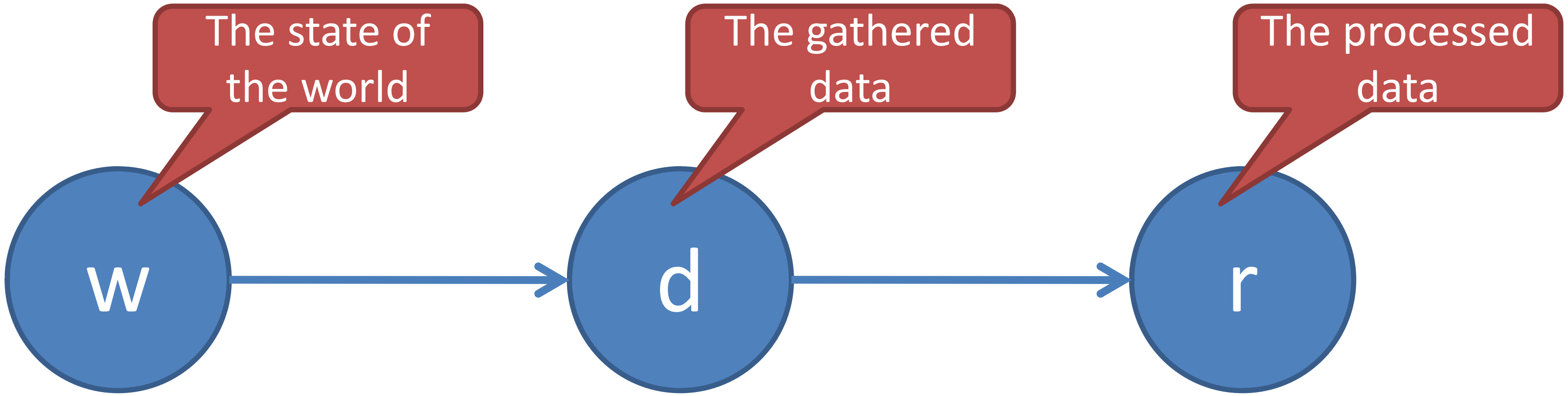


Image Representation

The data processing theorem



Markov chain

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

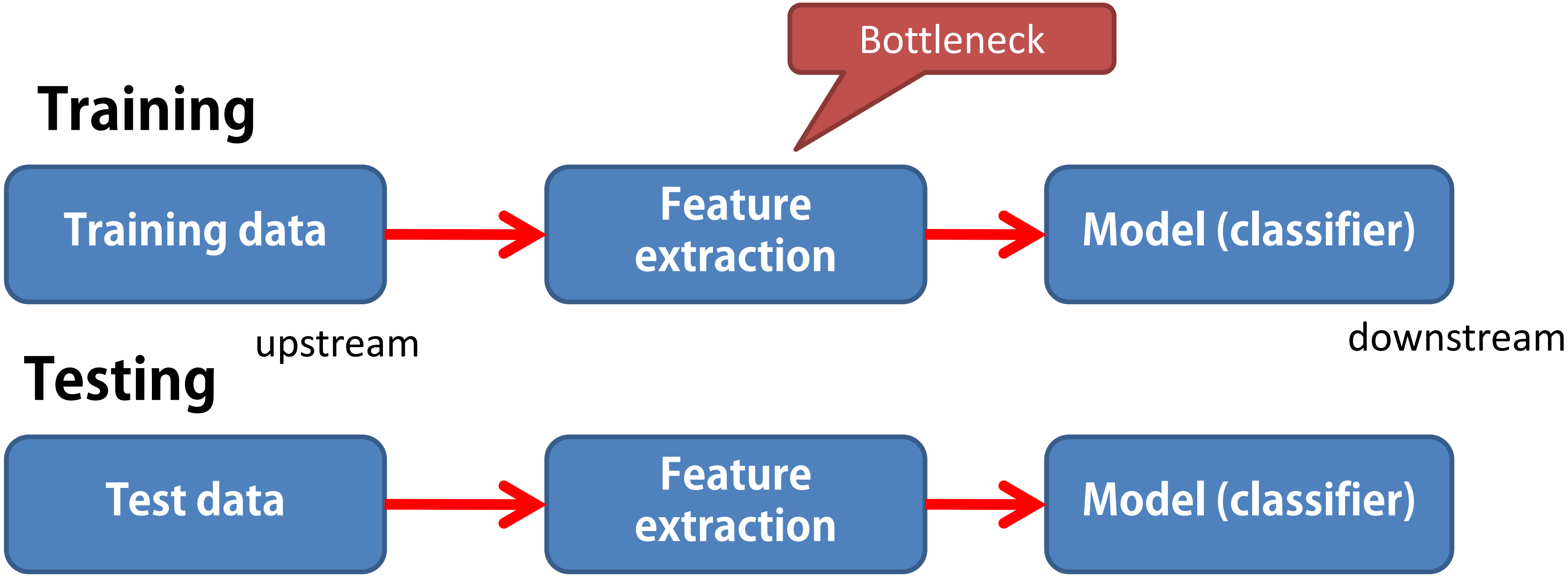
The average information

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

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

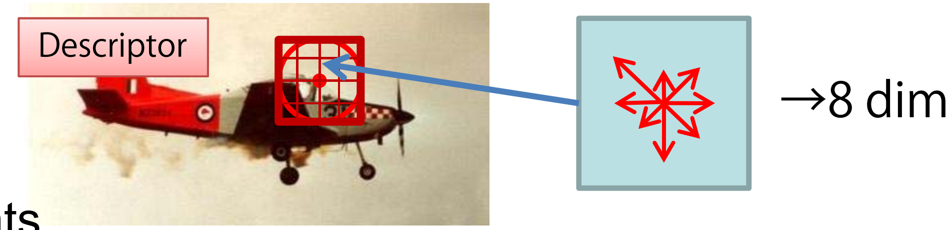
David J.C. MacKay. Information Theory, Inference, and Learning Algorithms. Cambridge University Press 2003.

Pipeline of Image Recognition



- Upstream process is more important than downstream process.
- Data and the feature extraction are more important than the classifier.
- Feature extraction is a bottleneck.
- If we have high quality data and a proper feature extraction method, even simple classifier can obtain high performance.

What's Image Feature?



General pipeline

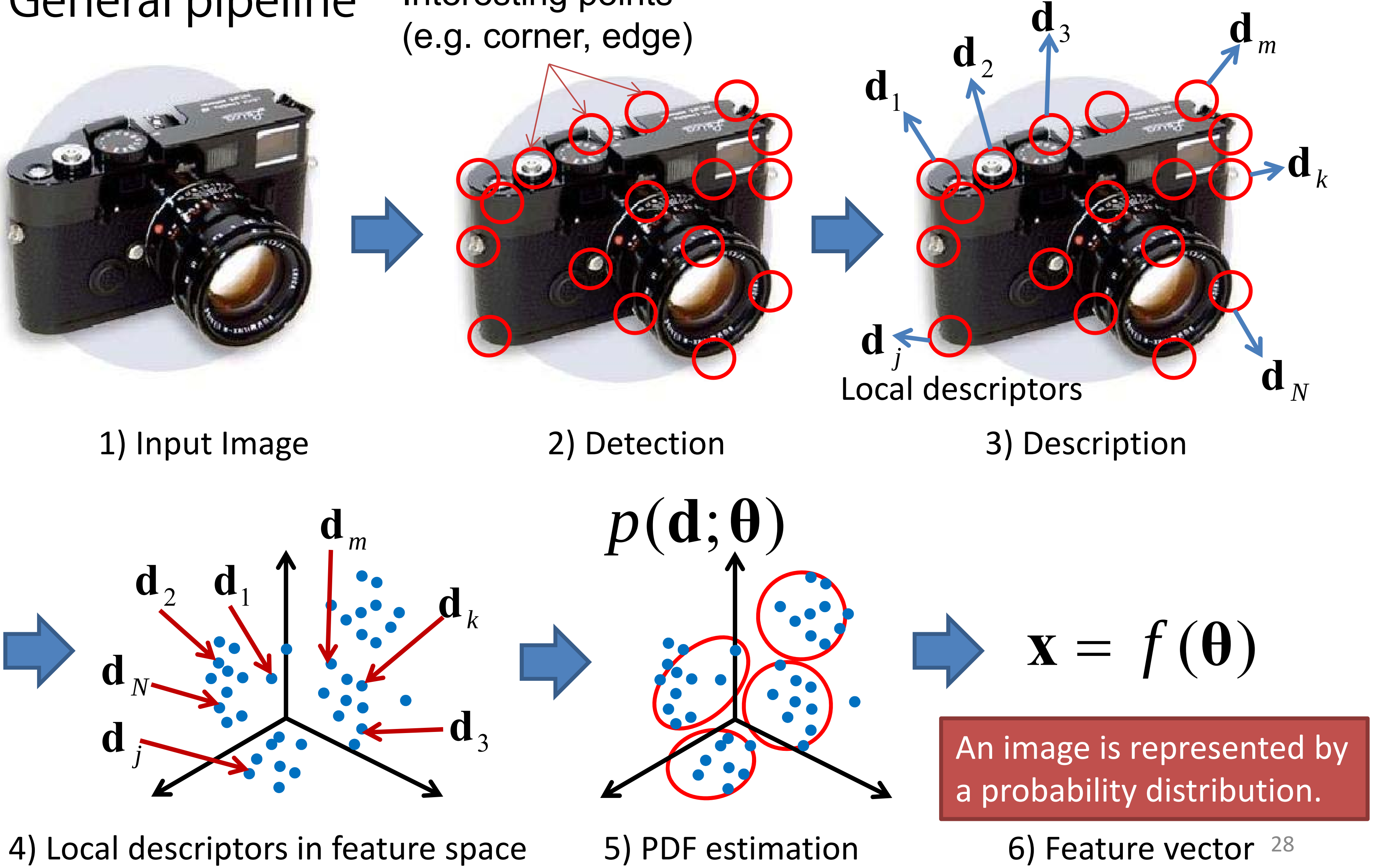
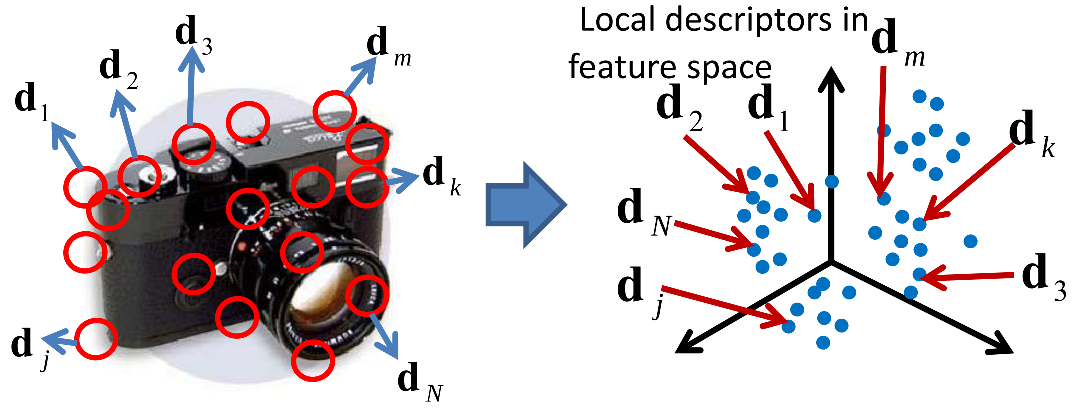


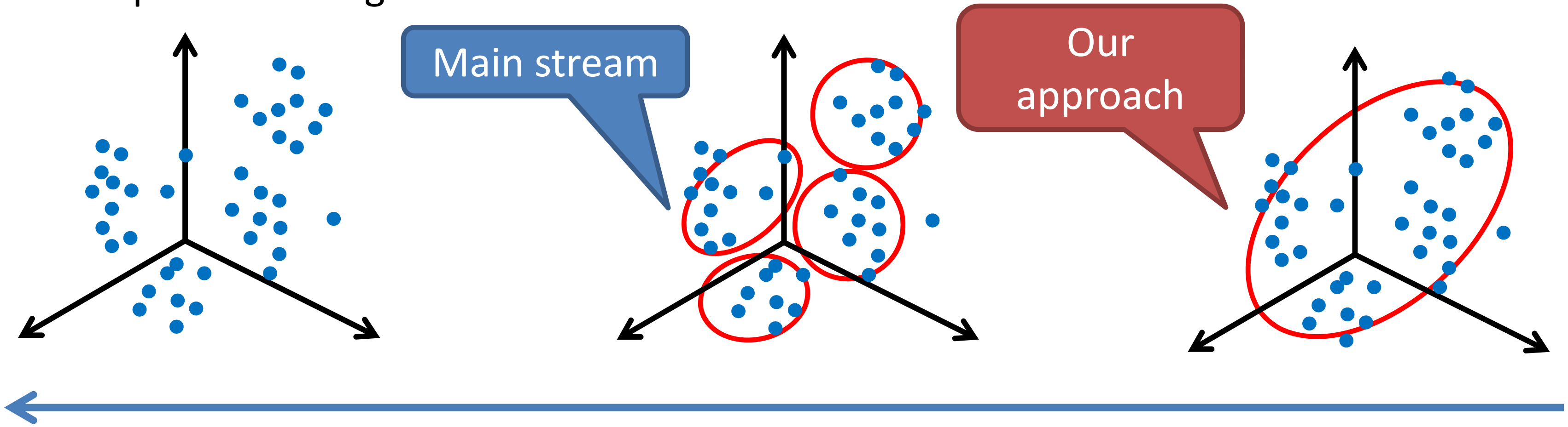
Image Representation



Descriptor matching

Codebook

Global feature



# of anchor points: large
Computational complexity: large

# of anchor points: small
Computational complexity: small

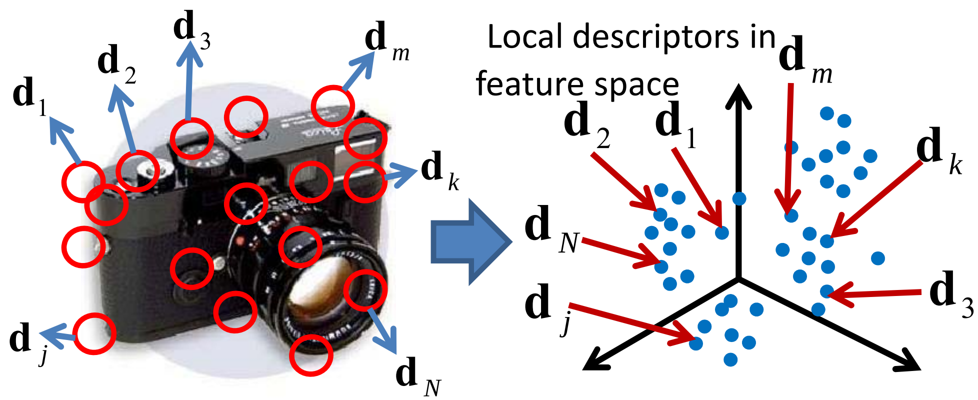
SVM-KNN
Naïve Bayes Nearest Neighbor
Graph Matching Kernel

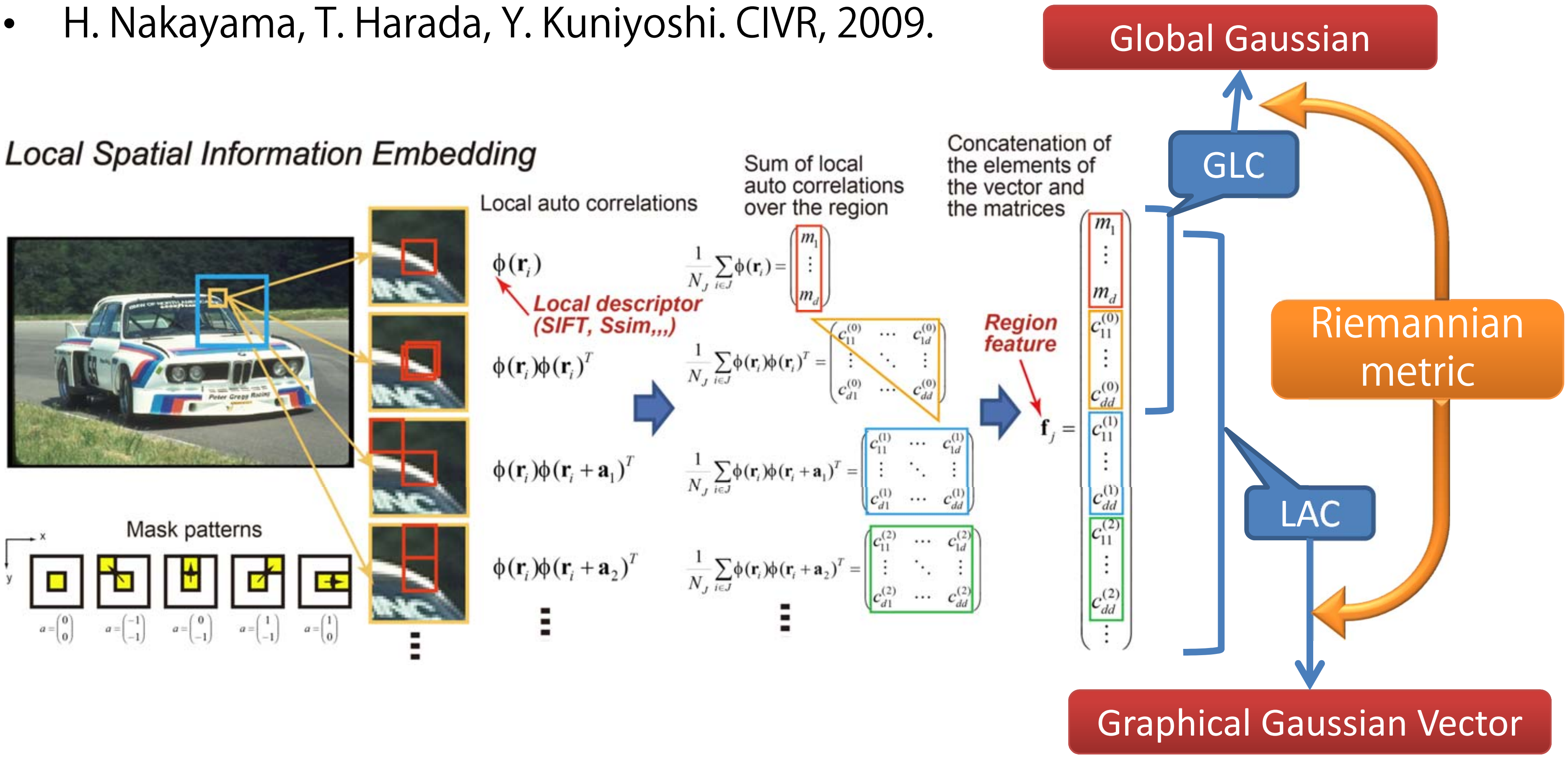
Bag of Visual Words
Gaussian Mixture Model
ScSPM, Super Vector, LLC
Fisher Vector

HLAC
GLC
Global Gaussian

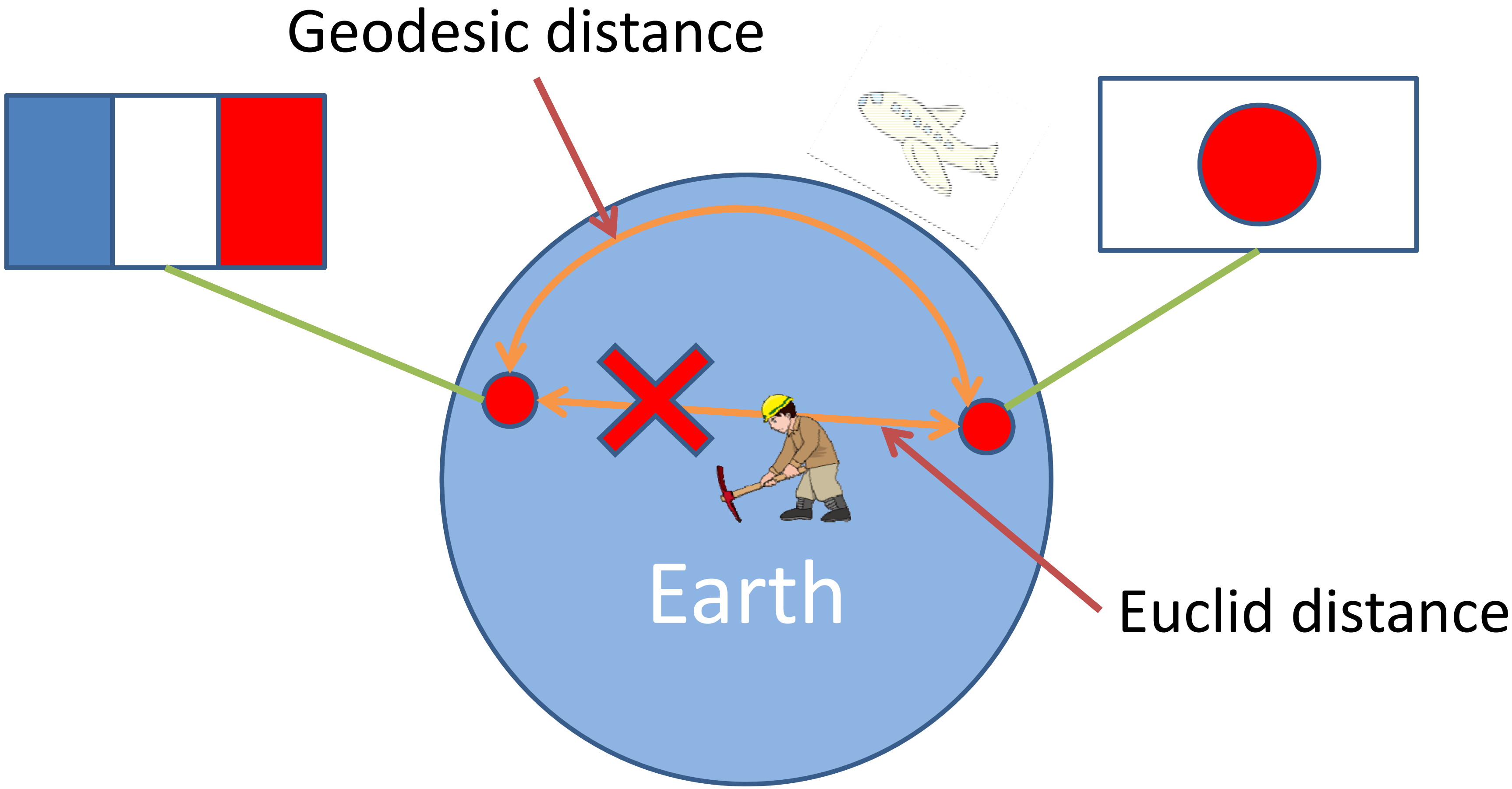
Image Feature Design

- 局所記述子の二次の統計量の有効活用
- 特徴間の適切な計量の埋め込み
- T. Harada, Y. Kuniyoshi. NIPS, 2012.
- T. Harada, H. Nakayama. ECCV ,2010.
- H. Nakayama, T. Harada, Y. Kuniyoshi. CVPR, 2010.
- H. Nakayama, T. Harada, Y. Kuniyoshi. CIVR, 2009.



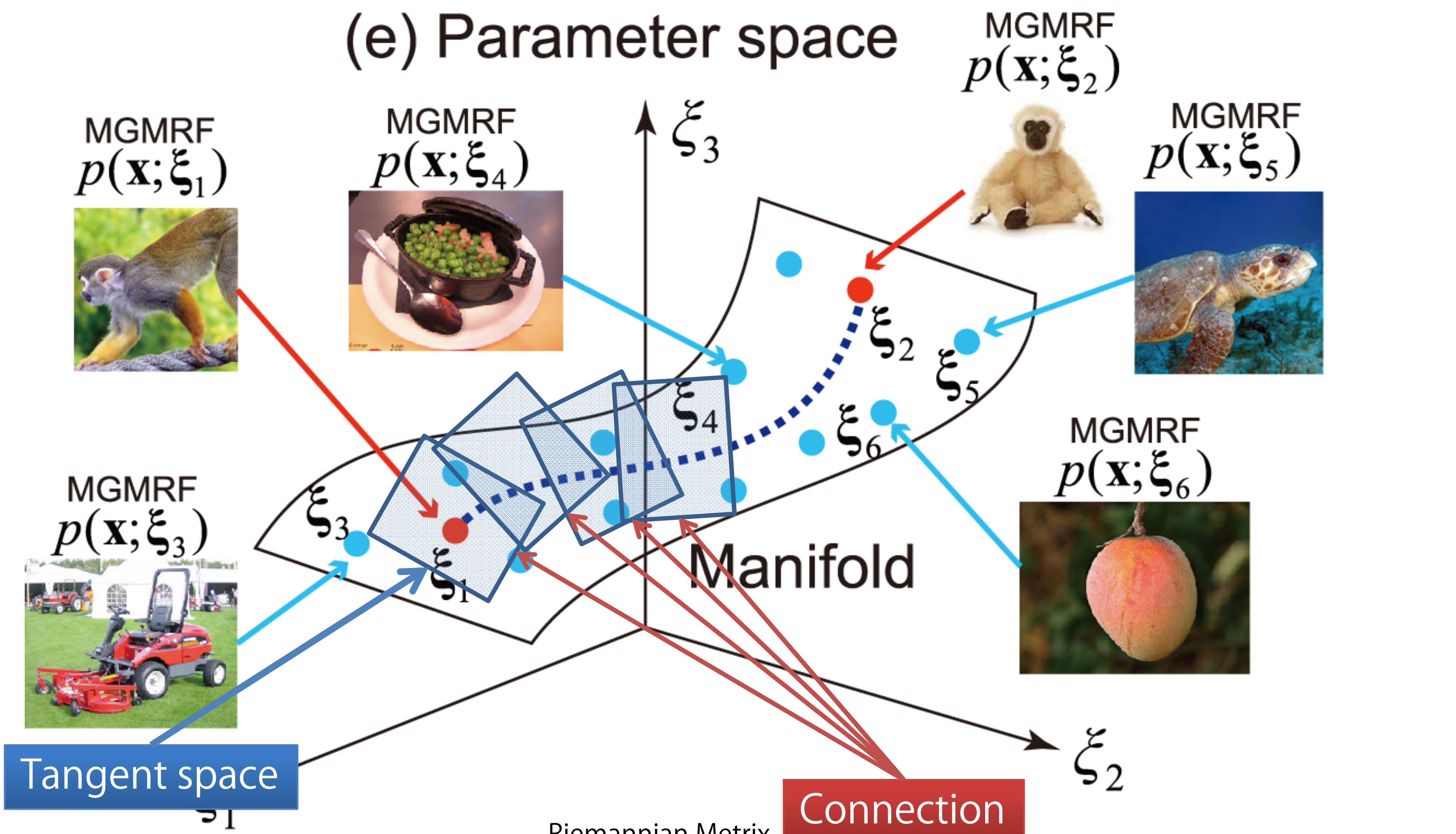


Distance between France and Japan



- Geodesic distance on the earth is more proper than Euclid distance.

Distance between Images



Riemannian Metrix

$$ds^2 = \mathrm{KL}[p(\boldsymbol{x};\boldsymbol{\eta}) : p(\boldsymbol{x};\boldsymbol{\eta}+d\boldsymbol{\eta})] = \frac{1}{2}d\boldsymbol{\eta}^{\top}\underline{G^*}\,d\boldsymbol{\eta}$$

Metric on tangent space: Fisher Information Matrix

Similarity measure

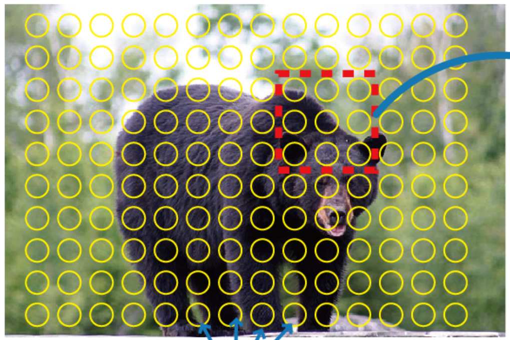
Feature → Better feature

局所特徴のモデル化

T. Harada. NIPS, 2012.

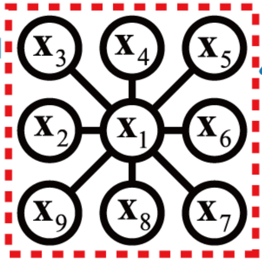
Face ?

(a) Densely sampled local descriptors



Local descriptors $\mathbf{x}_i \in \mathrm{R}^d$

(b) Multivariate Gaussian Markov Random Field (MGMRF)



$$\mathbf{x} = \left(\mathbf{x}_1^T \cdots \mathbf{x}_9^T\right)^T$$

(c) PDF (d) Feature Vector

$$p(\mathbf{x};\boldsymbol{\xi}) \rightarrow \boldsymbol{\xi}$$

Parameter of MGMRF

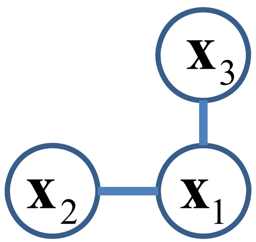
Gaussian MRF

指数型分布族 $p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}^\top \phi(\boldsymbol{x}) - \Phi(\boldsymbol{\theta}) + C(\boldsymbol{x})\right)$

θ 座標系 自然パラメータ $\boldsymbol{\theta} = (h_i, i \in \mathcal{V}) \cup (-\frac{1}{2}J_{ii}, i \in \mathcal{V}) \cup (-J_{ij}, \{i,j\} \in \mathcal{E})$

η 座標系 期待値パラメータ $\boldsymbol{\eta} = \mathbb{E}[\phi(\boldsymbol{x})] = (\mu_i, i \in \mathcal{V}) \cup (P_{ii} + \mu_i^2, i \in \mathcal{V}) \cup (P_{ij} + \mu_i\mu_j, \{i,j\} \in \mathcal{E})$

【例】



$$\mathbf{x} = \left(\mathbf{x}_1^T\, \mathbf{x}_2^T\, \mathbf{x}_3^T\right)^T$$

$E[\mathbf{x}] = 0$ とする

$P_{ii},\ i \in V$

異なる位置の局所特徴の相関

$P_{ij},\ \{i,j\} \in E$

$P_{ij},\ \{i,j\} \notin E$

$$P = E[\mathbf{x}\mathbf{x}^T] = \begin{pmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & 0 \\ P_{31} & 0 & P_{33} \end{pmatrix}$$

$$P_{ij} = E[\mathbf{x}_i \mathbf{x}_j^T]$$

２と３の間にはエッジはないので相関０

$$\boldsymbol{\eta} = \begin{pmatrix} \mathrm{upper}[P_{11}] \\ \mathrm{upper}[P_{22}] \\ \mathrm{upper}[P_{33}] \\ \mathrm{vec}[P_{12}] \\ \mathrm{vec}[P_{13}] \end{pmatrix}$$

局所自己相関特徴の一般形

GMRFにおけるFisher情報行列

T. Harada, and Y. Kuniyoshi. Graphical Gaussian Vector for Image Categorization. NIPS, 2012.

- 全ノードが接続されたグラフ構造におけるFisher情報行列（FIM）を計算
  - Full GaussianのFIMの計算

$$F_{ij}^{*}(\boldsymbol{\eta}) = \frac{\partial \boldsymbol{\mu}^{T}}{\partial \boldsymbol{\eta}_i} P^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}_j} + \frac{1}{2}\mathrm{tr}\left(P^{-1} \frac{\partial P}{\partial \boldsymbol{\eta}_i} P^{-1} \frac{\partial P}{\partial \boldsymbol{\eta}_j} \right)$$

- Full Gaussian FIMをグラフ部，非グラフ部に分解

J. K. Johnson. Convex Relaxation Methods for Graphical Models: Lagrangian and Maximum Entropy Approaches. PhD thesis, MIT, 2008.

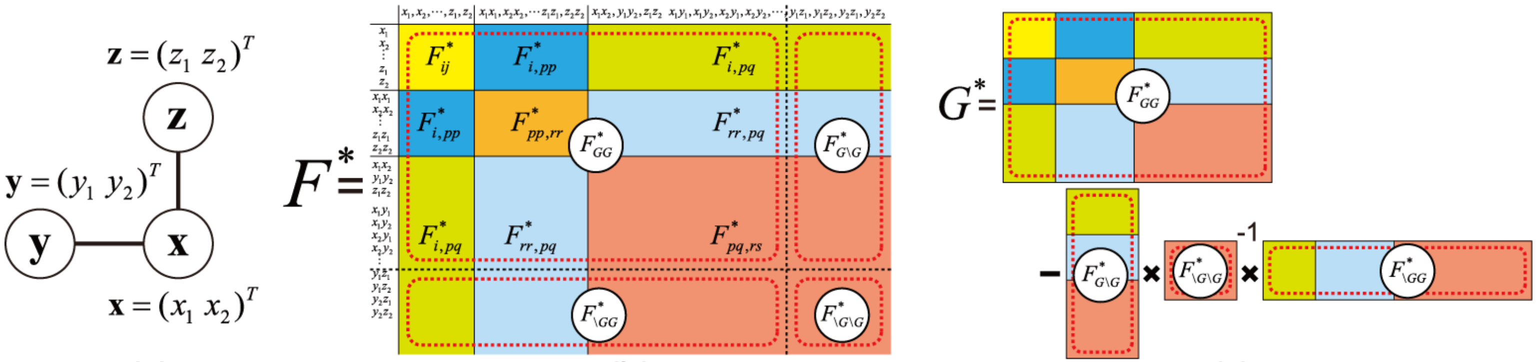
$$F^{*}(\boldsymbol{\eta}) = \begin{pmatrix} F^{*}_{\mathcal{G},\mathcal{G}}(\boldsymbol{\eta}) & F^{*}_{\mathcal{G},\backslash\mathcal{G}}(\boldsymbol{\eta}) \\ F^{*}_{\backslash\mathcal{G},\mathcal{G}}(\boldsymbol{\eta}) & F^{*}_{\backslash\mathcal{G},\backslash\mathcal{G}}(\boldsymbol{\eta}) \end{pmatrix}$$

シューアの補元

- GMRFにおけるFIMを計算

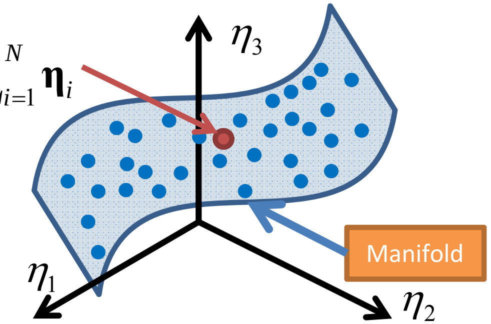
$$G^{*}(\boldsymbol{\eta}) = F^{*}_{\mathcal{G},\mathcal{G}}(\boldsymbol{\eta}) - F^{*}_{\mathcal{G},\backslash\mathcal{G}}(\boldsymbol{\eta}) \left(F^{*}_{\backslash\mathcal{G},\backslash\mathcal{G}}(\boldsymbol{\eta}) \right)^{-1} F^{*}_{\backslash\mathcal{G},\mathcal{G}}(\boldsymbol{\eta})$$

【例】



(a) (b) (c)

$$\boldsymbol{\eta}_c = \sum_{i=1}^{N} \boldsymbol{\eta}_i$$



期待値パラメータの値によってFIM異なる値を示すために扱いにくい．そこで我全ての訓練データの中心点における接空間を用いてこの空間を近似する．

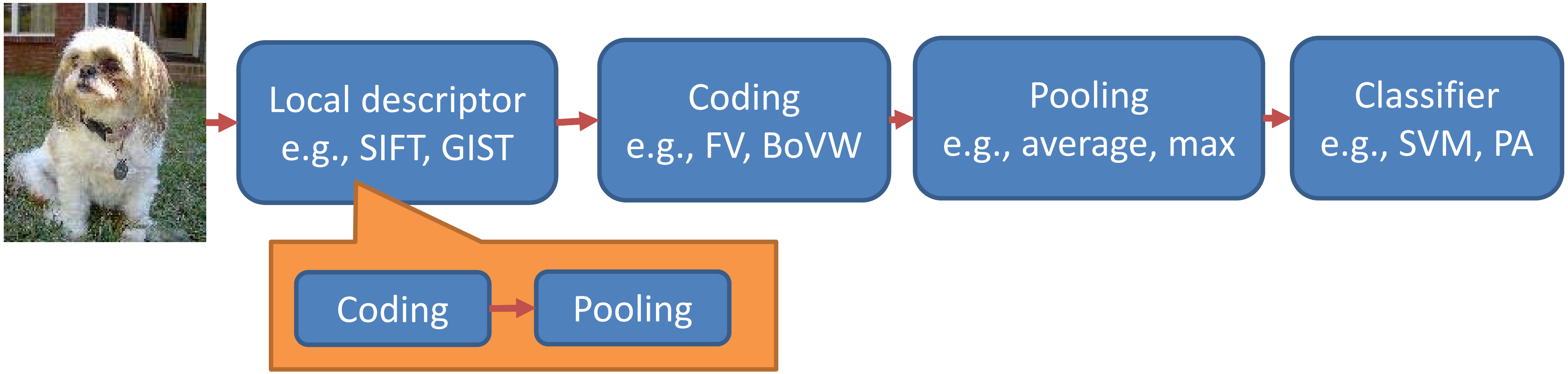
一つの計量ですべての空間を表現可能と大胆に仮定

$$G^{*}(\boldsymbol{\eta}) \approx G^{*}(\boldsymbol{\eta}_c)$$

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

Deep Learning

Image Recognition Pipeline



- Conventional image recognition pipeline is already deep!
- Problem
  - Local descriptors are handcrafted!
  - Coding method and classifier are learnt separately.
- Deep learning methods have been developed, but they have succeeded in only small networks.

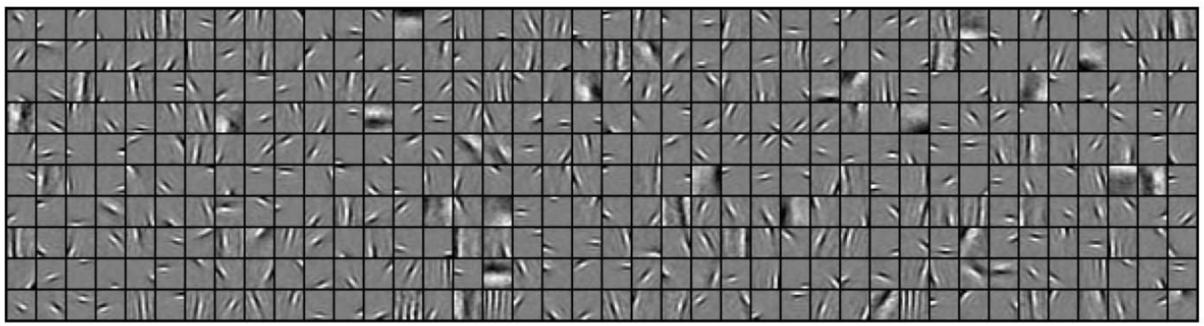
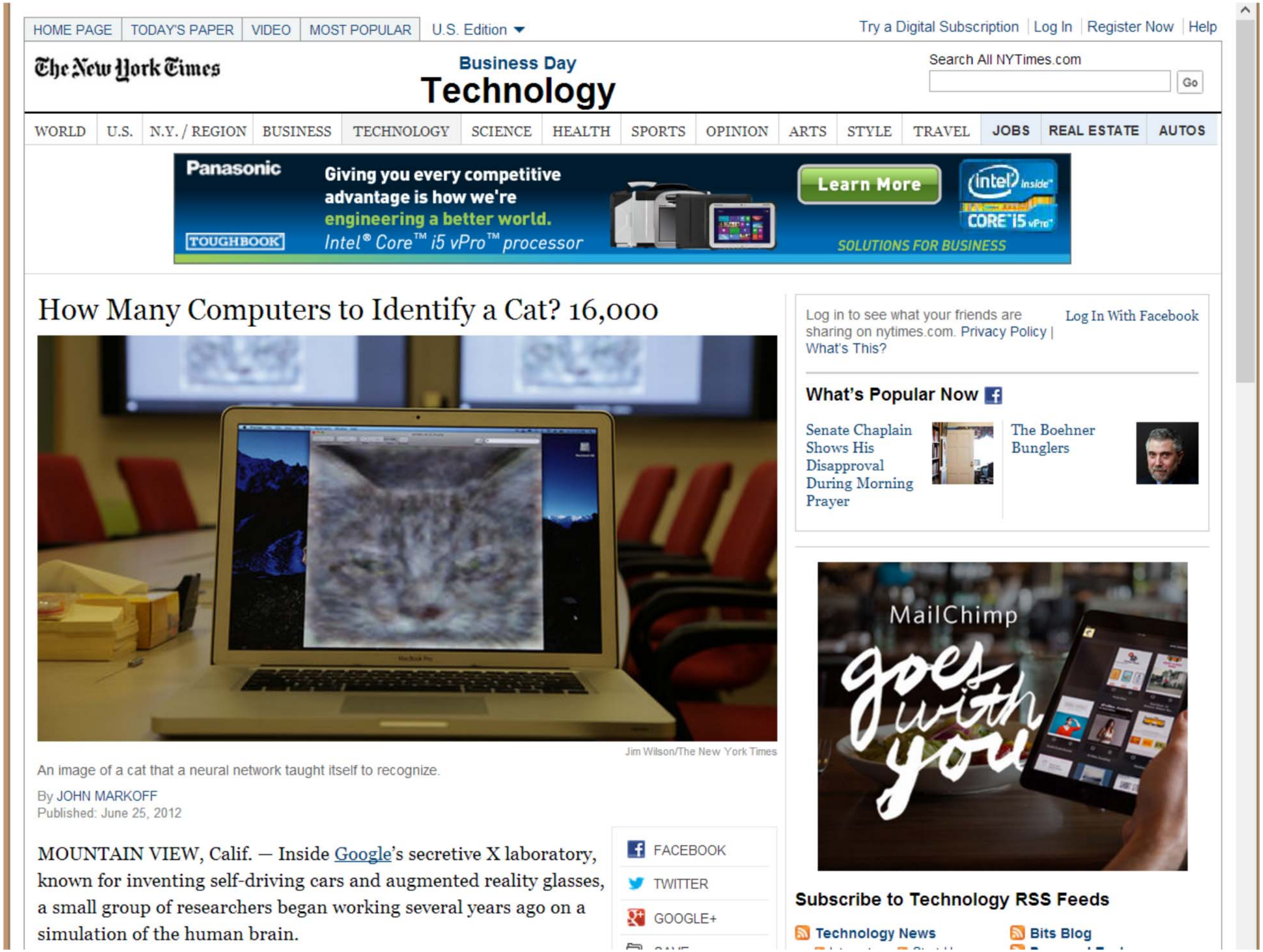


Figure 3: 400 first layer bases learned from the van Hateren natural image dataset, using our algorithm.
Honglak Lee, Chaitanya Ekanadham, Andrew Y. Ng. Sparse deep belief net model for visual area V2. NIPS, 2007.

It basically invented the concept of a cat.



HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | U.S. Edition

Try a Digital Subscription | Log In | Register Now | Help

The New York Times

Business Day

Technology

Search All NYTimes.com

Go

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION | ARTS | STYLE | TRAVEL | JOBS | REAL ESTATE | AUTOS

How Many Computers to Identify a Cat? 16,000

Jim Wilson/The New York Times

An image of a cat that a neural network taught itself to recognize.

By JOHN MARKOFF
Published: June 25, 2012

FACEBOOK
TWITTER
GOOGLE+

MOUNTAIN VIEW, Calif. — Inside Google's secretive X laboratory, known for inventing self-driving cars and augmented reality glasses, a small group of researchers began working several years ago on a simulation of the human brain.

Log in to see what your friends are sharing on nytimes.com. Privacy Policy | What's This?

Log In With Facebook

What's Popular Now

Senate Chaplain Shows His Disapproval During Morning Prayer

The Boehner Bunglers

Subscribe to Technology RSS Feeds

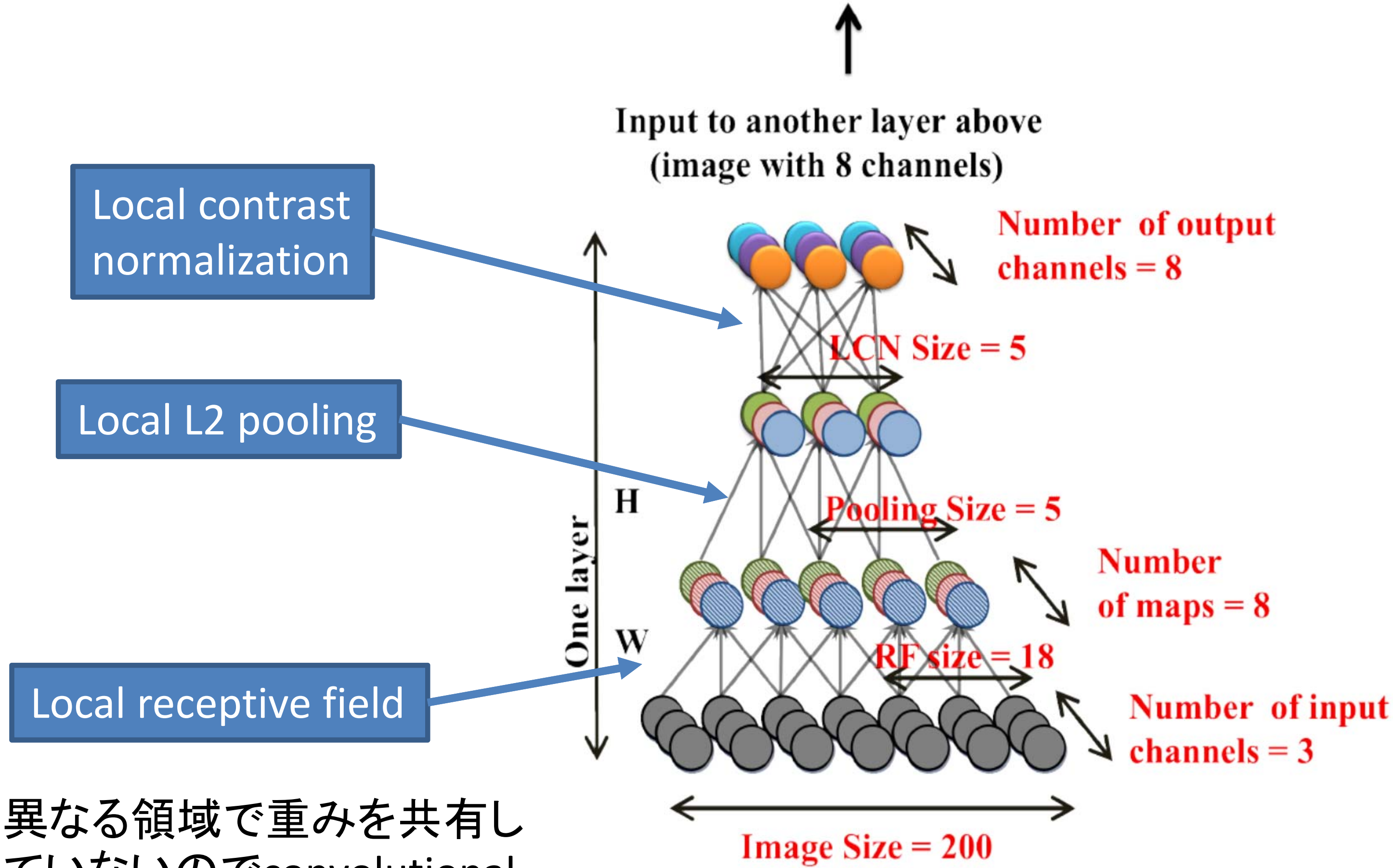
Technology News

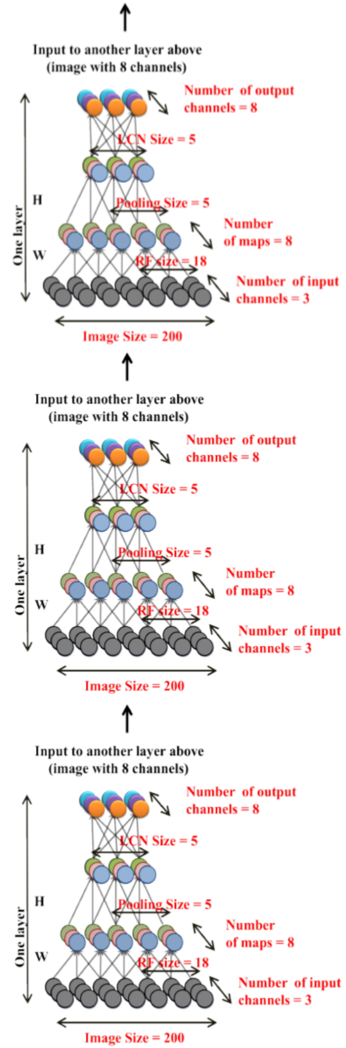
Bits Blog

http://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html?pagewanted=1&_r=4&ref=technology

構造

Q. V. Le et al. Building High Level Features Using Large Scale Unsupervised Learning. ICML 2012.





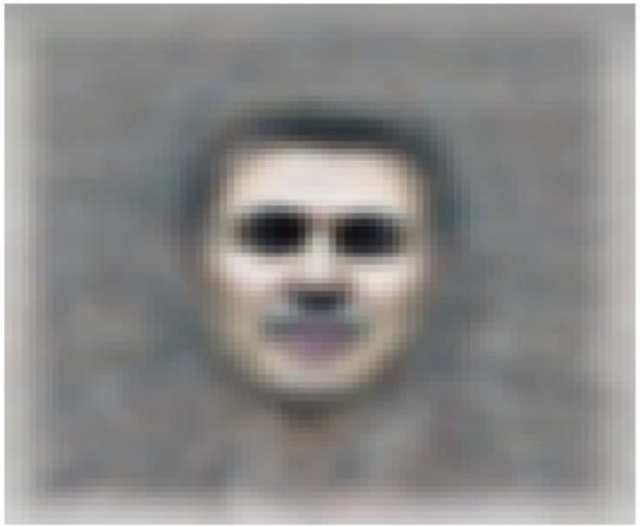
異なる領域で重みを共有していないのでconvolutional NNではない

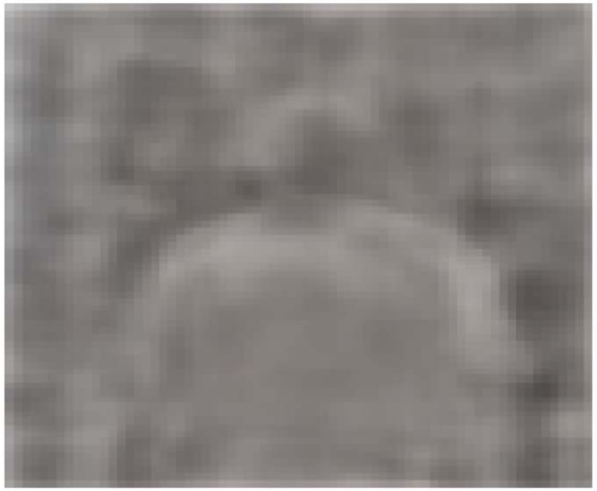
10億のパラメータを持つ．ただし人はこれよりも10^6倍大きなニューロンとシナプスを持つ

実験結果

Q. V. Le et al. Building High Level Features Using Large Scale Unsupervised Learning. ICML 2012.

人顔，猫顔，上半身に最も反応するニューロンを選択し，その振る舞いを観察





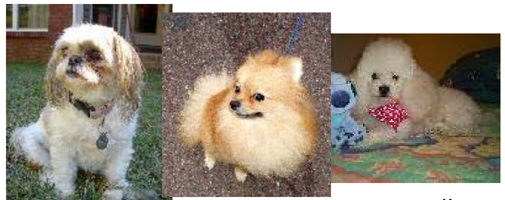
識別実験では，ネットワークの上位層に教師付きのone-vs-all logistic classifierを追加

大規模データを利用した画像認識

- ILSVRC (ImageNet Large Scale Visual Recognition Challenge)
  - 大規模なデータを利用した，国際的画像認識のコンペティション
  - http://www.image-net.org/challenges/LSVRC/2012/index
  - 現在最も困難な画像認識タスク

- Task 1
  - 120万枚の画像を学習して，1000クラスの画像を識別

- Task 2
  - 画像内に1000クラスの物体がどこあるのか検出
- Task 3
  - 120の犬の種類を当てるTask 1より分類が困難な識別タスク．



Shih-Tzu　Pomeranian　toy poodle

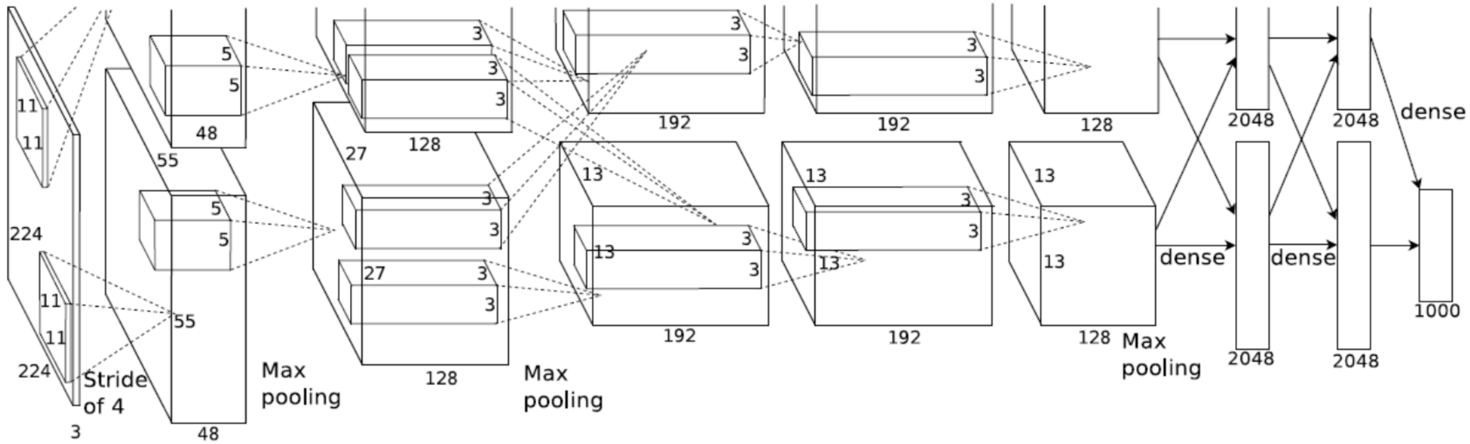
2位　Task 1　このチーム

Team	Flat Error
1) SuperVision Univ. of Toronto	0.153
2) ISI (ours) Univ. of Tokyo	0.262
3) OXFORD_VGG Univ. of Oxford	0.270

1位　Task 3

Team	mAP
1) ISI (ours) Univ. of Tokyo	0.323
2) XRCE/INRIA Xerox Research Centre Europe/INRIA	0.310
3) Uni Jena Univ. Jena	0.246

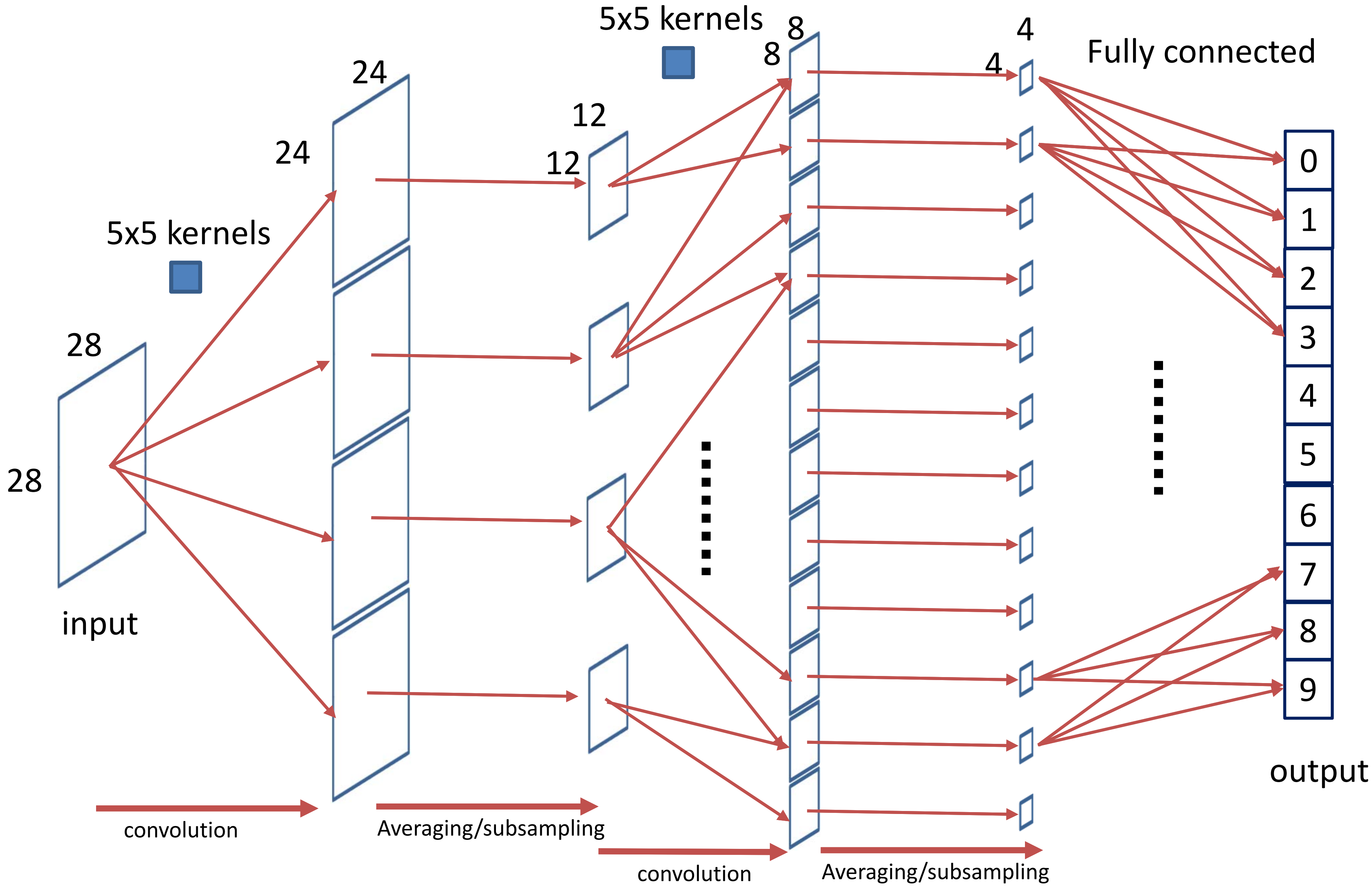
Deep Convolutional Neural Networks



- Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. NIPS, 2012.
- ILSVRC2012でトップになったシステム
- 特徴
  - 5 convolutional layers + 3 fully-connected layers
  - ReLU nonlinearity
  - Multiple GPGPU
  - Local response normalization
  - Overlappling pooling

Example of CNNs Architecture

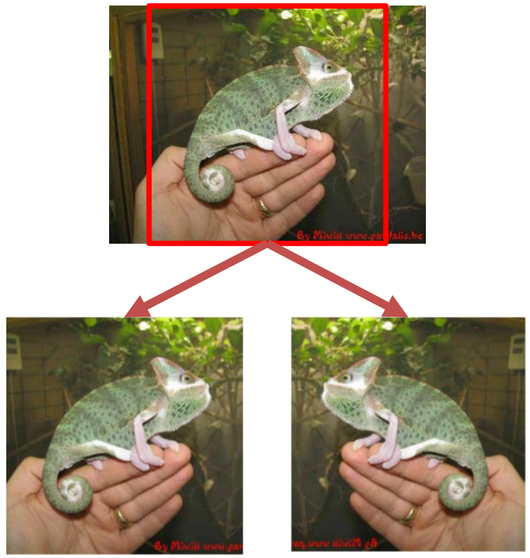
– Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In David Touretzky, editor, Advances in Neural Information Processing Systems 2 (NIPS*89), Denver, CO, 1990. Morgan Kaufman.



Reducing Overfitting

- Dropout
  - G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R.R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580, 2012.
  - テストセットのエラーを低減する良い手法として，多くの異なるネットワークの予測の平均を用いる
  - トレーニングデータが提示される毎に，隠れ層のニューロンの出力を1/2の確率で0とする.
    - トレーニングデータ毎に異なるネットワークアーキテクチャが選択される. ただしネットワークの重みは共有している.
  - 1st, 2nd convolutional layersに適用
  - テスト時には平均ネットワーク利用
  - N個の隠れ層ユニットを持つとすると，2^N個のネットワークによって予測されるラベルの確率の平均を用いてクラスラベルを予測していることになる.

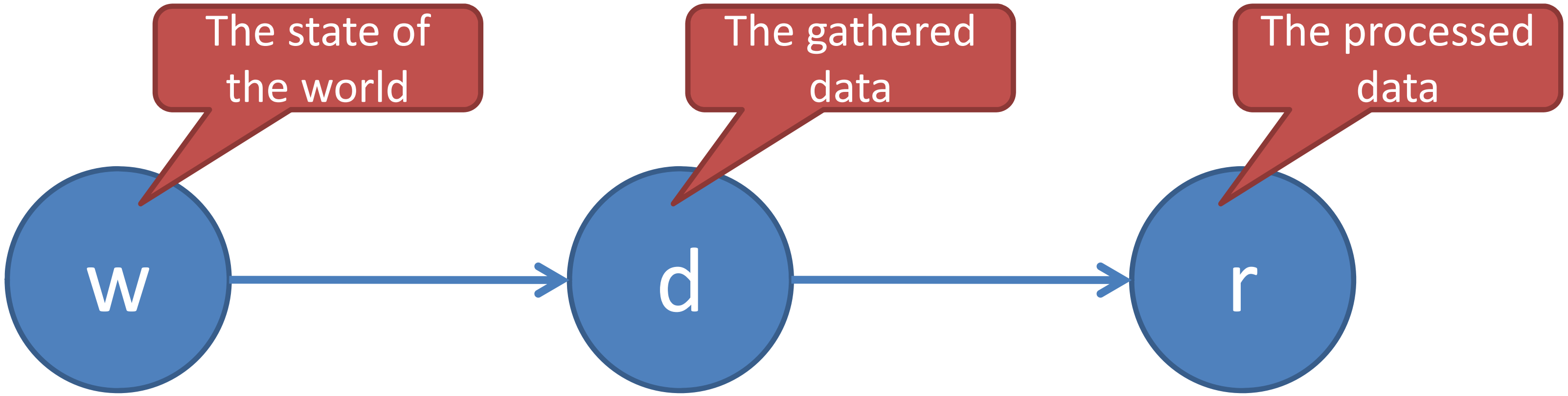
Data Augmentation



- トレーニングデータを2048倍に水増し
  - データを平行移動して拡張
  - 水平の鏡像画像を用いて拡張
  - Deep CNNsのoverfittingを回避する
- RGBチャンネルの輝度を変化させデータを水増し
  - 照明変化に不変にする
  - Top1 errorで1%の性能向上
  - RGBのピクセル値にPCAをかける
  - 各画像のRGBピクセルに以下の値を加える.

$$\begin{bmatrix} I_{xy}^R \\ I_{xy}^G \\ I_{xy}^B \end{bmatrix} = \begin{bmatrix} I_{xy}^R \\ I_{xy}^G \\ I_{xy}^B \end{bmatrix} + \underbrace{[\boldsymbol{p}_1 \quad \boldsymbol{p}_2 \quad \boldsymbol{p}_3]}_{\text{Eigenvectors}} \underbrace{\begin{bmatrix} \alpha_1\lambda_1 \\ \alpha_2\lambda_2 \\ \alpha_3\lambda_3 \end{bmatrix}}_{\text{Eigenvalues}}$$

The data processing theorem



Markov chain

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

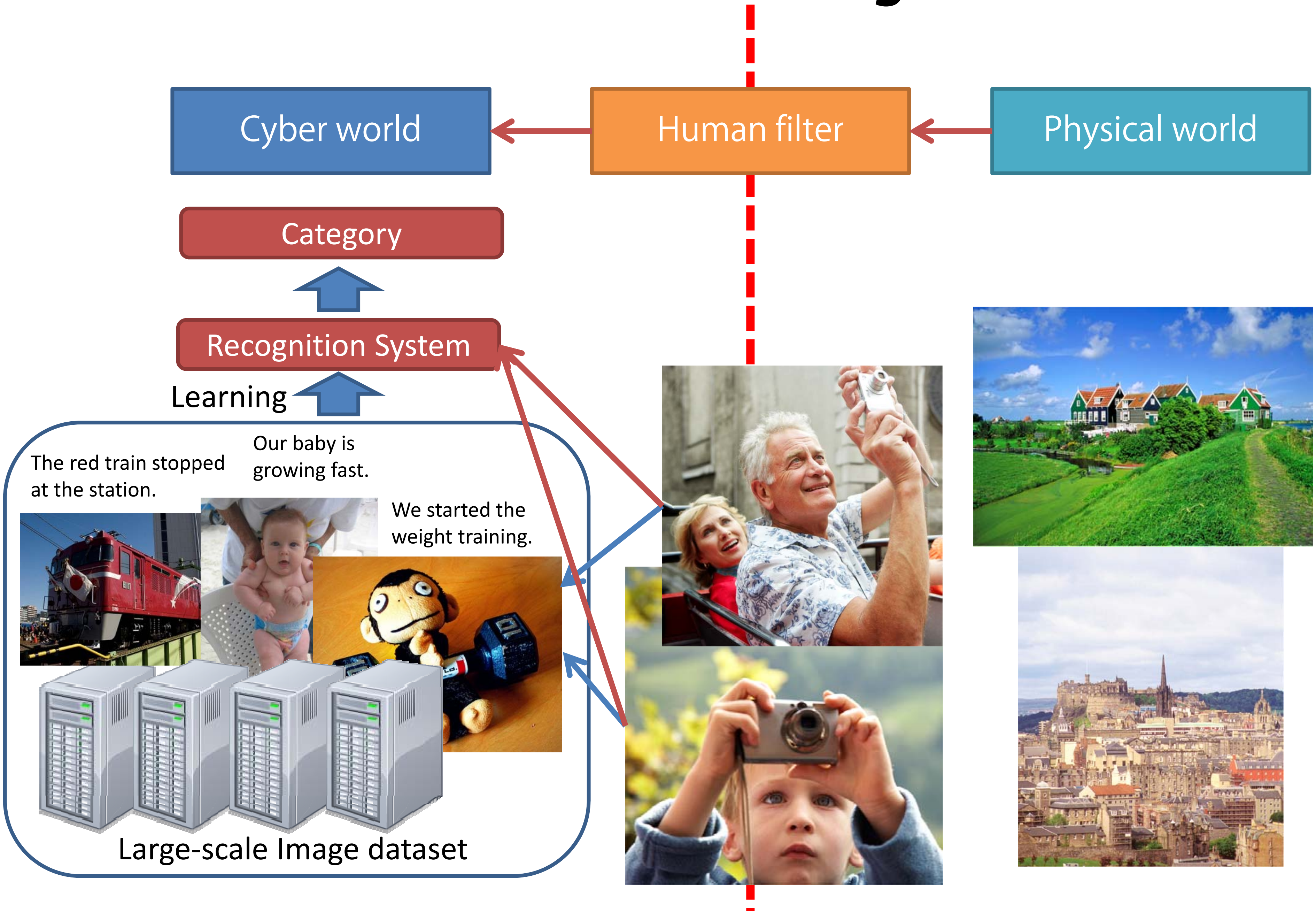
The average information

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

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.

Framework of Real-World Recognition



Journalist Robot

R. Matsumoto, H. Nakayama, T. Harada and Y. Kuniyoshi. Journalist Robot: Robot System Making News Articles from Real World. IROS, 2007.

The web data is huge, while the real-world is infinite!

Hardware

- Segway
- LRF
- Camera
  + Pan-tilt head
- Microphone
  + Arm
- Laptop PC
- Battery

Process Flow & software modules

Crawl

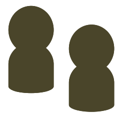
Anomaly detected

Approach to anomaly

Investigate

Publish

News, events.



[twitter.com]

SLAM | Human detection

Anomaly detection | Verbal interviewing

Results

News article generated (in Japanese)

What is this strange thing?

2011/02/12 23:34:11

Witness said, “Practicing poster session for coming conference. It is about a robot finding news”.

The picture taken by the system near the abnormal object.

- Picture for the article
- Dictation of the interview
- Accessible by web browser

Posting to a microblogging system

journalistrobot I found: http://localhost/zoomed_news_image.png
Witness said, “Practicing poster session for coming conference. It is about a robot finding news”.

about 19 minutes ago from api

In twitter client:

All Friends journalistrobot (localhost)

I found: http://localhost/zoomed_news_image.png
Witness said, “Practicing poster session for coming conference. It is about a robot finding news”.

journalistrobot, [+]
Sat 12 Feb 23:34 via api

The followers of the system gets easy access to the news.

まとめ

- まとめ
  - 大規模な画像データを用いた実世界認識システムを紹介
    - 高精度画像認識システム
    - 画像からの文章生成システム
  - 画像認識性能向上には，データ，特徴抽出，モデルの順に高い質が必要であることを述べた.
  - 高い表現能力を維持しつつ効率的な画像特徴の設計方法に関して紹介した.
- 今後（より実世界への歩み寄り）
  - データ収集と認識システム統合
  - ウェアラブルデバイスとの融合
  - ユーザインタフェースとの融合
  - **実世界の面白いがわかる知能！**
  - **空間の定義されていない世界に空間をどう定義するのか？**