

知能情報論 画像認識の（簡単な）歴史

2016年4月20日

東京大学 大学院情報理工学系研究科
原田達也

Marvin Minsky



- マービン・ミンスキ (Marvin Minsky, 1927-2016) は、アメリカ合衆国のコンピュータ科学者であり、認知科学者。専門は人工知能(AI)であり、MIT人工知能研究所の設立者の1人。初期の人工知能研究を行い、AIや哲学に関する著書でも知られ、「人工知能の父」と呼ばれる。現在ダートマス会議として知られる、"The Dartmouth Summer Research Project on Artificial Intelligence (1956)" の発起人の一人。

<http://ja.wikipedia.org/wiki/%E3%83%9E%E3%83%BC%E3%83%93%E3%83%B3%E3%83%BB%E3%83%9F%E3%83%B3%E3%82%B9%E3%82%AD%E3%83%BC>

In 1966,

Spend the summer linking a camera
to a computer and getting the
computer to describe what it saw.

Undergraduate student
Gerald Jay Sussman

Richard Szeliski. Computer Vision: Algorithms and Applications
(Texts in Computer Science). Springer-Verlag.

金出武雄



<http://www.dh.aist.go.jp/members/kanade.php>

- 金出 武雄（かなで たけお、1945年10月24日 - ）は、日本の工学者。兵庫県氷上郡春日町（現・丹波市）出身。京都大学大学院博士課程修了。工学博士。現米カーネギーメロン大学U.A. and Helen Whitaker University Professor。同大ロボティクス研究所所長を1992年から2001年まで10年に亘って務めた。専門はロボット工学、画像認識。

<http://ja.wikipedia.org/wiki/%E9%87%91%E5%87%BA%E6%AD%A6%E9%9B%84>

コンピュータビジョンの難しさ

金出武雄. 知能ロボットの技術：人工知能からのアプローチ（前編）：4. ロボット視覚. 情報処理, Vol.44, No.11, pp.1130--1137, 2003.

人間はほとんど意識なしにできるよう видる所以、その難しさ自体が認識されない面がある。

- 電気工学者
 - 「それは、データが多いからだ。NTSCというような低解像度のビデオでも、1秒間に20メガバイトのデータを作り出す。したがって、きわめて高速の計算機でなければ、簡単な画像処理すらできない。だからビジョンは難しい。」
- 幾何学的発想の人
 - 「ビジョンの難しいのは幾何学的に縮退しているからだ。」
- 物理学者
 - 「いや、それもそうだが、画像の測定値というものはいろいろな情報が輻輳してきた物である。」
- 人工知能研究者
 - 「それは、要するに「知識」というものを使わなければならないからだ。」

初期のビジョン研究では人が使っている「知識」に最も注目したのは自然であった。

初期の人工知能的アプローチ

金出武雄. 知能ロボットの技術：人工知能からのアプローチ（前編）：4.ロボット視覚.
情報処理, Vol.44, No.11, pp.1130–1137, 2003.



- 1970年～1980年中頃
 - Let's-program-what-I-think-I-am-doing approach
 - 自分がしていると思われることをプログラムしてみようアプローチ
 - 「緑色の明るい領域」
 - 幾何学的, 工学物理的な曖昧さが除かれた後の話
 - 「道路の上にあるから車である」
 - 物体に対する領域が正しく取り出されてから初めて意味のあるものである.
 - 信号とシンボルの不一致！
- 1980年～
 - 人工知能的・発見的プログラミングによる一般ビジョンシステムの研究は陰を潜め, ビジョン研究はシグナル的なもの特に幾何学的・物理学的・工学的な焦点を当てることで大きな進歩を遂げた.

2003年までの画像認識研究

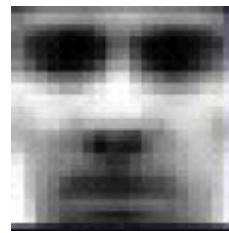
金出武雄. 知能ロボットの技術：人工知能からのアプローチ（前編）：4.ロボット視覚.
情報処理, Vol.44, No.11, pp.1130–1137, 2003.

- この20年間ビジョン研究は主に、形再現の問題、色・テクスチャ・動きといったビジョンの物理的側面を扱う個々のモジュールとその応用システムに大きな成果をあげたものの、「認識」、特に一般のシーンの認識という物理信号とシンボルの世界の融合問題を避けてきた。認識としてされたもののはほとんどが顔といった特定の物体の分類システムである。
- 長い間、認識、理解、知識といった面から遠ざかっていたロボットビジョンの研究が次の段階に進むには、新しい道具が一部にしろ手にある今、そういう研究にもう一度取り組み始める必要があるのではなかろうか。

顔検出認識のチュートリアル

- ICPR2004
 - Recent Advances in Face Detection
 - Ming Hsuan Yang
 - myang@honda-ri.com
 - <http://www.honda-ri.com>
 - http://vision.ai.uiuc.edu/mhyang
 - Honda Research Institute
 - Mountain View, California, USA

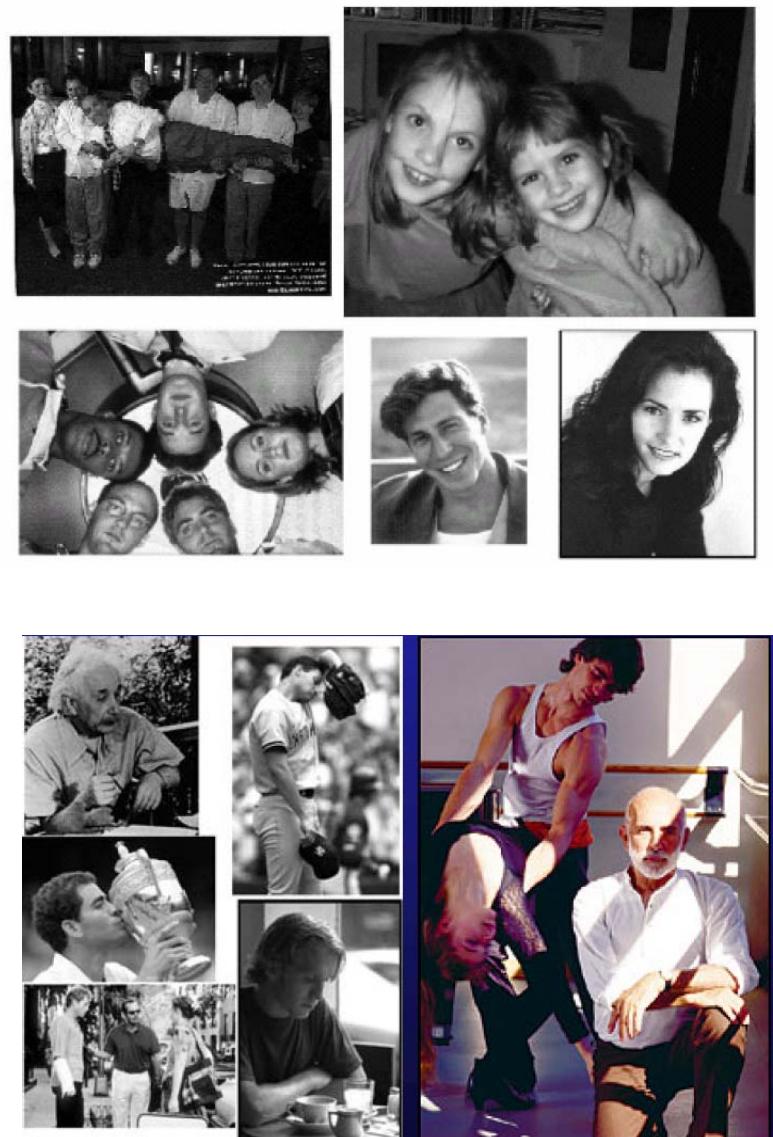
顔検出の問題



- 19x19の顔画像のサムネールを考える
- グレースケールだと $256^{361} = 2^{2888}$ 通りの組み合わせ
- 世界の人口：約 2^{32} 人
- 小さなサムネールでもとんでもなく高次元の問題となる。

顔認識の難しさ

- 顔の向き
 - 正面, 下向き, 横向き, 見上げる
- 顔の要素の有無
 - 髪の毛, ひげ, めがね, マスク
- 顔の表情
 - 喜怒哀楽
- オクルージョン
 - なんらかの物体の陰になっているなど
- 傾き
 - カメラの姿勢によって, 写っている顔の向きが変化する
- 撮影条件
 - 照明条件, カメラ特性, 解像度



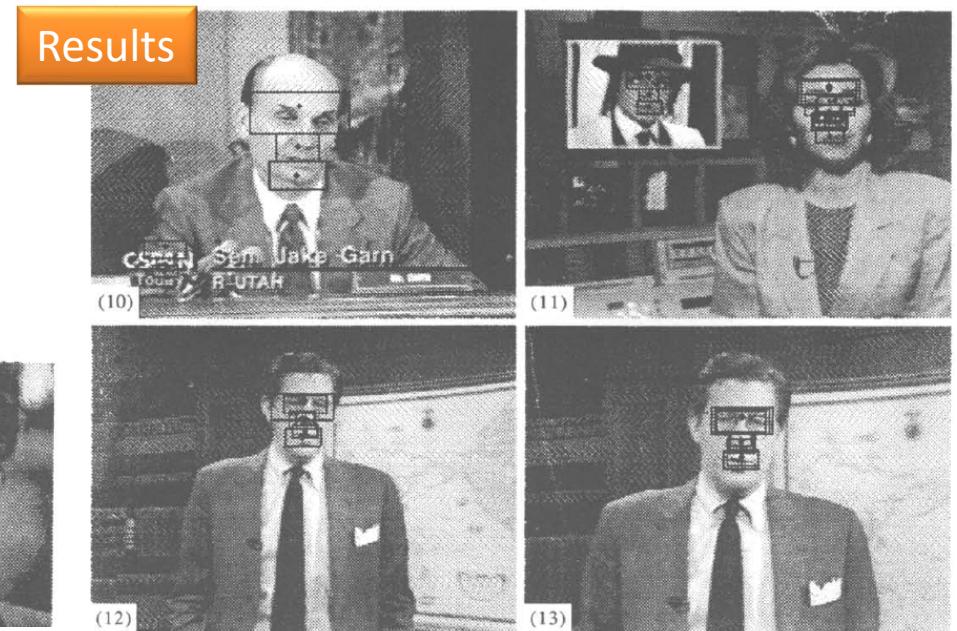
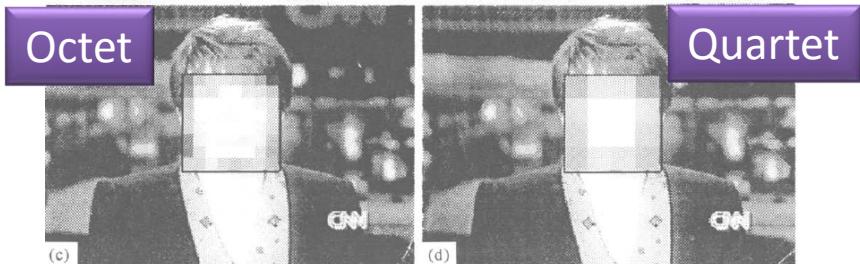
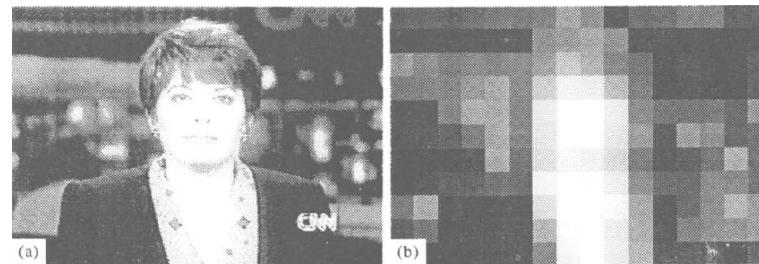
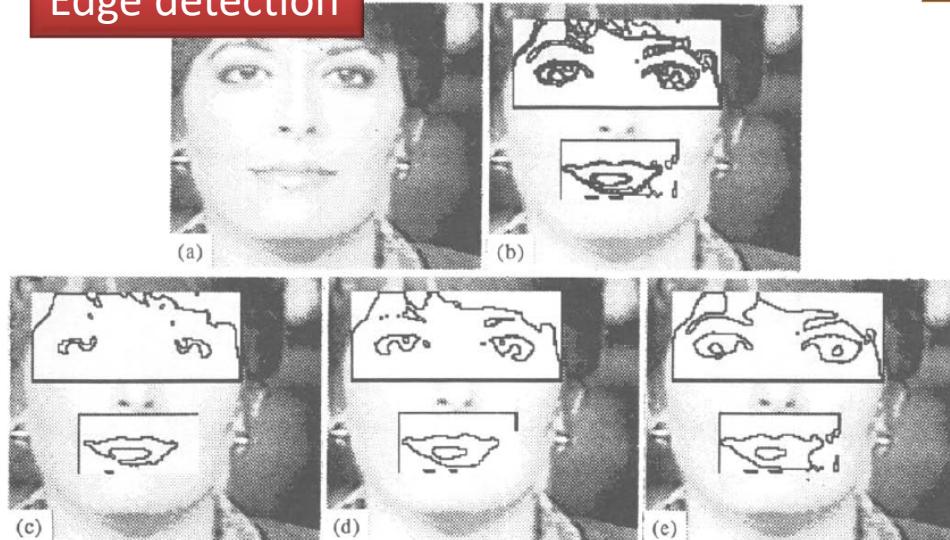
顔検出の仕組みの例

- 知識ベースの手法
 - 典型的な顔を構成する要素は何かを人が考えて表現する
- 特徴量ベースの手法
 - 姿勢, 見え方, 照明条件が変化したとしてもなるべく変化しない顔の構造的特徴を発見する
- テンプレートベースの手法
 - 顔や顔の要素を表現する典型的なパターンを蓄えておき, それらと比較する
- アピアランスベースの手法
 - 顔の代表的な変動をとらえるモデルを訓練データセットから学習する

知識ベースの手法1

- 人が設定したルールに従い“顔”とは何かをコーディングする手法
- G. Yang and T. Huang. Human Face Detection In A Complex Background. Pattern Recognition, Vol. 27, No. 1, pp.53-63, 1994.
- アルゴリズム
 - 多重解像度の画像、4セルの領域が同じ輝度を顔の候補領域とする。
 - 領域内の輝度を正規化する。
 - エッジ抽出し、まゆ毛や口を探す。

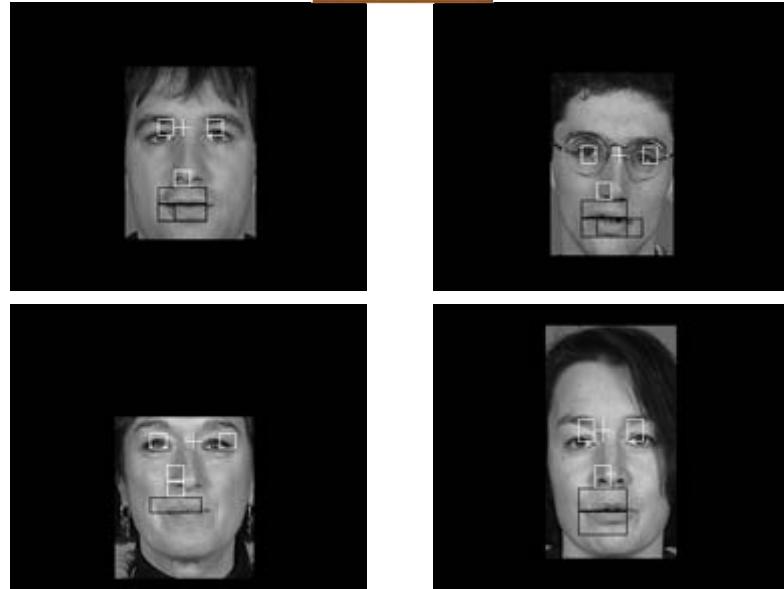
Edge detection



知識ベースの手法2

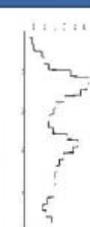
Results

- 人が設定したルールに従い“顔”とは何かをコーディングする手法
- Constantine Kotropoulos, and Ioannis Pitas. Rule-Based Face Detection in Frontal Views. ICASSP, 1997.
- アルゴリズム
 - 水平軸, 垂直軸に投影して顔候補を探す
 - 眉毛や鼻を探して, 検証する



複雑背景や複数人では
うまくいかないよね。

顔の境界は谷になるはず, , ,



(a)

(b)

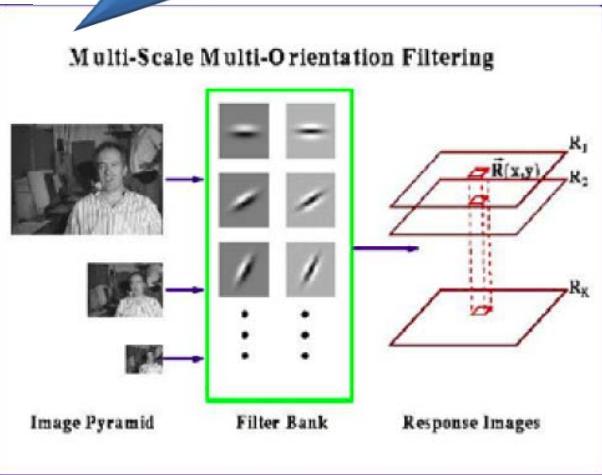
(c)

Fig. 3. (a) and (b) $n = 8$. (c) $n = 4$. Horizontal and vertical profiles. It is feasible to detect a single face by searching for the peaks in horizontal and vertical profiles. However, the same method has difficulty detecting faces in complex backgrounds or multiple faces as shown in (b) and (c).

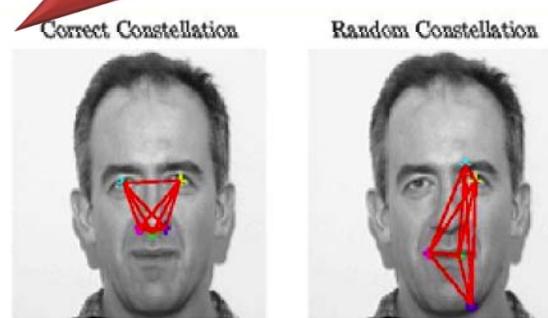
特徴量ベース手法1

- 初めに顔の部分的な（不变）特徴量を発見し、その後に全体の構成から検証する
- T.K. Leung, M.C. Burl, and P. Perona. Finding faces in cluttered scenes using random labeled graph matching. ICCV, 1995.

顔の特徴部位のテンプレートと入力画像との類似度を計る



特徴点をノードとするグラフを作成し、顔らしいグラフを選択



- Facial features must occur in a specific spatial arrangement.
- Form constellations from the candidate feature locations.
- Find constellation that have the appropriate structure.

複数の向きとスケールのガウスフィルタの出力平均

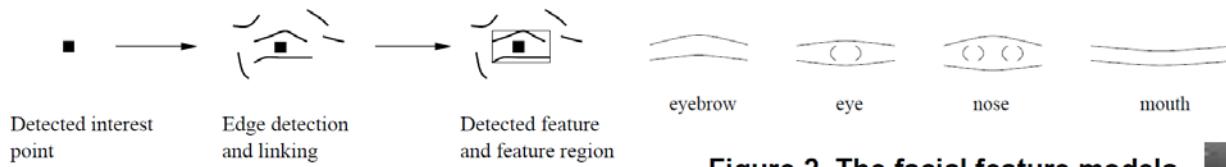
2つの特徴点が分かれれば他の特徴点が推定できる



Given two feature positions, locations and the amount of uncertainty of the others can be estimated.

特徴量ベース手法2

- 初めに顔の部分的な（不变）特徴量を発見し、その後に全体の構成から検証する
- K. C. Yow, and R. Cipolla. Detection of Human Faces under Scale, Orientation and Viewpoint Variations. In FG, 1996.



Results

Figure 2. The facial feature models.

Figure 3. Preattentive feature select



特徴点を抽出し、そのまわりのエッジを検出

モデルを当てはめ、
パーツを検出

顔モデルを当てはめ、
顔候補を選出

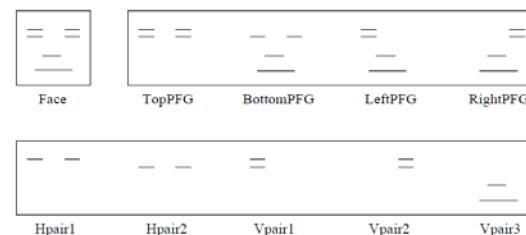


Figure 1. The face model and the face groups.

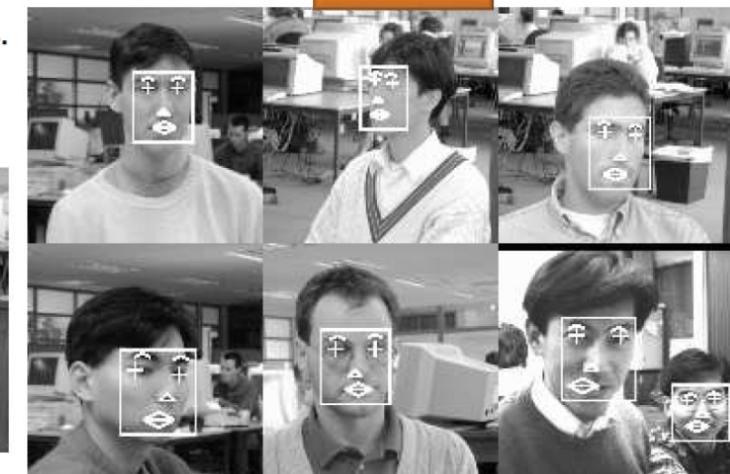


Figure 6. Reinforcement belief network.

テンプレートベースの手法

- 顔のテンプレートを持ち、入力画像との相関を計算する
- Ration Template
 - Pawan Sinha. Perceiving and recognizing three-dimensional forms. Doctoral Thesis, MIT, 1995.
 - 目の明るさは周囲に比較して暗いなど。
- A. Lanitis, C.J. Taylor, T.F. Cootes, and T. Ahmed. Automatic Interpretation of Human Faces and Hand Gestures Using Flexible Models. In International Workshop on Automatic Face- and Gesture-Recognition, 1995.

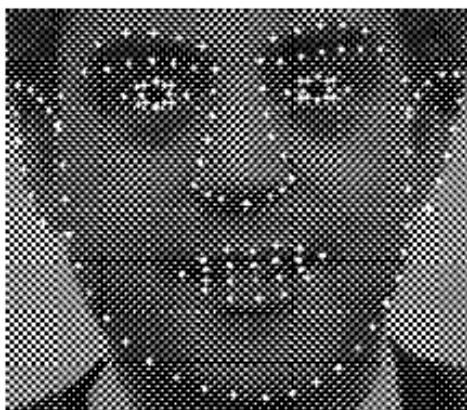


Fig. 1: Shape model points

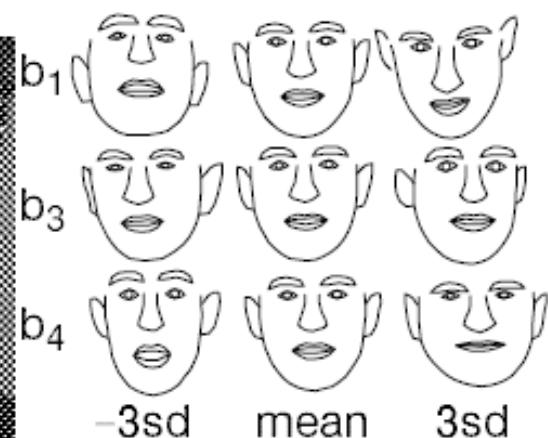
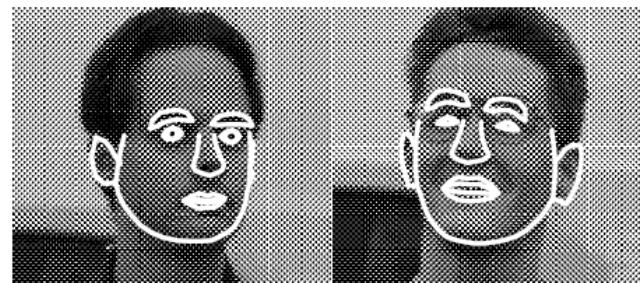
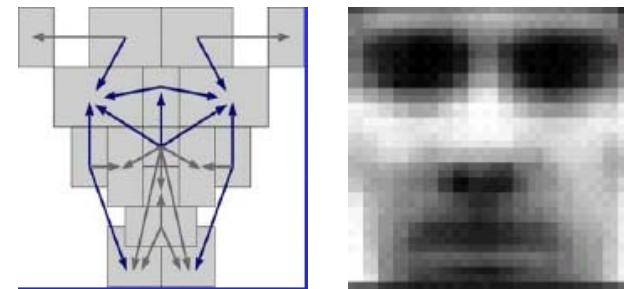


Fig. 2: The main modes of shape variation



Turning = -20 degrees
Nodding = 5 degrees
Turning = 20 degrees
Nodding = 10 degrees
Fig. 9: Pose recovery examples

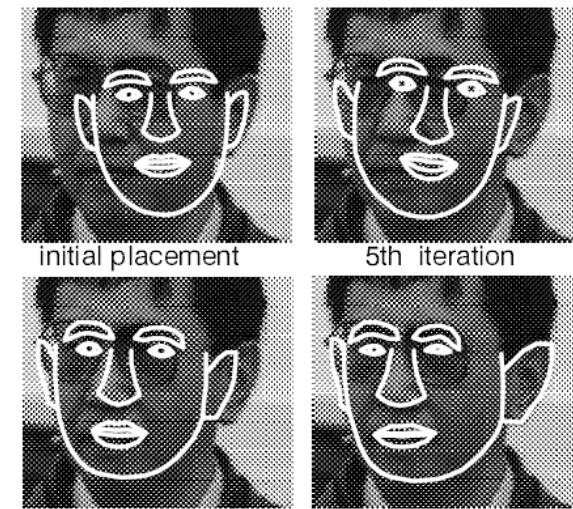
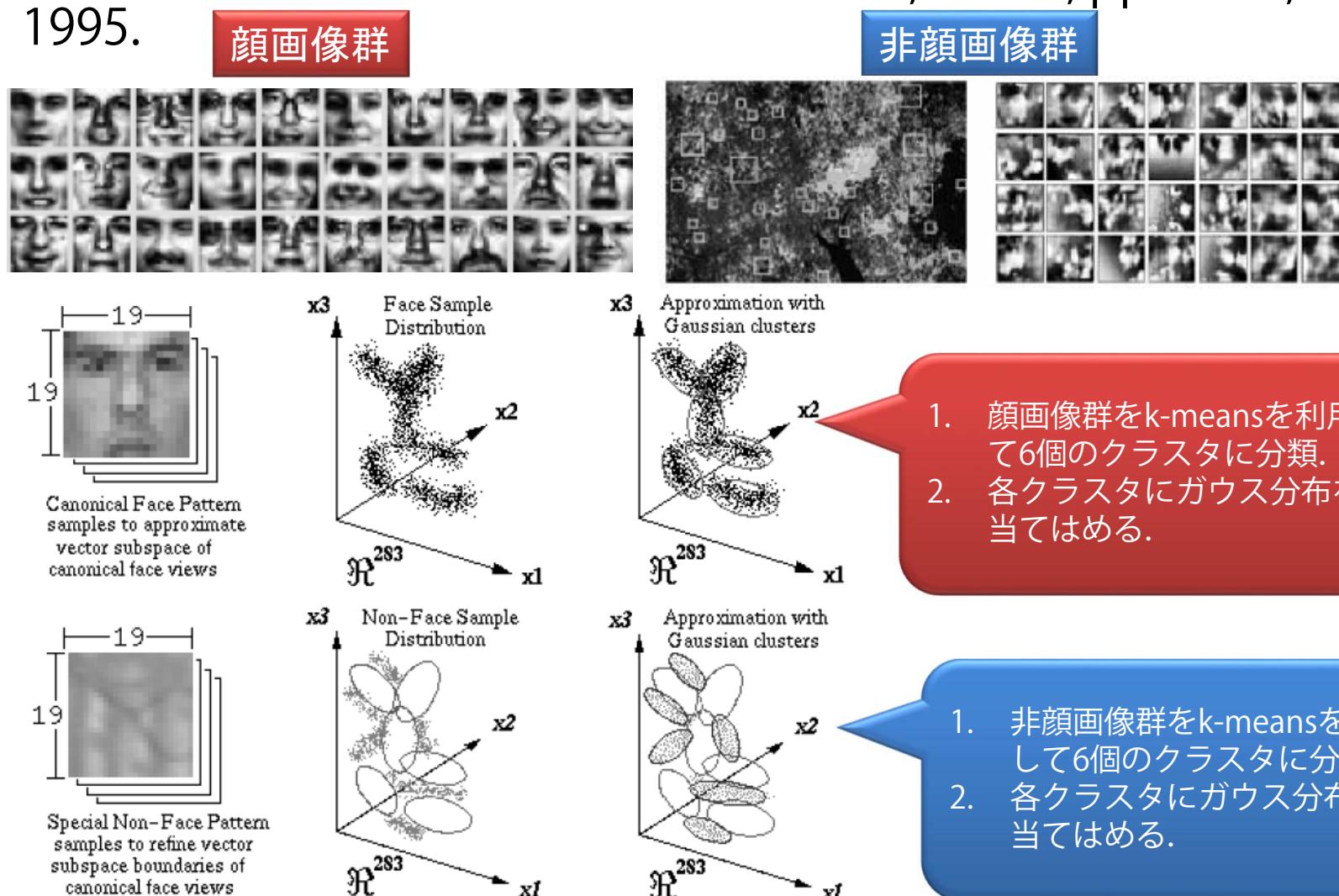


Fig. 6: Fitting the shape model

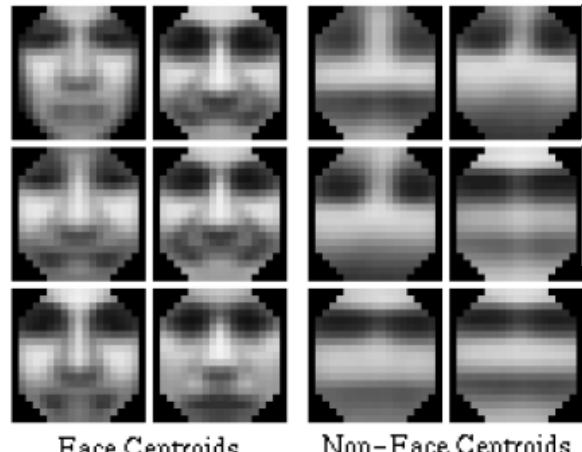
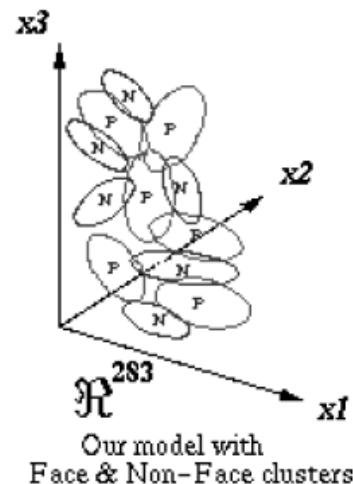
アピアランスベースの手法1

- 顔画像群（非顔画像群）を用いて、顔画像識別器を作成する
- K. K. Sung, and T. Poggio. Example Based Learning for View-Based Human Face Detection. IEEE PAMI, Vol.20, pp.39-51, 1995.



アピアランスベースの手法1

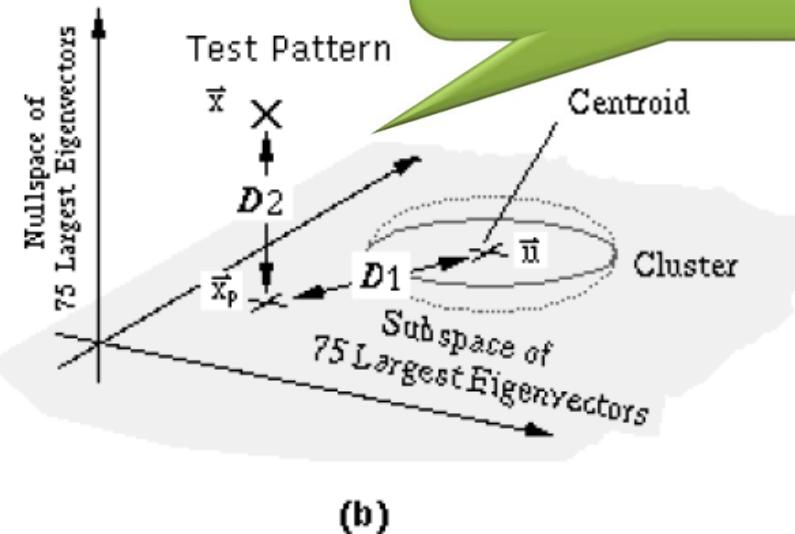
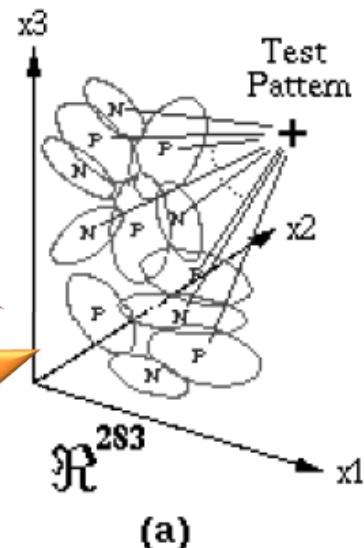
- 顔画像群（非顔画像群）を用いて、顔画像識別器を作成する
- K. K. Sung, and T. Poggio. Example Based Learning for View-Based Human Face Detection. IEEE PAMI, Vol.20, pp.39-51, 1995.



顔画像群が6クラスタ、
非顔画像群が6クラスタ
で、合計12クラスタ
ができる。

テスト画像と
12クラスタと
の距離を測る。

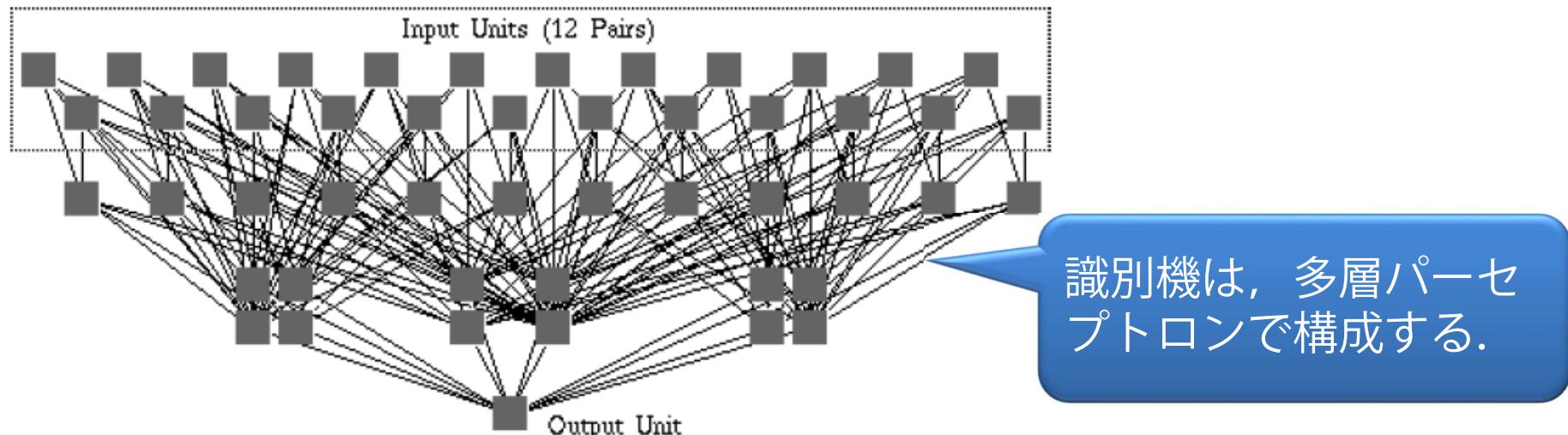
各クラスタについて2
種の距離を測るため、
12x2=24次元の特徴ベ
クトルとなる。



各クラスタはPCAで次元削減され
る。計量は部分空間への最短距
離 (D_2) と部分空間におけるマ
ハラノビス距離 (D_1)。

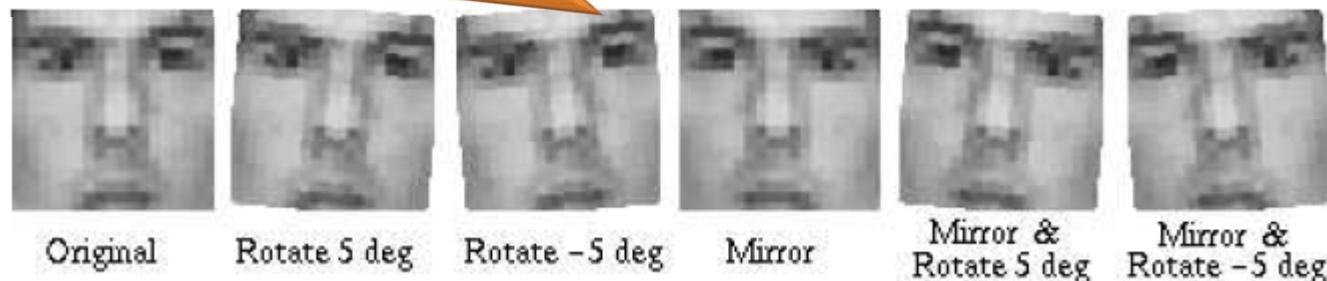
アピアランスベースの手法1

- 顔画像群（非顔画像群）を用いて、顔画像識別器を作成する
- K. K. Sung, and T. Poggio. Example Based Learning for View-Based Human Face Detection. IEEE PAMI, Vol.20, pp.39-51, 1995.



識別機は、多層パーセプトロンで構成する。

顔サンプルに回転、反転の変換を加えることで、人工的にサンプルを増やす。



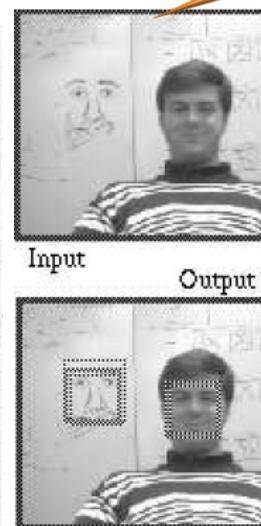
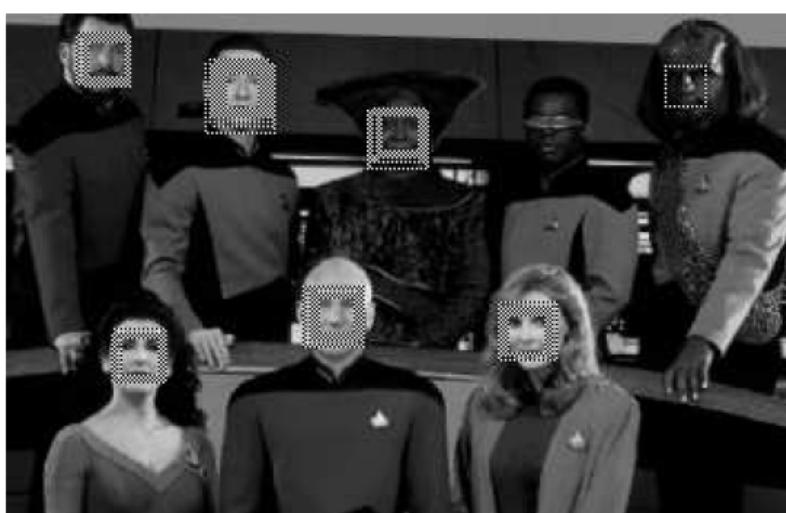
アピアランスベースの手法1

- 顔画像群（非顔画像群）を用いて、顔画像識別器を作成する
- K. K. Sung, and T. Poggio. Example Based Learning for View-Based Human Face Detection. IEEE PAMI, Vol.20, pp.39-51, 1995.

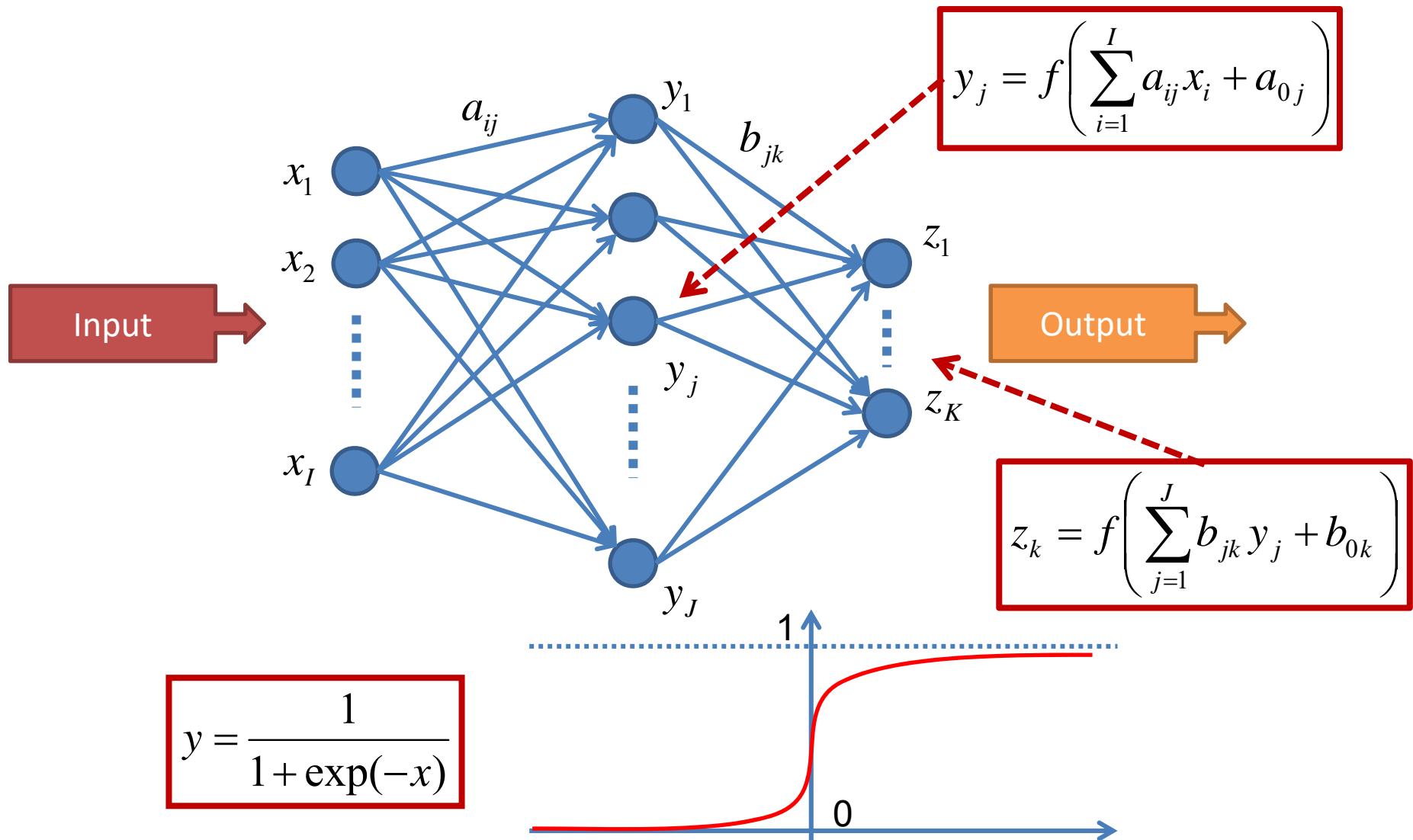


識別器で誤ったサンプル群を、非顔画像群として加えることで、識別性能をブーストさせる。

実験結果



多層パーセプトロン



- ロジスティック回帰モデルを多層化したもの
- 十分な数の隠れユニットがあればどんな連続関数も任意の精度で近似可能
- 誤差逆伝播により効率的に学習可能

誤差逆伝播学習

- 2乗誤差

$$\varepsilon^2 = \sum_{p=1}^P \| t_p - z_p \|^2$$

- 二乗誤差の結合加重に関する偏微分

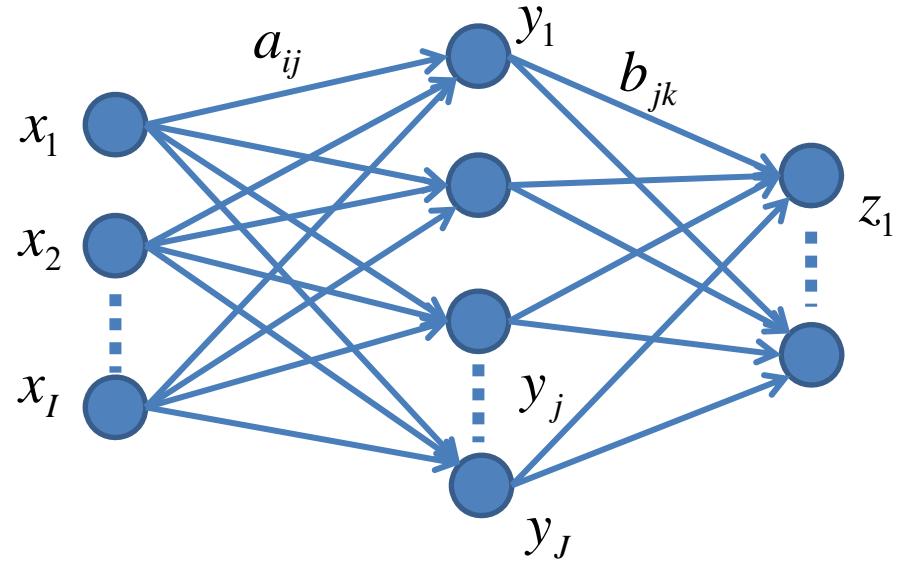
$$\frac{\partial \varepsilon^2}{\partial a_{ij}} = \sum_{p=1}^P \left(\left(-2 \sum_{k=1}^K (t_{pk} - z_{pk}) b_{jk} \right) y_{pj} (1 - y_{pj}) x_{pi} \right)$$

$$\frac{\partial \varepsilon^2}{\partial b_{jk}} = \sum_{p=1}^P (-2(t_{pk} - z_{pk}) y_{pj})$$

- 最急降下法による結合加重の更新

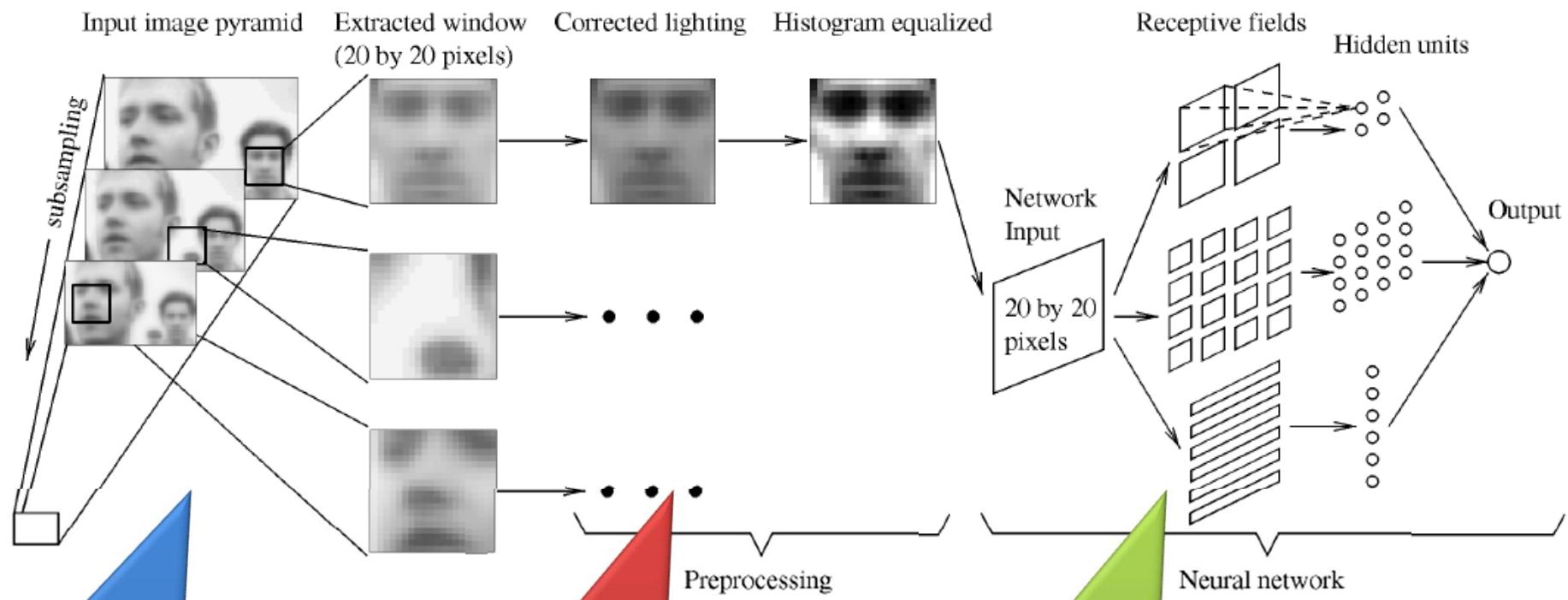
$$a_{ij} \leftarrow a_{ij} - \alpha \frac{\partial \varepsilon^2}{\partial a_{ij}}$$

$$b_{jk} \leftarrow b_{jk} - \alpha \frac{\partial \varepsilon^2}{\partial b_{jk}}$$



アピアランスベースの手法2

- 顔画像群（非顔画像群）を用いて、顔画像識別器を作成する
- H. A. Rowley, S. Baluja, and T. Kanade. Neural Network-Based Face Detection. IEEE PAMI, Vol.20, No.1, pp.23-38, 1998.



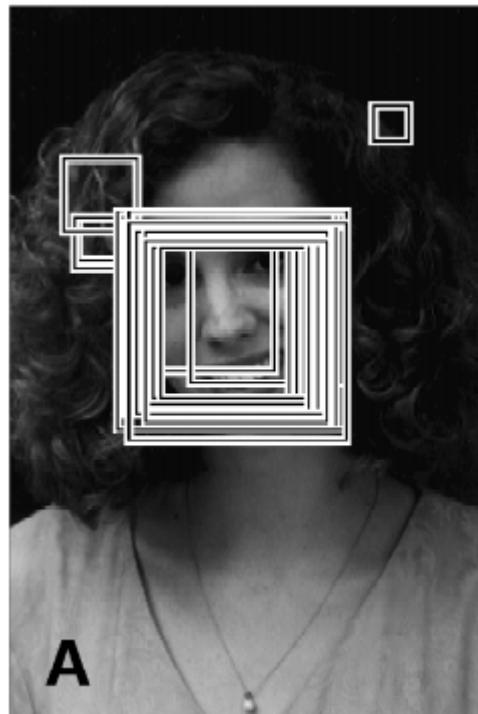
複数のスケールに対応させるために画像ピラミッドを構成。

照明条件の補正、輝度のヒストグラムを正規化する。

多層パーセプトロンを用いて顔識別器を構成。

アピアランスベースの手法2

- 顔画像群（非顔画像群）を用いて、顔画像識別器を作成する
- H. A. Rowley, S. Baluja, and T. Kanade. Neural Network-Based Face Detection. IEEE PAMI, Vol.20, No.1, pp.23-38, 1998.



シングルネットワーク結果.

【問題】

- 正しい顔の位置に重複して検出される。
- 誤って検出されることがある。

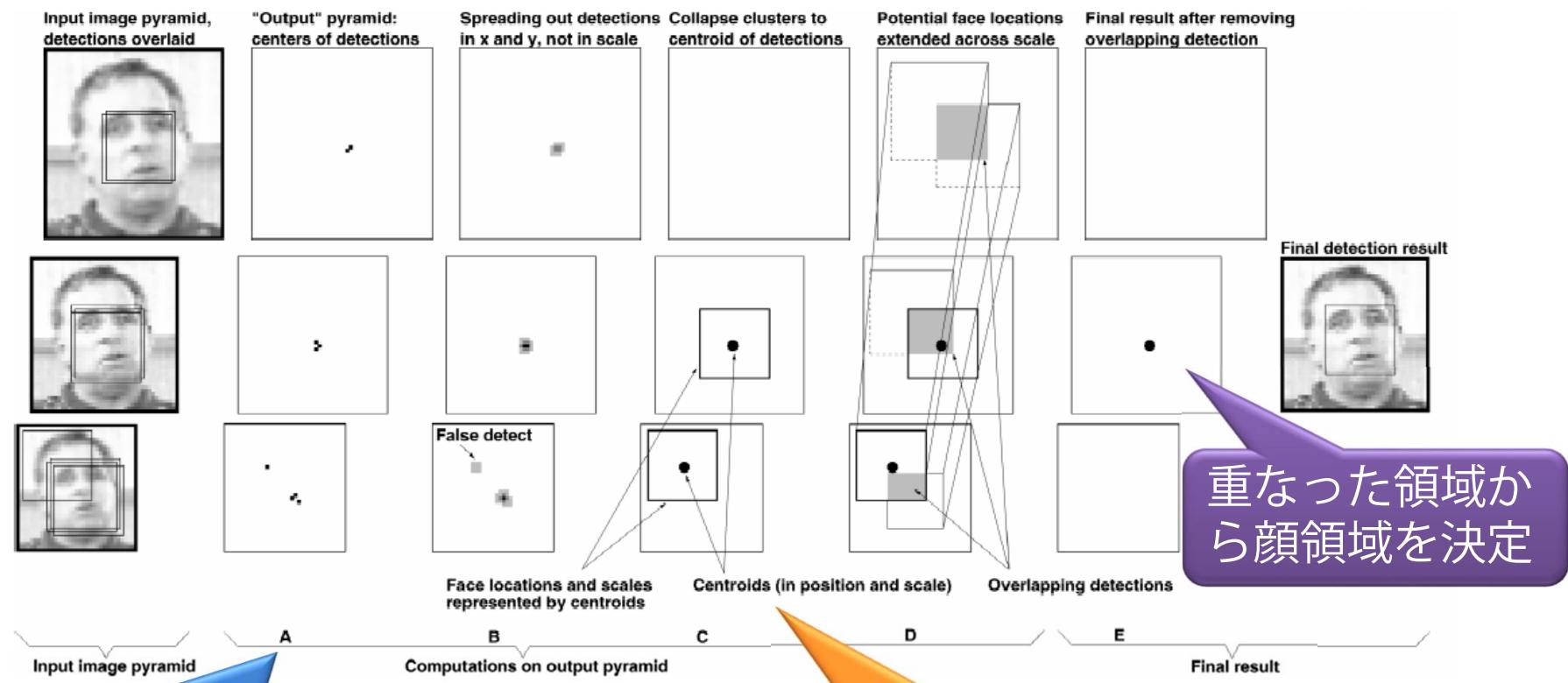


【解決策】

- ヒューリスティクスの活用。
- 複数ネットワークの利用。

アピアランスベースの手法2

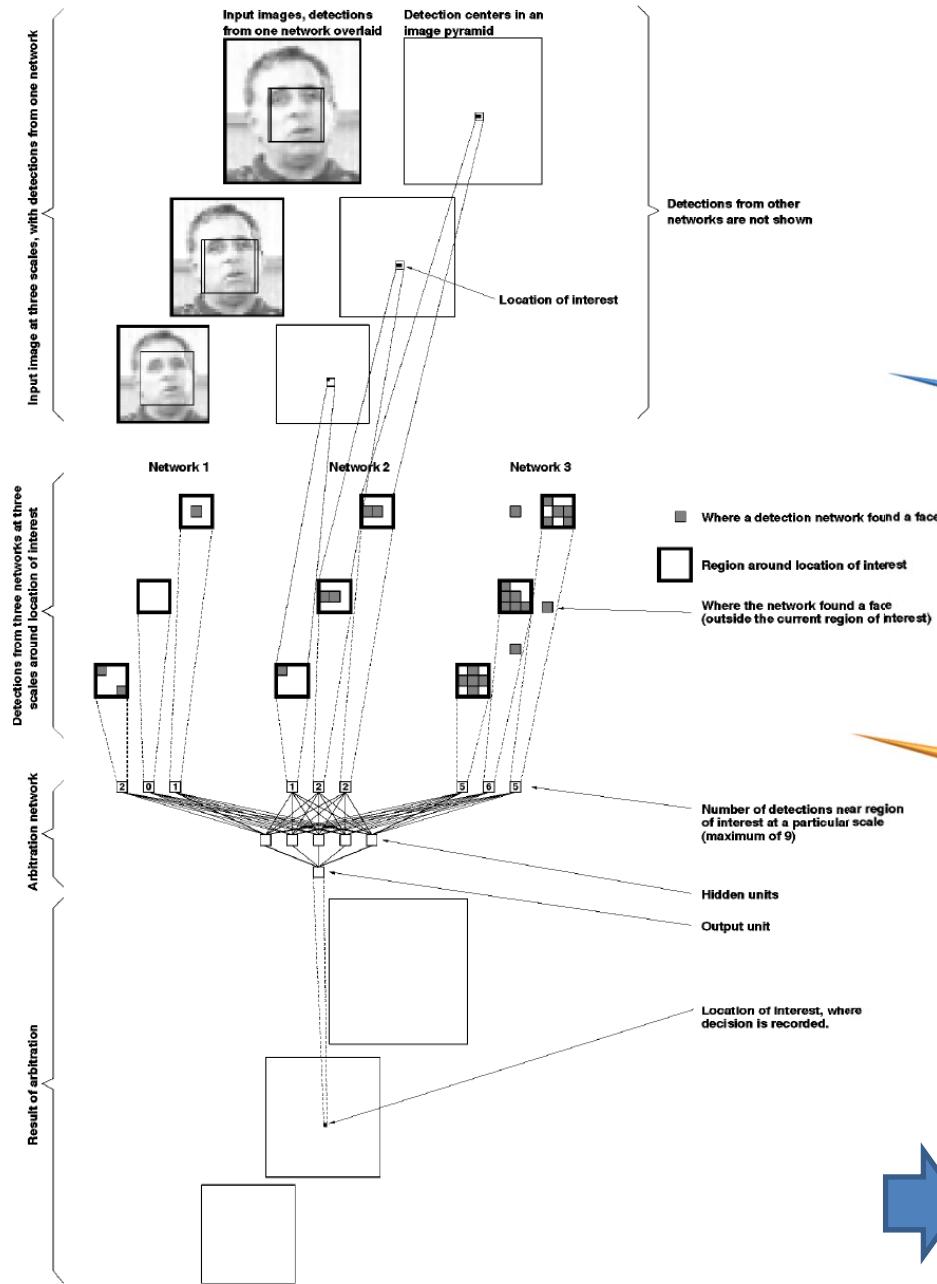
- 顔画像群（非顔画像群）を用いて、顔画像識別器を作成する
- H. A. Rowley, S. Baluja, and T. Kanade. Neural Network-Based Face Detection. IEEE PAMI, Vol.20, No.1, pp.23-38, 1998.



「顔部分は重複して検出され、非顔部分はあまり重複して検出されない。」といったヒューリスティクスを活用。

重複して検出した回数をカウントし、閾値以上ならその重心を計算。

アピアランスベースの手法2



- 顔画像群（非顔画像群）を用いて、顔画像識別器を作成する
- H. A. Rowley, S. Baluja, and T. Kanade. Neural Network-Based Face Detection. IEEE PAMI, Vol.20, No.1, pp.23-38, 1998.

【複数のネットワークを学習】

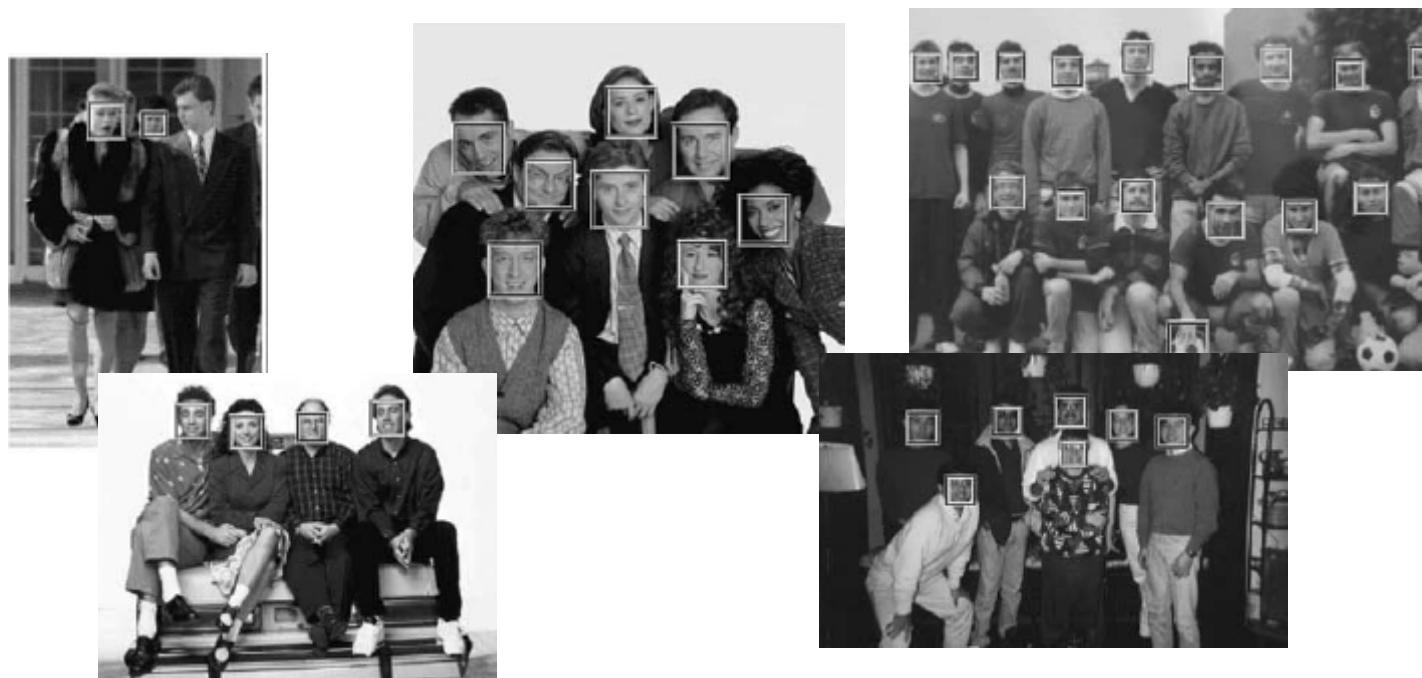
- 1) ランダム初期重み
- 2) ランダム非顔画像群
- 3) 非顔画像の提示順序入れ替え

複数ネットワークの出力結果を多層パーセプトロンで統合。

性能は、ヒューリスティクスの活用、
複数ネットワーク利用もほぼ同じ。

Viola Jones Face Detector

- リアルタイム物体検出手法
- 訓練は遅いが検出は非常に高速
- 主要なアイデア
 - 高速な特徴評価のための積分画像
 - 特徴抽出のためのBoosting
 - 非顔領域を高速に排除するためのAttentional cascade

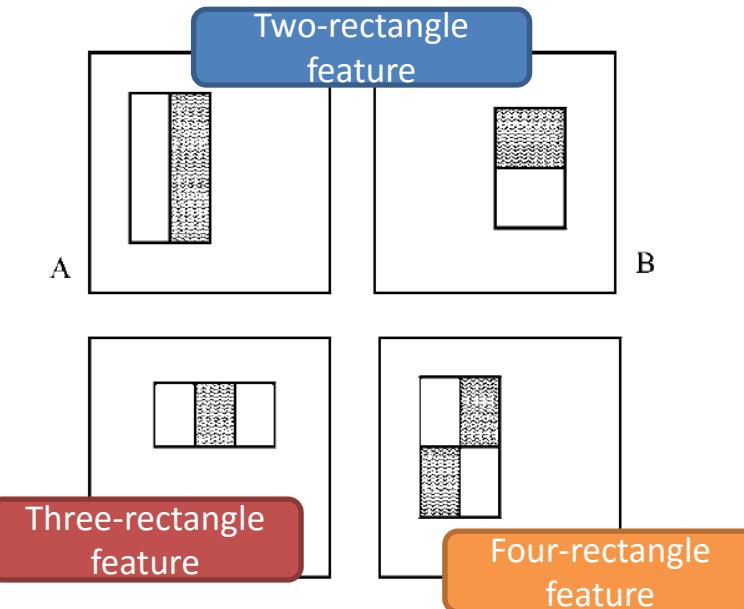


Viola and Jones. Robust Real-Time Face Detection. International Journal of Computer Vision 57(2), 137–154, 2004.

画像特徴

Rectangle Features

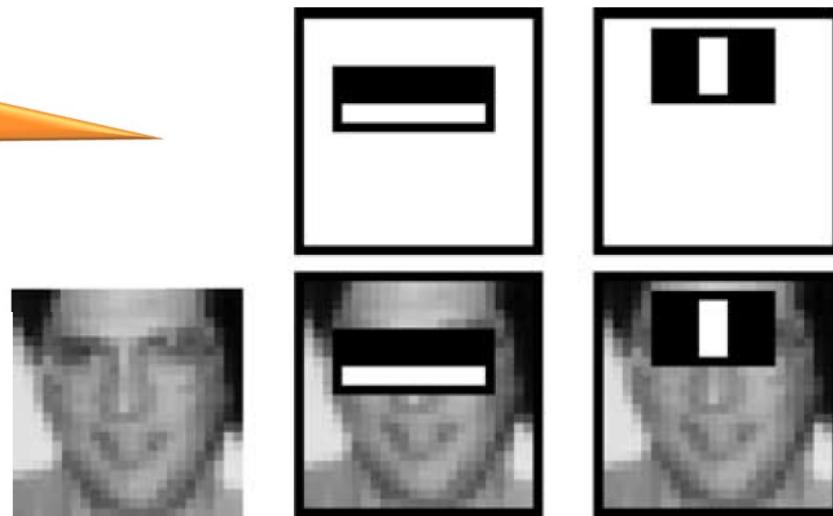
- Haar基底関数に類似
- 3種類の特徴
 - Two rectangle feature
 - Three rectangle feature
 - Four rectangle feature



特徴量の値 = \sum (白い領域のピクセルの値) - \sum (黒い領域のピクセルの値)

ちょうど目の位置と重なれば高い値が出力されると期待できる

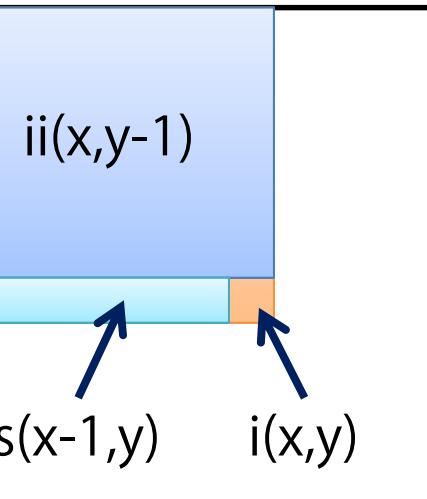
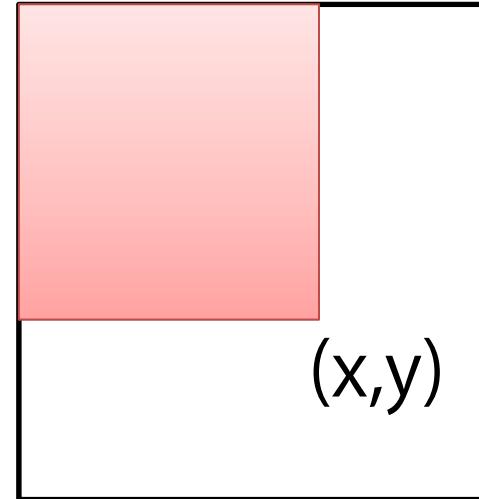
→ うまいRectangle featureは弱い顔識別機と理解できる。



積分画像

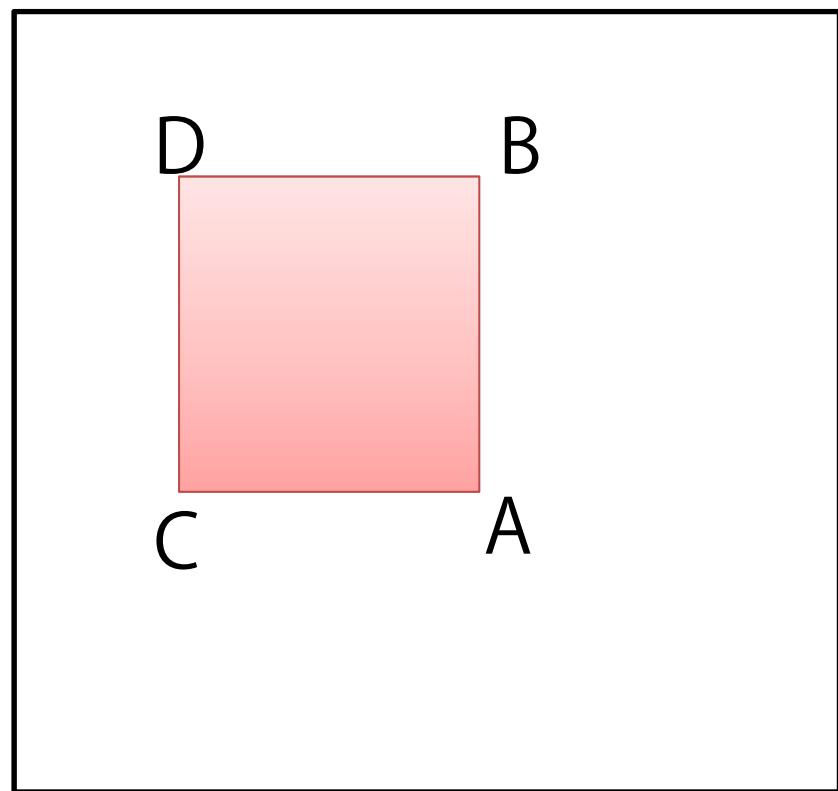
- 積分画像 (integral image) は原点(0,0)と(x,y)を対角の頂点とする長方形で囲まれる領域内のピクセル値の和
- 一度のパスで高速計算可能
- 行のピクセル値の和
 - $s(x, y) = s(x-1, y) + i(x, y)$
- 積分画像
 - $ii(x, y) = ii(x, y-1) + s(x, y)$

(0,0)



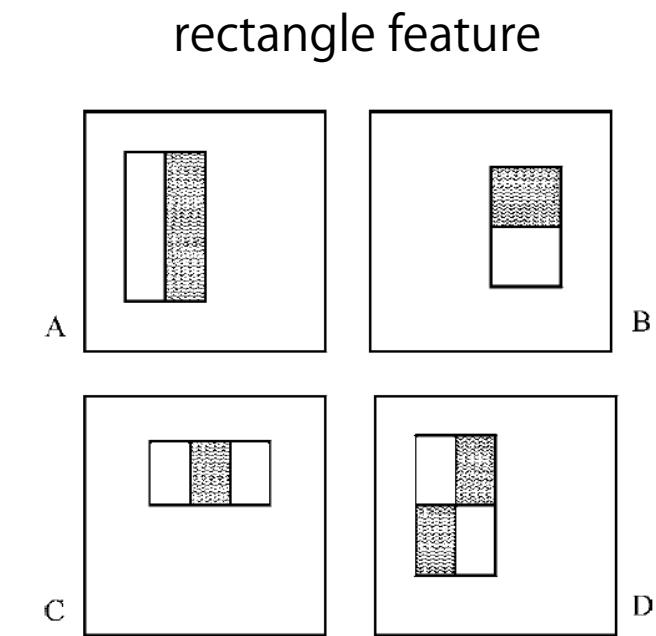
長方形領域内の和の計算

- A, B, C, Dをその位置における積分画像の値とする.
- 長方形領域内の画素値の合計
 - $A - B - C + D$
- 3加算だけで任意の長方形領域内の画素値の和を計算可能

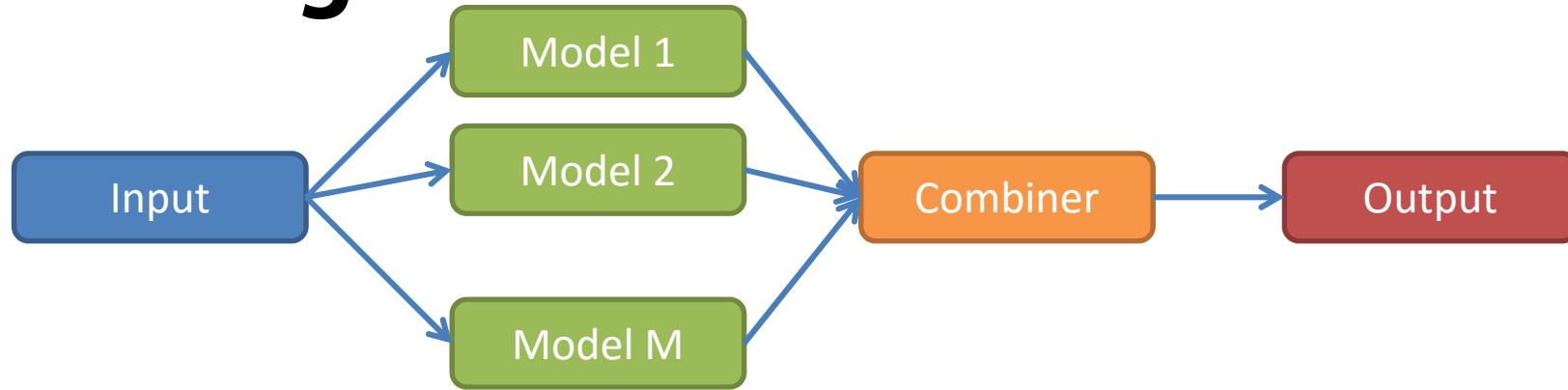


特徴選択

- 24×24の検出領域において、可能なrectangle featureの数は約160,000
- 検出時に全てのrectangle featureを評価するのは非現実的
- 全てのfeatureを利用せず、その一部のみを利用して良い識別器を構築するはどうすればよいか?
 - Boostingの利用



Boosting



- Boostingとは
 - 複数の基本となる学習機械を組み合わせることで、精度の高い学習機械を構成する手法.
 - 以前に学習した学習機械の結果を用いて間違いの多い訓練サンプルに重み付けを行い、この重み付けされたサンプルで新たに学習機械を学習する.
 - 逐次的に生成される学習機械を統合して一つの学習機械とする.
- AdaBoostにおける訓練
 1. 各訓練サンプルに等価な重みを与える
 2. For $m=1:M$
 - A) 訓練エラーを最小とする弱学習器を選択する
 - B) 現在の弱学習器によって誤識別された訓練サンプルの重みを増加
 3. 全ての弱学習器を重み付け線形結合し、最終的な識別器を構成する
 - 各弱学習器の重み付けは、識別精度に比例

AdaBoostのアルゴリズム

- 各訓練サンプルに等価な重みを与える

$$\left\{ w_n^{(1)} = \frac{1}{N} \right\}_{n=1}^N$$

- For $m=1:M$

- 訓練エラーを最小とする弱学習器を選択する

$$J_m = \sum_{n=1}^N w_n^{(m)} I(h_m(x_n) \neq y_n)$$

$$I(h_m(x_n) \neq y_n) = \begin{cases} 1 & \text{if } h_m(x_n) \neq y_n \\ 0 & \text{otherwise} \end{cases}$$

- 現在の弱学習器によって誤識別された訓練サンプルの重みを増加

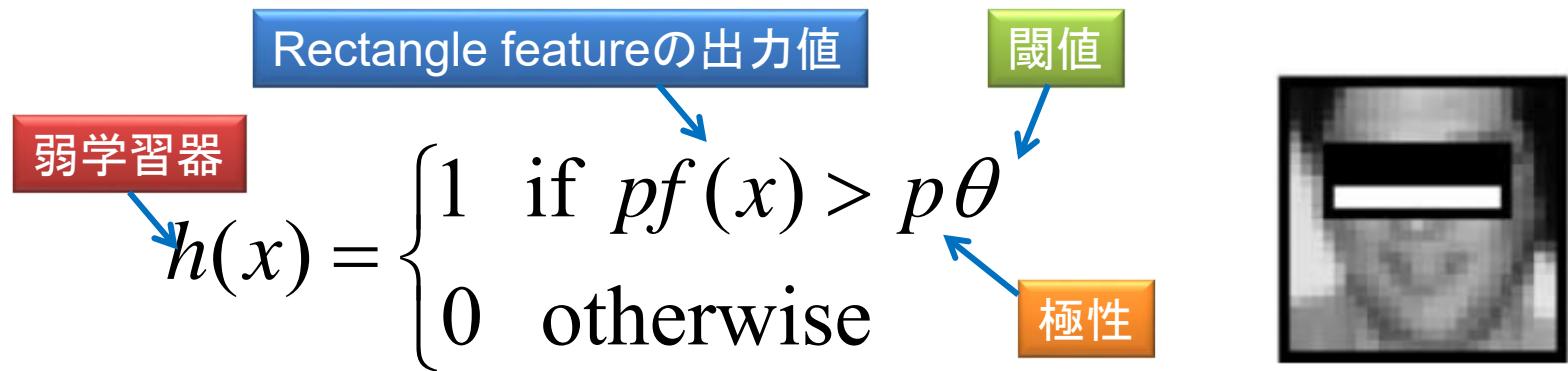
$$\begin{aligned} \epsilon_m &= \frac{\sum_{n=1}^N w_n^{(m)} I(h_m(x_n) \neq y_n)}{\sum_{n=1}^N w_n^{(m)}} & w_n^{(m+1)} &= w_n^{(m)} \exp\{\alpha_m I(h_m(x_n) \neq y_n)\} \\ \alpha_m &= \ln \left(\frac{1 - \epsilon_m}{\epsilon_m} \right) \end{aligned}$$

- 全ての弱学習器を重み付け線形結合し、最終的な識別器を構成する
 - 各弱学習器の重み付けは、識別精度に比例

$$H(x) = \operatorname{sign} \left(\sum_{m=1}^M \alpha_m h_m(x) \right)$$

顔検出におけるBoosting

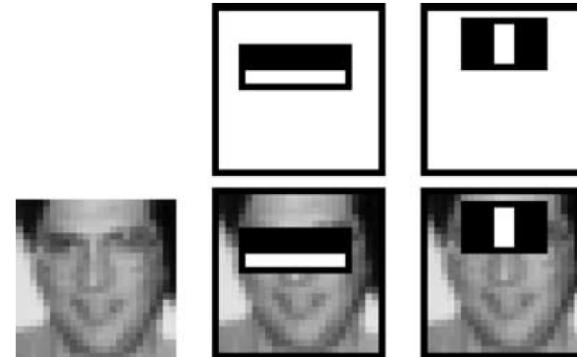
- Rectangle featureに基づく弱学習器の定義
 - 弱学習器とrectangle featureが1対1対応



- For $m=1:M$
 - 各訓練サンプルに、弱学習器を適用
 - 各弱学習器に最良の閾値を選択
 - 最良の弱学習器と閾値の組合せを選択
 - 訓練サンプルに再重みづけ

Boostingによって選択された特徴

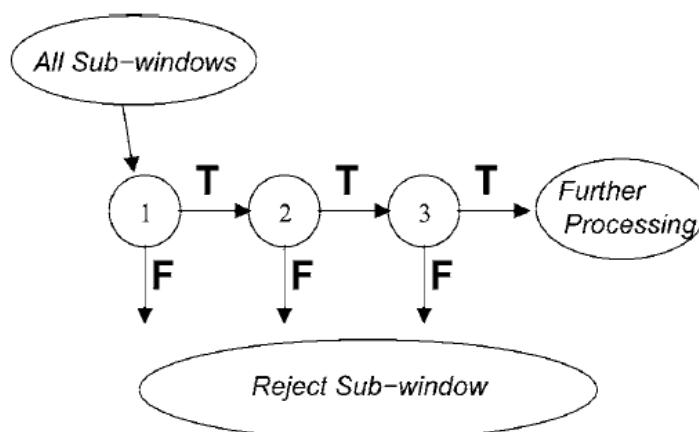
- Boostingによって選択された1番目, 2番目のrectangle feature
 - 100%の検出率
 - 50%のfalse positive



- 200個のfeatureを用いることで95%の検出率, 1/14,084のfalse positive
- 性能が不十分
 - False positiveを1/1,000,000未満としたい
 - Featureを増やせばfalse positiveを低下させられるが計算時間がかかる

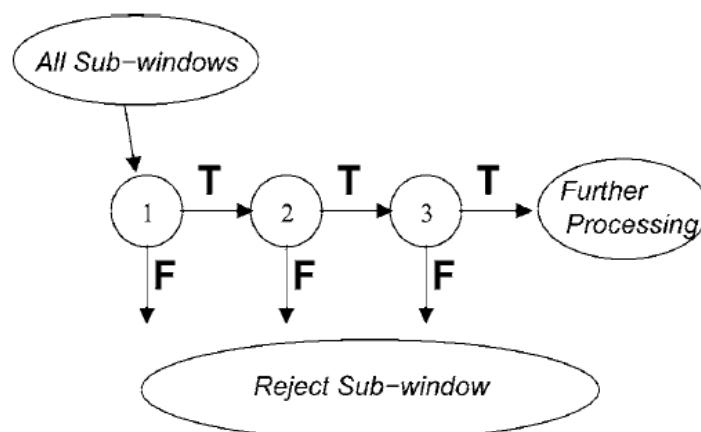
Attentional cascade

- ほとんどのサブウィンドウが非顔画像であるので、非顔画像を素早く判定することが高速化の鍵
- ほとんど全ての顔画像を検出し、多くの非顔画像を拒絶する簡単な識別器から開始
 - 顔画像は必ず顔と判断するが、顔でない画像も顔として判断される可能性がある
- 第1識別器で顔と判断されたデータは、第2識別器に入力される
 - 第2識別器は第1識別器よりも複雑で顔でない画像を顔と判断してしまう率 (false positive) が低い
- 各識別器で顔でないと判断された場合、即座に顔でないと決断



Attentional cascadeの学習

- 各ステージに目標検出率と目標false positive率を設定
- 目標検出率、目標false positive率が達成されるまで現在のステージに特徴を追加
 - AdaBoost閾値を下げる必要性
- 総合的なfalse positive率が十分低くない場合、別のステージを追加
- 次のステージのnegative訓練サンプルとして、現在のステージでfalse positiveと判定されたデータを利用



訓練データと学習

- 訓練データ
 - 5000枚顔画像
 - 顔画像は正規化
 - スケール, 移動
 - 顔画像には多様性
 - 複数人, 照明, 姿勢
- 数週間の訓練時間
 - 466 MHz Sun workstation
- 38 レイヤー, 6061 features
- 検出時間 : 15Hz
 - 700 Mhz Pentium III
 - 384 x 288 pixel



Viola Jones 顔検出まとめ

- Rectangle features
- 高速な特徴評価のための積分画像
- 特徴抽出のためのBoosting
- 非顔領域を高速に排除するためのAttentional cascade

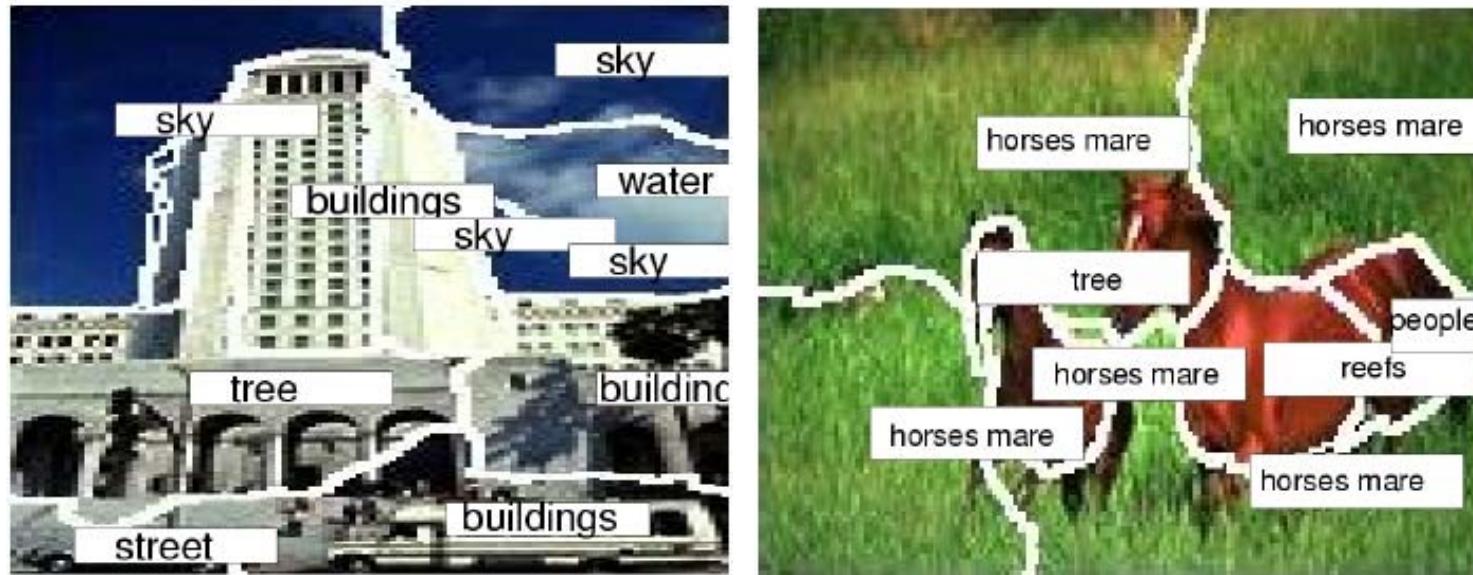
2003年までの画像認識研究

金出武雄. 知能ロボットの技術：人工知能からのアプローチ（前編）：4.ロボット視覚. 情報処理, Vol.44, No.11, pp.1130–1137, 2003.

- この20年間ビジョン研究は主に、形再現の問題、色・テクスチャ・動きといったビジョンの物理的側面を扱う個々のモジュールとその応用システムに大きな成果をあげたものの、「認識」、特に一般のシーンの認識という物理信号とシンボルの世界の融合問題を避けてきた。認識としてされたもののはほとんどが顔といった特定の物体の分類システムである。
- 長い間、認識、理解、知識といった面から遠ざかっていたロボットビジョンの研究が次の段階に進むには、新しい道具が一部にしろ手にある今、そういう研究にもう一度取り組み始める必要があるのではなかろうか。

一般画像認識 (image annotation) のモデル

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



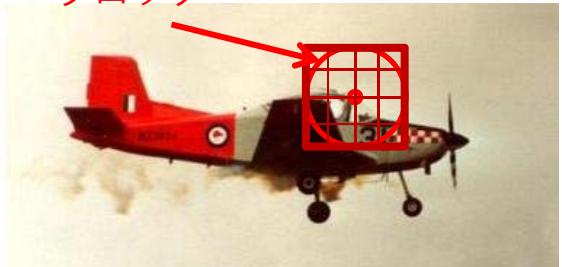
From Duygulu et al., ECCV2002.

2003年周辺の画像認識に関する重要な研究

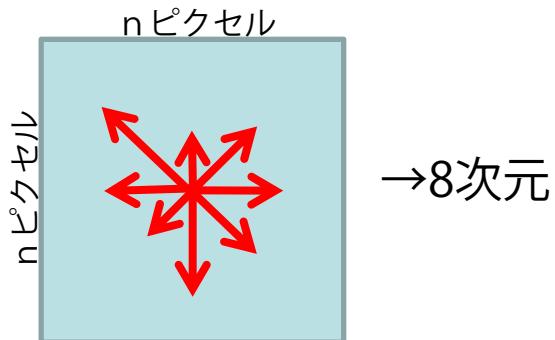
- 1999年 デファクトスタンダードの局所特徴の誕生：SIFT
 - D. G. Lowe. Object recognition from local scale-invariant features. ICCV, 1999.
 - D. G. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 60(2):91–110, 2004.
- 2003年 デファクトスタンダードの画像表現の誕生：Bag of Visual Words
 - J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. ICCV, 2003.
 - G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. Int. Workshop on Stat. Learning in Comput. Vision, 2004.
- 2004年 デファクトスタンダードの画像認識用データセットの誕生：Caltech101
 - L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. *CVPR Workshop on Generative-Model Based Vision*, 2004.

局所特徴量(SIFT)

ブロック



1. ブロックを表現するベクトル



3. 16ブロックの勾配ヒストグラムをまとめて一つのベクトルとする

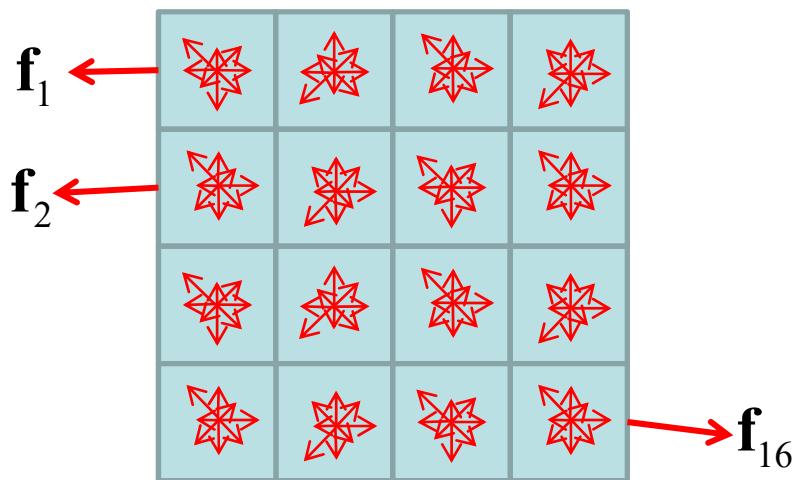
$$\mathbf{f}^T = \left(\mathbf{f}_1^T \ \mathbf{f}_2^T \ \cdots \ \mathbf{f}_{16}^T \right)$$

つまり、8次元×16ブロック=128次元のベクトルとなる

(注) ここまでプロセスをSIFT descriptorと呼ぶ。

SIFTとは特徴点検出+SIFT descriptorのことで、両者は最近では区別される

2. 着目領域内の16ブロック全てに勾配ヒストグラムを計算する



4. 得られたベクトルを正規化する

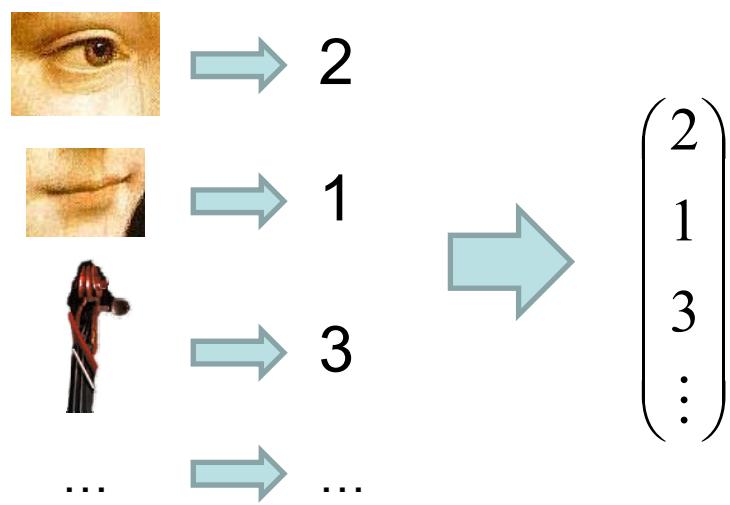
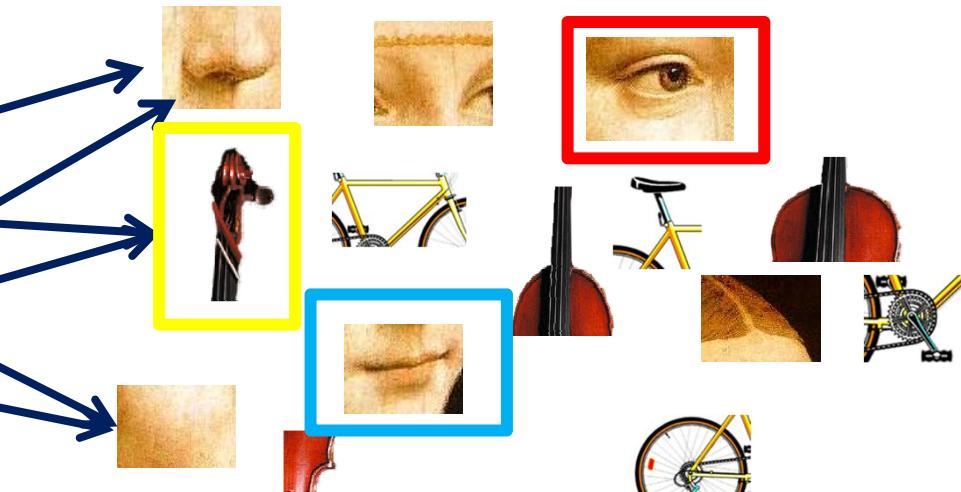
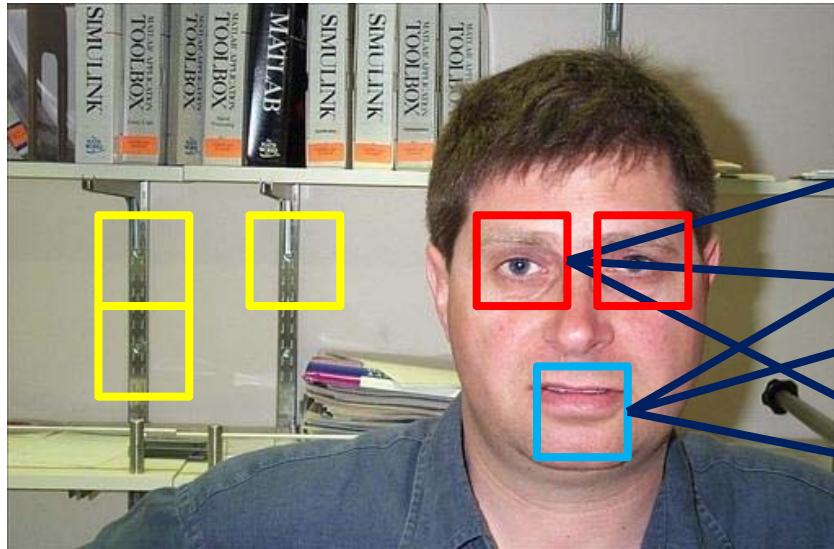
$$\mathbf{f}' = \frac{\mathbf{f}}{\|\mathbf{f}\|}$$

局所領域での正規化を行っているので
照明変化に頑健になる

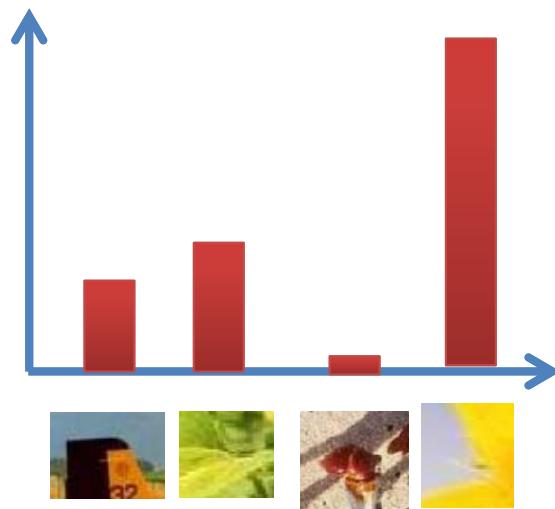
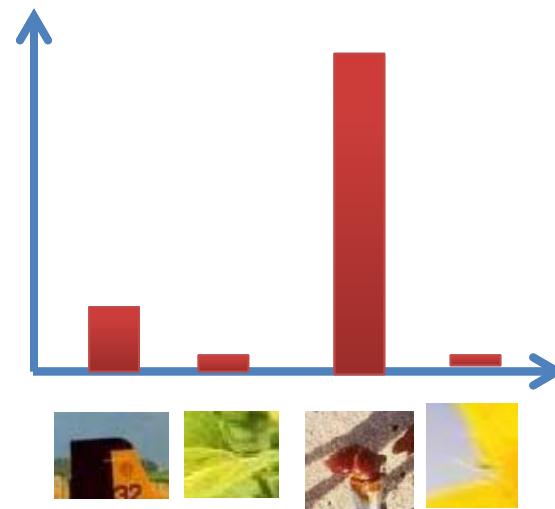
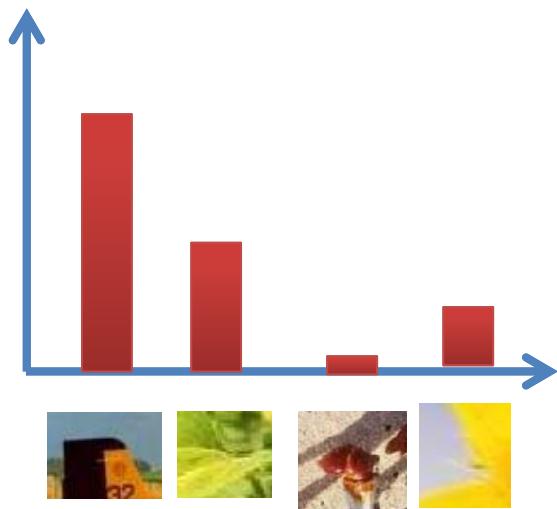
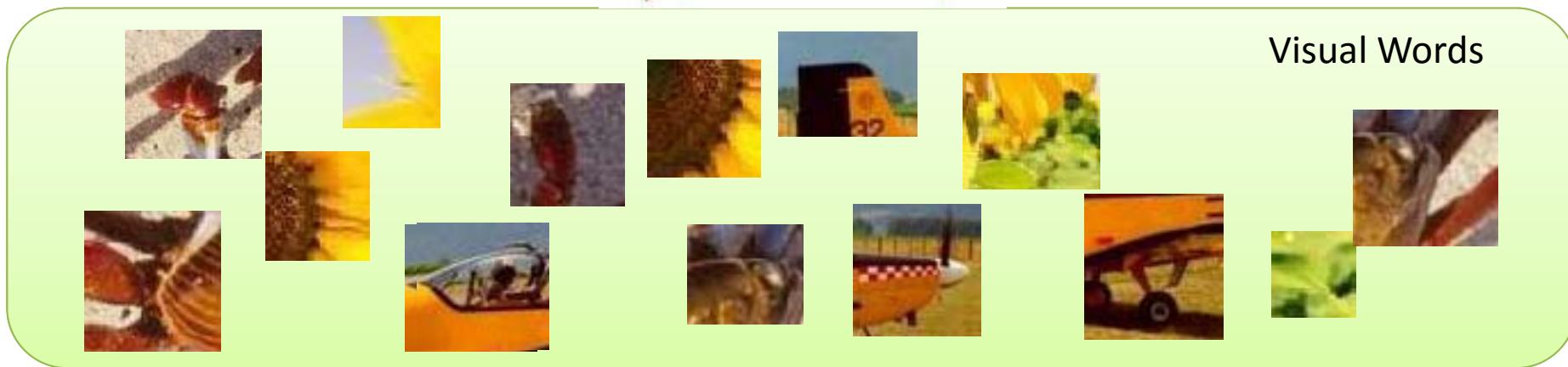
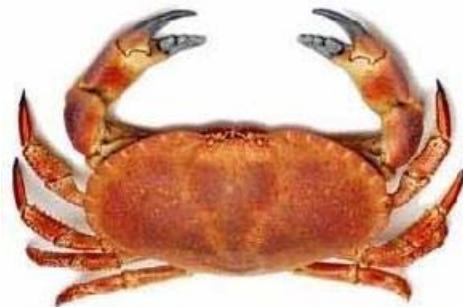
Visual Words?

Li Fei Fei, cvpr07 tutorial
より抜粋

Visual words

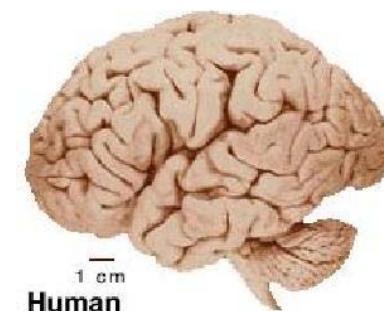
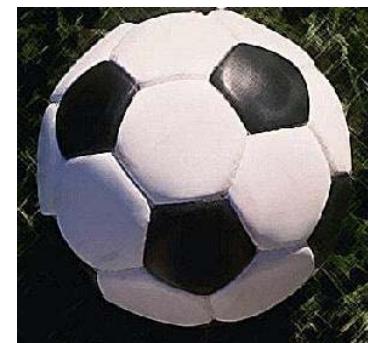


Bag of Visual Words?



Caltech-101

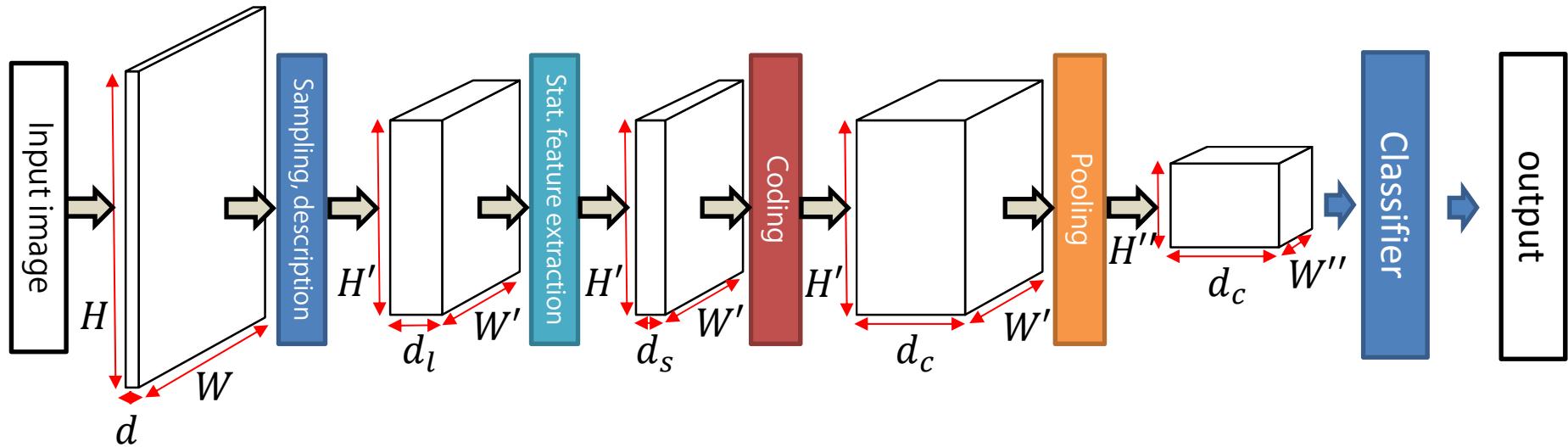
- Pictures of objects belonging to 101 categories. About 40 to 800 images per category. Most categories have about 50 images. Collected in September 2003 by Fei-Fei Li, Marco Andreetto, and Marc 'Aurelio Ranzato. The size of each image is roughly 300 x 200 pixels.
- http://www.vision.caltech.edu/Image_Datasets/Caltech101/



画像識別手法の大まかな歴史2004-2012

- 2004~ BoVW + (kernel) SVM
 - 画像認識へ自然言語処理のアイデアの適用
- 2005~ BoVW + probabilistic topic model
 - 画像認識へ自然言語処理のアイデアの適用
- 2007~ BoVW + Multiple Kernel Learning
 - 複数の異なる特徴の融合
- 2009~ Sparse coding + Linear SVM
 - 大規模データへの適用
 - 特徴空間での多様体を考慮したコーディング
- 2007~ Fisher vector + Linear SVM
 - 大規模データへの適用
 - カーネル近似

Standard Visual Recognition Pipeline



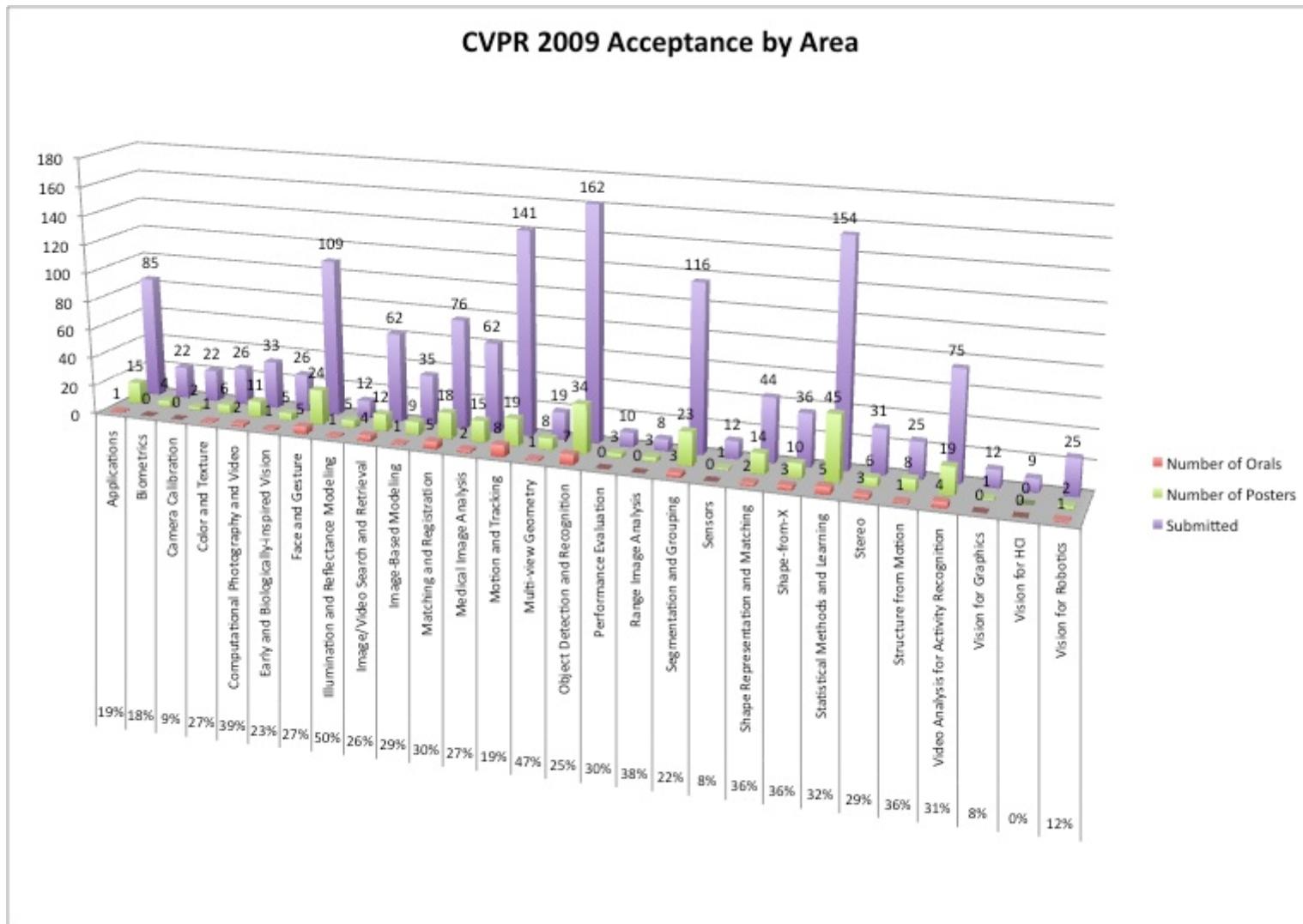
2004~2012の間に画像認識の基本的なパイプラインが完成

ビジョン系の国際学会

- ICCV
 - International Conference on Computer Vision
 - 隔年開催
- ECCV
 - European Conference on Computer Vision
 - 隔年開催
- CVPR
 - IEEE Conference on Computer Vision and Pattern Recognition
 - 毎年開催
- BMVC
 - British Machine Vision Conference
 - 毎年開催
- ACCV
 - Asian Conference on Computer Vision
 - 每年開催

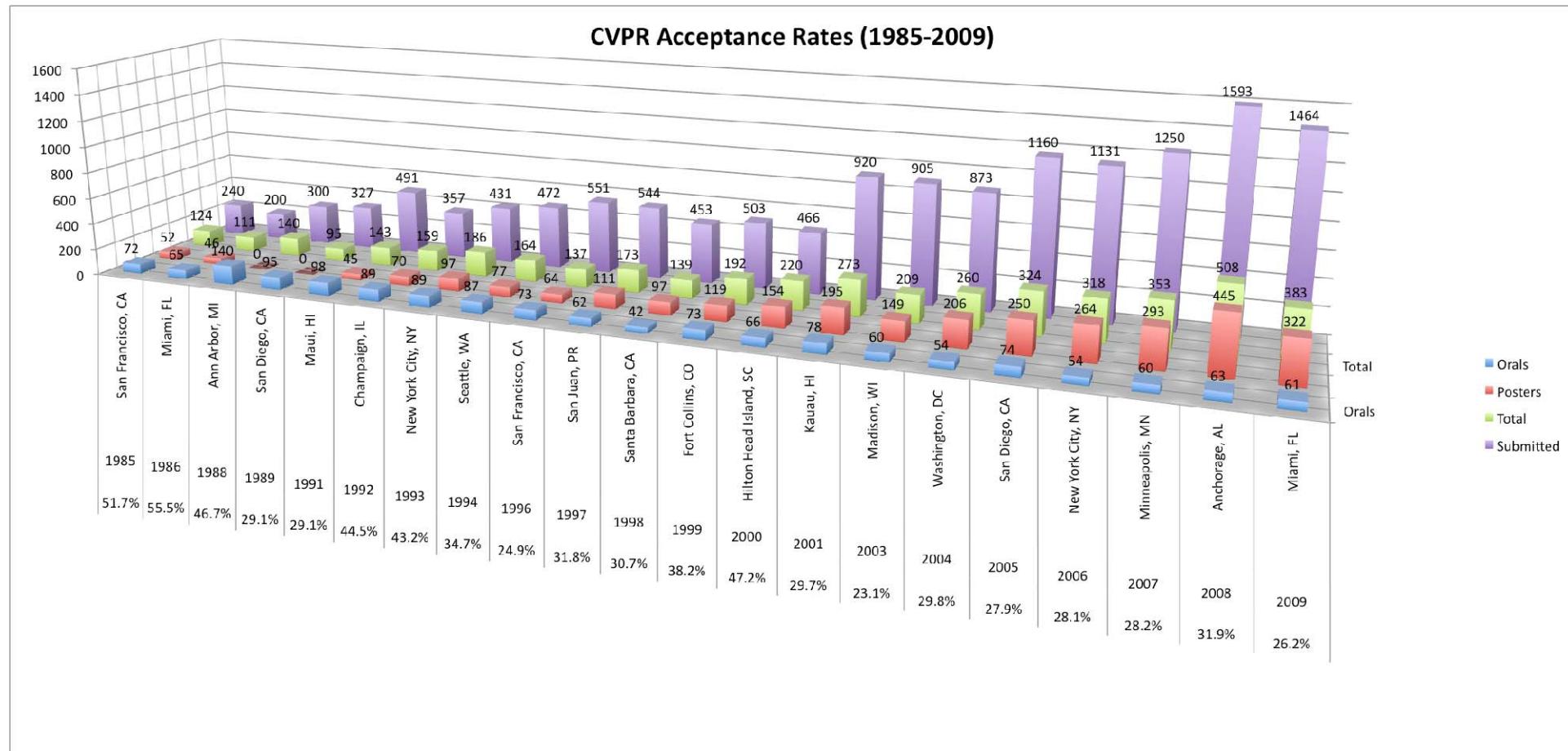
CVPR2009に見るトレンド

- <http://www.cvpr2009.org/stats>

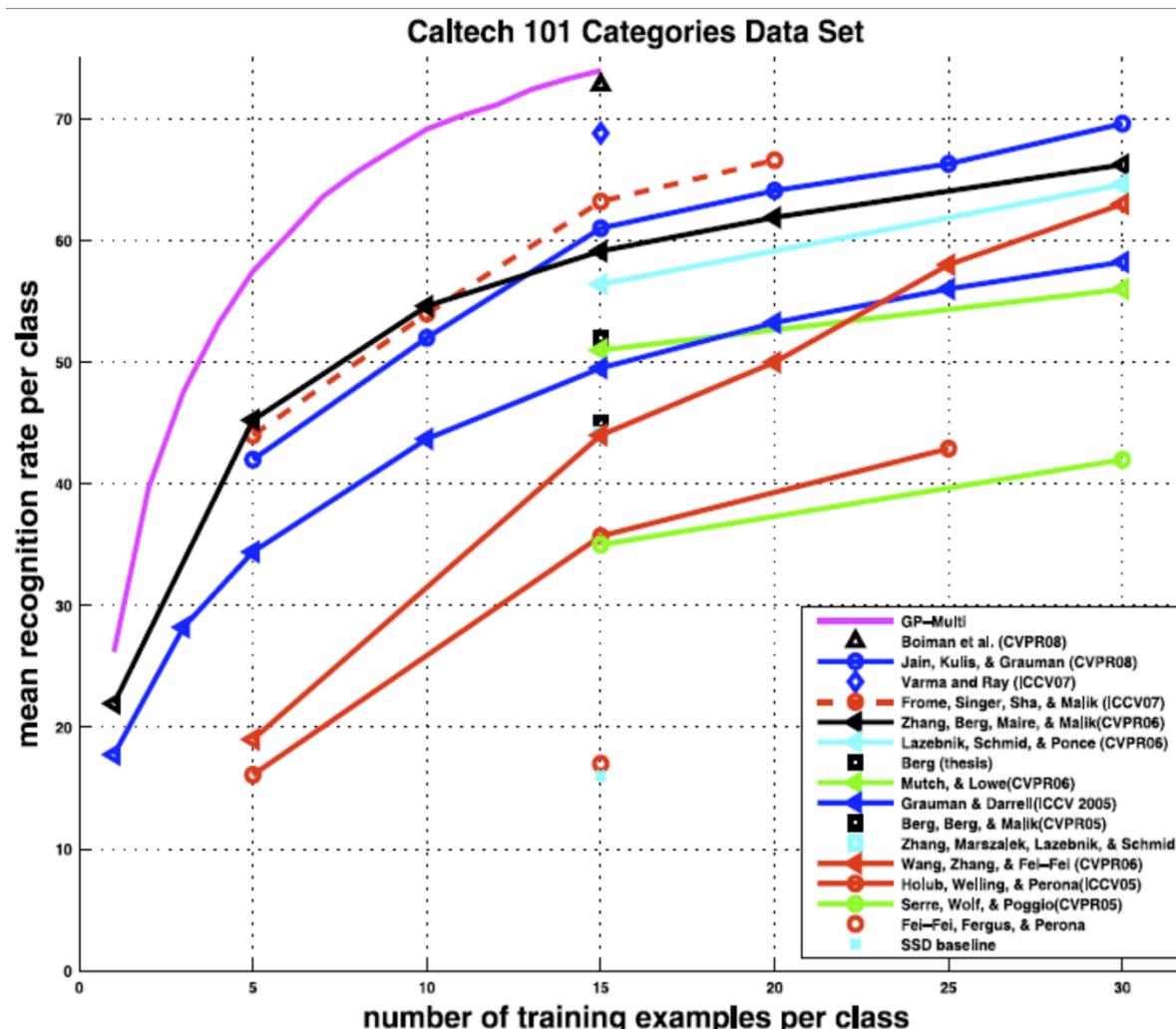


	Area	Number of ORALS Accepted	%	Number of Posters Accepted	%	Number of Accepted Papers	%	Number of Papers Submitted
Applications		1	1.1%	15	17.2%	16	18.4%	87
Biometrics		0	0.0%	4	16.0%	4	16.0%	25
Camera Calibration		0	0.0%	2	9.1%	2	9.1%	22
Color and Texture		1	3.8%	6	23.1%	7	26.9%	26
Computational Photography and Video		2	6.1%	11	33.3%	13	39.4%	33
Early and Biologically-inspired Vision		1	3.8%	5	19.2%	6	23.1%	26
Face and Gesture		5	4.6%	24	22.0%	29	26.6%	109
Illumination and Reflectance Modeling		1	8.3%	5	41.7%	6	50.0%	12
Image/Video Search and Retrieval		4	6.5%	12	19.4%	16	25.8%	62
Image-Based Modeling		1	2.9%	9	25.7%	10	28.6%	35
Matching and Registration		5	6.6%	18	23.7%	23	30.3%	76
Medical Image Analysis		2	3.2%	15	24.2%	17	27.4%	62
Motion and Tracking		8	5.7%	19	13.5%	27	19.1%	141
Multi-view Geometry		1	4.3%	8	34.8%	9	39.1%	23
Object Detection and Recognition		7	4.3%	34	20.7%	41	25.0%	164
Performance Evaluation		0	0.0%	3	30.0%	3	30.0%	10
Range Image Analysis		0	0.0%	3	37.5%	3	37.5%	8
Segmentation and Grouping		3	2.6%	23	19.8%	26	22.4%	116
Sensors		0	0.0%	1	8.3%	1	8.3%	12
<hr/>								
Shape Representation and Matching		2	4.5%	14	31.8%	16	36.4%	44
Shape-from-X		3	8.3%	10	27.8%	13	36.1%	36
Statistical Methods and Learning		5	3.2%	45	29.2%	50	32.5%	154
Stereo		3	9.7%	6	19.4%	9	29.0%	31
Structure from Motion		1	3.4%	8	27.6%	9	31.0%	29
Video Analysis for Activity Recognition		4	5.3%	19	25.3%	23	30.7%	75
Vision for Graphics		0	0.0%	1	8.3%	1	8.3%	12
Vision for HCI		0	0.0%	0	0.0%	0	0.0%	9
Vision for Robotics		1	4.0%	2	8.0%	3	12.0%	25
Total		61	4.2%	322	22.0%	383	26.2%	1464

CVPRの論文投稿数の変遷

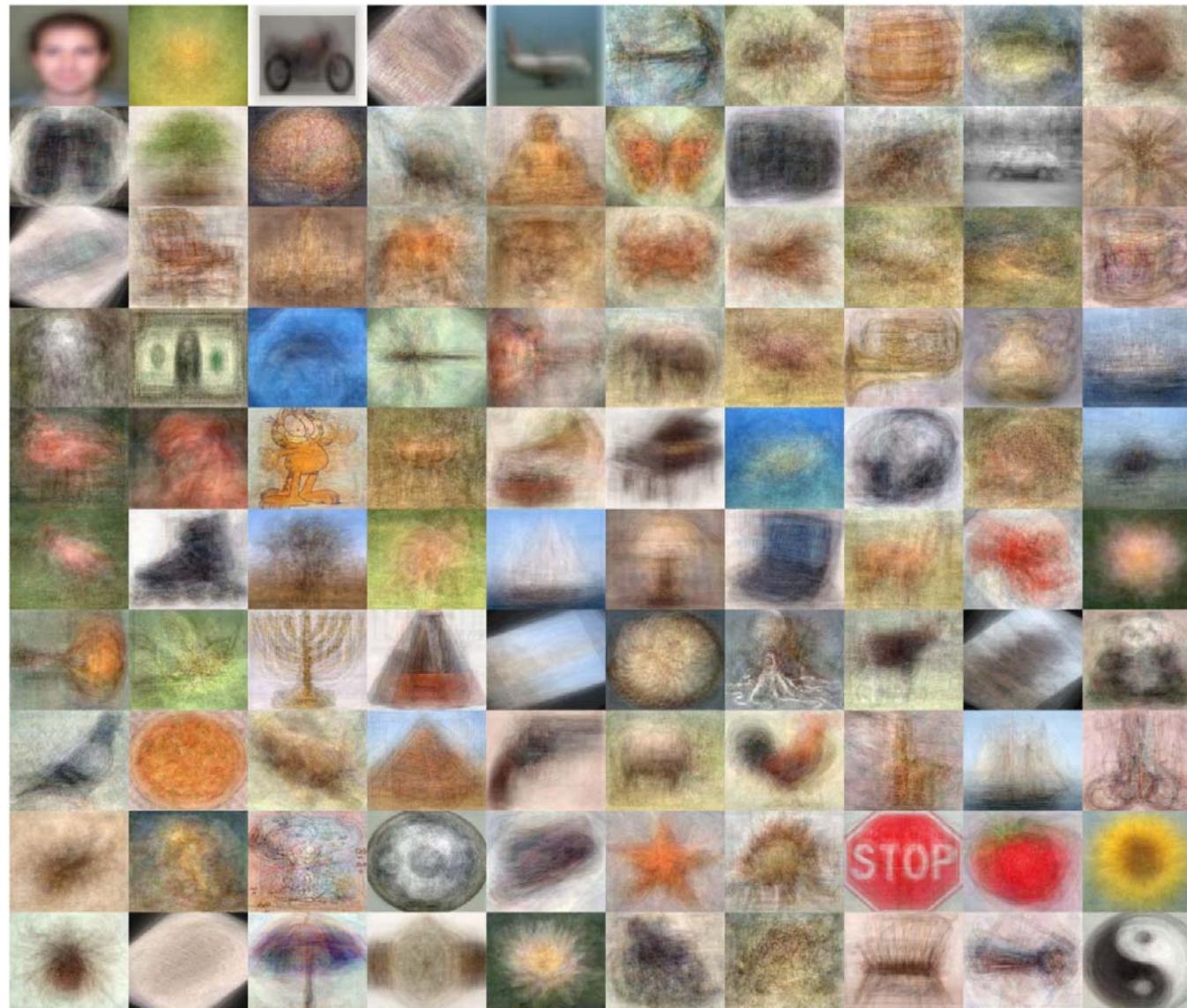


Caltech101における性能の変遷



Gaussian Processes for Object Categorization. A. Kapoor, K. Grauman, R. Uratsun, and T. Darrell. In International Journal of Computer Vision (IJCV), Vol. 88, No. 2, 2010.

Dataset Bias



http://www.vision.caltech.edu/Image_Datasets/Caltech101/averages100objects.jpg

The rise of the modern dataset

データセットの発展：不公平と偏りに対する争いの物語

- COIL-100 dataset
 - 当時のモデルベースの考え方に対する反発
 - データドリブンなアピアランスモデルの採用
- 15 Scenes dataset, Corel Stock Photo
 - シンプルな背景への反発
 - 見た目の複雑さの採用
- Caltech101
 - Corelのようなプロフェッショナルが撮影した画像に対する反発（一部）
 - インターネット画像のwildnessの採用
- MSRC, LabelMe
 - 1つの物体があるというメンタリティへの反発
 - 多くの物体がある複雑なシーンの採用
- PASCAL VOC
 - 以前のトレーニングとテスト基準への反発
- Tiny Images, ImageNet, SUN09
 - 実世界の複雑さに対して小さすぎるデータセットの学習とテストの不適さに対する反発

Antonio Torralba, Alexei A. Efros.
Unbiased Look at Dataset Bias.
CVPR, 2011.



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.

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

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.

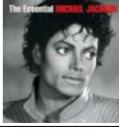
	2.4 M	80M	2B		2.4M	80M	2B
	(no results)	(no results)	<i>prison break</i> , sarah callies , sara tancredi , looking		(no results)	house paint, color	<i>house</i> , paint , wanta- toos, house painting, hardwood floor, interior design
	michael jackson	michael jackson , <i>rock pop</i>	michael jackson ,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

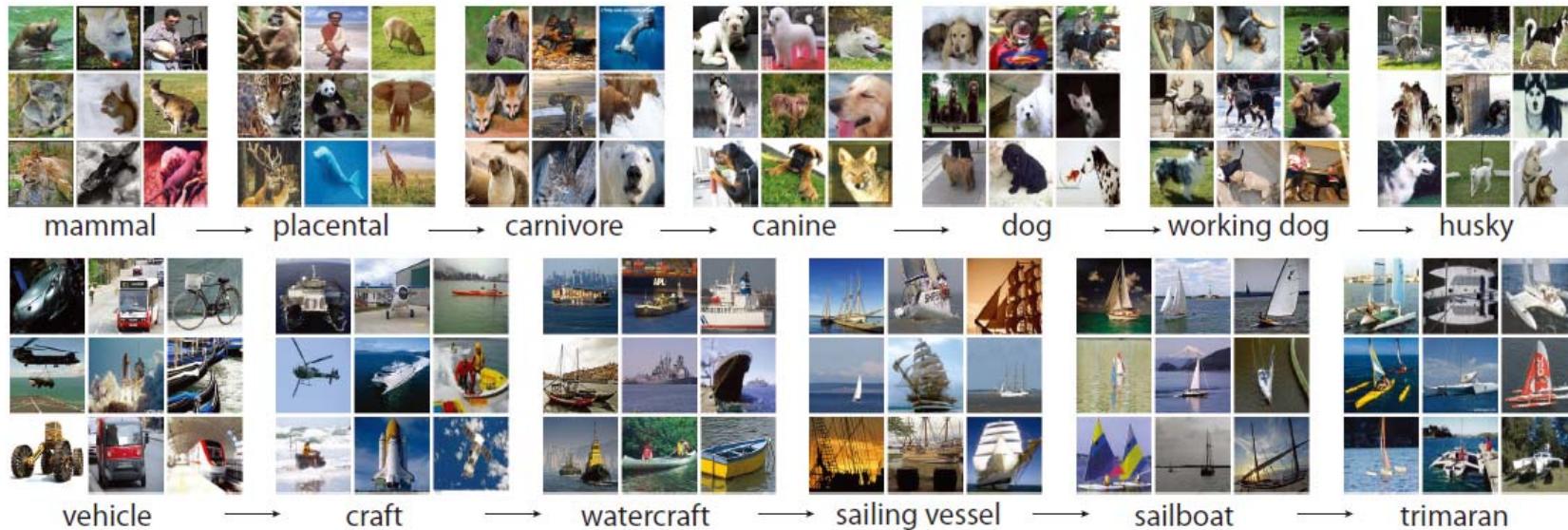


Figure 1: A snapshot of two root-to-leaf branches of ImageNet: the **top** row is from the mammal subtree; the **bottom** row is from the vehicle subtree. For each synset, 9 randomly sampled images are presented.

IMAGENET Large Scale Visual Recognition Challenge 2010 (ILSVRC2010)

Held as a "taster competition" in conjunction with PASCAL Visual Object Classes Challenge 2010 (VOC2010)

[Registration](#) [Download](#) [Introduction](#) [Data](#) [Task](#) [Development kit](#) [Timetable](#) [Features](#) [Submission](#) [Citation^{new}](#) [Organizers](#)
[Contact](#)

News

- September 2, 2014: [A new paper](#) which describes the collection of the ImageNet Large Scale Visual Recognition Challenge dataset, analyzes the results of the past five years of the challenge, and even compares current computer accuracy with human accuracy is now available. *Please cite it when reporting ILSVRC2010 results or using the dataset.*
- For latest challenge, please visit [here](#).
- September 16, 2010: Slides for [overview of results](#) are available, along with slides from the two winning teams:

Winner: NEC-UIUC

Yuanqing Lin, Fengjun Lv, Shenghuo Zhu, Ming Yang, Timothee Cour, Kai Yu (NEC). LiangLiang Cao, Zhen Li, Min-Hsuan Tsai, Xi Zhou, Thomas Huang (UIUC). Tong Zhang (Rutgers).

[PDF] **NB: This is unpublished work. Please contact the authors if you plan to make use of any of the ideas presented.**

Honorable mention: XRCE

Jorge Sanchez, Florent Perronnin, Thomas Mensink (XRCE)

[PDF] **NB: This is unpublished work. Please contact the authors if you plan to make use of any of the ideas presented.**

- September 3, 2010: [Full results](#) are available. Please join us at the [VOC workshop](#) at ECCV 2010 on 9/11/2010 at Crete, Greece. At the workshop we will provide an overview of the results and invite winning teams to present their methods. We look forward to seeing you there.
- August 9, 2010: Submission deadline is extended to [4:59pm PDT, August 30, 2010](#). There will be no further extensions.
- August 8, 2010: [Submission site](#) is up.
- June 16, 2010: Test data is available for [download!](#).
- May 3, 2010: Training data, validation data and development kit are available for [download!](#).
- May 3, 2010: [Registration](#) is up!. Please register to stay updated.
- Mar 18, 2010: We are preparing to run the ImageNet Large Scale Visual Recognition Challenge 2010 (ILSVRC2010)

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.

2位

Task 1

1位

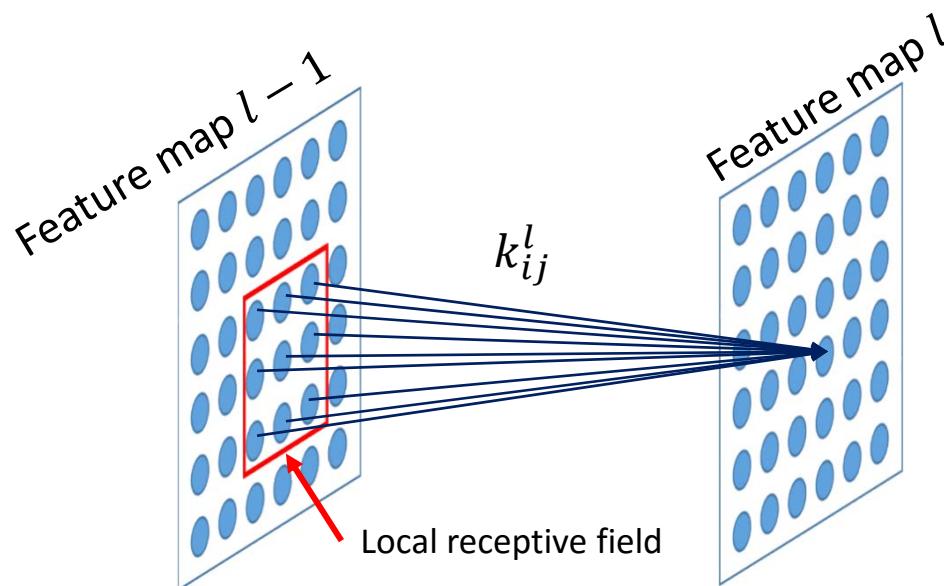
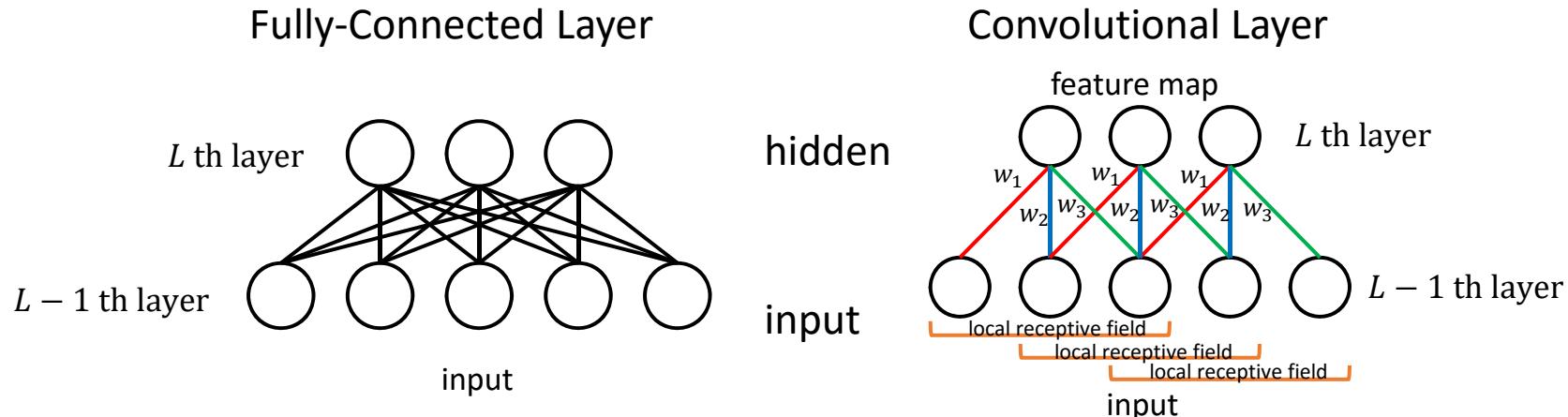
Task 3



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

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

Convolutional and Fully-Connected Layer

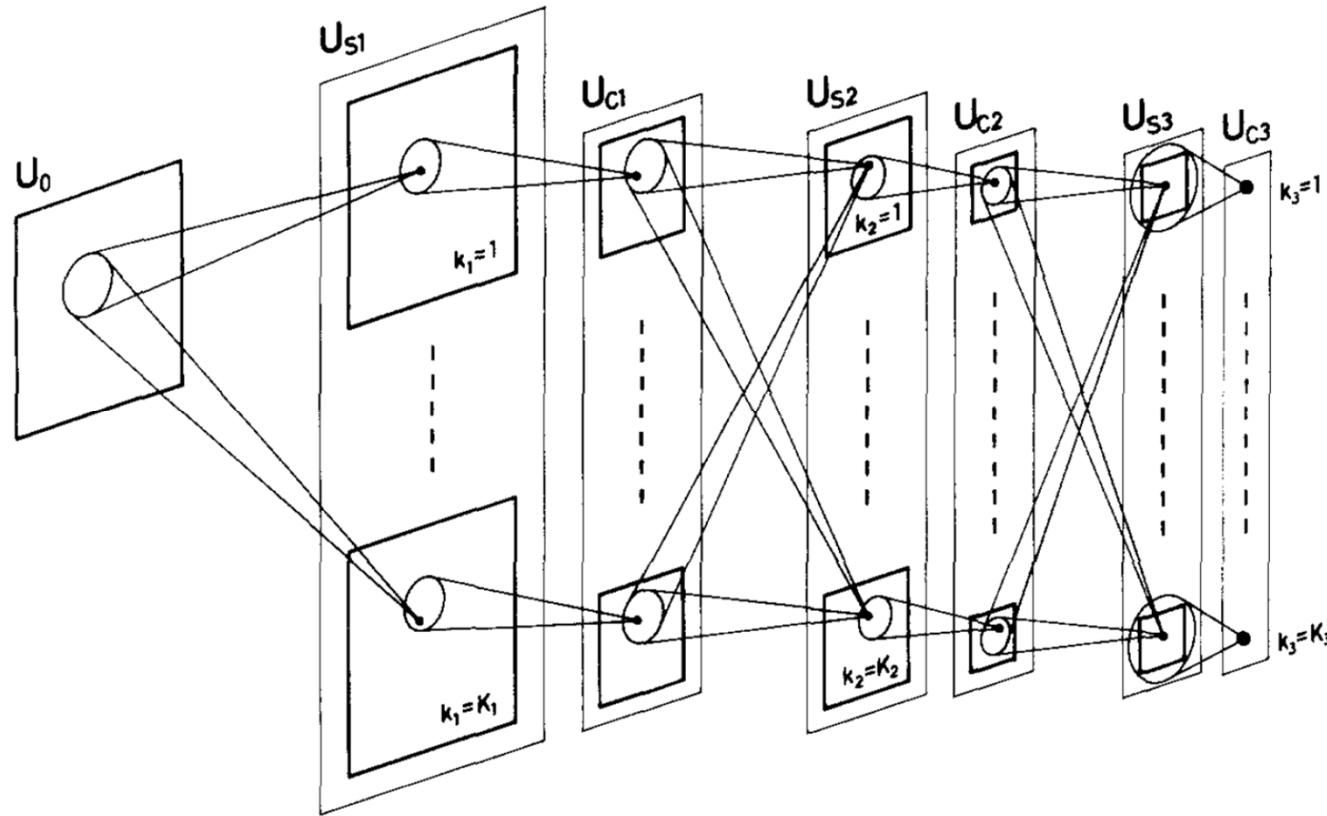


Logistic sigmoid function
 $f(x) = (1 + e^{-\beta x})^{-1}$

Hyperbolic tangent function
 $f(x) = \text{atanh}(bx)$

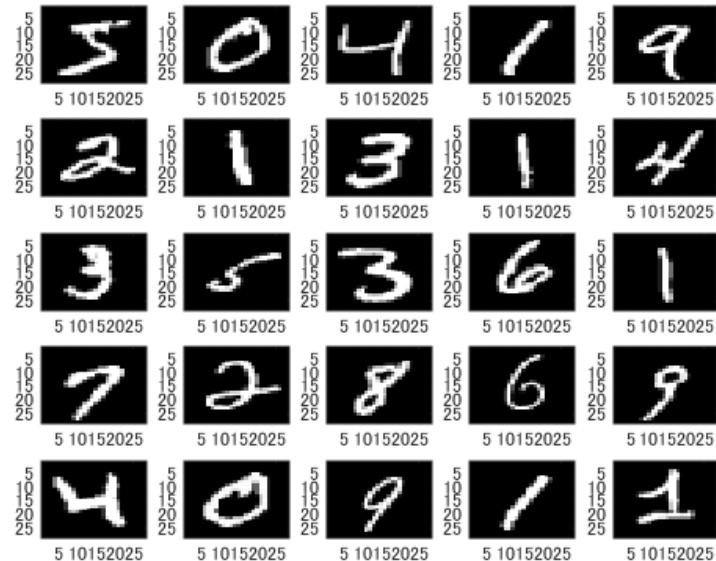
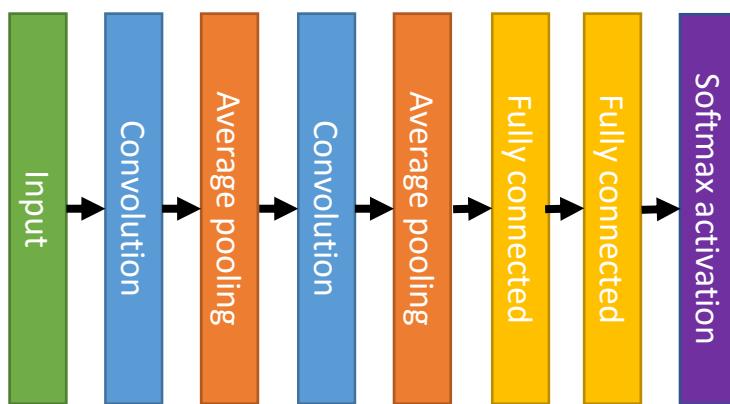
$$x_j^l = f \left(\sum_{i \in M_j} x_i^{l-1} * k_{ij}^l + b_j \right)$$

Neocognitron



Kuniyuki Fukushima. Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. Biol. Cybernetics 36, 193 202 (1980)

Convolutional Neural Networks



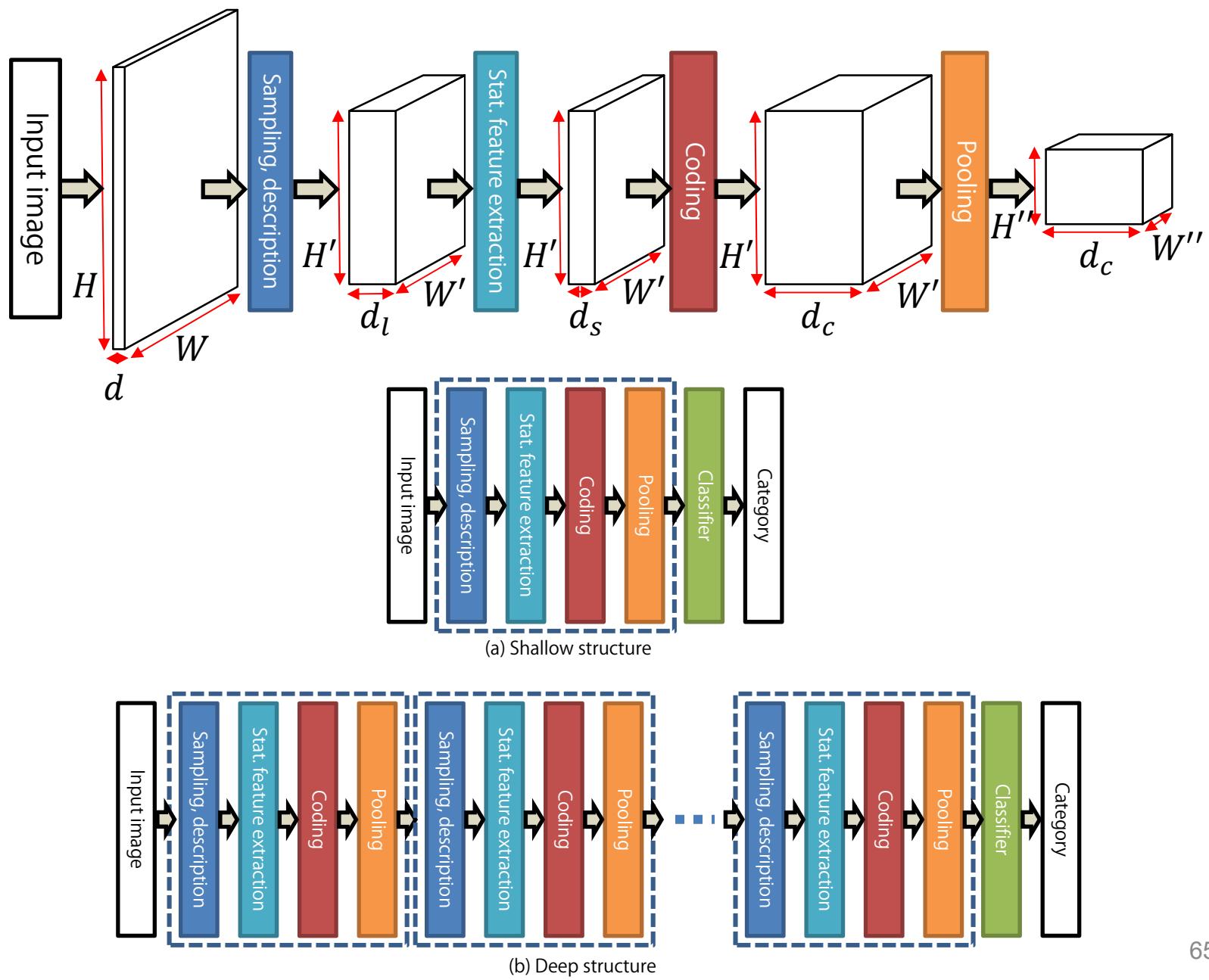
(a) LeNet-5, 1989

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. NIPS 1989.

Modules of Convolutional Neural Networks

- Convolution layer
- fully-connected layer
- Pooling layer

Mapping Function for Visual Recognition



IMAGENET Large Scale Visual Recognition Challenge (ILSVRC) 2010-2013

~~20 object classes~~ ~~22,591 images~~

200 object classes

1000 object classes

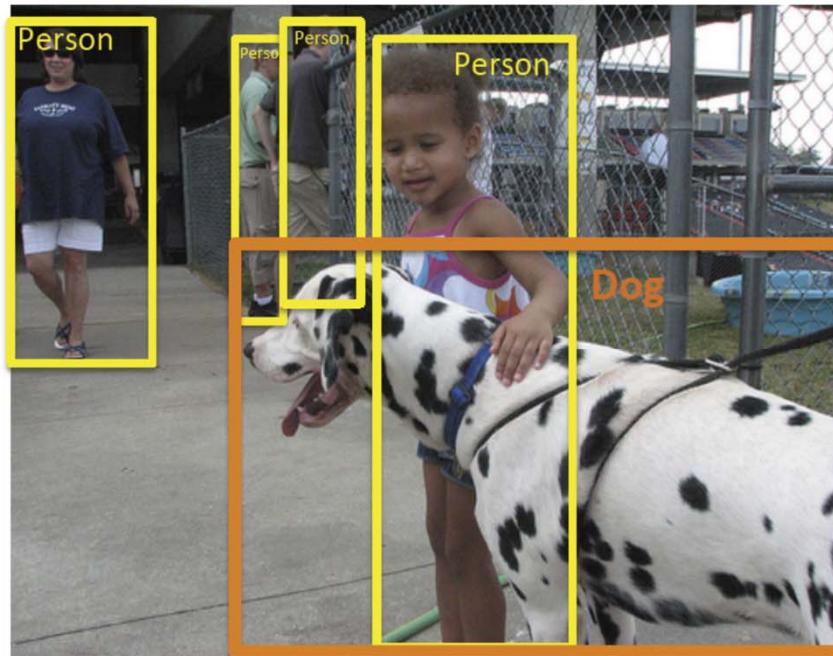
456,191 images

1,431,167 images

NEW

DET

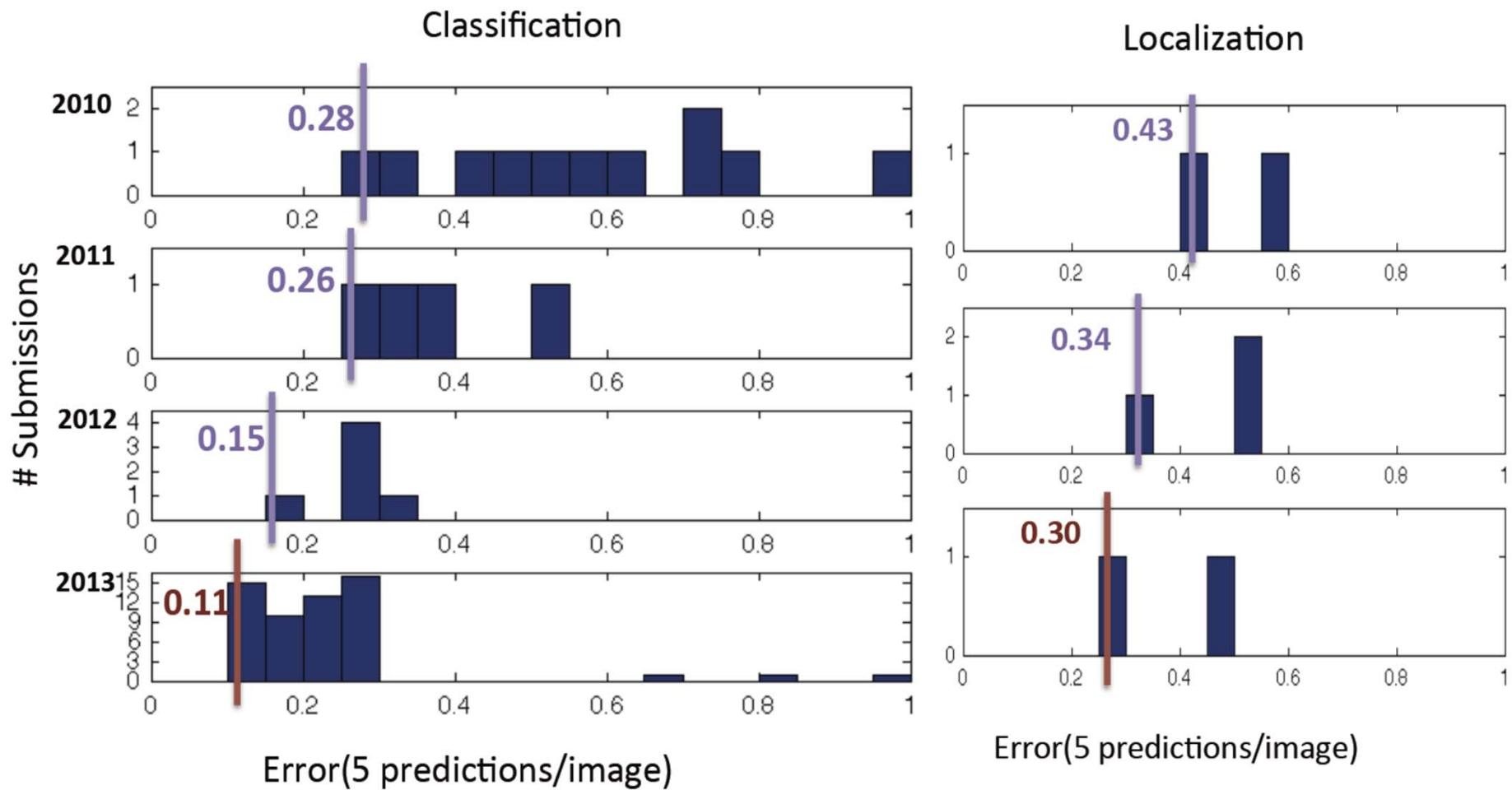
CLS-LOC



<http://image-net.org/challenges/LSVRC/{2010,2011,2012,2013}>

http://www.image-net.org/challenges/LSVRC/2013/slides/ILSVRC2013_12_7_13_clsloc.pdf

ILSVRC over the years



最新の画像識別性能

multi-model results

	Team	Top 5 error [%]	
post-competition	Google	4.9	Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift
	MSRA, PReLU-nets	4.94	Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification
	Human	5.1	
	Baidu	5.98	
	VGG (arXiv v5)	6.8	
in competition ILSVRC 14	GoogLeNet	6.66	
	VGG (Oxford)	7.32	
	MSRA, SPP-nets	8.06	

ILSVRC2015

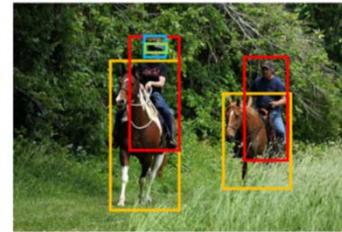
- Main competitions
 - Object detection
 - 200 fully labeled categories and 0.4 million images



Win bottle Table Chair



Person



horse person helmet sunglasses

http://image-net.org/challenges/talks/ILSVRC+MSCOCO_12_17_15_detection.pdf

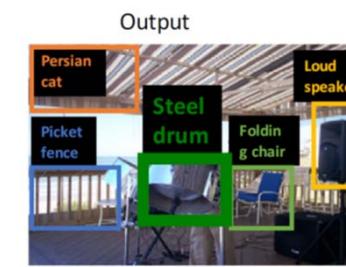
- Object classification and localization
 - 1000 categories and 1.2 million images



Output:
Scale
T-shirt
Steel drum
Drumstick
Mud turtle



Steel drum



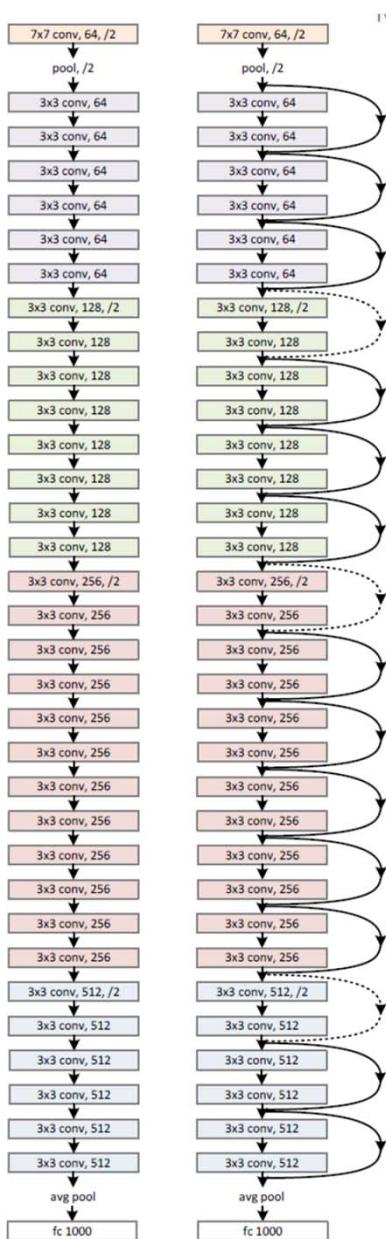
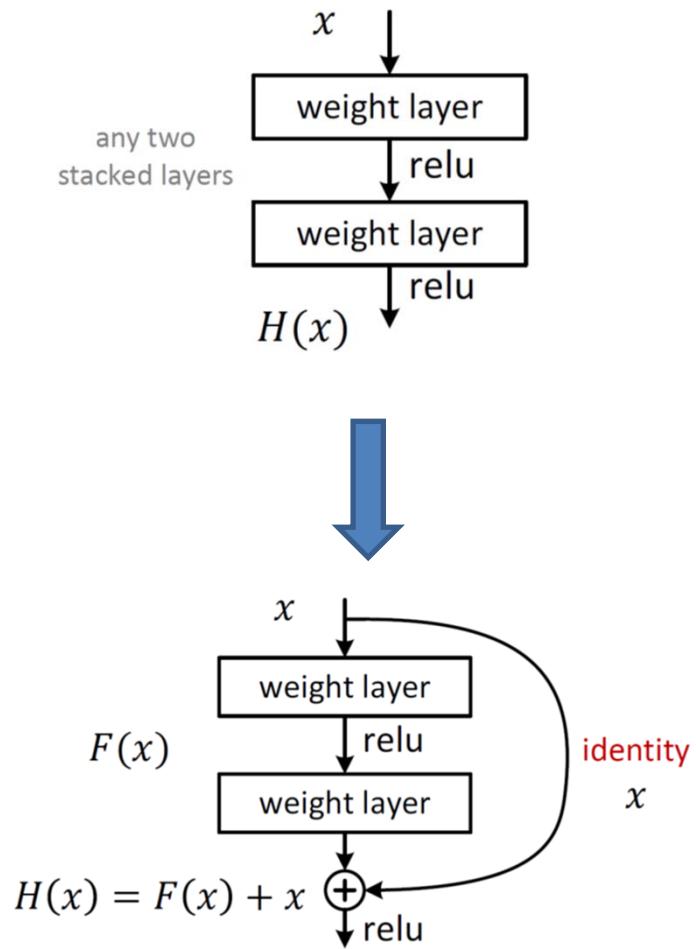
Output

http://image-net.org/challenges/talks/ILSVRC2015_12_17_15_clsloc.pdf

- Taster competitions
 - Object detection from video
 - Fully annotated 30 object classes across 5,354 snippets
 - Scene classification
 - 401 scene categories, 8.1M train, 20k val, 381k test

MSRA's method

Residual Network



Similar methods in shallow networks

Fisher Vector [Perronnin et al 2007]

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

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

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

VLAD [Jegou et al 2010]

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

MSRA's method

- http://image-net.org/challenges/talks/ilsvrc2015_deep_residual_learning_kaiminghe.pdf
- **1st places in all five main tracks**
 - ImageNet Classification: “*Ultra-deep*” (quote Yann) **152-layer** nets
 - ImageNet Detection: **16%** better than 2nd
 - ImageNet Localization: **27%** better than 2nd
 - COCO Detection: **11%** better than 2nd
 - COCO Segmentation: **12%** better than 2nd

MSRA's method

- http://image-net.org/challenges/talks/ilsvrc2015_deep_residual_learning_kaiminghe.pdf

