一般物体認識のための局所特徴量 (SIFT と HOG)
Gradient-Based Local Features for Generic Object Recognition

藤吉弘亘 †
Hironobu Fujiyoshi†

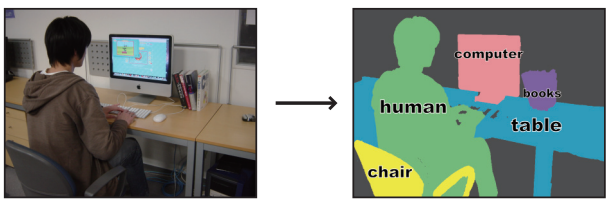
† 中部大学 工学部 情報工学科
†Dept. of Computer Science, Chubu University

Abstract: Generic Object Recognition is one of the biggest problem in the research area of computer vision. Gradient-based local features such as SIFT and HOG are commonly used for object dection and object categorization. Scale-Invariant Feature Transform(SIFT) is an approach for detecting and extracting local feature descriptors that are reasonably invariant to changes in illumination, image noise, rotation, scaling, and small changes in viewpoint. Because the SIFT algorithm can describe characteristics of feature points that are invariant to scale and rotation changes, it has been used for image matching such as image mosaicing and generic object recognition. In this paper, we describe the SIFT algorithm and introduce applications that use it. We also describe another algorithm called "Histograms of Oriented Gradients(HOG)" which is based on gradient feature extraction similar to the SIFT algorithm. We also introduce an example of how the HOG algorithm can be used for people detection.

1 はじめに

今日のデジタルカメラの普及に伴い，デジタル化された写真や動画像が我々の日常に大量に存在し，それらのデジタル画像にはさまざまなシーンや物体が撮影されている．このように実世界シーンを制約のない状態で撮影された画像中から，どのような物体が画像内に存在するか，どのようなシーンであるかを一般的な名称でコンピュータが認識することを一般物体認識と呼ぶ [1]．コンピュータの発展により大量のデータを高速に処理することが可能となり，さらに機械学習の分野で研究された学習手法が一般物体認識へ適用できるようになったため，一般物体認識の研究が躍進した．中でも，Viola & Jones の顔検出手法 [3] や Bag-of-Keypoints による画像分類 [2] 等のように統計的機械学習法を用いた一般物体認識の有効なアプローチが提案されている．また，各手法の客観的評価を行うために共通データセットが構築され，近年では，一般物体認識の研究が一層盛んに取り組まれている．

画像分類で用いられている Bag-of-Keypoints というアプローチでは，局所特徴量として勾配情報から算出される SIFT(Scale Invariant Feature Transform) が用いられている．SIFT は，スケール変化，回転変化に不変な特徴量を記述することができるため，特定物体認識だけでなく画像分類の特徴量としても有効である．一方，歩行者認識等の物体検出においても，局所特徴量として勾配情報から算出する勾配方向ヒストグラム HOG（Histogram of Oriented Gradients) が用いられている．本チュートリアルでは，一般物体認識の入力特徴として用いられている勾配情報から得られる局所特徴量として SIFT と HOG のアルゴリズムについて紹介し，それぞれ画像分類と物体検出への利用方法について述べる．



(a) 入力画像　(b) 認識結果

Figure 1: 一般物体認識

2 一般物体認識とは

一般物体認識とは，制約のない実世界シーンの画像に対して，コンピュータがその中に含まれる物体を一般的な名称で認識することである．一般物体認識の最終的な目標は図 1 に示すようにセグメンテーションされた各領域に対して物体カテゴリのラベルを付けることであるといえる．

2.1 一般物体認識の難しさ

一般物体認識を困難にしている大きな要因として，一般的な名称を指すカテゴリが膨大であることと，同一物体カテゴリ内のバリエーションの多様さが挙げられる．一般的な名称によるカテゴリ分類を行う場合，そのカテゴリ数は 10,000 から 30,000 になるといわれている．またカテゴリ内にはサブクラスが存在する場合がある．例えば,「車」というカテゴリには「セダン」「ワゴン」「トラック」のような形状の違い,「ヘッドライト」「フロントガラス」のような部位の違いによるサブクラスに分けられると考えられる．そのため，どこまでを対象のカテゴリとするのかを明確に定義することが困難である．

一方，同一の物体においても視点の変化，照明変化，オクルージョン，スケール変化，物体の変形などによりアピアランスが大きく変化するため，このような条件下で認識することは難しい問題である．また，認識対象の物体が画像中のどこに存在しているかを検出する必要もある．一般物体認識問題では，アピアランスの変化に不変で，かつ，

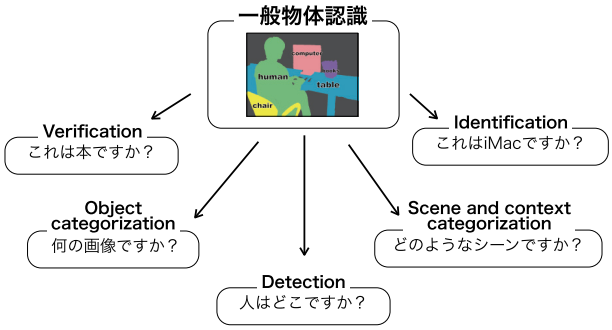


Figure 2: 一般物体認識の簡略化

物体以外の背景情報に影響を受けない特徴抽出法と識別器の作成が重要となる．

2.2 一般物体認識の簡略化

一般物体認識は不確定な要素が多いため，非常に難しい問題である．そこで，一般物体認識問題にそれぞれの制約を設けることで問題を簡略化する．簡略化した一般物体認識問題の例を以下に示す (図 2)．

- Verification（物体照合）
  画像中のある物体を指し，対象物体のカテゴリであるかを照合する問題
- Object categorization（画像分類）
  画像中の物体が何の物体カテゴリであるかを分類する問題
- Detection（特定の物体検出）
  特定物体が画像中のどこにあるかを求める問題
- Identification（特定物体認識）
  画像中のある物体の固有名詞を答える問題
- Scene and context categorization（シーン・カテゴリ認識）
  場所や天気などどのようなシーンであるかを認識する問題

Detection（特定の物体検出）においては，カテゴリの多様さを抑制するため，人や顔等の特定のカテゴリのみを対象とし，対象カテゴリの物体が画像中のどこにあるか（ローカライゼーション）を答える問題としている．一方，Object categorization（画像分類）では，対象物体のローカライゼーションを無視して，画像中に認識対象の物体が大きく撮影されているという条件下で，その物体カテゴリを選択する問題（カテゴライゼーション）としている．

このように一般物体認識では，ローカライゼーションとカテゴライゼーションの 2 つの制約を導入することで，現状では問題を簡略化して解いていることになる．この二つを同時に解決するアプローチを実現できれば一般物体認識を解決することになるため，ローカライゼーション問題とカテゴライゼーション問題は一般物体認識問題において重要な課題であるといえる．

本稿では，近年研究が盛んに行われているローカライゼーション問題である Detection(特定の物体検出) と，カテゴライゼーション問題である Object categorization(画像分類) を対象とし，画像分類で用いられている SIFT と物体検出で用いられている HOG について述べる．

3 SIFT と画像分類

本章では，一般物体認識におけるカテゴライゼーション問題に用いられる局所特徴量である SIFT[4] とその画像分類への適用手法について述べる．

3.1 SIFT のアルゴリズム

SIFT の処理は，特徴点 (以下，キーポイントと呼ぶ) の検出 (detection) と特徴量の記述 (description) の 2 段階からなり，各処理は以下の流れとなる．

detection
1. スケールとキーポイント検出
2. キーポイントのローカライズ

description
3. オリエンテーションの算出
4. 特徴量の記述

1. スケールとキーポイント検出では，DoG 処理によりスケールとキーポイントを検出し，2. キーポイントのローカライズでは，1. で検出されたキーポイントから特徴点として向かない点を削除し，その後サブピクセル推定を行う．3. オリエンテーションの算出では，回転に不変な特徴を得るためにキーポイントのオリエンテーションを求める．4. 特徴量の記述では，3. で求めたオリエンテーションに基づいてキーポイントの特徴量を記述する．以下に各処理の詳細を述べる．

3.1.1 スケールとキーポイント検出

第 1 段階のキーポイント検出では，DoG 処理を用いてスケールスペースにおける極値探索をすることで，キーポイントの位置とスケールを決定する．

Difference-of-Gaussian 処理　キーポイント候補点は，スケールの異なるガウス関数 $G(x,y,\sigma)$ と入力画像 $I(u,v)$ を畳み込んだ平滑化画像 $L(u,v,\sigma)$ の差分 (DoG 画像) から求める．それぞれ以下の式により求める．

$$L(u,v,\sigma) = G(x,y,\sigma) * I(u,v) \quad (1)$$

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \quad (2)$$

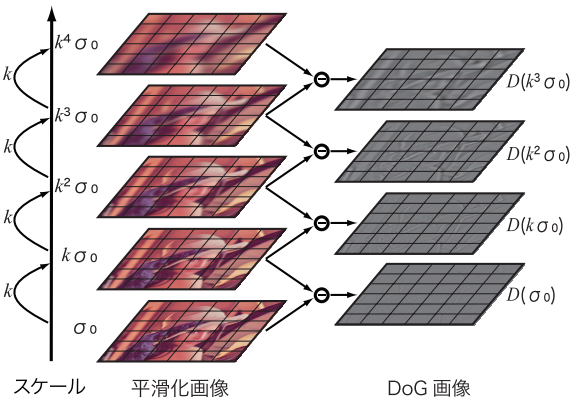


Figure 3: DoG 処理の流れ

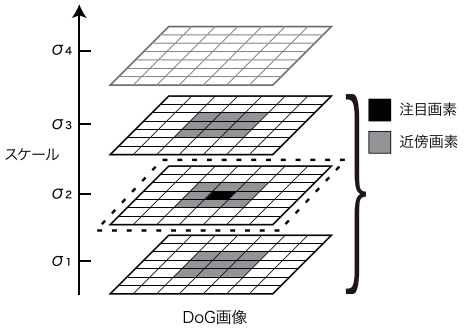


Figure 4: 極値検出の流れ

DoG の結果の画像を $D(u,v,\sigma)$ とすると，DoG 画像は次式で求まる．

$$D(u,v,\sigma) = (G(x,y,k\sigma) - G(x,y,\sigma)) * I(u,v)$$
$$= L(u,v,k\sigma) - L(u,v,\sigma) \quad (3)$$

この処理を σ_0から k 倍ずつ大きくした 異なるスケール間で行い，図 3 に示すような複数の DoG 画像を求める．σ が一定の割合で増加し続けると，ガウシアンフィルタのウィンドウサイズが大きくなり，処理できない端領域の拡大と計算コストの増加という問題が発生する．この問題に対し，SIFT では画像のダウンサンプリングにより σ の変化の連続性を保持した平滑化処理を実現している．

DoG 画像からの極値検出　DoG は異なるスケールによる平滑化画像の差分のため，DoG の値が大きくなる σ では，スケールの変化領域にエッジ等の情報量を多く含んでいるといえる．そこで，DoG 画像から極値を検出し，キーポイントとスケールを決定する．極値の検出は，図 4 のように DoG 画像 3 枚一組で行う．DoG 画像 (図 4 中の点線で囲まれた画像) の注目画素 (図 4 中の黒色領域) と，その周りの 26 近傍 (図 4 中の灰色領域) を比較し，極値であった場合，その画素をキーポイント候補点として検出する．このような極値検出は，σ の値の小さい DoG 画像から行う．一度極値が検出された画素は，より大きなスケールで極値が検出されてもキーポイント候補点としない．この処理をスケールの異なる DoG 画像の全画素に対して行う．

3.1.2　キーポイントのローカライズ

3.1.1 により検出されたキーポイント候補点の中には，DoG 出力値が小さい点 (low contrast) やエッジ上の点が含まれており，これらの点はノイズや開口問題に影響を受け易いという問題がある．そこで，キーポイント候補点の中から， 主曲率とコントラストにより安定したキーポイントに絞り込む．さらに，キーポイントのサブピクセル推定により位置とスケールを算出する．図 5(a) は検出された全キーポイント候補点を表している．図中の円の中心がキーポイント位置，円の半径がキーポイントの持つスケールである．図 5(b) では，主曲率によりドア等のエッジ上の点が削除されていることがわかる．

3.1.3　オリエンテーションの算出

検出したキーポイントに対して，第 2 段階の処理である特徴量の記述を行う．まず，検出された各キーポイントのオリエンテーションを求める．オリエンテーションはキーポイントにおける方向を表し，特徴量記述の際にオリエンテーションにより向きの正規化を行うことで，回転に不変となる．キーポイントのオリエンテーションを求めるには，まずキーポイントが検出された平滑化画像 $L(u,v)$ の勾配強度 $m(u,v)$ と勾配方向 $\theta(u,v)$ を以下の式により求める．

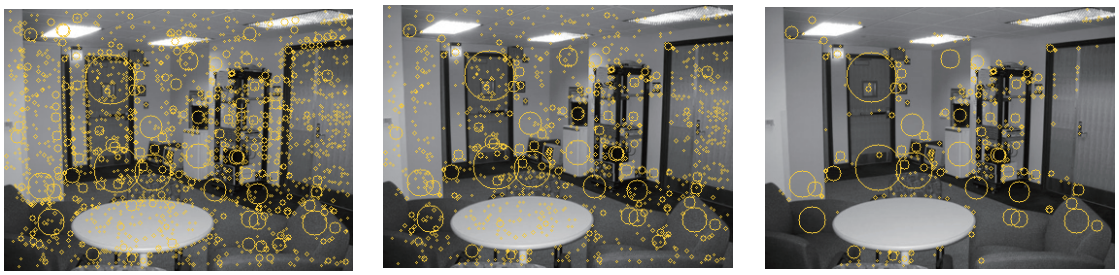
$$m(u,v) = \sqrt{f_u(u,v)^2 + f_v(u,v)^2} \quad (4)$$
$$\theta(u,v) = \tan^{-1}\frac{f_v(u,v)}{f_u(u,v)} \quad (5)$$

$$\begin{cases} f_u(u,v) = L(u+1,v) - L(u-1,v) \\ f_v(u,v) = L(u,v+1) - L(u,v-1) \end{cases} \quad (6)$$

局所領域における勾配強度 $m(x,y)$ と勾配方向$\theta(x,y)$ から図 6 に示すような重み付方向ヒストグラム h を以下の式により作成する．

$$h_{\theta'} = \sum_x \sum_y w(x,y)\cdot\delta[\theta',\theta(x,y)] \quad (7)$$
$$w(x,y) = G(x,y,\sigma)\cdot m(x,y) \quad (8)$$

ここで，h_θ は，全方向を 36 方向に量子化したヒストグラムである．$w(x,y)$ はある局所領域の画素 (x,y) での重みであり，キーポイントが持つスケールサイズのガウス窓 $G(x,y,\sigma)$ と勾配強度 $m(x,y)$ から求める．δ は Kronecker のデルタ関数で，勾配方向 $\theta(x,y)$ が量子化した方向 θ' に含まれるとき 1 を返す．また，このときのガウス窓にはキーポイントが持つスケールを用いる．ガウス窓による重み付けにより，キーポイントに近い特徴量がより強く反映される．この 36 方向のヒストグラムの最大値から 80% 以上となるピークをキーポイントのオリエンテーションとして割り当てる．



(a) 全キーポイント候補点
(キーポイント候補 1895 点)

(b) エッジ上のキーポイントを削除
(キーポイント候補 1197 点)

(c) 低コントラストのキーポイントを削除
(キーポイント候補 421 点)

Figure 5: キーポイント候補点の絞り込み

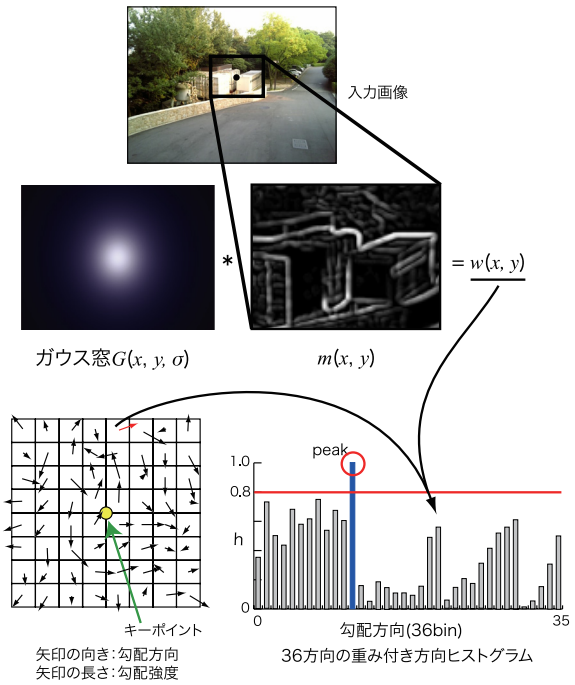


Figure 6: ヒストグラム作成の流れ

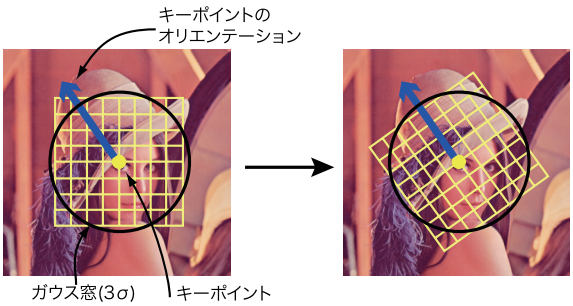


Figure 7: 特徴量を記述する領域

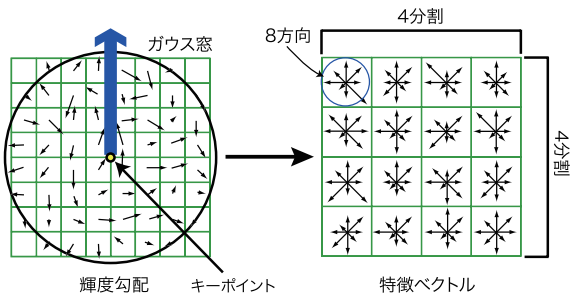


Figure 8: ブロックごとの特徴量記述

3.1.4 特徴量の記述

検出したオリエンテーションを基に，SIFT descriptor により 128 次元の特徴量を記述する．まず，図 7 に示すようにキーポイントのオリエンテーション方向に回転する．特徴量の記述には，キーポイント周辺領域の持つ勾配情報を用いる．使用する勾配情報は，キーポイントを中心とし，そのキーポイントが持つスケールを半径とした円領域内から求める (図 7 中のガウス窓内の領域)．周辺領域を一辺を 4 ブロックの計 16 ブロックに分割し，図 8 に示すようにブロックごとに 8 方向 (45 度ずつ) の勾配方向ヒストグラムを作成する．この勾配方向ヒストグラムは，キーポイントのオリエンテーションを算出したときに作成したヒストグラムと同様の手法で求める．

図 8 の例では $4 \times 4 = 16$ ブロックに各 8 方向のヒストグラムを作成するため，$4 \times 4 \times 8 = 128$ 次元の特徴ベクトルとしてキーポイントの特徴を記述する．このように，キーポイントが持つオリエンテーション方向に座標軸をあわせた領域で特徴を記述するため，回転に不変な特徴量となる．また，128 次元の各特徴ベクトルの長さはベクトルの総和で正規化する．これにより，キーポイントは照明変化に対して影響の少ない特徴量となる．

3.1.5 SIFT を用いた対応点探索

異なる画像間で抽出された各キーポイントの SIFT 特徴量を比較することで，画像間の対応点探索が可能となる．あるキーポイント 1 点に対して，異なる画像中に含まれる全キーポイントとの特徴量間のユークリッド距離 d を算出し，その中で最も d が最小となる点同士を対応点として検出する．図 9 に同じ方向から距離を変えて撮影した画像のマッチング例を示す．このように，SIFT 特徴量を用いると，スケール変化に影響を受けず，対応点の検出が可能であることがわかる．

3.2 Bag-of-Keypoints による画像分類

SIFT は局所領域のマッチングを頑健に行うことができるため，図 10(a) のように同一人物における対応点を SIFT

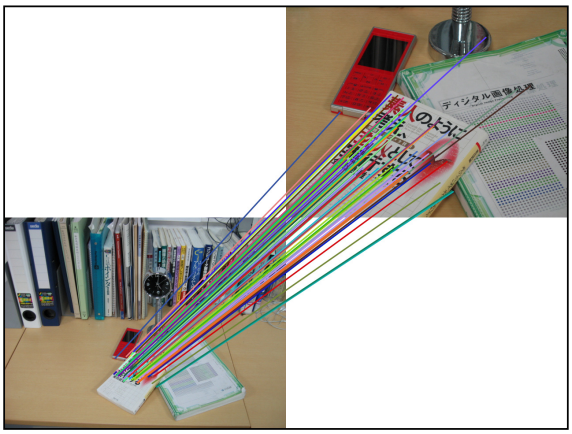


Figure 9: 対応点探索例

Figure 10: SIFT による顔画像間のマッチング例

により求めることが可能である．そのため，特定物体の同定には有効である．しかし，図 10(b) のように異なる人物においては SIFT による対応点を求めることができないため，一般物体認識におけるカテゴリ分類には，SIFT 特徴量をそのまま使用することは困難である．そこで，SIFT 特徴量をベクトル量子化することによって，keypoint を visual word として捉える Bag-of-Keypoints[2] というアプローチが提案されている．

3.2.1 Bag-of-Keypoints

Bag-of-Keypoints は，文書分類手法である Bag-of-words [6] を画像に適用した手法であり，Bag-of-words で文章を単語の集合と見なし，単語の語順を無視してその頻度で文章の分類を行うのと同様に，画像を局所特徴量 (keypoint) の集合と見なし，その位置情報を無視して画像の認識を行う．図 11 に顔とバイクの画像の visual word による表現例を示す．SIFT 特徴量をベクトル量子化した visual word を用いることで，入力画像における対象物の拡大縮小や，回転に影響を受けない画像分類が可能となる．

図 12 に SIFT を用いた Bag-of-Keypoints の流れを示す．Bag-of-Keypoints では，事前に各物体カテゴリの学習画像から SIFT 特徴量を抽出し，局所特徴量をベクトル量子化

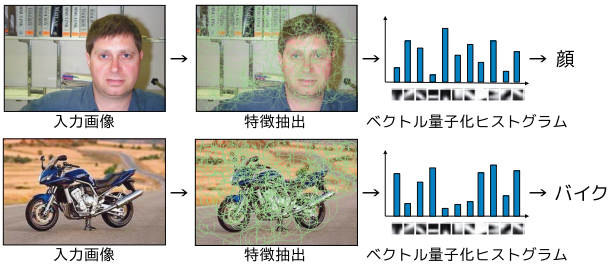


Figure 11: visulal word 表現

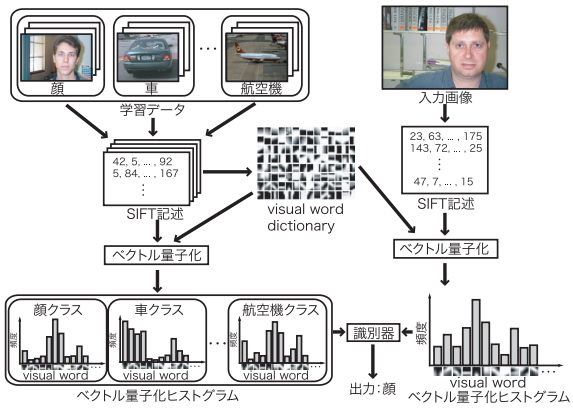


Figure 12: Bag-of-Keypoints の流れ

する．このベクトル量子化された特徴量は visual word や visual alphabet と呼ばれ，物体カテゴリ毎の visual word のヒストグラムを用いて学習画像群から識別器を作成しておく．1 枚の未知入力画像から得られた visual word のヒストグラムをその未知入力画像の特徴量として，各物体カテゴリの識別器に入力し判定する．

3.2.2 Bag-of-Keypoints の拡張

画像分類に有効な Bag-of-Keypoints は様々な拡張手法が提案されている．Manning ら [2] はキーポイントの検出に Affine Invariant keypoint [7] を利用し，領域をアフィン変換して SIFT descriptor により特徴量を記述することで，アフィン変換に頑健な特徴量に基づくカテゴリ分類を実現している．Fei-Fei ら [8] は風景画像など 13 クラスの画像分類法を提案している．この手法では，自然風景シーンではエッジやコーナーといった特徴点の抽出が困難であるため，画像を等間隔に分割し，ランダムに決定したスケールで SIFT 特徴量を記述している．キーポイントをグリッドに分割することで，DoG 極値を用いる場合よりも高精度な分類が可能であることを報告している．Agarwal らは，画像中の visual word のヒストグラムを局所領域で作成し上位階層の特徴量を計算し，これを繰り返すことで階層的な特徴量の記述を行う Hyperfeature[9] を提案している．Nagahashi ら [10] は，visual word を構造ごとに分割した領域で作成することで，識別率を向上させている．Sivic ら [11] は，ビデオ中の各フレームに含まれる visual word を利用し，テキスト検索の手法を応用することで，視点の異

なる同一シーンを高速に検索できる Video Google を提案している．このように，Bag-of-Keypoints アプローチにおける特徴量として，スケールと回転に不変な特徴量を得ることができる SIFT が用いられている．

4　HOG と物体検出

特定の物体検出のための局所特徴量として，Histograms of Oriented Gradients(HOG)[12] が提案されている．HOG は，SIFT と同様に局所領域における輝度の勾配方向をヒストグラム化した特徴量である．SIFT と類似した特徴量の記述を行うが，SIFT は特徴点に対して特徴量を記述するのに対し，HOG ではある一定領域に対する特徴量の記述を行う．そのため，大まかな物体形状を表現することが可能であり，人検出 [12]-[13] や車検出 [14] 等の物体検出に用いられている．

4.1　HOG のアルゴリズム

HOG 特徴量を算出するためには，画像から輝度勾配を算出し，算出された勾配強度と勾配方向から輝度の勾配方向ヒストグラムを作成し，正規化を行う．以下に HOG 算出アルゴリズムについて述べる．

4.1.1　輝度勾配算出

各ピクセルの輝度から SIFT と同様に勾配強度 m と勾配方向 θ を次式より算出する．

$$m(u,v) = \sqrt{f_u(u,v)^2 + f_v(u,v)^2} \tag{9}$$

$$\theta(u,v) = \tan^{-1}\frac{f_v(u,v)}{f_u(u,v)} \tag{10}$$

$$\begin{cases} f_u(u,v) = I(u+1,v) - I(u-1,v) \\ f_v(u,v) = I(u,v+1) - I(u,v-1) \end{cases} \tag{11}$$

4.1.2　セルによるヒストグラム化

図 13 に示すように，算出された勾配強度 m と勾配方向 θ を用いて，5×5 ピクセルをセルとした領域において輝度の勾配方向ヒストグラムを作成する．輝度の勾配方向ヒストグラムは，$0^\circ - 180^\circ$ を 20° ずつに分割するため，9 方向の勾配方向ヒストグラムとなる．

4.1.3　ブロックによる正規化

各セルで作成した輝度の勾配方向ヒストグラムを 3×3 セルを 1 ブロックとして正規化を行う．i 行 j 列のセル (i,j) の特徴量 (9 次元) を $F_{ij} = [f_1, f_2, \cdots, f_9]$ とすると，k 番目のブロックの特徴量 (81 次元) は $V_k = [F_{i\,j}, F_{i+1\,j}, F_{i+2\,j}, F_{i\,j+1}, F_{i+1\,j+1}, F_{i+2\,j+1}, F_{i\,j+2}, F_{i+1\,j+2}, F_{i+2\,j+2}]$ と表すことができる．正規化後の特徴量を v としたとき，次式より正規化する．

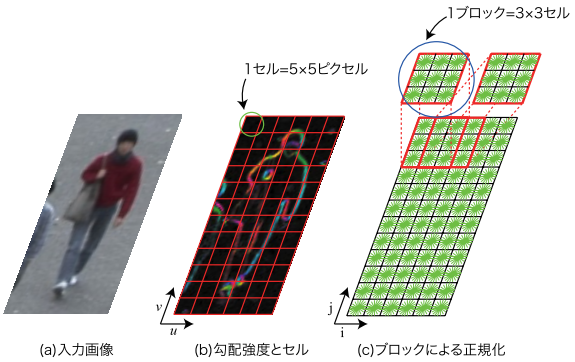


Figure 13: HOG で用いる領域の構造

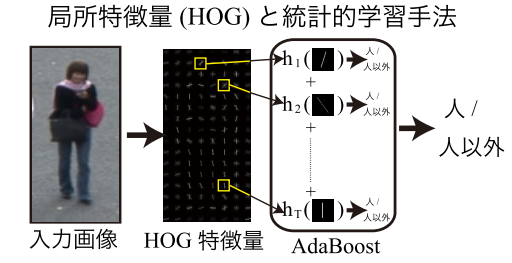
$$v = \frac{f}{\sqrt{||V_k||_2^2 + \epsilon^2}} \qquad (\epsilon = 1) \tag{12}$$

正規化は，図 13(c) のようにブロックを 1 セルずつ移動させることによって正規化を行う．そのため，特徴量 f は異なるブロックの領域によって何度も正規化される．入力画像を 30×60 ピクセルとした場合，横方向に 4 ブロック，縦方向に 10 ブロック，合計 40 ブロックに対して正規化を行う．各ブロックごとに正規化された HOG 特徴量は，40 ブロック $\times$ 81 次元 $= 3,240$ 次元となる．

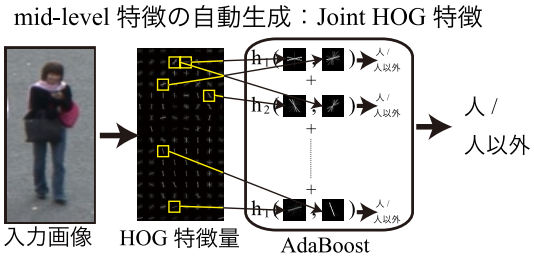
4.2　HOG と統計的学習手法による人検出

人は動きとともに形状が変化する非剛体な物体であるため，顔検出と比較して難しい問題である．また，画像中での人同士の重なりによるオクルージョンの発生や衣服の違い，照明や影の影響も検出を困難とする要因である．このような問題に対して，Dalal 等により局所領域における勾配方向をヒストグラム化した Histograms of Oriented Gradients(HOG) 特徴量と統計的学習手法を組み合わせた人検出法 [12] が提案された．HOG 特徴量は，照明の変動による影響が少なく，局所的な幾何学的変化に頑健であるため高精度な人検出を可能とした．近年では，統計的学習法を用いて HOG のような low-level 特徴量をどのように捉えるかが重要な課題となっている．

図 14(a) に統計的学習手法である AdaBoost による HOG 特徴量の捉え方を示す．学習後の AdaBoost の弱識別器により 1 個の HOG 特徴量が自動で選択され，最終的に多数ある弱識別器の重み付き多数決により人と人以外に判別することが可能となる．このような識別器を画像左上から検出ウィンドウをスケール変化させ複数回ラスタスキャンを行い，各検出ウィンドウに対して人であるかを識別する．人として識別されたウィンドウは，最後に Mean Shift ク



(a) HOG 特徴量 + AdaBoost



(b) Joint HOG 特徴量 + AdaBoost

Figure 14: 人検出における特徴量の捉え方

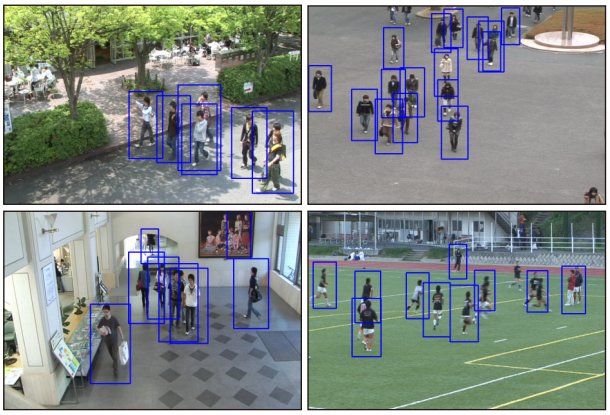


Figure 15: ラスタスキャン走査による人検出例

ラスタリング [15] 等によるウィンドウの統合処理を行う．図 15 に，複数の場所で撮影した映像に対する人検出例を示す．背景が複雑なテクスチャを持つ場合や，人と人との重なりによるオクルージョンが発生した状態においても，歩行者や立っている人を検出することが可能である．

4.3 局所特徴量とブースティング

物体検出に用いられる特徴量は HOG のように low-level 化しており，統計的学習手法であるブースティングとの組み合わせが有効である．ここでは，図 16 に示すように物体検出に用いられる識別器と特徴量の変遷を 3 つの世代に分け，どのように進化してきたかと今後のアプローチについて述べる [16].

4.3.1 識別器と特徴量の変遷

物体検出に用いられる特徴量は，第 1 世代の global な特徴量，第 2 世代の low-level な特徴量，第 3 世代の mid-level

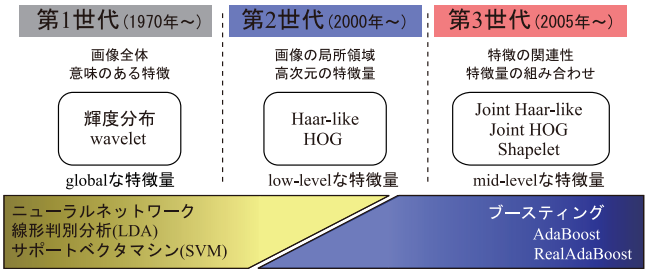


Figure 16: 識別器と特徴量の変遷

な特徴量に分けることができる．以下に各世代毎の特徴量について述べる．

第 1 世代　物体検出に用いられる第 1 世代の特徴量は，輝度分布や wavelet[17] など画像全体から得られるものや，形状の複雑度のような意味のある特徴量を研究者・開発者が決定していた．第 1 世代では識別器としてニューラルネットワークが用いられており，次元数が大きな特徴量では，ニューラルネットワークのサイズが大きくなり，構築が困難となる．そのため，入力ユニットの数を小さくする必要があり，低次元数の特徴量が用いられていた．

第 2 世代　2000 年以降に提案された第 2 世代では，Haar-like 特徴 [3] や HOG 特徴などの low-level な局所特徴量から，統計的学習法であるブースティングを用いて特徴選択を行うため，高次元 (数千～数十万パターン) の特徴量を扱うことが可能となった．これにより，第 1 世代では研究者が特徴量を決定していたが，第 2 世代では研究者の事前知識なく，多数の特徴量候補の中から共通する有効な特徴量を自動的に選択することができる．

第 3 世代　第 3 世代 (2005 年以降) では，第 2 世代の low-level 特徴量を AdaBoost の特徴選択を利用して組み合わせ，識別に有効な関連性や共起性を表現した Joint Haar-like[18] や Joint-HOG[19] や Shapelet[20] が登場した．これらは，low-level 特徴量である第 2 世代の特徴量をベースとし，1 つ 1 つでは意味のない low-level な特徴量に対して，関連性を考慮することで何らかの意味をもつ mid-level な特徴量に拡張している．

次節に，第 3 世代の特徴量として Joint-HOG について述べる．

4.3.2 2 段階 AdaBoost による Joint HOG 特徴

人には，形状の左右対称性や連続したエッジがあり，これらの特徴を捉えることで検出精度を向上させることができると考えられる．三井らは，人独特の形状を捉えるために，複数の HOG 特徴量を組み合わせた Joint HOG 特徴量と，2 段階に構築した AdaBoost による学習法 [19] を提案した．複数の low-level な特徴量である HOG 特徴量

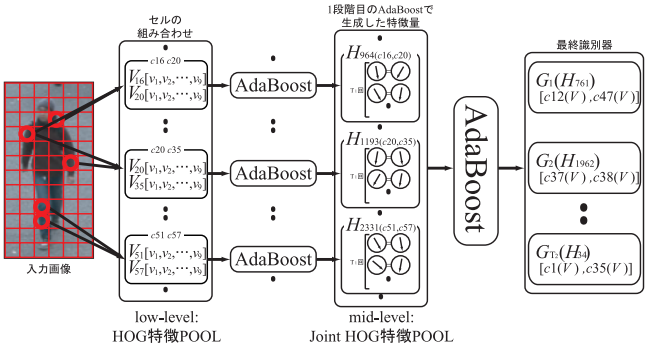


Figure 17: 2 段階 AdaBoost による Joint HOG

を AdaBoost により組み合わせることで mid-level な特徴量である Joint HOG 特徴の POOL を自動的に生成し，この Joint HOG 特徴を入力とした 2 段階目の AdaBoost により，識別に有効な Joint HOG 特徴を自動的に選択して最終識別器を構築する (図 17).

図 14(b) に，AdaBoost による Joint HOG 特徴の捉え方を示す．HOG 特徴量では，1 個の弱識別器が 1 個の HOG 特徴量を用いて識別したのに対し，Joint HOG 特徴では，1 個の弱識別器が位置の異なる 2 つの領域内に含まれる複数の HOG 特徴量を用いて識別を行う．これにより，従来の単一の HOG 特徴量のみでは捉えることができない物体形状の対称性や連続的なエッジを自動的に捉えることが可能となり，高精度な人検出を実現した．

5 おわりに

本稿では，一般物体認識において顔検出や人検出に代表される物体検出と Bag-of-Keypoints による画像分類手法と，その入力特徴量として SIFT と HOG について述べた．これらの特徴量は，統計的学習手法の進化が伴い有効なアプローチが提案されるようになった．今後は，局所特徴量間の共起やコンテキストの利用による認識能力の高精度化と，より多くの物体カテゴリを対象とするために一層の処理の高速化が期待される．一般物体認識はコンピュータビジョンにおける究極の問題であり，研究者・開発者間で問題を共有し，このような難しい問題にチャレンジしていくことによって本分野が大きく前進することを期待する．

参考文献

[1] 柳井啓司, “ 一般物体認識の現状と今後”, 情報処理学会論文誌. コンピュータビジョンとイメージメディア, vol. 48, no. 16, pp. 1-24, 2007.

[2] G. Csurka, C.R. Dance, L. Fan, and C. Bray, “Visual categorization with bags of keypoints”, Proc. of European Conference on Computer Vision (ECCV), pp. 1-22, 2004.

[3] P. Viola and M. Jones, “Robust Real-Time Face Detection”, Int. Journal of Computer Vision, 57(2), pp. 137-154, 2004.

[4] D. Lowe, “Distinctive image features from scale-invariant keypoints”, Proc. of International Journal of Computer Vision (IJCV), 60(2), pp. 91-110, 2004.

[5] M. Brown and D. G. Lowe, “Recognising panoramas”, Proc. of IEEE International Conference on Computer Vision (ICCV), pp. 1218-1225, Nice, France, October, 2003.

[6] C. D. Manning, and H. SchFutze, “Foundation of statistical natural language processing”, The MIT Press, 1999.

[7] K. Mikolajczyk, and C. Schmid, “An affine invariant interest point detector”, Proc. of European Conference on Computer Vision (ECCV), pp. 128-142, 2002.

[8] L. Fei-Fei, and P. Perona, “A bayesian hierarchical model for learning natural scene categories”, Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol 2, pp. 524 - 531, 2005.

[9] A. Agarwal, and B. Triggs, “Hyperfeatures – multilevel local coding for visual recognition”, Proc. of European Conference on Computer Vision (ECCV), vol. 1, pp 30-43, 2006.

[10] T. Nagahashi, H. Fujiyoshi, T. Kanade, “Object type classification using structure-based feature representation”, MVA2007：IAPR Conference on Machine Vision Applications, pp. 142-145, May, 2007.

[11] J. Sivic, and A. Zisserman, “Video google: A text retrieval approach to object matching in videos”, Proc. of IEEE International Conference on Computer Vision (ICCV), vol. 2, pp. 1470-1477, 2003.

[12] N. Dalal, B. Triggs, “Histograms of oriented gradients for human detection”, Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 886-893, 2005.

[13] 山内悠嗣, 藤吉弘亘, Bon-Woo Hwang, 金出武雄, “アピアランスと時空間特徴の共起に基づく人検出”, 第 10 回画像の認識・理解シンポジウム (MIRU2007), pp. 1492-1497, Jul, 2007.

[14] F. Han, Y. Shan, R. Cekander, H. S. Sawhney and R. Kumar, “A two-stage approach to people and vehicle detection with HOG-based SVM”, Proc. of Workshop on Performance Metrics for Intelligent Systems, pp. 133-140, 2006.

[15] D. Comaniciu and P. Meer, “Mean shift analysis and applications”, International Conference on Computer Vision, pp. 1197-1203, 1999.

[16] 山下隆義, “統計的学習法を用いた物体認識における特徴量の進化”, 第 14 回画像センシングシンポジウム SSII08, OR4-01, 2008.

[17] H. Schneiderman, T. Kanade , “A statistical method for 3d object detection applied to faces and cars”, IEEE Computer Vision and Pattern Recognition, pp. 746-751, 2000.

[18] T. Mita, T. Kaneko, B. Stenger, O. Hori, “Discriminative Feature Co-occurrence Selection for Object Detection”, IEEE Pattern Analysis and Machine Intelligence, vol. 30, no. 7, pp. 1257-1269, 2008.

[19] 三井相和, 山内悠嗣, 藤吉弘亘, “Joint HOG 特徴を用いた 2 段階 AdaBoost による人検出”, 第 14 回画像センシングシンポジウム SSII08, IN1-06, 2008.

[20] P. Sabzmeydani and G. Mori, “Detecting Pedestrians by Learning Shapelet Features”, IEEE Computer Vision and Pattern Recognition, pp1-8, 2007.

中部大学工学部情報工学科
愛知県春日井市松本町 *1200*
Phone: 0568-51-1111(4414), Fax: 0568-51-1540
E-mail: hf@vision.cs.chubu.ac.jp