Gradient ベースの特徴抽出
-SIFT と HOG-

藤吉 弘亘

中部大学 工学部 情報工学科

E-mail: hf@cs.chubu.ac.jp

あらまし Scale-Invariant Feature Transform(SIFT) は，特徴点の検出と特徴量の記述を行うアルゴリズムである．検出した特徴点に対して，画像の回転・スケール変化・照明変化等に頑健な特徴量を記述するため，イメージモザイク等の画像のマッチングや物体認識・検出に用いられている．本稿では，SIFT のアルゴリズムについて概説し，具体例として SIFT を用いたアプリケーションや応用手法への展開について紹介する．また，SIFT と同様に gradient ベースの特徴抽出法である Histograms of Oriented Gradients(HOG) のアルゴリズムとその応用例として人検出についても紹介する．

Gradient-Based Feature Extraction
-SIFT and HOG-

Hironobu Fujiyoshi

Dept. of Computer Science, Chubu University

E-mail: hf@cs.chubu.ac.jp

Abstract 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 HOG can be used for people detection.

1 はじめに

画像間の対応点を求めるために必要な局所特徴量を抽出するには，対象となる画像から特徴点を検出する必要がある．Harris らは，1988 年に特徴点としてコーナーを検出する手法 (Harris Corner Detector)[1] を提案した．Lindeberg はスケールスペースを用いることで画像の構造を解析し，blob の検出と自動スケール選択を行う手法 [2] を提案した (1994 年)．また，Schmid らは Harris corner detector によって検出された特徴点に対し，その点の画素値や微分値から算出した値を特徴量とし，画像の回転に頑健な局所特徴量を記述した [3](1997 年)．これにより，回転変化が生じても画像間のマッチングや認識を行うことが可能となった．しかし，Schmid らの手法に用いられている Harris corner detector は，画像のスケール変化に敏感であるため拡大・縮小等の異なるスケールの画像間ではマッチングが困難である．Lowe は Schmid らの局所領域の特徴量記述という考えを拡張し，スケールスペースを用いることで，画像のスケール変化や回転に不変な特徴量を記述する Scale-Invariant Feature Transform(SIFT) を提案した [4]．SIFT は，回転・スケール変化等に

不変な特徴量を記述するため，イメージモザイク等の画像のマッチングや物体認識に用いられている．さらに，PCA(Principal Component Analysis) を用いて勾配情報を部分空間へ射影しマッチング精度を向上させる PCA-SIFT や，SIFT の特徴量記述時における背景の影響を軽減する BSIFT 等の SIFT を拡張した手法が多く提案されている [20]-[23].

本稿では，SIFT のアルゴリズムについて概説し，SIFT を用いたアプリケーション例と SIFT の拡張手法について紹介する．また，SIFT に類似した gradient ベースの特徴抽出法である Histograms of Oriented Gradients(HOG)[24] についてもアルゴリズムを概説し，一般物体認識として，人検出の応用例を紹介する．

2 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．

2.1 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．

2.1.1 LoG によるスケール探索

特徴点のスケール探索には，ガウス関数が有効であることが Koenderink[6] や Lindeberg[2] により報告されている．Lindeberg は，ガウシアンカーネルを用いたスケールスペースとして Scale-normalized Laplacian-of-Gaussian(以下，LoG と呼ぶ) を提案している．LoG は，画像にスケール σ を変化させながら式 (1) で表される LoG オペレータ (図 1) を適用し，その極大位置を特徴点のスケールと決定する．

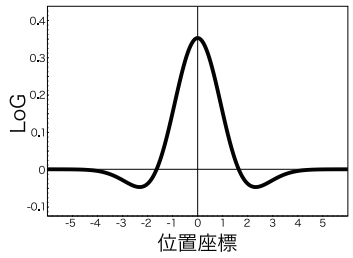


図 1: LoG オペレータ

$$LoG = f(\sigma) = -\frac{x^2+y^2-2\sigma^2}{2\pi\sigma^6}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \quad (1)$$

σ はガウシアンフィルタのスケール，x と y は注目画素からの距離である．LoG は計算コストが高いという問題があり，Lowe[4] によってより効率的な極値検出法として Difference-of-Gaussian(DoG) を用いる手法が提案されている．DoG と LoG の関係は次式の拡散方程式から導かれる．

$$\frac{\partial G}{\partial \sigma} = \sigma\nabla^2 G \quad (2)$$

G はガウス関数であり，右辺はガウス関数の 2 次微分 (LoG) である．式 (2) は次式のように表すことができる．

$$\frac{\partial G}{\partial \sigma} \approx \frac{G(x,y,k\sigma)-G(x,y,\sigma)}{k\sigma-\sigma} \quad (3)$$

式 (2) と式 (3) から次式が成り立つ．

$$\sigma\nabla^2 G = \frac{\partial G}{\partial \sigma} \approx \frac{G(x,y,k\sigma)-G(x,y,\sigma)}{k\sigma-\sigma} \quad (4)$$

したがって次式が得られる．

$$(k-1)\sigma^2\nabla^2 G \approx G(x,y,k\sigma)-G(x,y,\sigma) \quad (5)$$

ここで，$\sigma^2\nabla^2 G$ は LoG であるので，式 (5) は DoG が LoG の近似であることを表している．LoG と DoG から得られる結果がほぼ同じであるため，SIFT では計算効率の良い DoG を用いる．

2.1.2 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．

$$L(u, v, \sigma) = G(x, y, \sigma) * I(u, v) \quad (6)$$

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \quad (7)$$

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．

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

$$= L(u, v, k\sigma) - L(u, v, \sigma) \quad (8)$$

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 2 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．

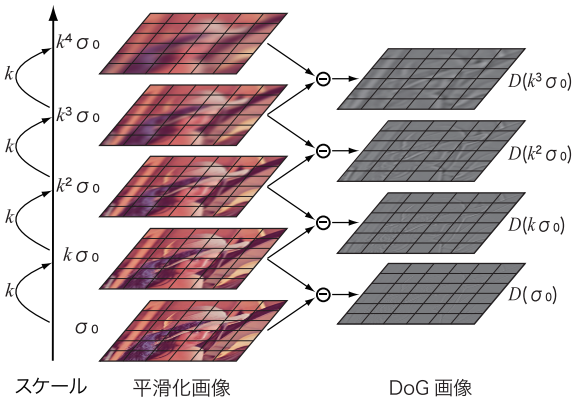


図 2: DoG 処理の流れ

2.1.3 σ の連続性を保持した平滑化処理

σ の連続性を保持した効率的な平滑化処理を図 3 に示す．はじめに，入力画像を初期値である σ_0 で平滑化を行い，平滑化画像 $L_1(\sigma_0)$ を得る．次に σ_0 を k 倍した値 $k\sigma_0$ で平滑化を行い $L_1(k\sigma_0)$ を得る．同様の処理により，σ の異なる複数の平滑化画像を得る．ここまでの処理 1 セットを 1 オクターブとする．

次に，複数生成された平滑化画像の中から $2\sigma_0$ で平滑化された画像 $L_1(2\sigma_0)$ を 1/2 のサイズにダウンサンプリングする．1 オクターブにおける処理回数については増加率 k の設定とともに後述する．1/2 のサイズにダウンサンプリングされた画像 $L_2(\sigma_0)$ と，$2\sigma_0$ で平滑化を行った画像 $L_1(2\sigma_0)$ には以下のような関係が成り立つ．

$$L_1(2\sigma_0) \approx L_2(\sigma_0) \quad (9)$$

この関係を利用することで，σ の最大値を制限することができるため，ガウシアンフィルタのウィンドウサイズによる計算量の増加を防ぐことができる．

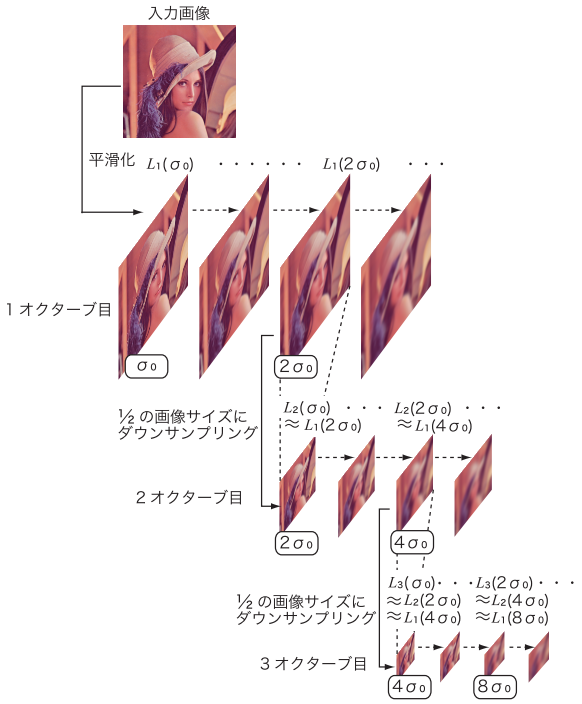


図 3: σ の連続性を保持した効率的な平滑化処理

2.1.4 増加率 k

σ の増加率 k は，1 オクターブにおけるスケールスペースの分割数により決定する．スケールスペースの分割数を s とした場合，1 オクターブでは，スケールスペースは σ_0 から $2\sigma_0$ まで増加するため，σ の増加率 k は $k = 2^{1/s}$ となる．図 4 に示す DoG 処理の例では，$s = 2$(分割数 2) であるため，$k = 2^{1/2} = \sqrt{2}$ となる．極値探索には DoG 画像を 3 枚 1 組で処理する必要があるため，s 枚の極値検出の対象となる画像を得るためには $s + 2$ 枚の DoG 画像が必要となる．さらに，$s + 2$ 枚の DoG 画像を得るためには

$s+3$ 枚の平滑化画像が必要になる．したがって，1 オクターブにおける平滑化の回数は $s+3$ 回となる．ここで求まる極値検出対象画像は次章で行う極値検出に用いるものである．文献 [5] では，実験により分割数 s=3，初期値 $\sigma_0 = 1.6$ のとき，最適なキーポイントを得ることができると報告されている．

1 枚の入力画像に対するオクターブ数は，入力画像のサイズに依存する．入力画像は，処理が進むにつれて 1/2 のサイズにダウンサンプリングされる．ダウンサンプリングを続けた結果，画像の一辺のサイズがある値以下になったとき処理を終了する．この値は任意で決定する．この値が大きければ，少ないオクターブ数になり，小さければ多いオクターブ数になる．例えば 640×480 ピクセルの画像に対して，ダウンサンプリング後の最小のサイズを 10 とした場合，6 回目のダウンサンプリングで一辺が 7 ピクセルとなり処理が終了するため，オクターブ数は 5 となる．

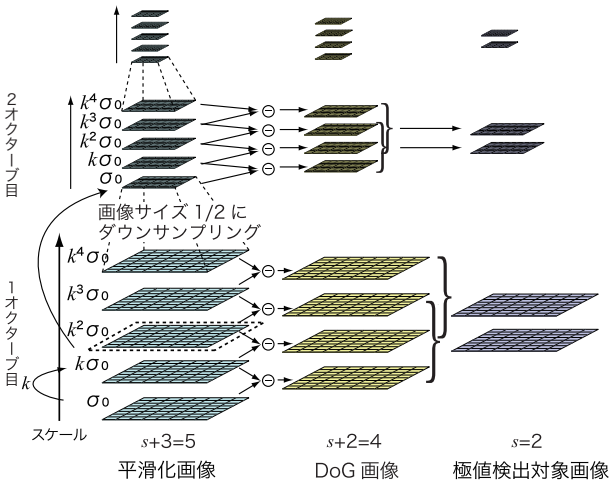


図 4: $s = 2(k = \sqrt{2})$ のときの DoG 処理例

2.1.5 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 5 のように DoG 画像 3 枚一組で行う．DoG 画像 (図 5 中の点線で囲まれた画像) の注目画素 (図 5 中の黒色領域) と，その周りの 26 近傍 (図 5 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．

次に，スケールスペースの極値の性質について述べる．画像中のある座標におけるスケール変化と DoG 出力値の推移を図 6 に示す．実線の円で示すスケールサイズのとき，右に示すグラフから DoG 出力が最大値 (極大値) となることがわかる．図 6(a) の原画像を 2 倍に拡大した (b) においても，実線で示すスケールにて DoG の値が最大となる．このとき，図 6(a) の DoG の極値を σ_1，図 6(b) を σ_2 とすると，$\sigma_2 = 2\sigma_1$ の関係が成り立つ．このように，画像サイズが 2 倍になると，DoG の極値探索により検出されたキーポイントのスケール σ も比例して 2 倍となる．SIFT は，特徴を最も含むスケール σ を自動的に決定するため，空間的に同範囲の領域から特徴量を記述することで，拡大・縮小に不変な特徴量となる．

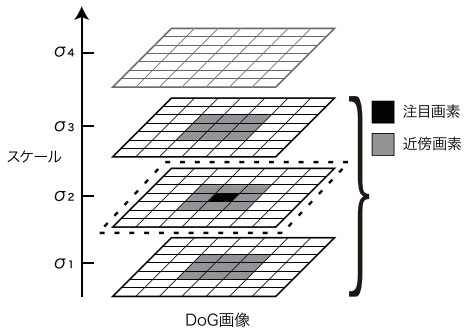


図 5: 極値検出の流れ

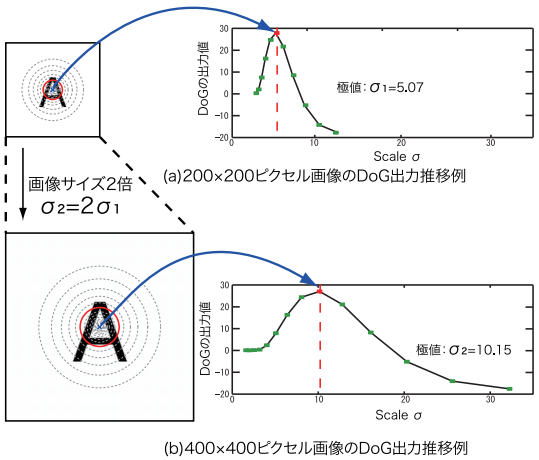
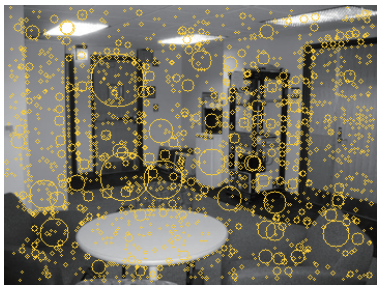
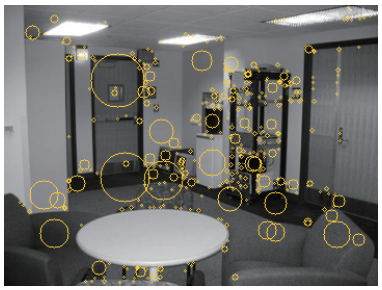


図 6: スケールと DoG 出力の関係



(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)

図 7: キーポイント候補点の絞り込み

2.2 キーポイントのローカライズ

2.1 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．

2.2.1 主曲率によるキーポイントの絞り込み

エッジ上に存在するキーポイント候補点の削除方法について述べる．キーポイント候補点における 2 次元ヘッセ行列 $\mathbf{H}$ を次式により計算し，主曲率を求める．

$$\mathbf{H} = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \tag{10}$$

行列内の導関数は，キーポイント候補位置での DoG 出力値の 2 次微分から得られる．ここで，ヘッセ行列から求められる第 1 固有値を α，第 2 固有値を $\beta(\alpha > \beta)$ とする．このときヘッセ行列の対角成分の和 $\mathrm{Tr}(\mathbf{H})$ と行列式 $\mathrm{Det}(\mathbf{H})$ は次のように計算できる．

$$\mathrm{Tr}(\mathbf{H}) = D_{xx} + D_{yy} = \alpha + \beta \tag{11}$$

$$\mathrm{Det}(\mathbf{H}) = D_{xx}D_{yy} - (D_{xy})^2 = \alpha\beta \tag{12}$$

さらに，γ を第 1 固有値と第 2 固有値の比率とし，$\alpha = \gamma\beta$ とすると次式のようになる．

$$\frac{\mathrm{Tr}(\mathbf{H})^2}{\mathrm{Det}(\mathbf{H})} = \frac{(\alpha+\beta)^2}{\alpha\beta} = \frac{(\gamma\beta+\beta)^2}{\gamma\beta^2} = \frac{(\gamma+1)^2}{\gamma} \tag{13}$$

この値は固有値そのものではなく，固有値 α, β の比率で決まる．したがって，固有値を求めずにエッジ上の点であるか判別することが可能となる．この値を次式に示すようにしきい値処理することで，不要なキーポイント候補点を削除する．

$$\frac{\mathrm{Tr}(\mathbf{H})^2}{\mathrm{Det}(\mathbf{H})} < \frac{(\gamma_{th}+1)^2}{\gamma_{th}} \tag{14}$$

式 (14) を満足するような点をキーポイント候補とする．しきい値は γ_{th} により決定する．この処理は，ハリスのコーナー検出に良く似たもので，固有値の比率がしきい値より大きい点，つまりエッジ上に存在する点が削除される．文献 [5] では $\gamma_{th} = 10$ を採用しており，しきい値は 12.1 となる．

図 7(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 7(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．

2.2.2 キーポイントのサブピクセル位置推定

3 変数 (x, y, σ) の 2 次関数をフィッティングすることで，キーポイント候補点のサブピクセル位置とスケールを算出する．ある点 $\mathbf{x} = (x, y, \sigma)^T$ での DoG 関数 $D(\mathbf{x})$ をテイラー展開すると次式のようになる．

$$D(\mathbf{x}) = D + \frac{\partial D}{\partial \mathbf{x}}^T \mathbf{x} + \frac{1}{2}\mathbf{x}^T \frac{\partial^2 D}{\partial \mathbf{x}^2}\mathbf{x} \tag{15}$$

式 (15) について $\mathbf{x}$ に関する偏導関数を求め，0 とすると次式が得られる．

$$\frac{\partial D}{\partial \mathbf{x}} + \frac{\partial^2 D}{\partial \mathbf{x}^2}\hat{\mathbf{x}} = 0 \tag{16}$$

このとき $\hat{\mathbf{x}}$ はキーポイント候補点 (極値) のサブピクセル位置を表している．この式を変形し次式を得る．

$$\frac{\partial^2 D}{\partial \mathbf{x}^2}\hat{\mathbf{x}} = -\frac{\partial D}{\partial \mathbf{x}} \tag{17}$$

この式は次のように表される．

$$\begin{bmatrix} \frac{\partial^2 D}{\partial x^2} & \frac{\partial^2 D}{\partial xy} & \frac{\partial^2 D}{\partial x\sigma} \\ \frac{\partial^2 D}{\partial xy} & \frac{\partial^2 D}{\partial y^2} & \frac{\partial^2 D}{\partial y\sigma} \\ \frac{\partial^2 D}{\partial x\sigma} & \frac{\partial^2 D}{\partial y\sigma} & \frac{\partial^2 D}{\partial \sigma^2} \end{bmatrix} \begin{bmatrix} x \\ y \\ \sigma \end{bmatrix} = - \begin{bmatrix} \frac{\partial D}{\partial x} \\ \frac{\partial D}{\partial y} \\ \frac{\partial D}{\partial \sigma} \end{bmatrix} \tag{18}$$

式 (18) をキーポイント候補点のサブピクセル位置 $\hat{\mathbf{x}}$ を得るために変形する．

$$\begin{bmatrix} x \\ y \\ \sigma \end{bmatrix} = - \begin{bmatrix} \frac{\partial^2 D}{\partial x^2} & \frac{\partial^2 D}{\partial xy} & \frac{\partial^2 D}{\partial x\sigma} \\ \frac{\partial^2 D}{\partial xy} & \frac{\partial^2 D}{\partial y^2} & \frac{\partial^2 D}{\partial y\sigma} \\ \frac{\partial^2 D}{\partial x\sigma} & \frac{\partial^2 D}{\partial y\sigma} & \frac{\partial^2 D}{\partial \sigma^2} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial D}{\partial x} \\ \frac{\partial D}{\partial y} \\ \frac{\partial D}{\partial \sigma} \end{bmatrix} \tag{19}$$

得られた式 (19) を解くことにより，キーポイント候補点のサブピクセル位置 $\hat{\mathbf{x}} = (x, y, \sigma)$ を得る．

2.2.3 コントラストによるキーポイントの絞り込み

サブピクセル位置での DoG 出力を算出し，コントラストによるキーポイントの絞り込みを行う．式 (19) は次のように表される．

$$\hat{\mathbf{x}} = -\frac{\partial^2 D}{\partial \mathbf{x}^2}^{-1}\frac{\partial D}{\partial \mathbf{x}} \tag{20}$$

式 (20) を式 (15) に代入すると次式が得られる．

$$D(\hat{\mathbf{x}}) = D + \frac{1}{2}\frac{\partial D}{\partial \mathbf{x}}^T \hat{\mathbf{x}} \tag{21}$$

D は DoG 関数であり，$\hat{\mathbf{x}}$ はサブピクセル位置を表しているため，式 (21) はサブピクセル位置での DoG 出力値となる．この DoG の値からキーポイント削除の判別を行う．文献 [5] では，しきい値として 0.03 を用いている．サブピクセル位置での DoG 出力の絶対値がしきい値より小さい場合 (つまり，コントラストが低い場合) ノイズに影響されやすいため削除する．図 7(c) にコントラストにより絞り込まれたキーポイントを示す．

2.3 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．

$$m(u,v) = \sqrt{f_u(u,v)^2 + f_v(u,v)^2} \tag{22}$$

$$\theta(u,v) = \tan^{-1}\frac{f_v(u,v)}{f_u(u,v)} \tag{23}$$

$$\begin{cases} f_u(u,v) = L(u+1,v) - L(u-1,v) \\ f_v(u,v) = L(u,v+1) - L(u,v-1) \end{cases} \tag{24}$$

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 8 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．

$$h_{\theta'} = \sum_x \sum_y w(x,y) \cdot \delta\left[\theta', \theta(x,y)\right] \tag{25}$$

$$w(x,y) = G(x,y,\sigma) \cdot m(x,y) \tag{26}$$

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．

図 8 の例ではキーポイントに割り当てられるオリエンテーションは 1 方向のみであるが，図 9 のようにコーナーのようなキーポイントでは 2 方向となり，2 つのオリエンテーションを持つ．このように，1 つのキーポイントに対して複数のオリエンテーションが割り当てられる場合がある．

2.4 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 10 に示すようにキーポイントのオリエンテーショ

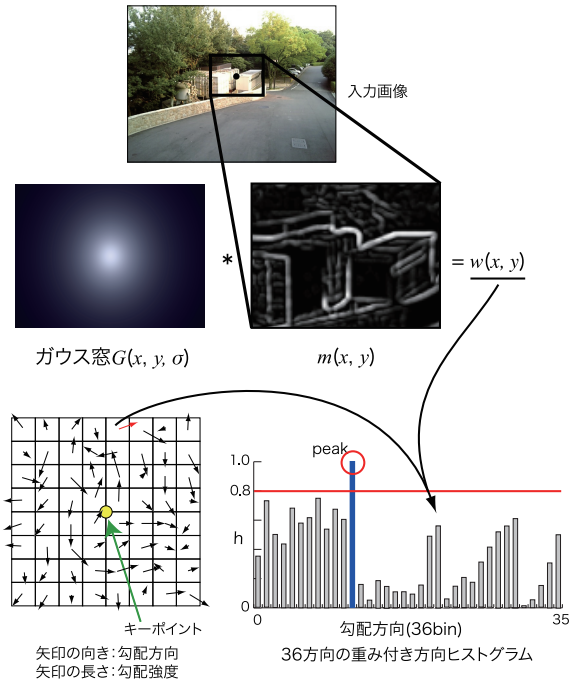


図 8: ヒストグラム作成の流れ

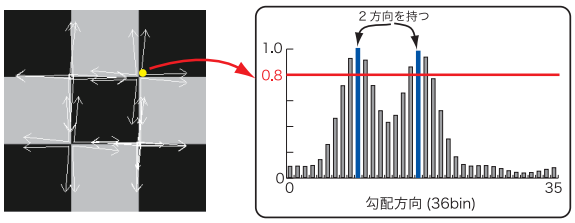


図 9: 2 方向のオリエンテーションを持つキーポイント

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 10 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 11 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 11 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．

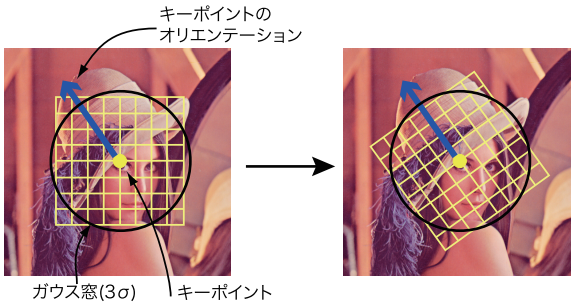


図 10: 特徴量を記述する領域

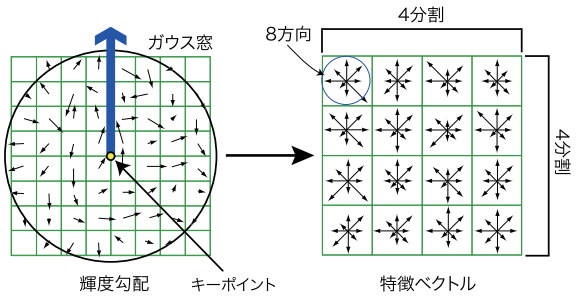
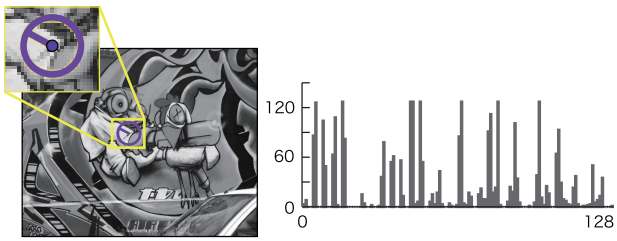


図 11: ブロックごとの特徴量記述

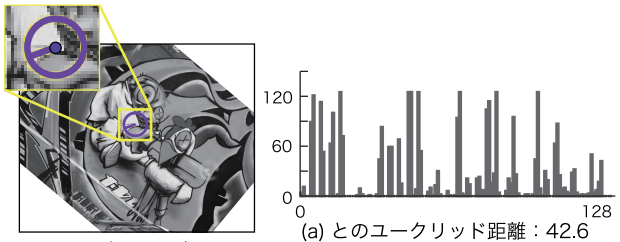
2.5 画像変化に対する SIFT 特徴量

画像の回転やスケール変化に対する SIFT 特徴量の影響について検討する．まず，参照用画像から SIFT によりキーポイントを検出する．次に，参照用画像に対して，回転・スケール変化・照明変化・アフィン変化・JPEG 圧縮の 5 種類の画像を作成する．参照用画像から検出されたあるキーポイント 1 点に注目し，画像変化に対する特徴量の影響について比較する．

図 12 に各変化に対する SIFT 特徴量 (128 次元) を示す．図 12 中の円の中心がキーポイントの位置であり，円はスケールであり特徴を記述する範囲である．また，円の中心から伸びる直線の方向は，そのキーポイントのオリエンテーションである．図 12(b), (c), (d), (e) に示す回転，スケール変化，照明変化，JPEG 圧縮 (ノイズの付加) が発生した画像



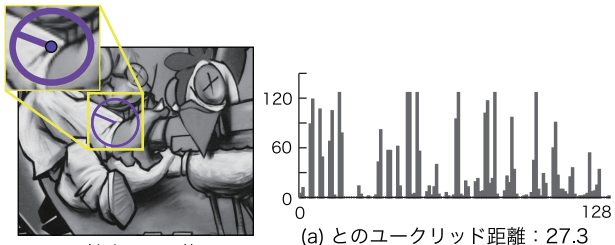
(a) 参照用画像



回転：45 度

(a) とのユークリッド距離：42.6

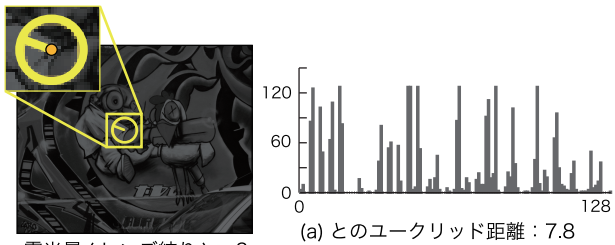
(b) 回転



拡大：2.0 倍

(a) とのユークリッド距離：27.3

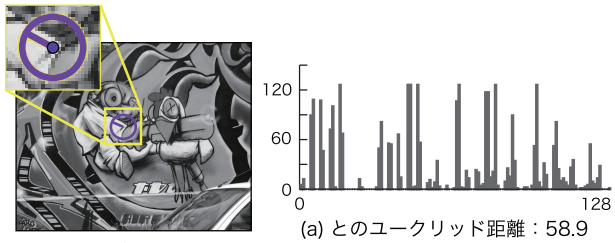
(c) スケール変化



露光量 (レンズ絞り) : -3

(a) とのユークリッド距離：7.8

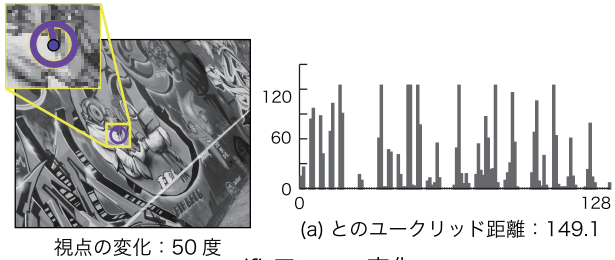
(d) 照明変化



圧縮率：60%

(a) とのユークリッド距離：58.9

(e) JPEG 圧縮



視点の変化：50 度

(a) とのユークリッド距離：149.1

(f) アフィン変化

図 12: 画像変化に対する SIFT 特徴量

と原画像とのユークリッド距離はどれも小さく，同じような特徴量を記述していることがわかる．しかし，図 12(f) に示すアフィン変化の場合は，画像に歪みの変化が含まれるため，スケールと方向を正規化して特徴を記述するだけでは不十分であり，ユークリッド距離が大きくなる．微小なアフィン変化に対してはある程度頑健ではあるが，不変ではないということに注意する必要がある．この問題に対し，Mikolajczyk らはアフィン変化に対応した領域検出器を用いることで，SIFT と同様の特徴記述を行う手法を提案している [15].

3 SIFT を用いたアプリケーション

scholar.google.com[7] で SIFT 論文 [5] の引用件数を調査したところ，911 件であった[1]．そのうち 291 件が SIFT を用いたアプリケーションに関する論文であり，その応用は以下の 4 つに大別できる．

- 対応点探索による画像のマッチング (142 件，49%)
- 特定画像を用いた物体認識 (71 件，24%)
- 画像分類 (73 件，25%)
- 特徴点追跡 (5 件，2%)

括弧内の数値は，SIFT を参照している論文数と，合計引用数 291 に対する各アプリケーションの割合である．本章では，上記の 4 つの SIFT を用いたアプリケーションについて述べる．

3.1 対応点探索による画像のマッチング

SIFT により異なる画像間で抽出された各キーポイントの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．以下に，対応点探索の流れを示す．

画像 $I1$ 中のあるキーポイント k_{I1} と画像 $I2$ 中のあるキーポイント k_{I2} の特徴量をそれぞれ $\mathbf{v}^{k_{I1}}$, $\mathbf{v}^{k_{I2}}$ とすると，特徴量間のユークリッド距離 d は次式により算出される．

$$d(\mathbf{v}^{k_{I1}}, \mathbf{v}^{k_{I2}}) = \sqrt{\sum_{i=1}^{128}(v_i^{k_{I1}} - v_i^{k_{I2}})^2} \qquad (27)$$

ここで，SIFT 特徴の次元数は 128 次元である．図 13 に示すように，あるキーポイント 1 点に対して，

[1] 2007 年 5 月 22 日時点での検索件数

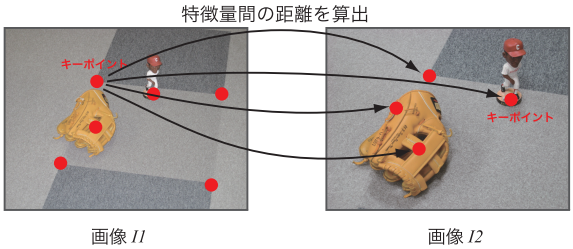


図 13: 対応点探索

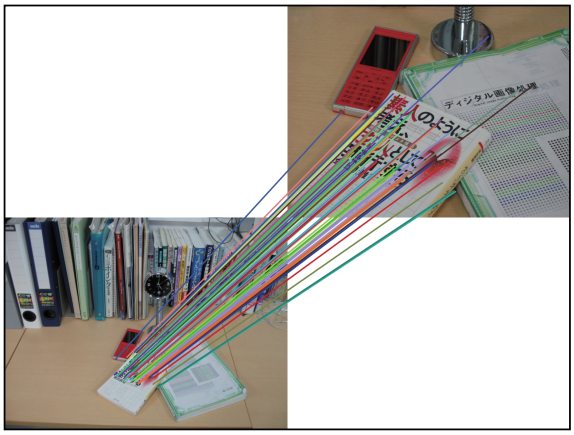


図 14: 対応点探索例

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 14 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． Autostitch[8], [9] は，SIFT を利用したモザイク画像を自動生成するフリーウェアである．Autostitch では，図 15 に示す 1600 × 1200 画素の 3 枚の画像からモザイク画像を約 5 秒で生成する．

3.2 特定画像を用いた物体認識

あらかじめ参照画像から SIFT キーポイントを求めておき，入力画像と対応点探索を行うことで，物体認識が可能となる．Lowe[5] が提案した手法では，テンプレートから検出された各対応点の持つ 2 次元座標，スケール，オリエンテーションの 4 つのパラメータから一般化ハフ変換を用いて投票を行う．3 点以上の投票点からテンプレートと入力画像間のアフィンパラメータを算出する．同様に高木らは，車

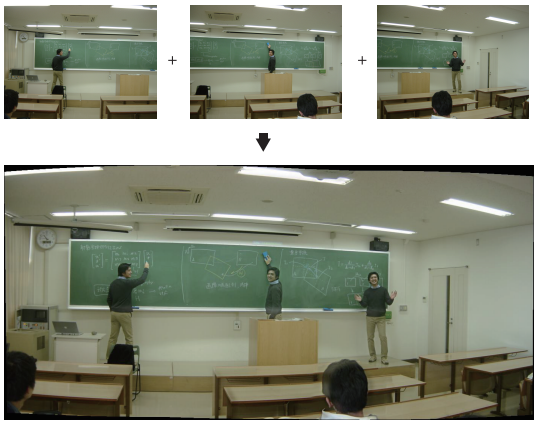


図 15: Autostitch で生成したモザイク画像

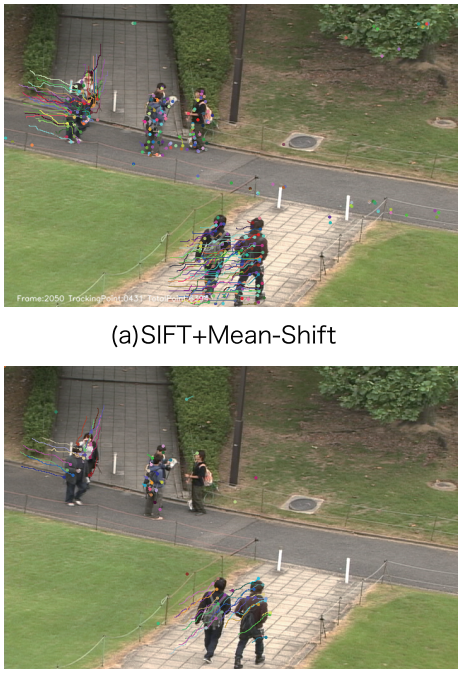
載カメラによって撮影された前方実環境画像から，SIFT 特徴量を用いてイラストパターンの道路標識と実画像のマッチングにより標識を認識する手法を提案している [10]．得られた対応点のスケール，オリエンテーションから，中心位置座標に投票して標識の認識を実現している．図 16 に標識検出・認識結果例を示す．

図 16: 道路交通標識の認識例

3.3 特徴点追跡

従来，特徴点追跡には KLT 法 [11] が用いられている．KLT 法は，局所領域における各点の動きは同一であると仮定し，弛緩法により目的関数を最小化する手法である．微小時間における領域は，平行移動のみしかしない，照明の変化による輝度値の変化がない，という状態を仮定して特徴点の移動先を求める．そのため，対象物体の運動に回転やスケール変化を含む場合や，照明の変化による輝度値の変化が激しい場合，特徴点の追跡に失敗することがある．一方，SIFT は画像の回転・スケール変化・照明変

化に頑健な特徴量を記述することが可能である．したがって，SIFT により抽出された特徴点を追跡対象に用いることで，KLT 法では困難であった回転・スケール変化・照明変化が含まれる場合でも頑健に追跡を行うことが可能である．SIFT を用いた手法として，都築らは SIFT 特徴量の類似度を重みとした Mean-Shift 探索による特徴点追跡を提案し，非剛体の追跡へ応用している [12]．非剛体の追跡例として，図 17 に SIFT を用いた Mean-Shift 探索による人の追跡と KLT 法による人の追跡の結果を示す．図中の各点は特徴点の軌跡を表している．SIFT を用いた特徴点追跡法は KLT 法に比べ，より多くの点を長時間にわたり追跡できていることがわかる．



(a)SIFT+Mean-Shift

(b)KLT 法

図 17: SIFT を用いた Mean-Shift 探索による特徴点追跡

3.4 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，特定物体の同定には有効である．しかし，一般物体認識問題などのクラス分類には，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 を用いた一般物体認識・画像分類手法として，Bag-of-Keypoints[14] というアプローチが提案されている．Bag-of-Keypoints は，文書分類手法である Bag-of-words [13] 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．具体的には，事前にすべての学習画像から SIFT 特徴量を抽出し，その局所特徴量をベクトル量子化する．このベクトル量子化された特徴量は visual word や visual alphabet などと呼ばれる．1 枚の入力画像から得られた visual word のヒストグラムをその画像の特徴量として，識別器を構築する．図 18 に Bag-of-Keypoints の流れを示す．

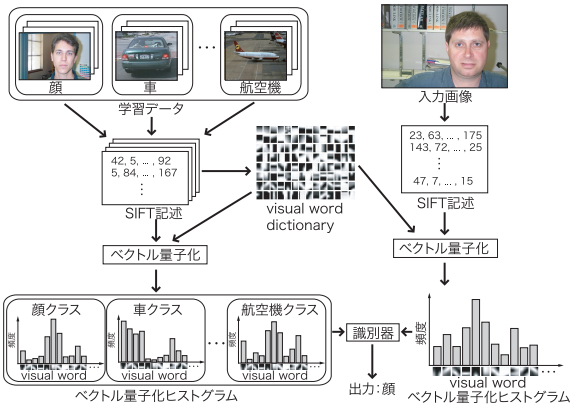


図 18: Bag-of-Keypoints の流れ

Manning ら [14] はキーポイントの検出に Affine Invariant keypoint [15] 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物体認識を実現している．Fei-Fei ら [16] 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 [17] 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 を提案している．Nagahashi ら [18] 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 [19] 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，SIFT が用いられている．

4 SIFT の拡張

SIFT を拡張した手法が多く提案されている [20]-[23]．PCA-SIFT[20] は，SIFT によって検出された局所領域から得られる勾配情報を，PCA を用いて次元圧縮を行い特徴量を記述する手法である．BSIFT[21] は，背景情報 (Background) の影響を軽減して，対象とする物体の SIFT 特徴を記述する手法である．CSIFT[22] は，照明変化による物体の見えの変化の影響を除去し，照明変化に頑健な特徴を記述する手法である．さらに，SIFT を N 次元 (3 次元や 3 次元+時間) に拡張した N-SIFT[23] が提案されている．本章では，SIFT の拡張として PCA-SIFT[20] と BSIFT[21] について述べる．

4.1 PCA-SIFT

PCA-SIFT は，SIFT で検出した局所領域の勾配情報に対して主成分分析 (PCA) を適用し，SIFT 特徴の頑健性や識別性を向上させる手法である．PCA-SIFT では特徴点の検出とオリエンテーションの算出までの処理は SIFT と全く同じであり，特徴量記述のみ異なる．

特徴量記述

SIFT では，キーポイントのスケールに対応した領域を 4×4 のブロックに分割し，ブロックごとに 8 方向の方向ヒストグラムを作成することで 128 次元の特徴量を記述する．一方，PCA-SIFT では，キーポイントのスケールに対応した領域を 41×41 のパッチにリサンプリングする．リサンプリングしたパッチの水平・垂直方向の勾配を算出し，39×39×2=3,042 次元の特徴量を得る．求めた 3,042 次元の特徴量に PCA を適用し，次元圧縮を行う．PCA に用いる射影行列は，あらかじめ学習画像を用いて算出しておく．文献 [20] では，21,000 個のキーポイントから算出したパッチを用いて射影行列を求めている．圧縮する次元数は，実験により 36 次元が最も有効であることが報告されている．

ノイズが加えられた画像，回転・スケール変化がある画像，視点の変化がある画像，輝度を低減させた画像に対して，SIFT と PCA-SIFT の比較実験の結果，どの場合でも PCA-SIFT の性能が SIFT の性能を上回っていることが報告されている．また，PCA-SIFT の特徴量は低次元に圧縮されているため，SIFT よりも高速にマッチングを行うことが可能である．

4.2 BSIFT(Background and Scale Invariant Feature Transform)

SIFT は対象物体以外の背景画素を含んだ特徴量記述を行うため，背景の影響を受けやすいという問題がある．これは特徴点検出と特徴量記述に使用するガウス窓が，物体領域と背景領域を含むからである．Stein ら [21] は，物体と背景の境界情報を用いることで，物体領域のみの情報による特徴点の検出と特徴量を記述する手法である BSIFT を提案している．BSIFT では，境界情報が重要であり，物体の境界情報が既知であるか，もしくはデプス画像等から境界情報が得られた画像を対象とする．

特徴点検出

SIFT は DoG の極値を特徴点として検出する．しかし，この方法では背景情報を含んだまま特徴点の検出を行う．そのため，同一物体でも異なる位置に特徴点が検出される場合がある．そこで，BSIFT では DoG のための平滑化として，ガウス関数ではなく以下の式を用いる．

$$I^{k+1}(u,v) \leftarrow I^k(u,v) + \tau \nabla^2 I^k(u,v) \tag{28}$$

$$\nabla^2 I(u,v) = \frac{\partial^2 I}{\partial u^2} + \frac{\partial^2 I}{\partial v^2} \tag{29}$$

k は繰り返し回数である．式 (29) は，ラプラシアンフィルタであり，式 (28) は 2 次微分を用いた平滑化である．式 (28) を用いることで，物体の境界に注目した平滑化を行うことができる．ここで τ を 0 に近づけたとき，$\sigma = \sqrt{2k\tau}$ でのガウシアンフィルタによる平滑化と同等の結果となる．文献 [21] では，$\tau = 0.2$ としている．

特徴量記述

SIFT による特徴量の記述は，キーポイントを中心としたガウス分布を重みとして輝度勾配ヒストグラムを作成する．しかし，この方法では背景画素の情報も含んだ特徴を記述してしまう．特に境界付近においては背景の影響が大きいため，マッチング失敗

の原因となる．BSIFT では，キーポイントを中心としたガウス分布と境界情報を用いて距離変換した値を掛け合わせて重み分布とする (図 19)．したがって，背景領域の影響を軽減した SIFT 特徴量を得ることができる．

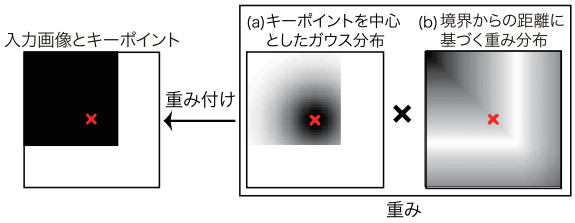


図 19: BSIFT で用いる重み分布

5 Histograms of Oriented Gradients

一般物体認識のための gradient ベースの特徴量として，Histograms of Oriented Gradients(HOG)[24] 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 [24]-[28] や車検出 [26] 等の一般物体認識等に用いられている．

5.1 HOG 特徴量の算出

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．

5.1.1 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．

$$m(u,v) = \sqrt{f_u(u,v)^2 + f_v(u,v)^2} \quad (30)$$

$$\theta(u,v) = \tan^{-1}\frac{f_v(u,v)}{f_u(u,v)} \quad (31)$$

$$\begin{cases} f_u(u,v) = I(u+1,v) - I(u-1,v) \\ f_v(u,v) = I(u,v+1) - I(u,v-1) \end{cases} \quad (32)$$

5.1.2 セルによるヒストグラム化

図 20 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，$0° - 180°$ 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．

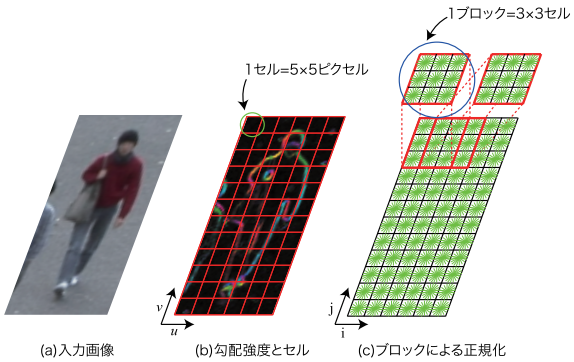


図 20: HOG で用いる領域の構造

5.1.3 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 $\mathrm{V}_k = [\mathrm{F}_{i\ j}, \mathrm{F}_{i+1\ j}, \mathrm{F}_{i+2\ j}, \mathrm{F}_{i\ j+1}, \mathrm{F}_{i+1\ j+1}, \mathrm{F}_{i+2\ j+1}, \mathrm{F}_{i\ j+2}, \mathrm{F}_{i+1\ j+2}, \mathrm{F}_{i+2\ j+2}]$ 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．

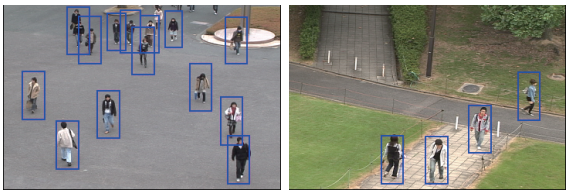
$$v = \frac{f}{\sqrt{||\mathrm{V_k}||_2^2 + \epsilon^2}} \qquad (\epsilon = 1) \qquad (33)$$

正規化は，図 20(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 × 81 次元 $= 3,240$ 次元となる．

5.2 HOG を用いた一般物体認識

HOG 特徴量は，回転やスケール変化に不変ではないが，局所的な幾何学的変化と明度変化に不変であり，図 21(a) に示すように，大まかな形状特徴量を表現可能なため，一般物体認識に用いられている．Dalal らは，HOG 特徴量を算出し，SVM(Support Vector Machine) を用いて人検出する方法を提案した [24]．HOG は，PCA-SIFT や Shape Contexts を特徴量とした手法よりも高精度な人検出が可能であることが報告されている．図 21(b) に HOG を用いた人検出の結果を示す．さらに，アピアランス特徴である HOG に動きの情報としてオプティカルフローを用いた人検出法 [25] や，HOG と時空間特徴の共起を表現することで人検出精度を向上させる手法 [28] が提案されている．

(a)HOG 特徴量抽出例



(b)HOG を用いた人検出例

図 21: HOG 特徴量とその応用例

6 まとめ

本稿では，gradient ベースの特徴抽出法である SIFT と HOG のアルゴリズムとその応用例について紹介した．SIFT は画像の回転・スケール変化・照明変化等に頑健である特徴量を記述することが可能であり，物体認識や画像のマッチング等に用いられている．SIFT は特徴点の局所的な特徴を表すため，特定物体の認識や同定に利用されているが，形状が異なる人等の一般物体認識への利用は困難である．一方，HOG は回転やスケールの影響を受けるが，領域に対して大まかな形状の特徴量を記述するため，人や車等の一般物体認識に用いられている．

SIFT については，SIFT の Web サイト [31] から，実行形式ファイルを入手することが可能である．また，Vedaldi の Web サイトでは，C++で実装した SIFT++[32] や MATLAB で実装した SIFT[33] のソースコードが公開されている．SIFT の拡張として紹介した PCA-SIFT についても，Ke らの Web サイト [34] からソースコードを入手可能である．また，SIFT は DoG の計算等のため計算コストが高いという問題があるが，GPU による実装も検討されている [35]．

参考文献

[1] C. Harris and M. Stephens, “A combined corner and edge detector”, Proc. of Fourth Alvey Vision Conference, pp. 147-151.

[2] T. Lindeberg, “Scale-space theory: A basic tool for analysing structures at different scales”, Proc. of Journal of Applied Statistics, 21(2), pp. 224-270.

[3] C. Schmid and P. Mohr, “Local grayvalue invariants for image retrieval”, Proc. of IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), vol. 19, no. 5, pp.530-534, May, 1997.

[4] D. G. Lowe, “Object recognition from local scale-invariant features”, Proc. of IEEE International Conference on Computer Vision (ICCV), pp. 1150-1157, 1999.

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

[6] J. J. Koenderink, “The structure of images”, Proc. of Biological Cybernetics, vol. 50, pp. 363-370, 1984.

[7] `http://scholar.google.com/`

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

[9] `http://www.cs.ubc.ca/~mbrown/autostitch/autostitch.html`

[10] 高木雅成, 藤吉弘亘, “SIFT 特徴量を用いた交通道路標識認識”, 第 13 回画像センシングシンポジウム SSII07, LD2-06, 2007.

[11] C. Tomasi and T. Kanade, “Detection and tracking of point features”, Technical report, CMU-CS-91-132, 1991.

[12] 都築勇司, 藤吉弘亘, 金出武雄, “SIFT 特徴量に基づく Mean-Shift 探索による特徴点追跡”, 情報処理学会 研究報告 CVIM157, pp. 101-108, 2007.

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

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Y. Ke, R. Sukthankar, “PCA-SIFT: A more distinctive representation for local image descriptors”, Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 511-517, 2004.

[21] A. Stein, M. Herbert, “Incorporating background invariance into feature-based object recognition”, Proc. of IEEE Workshop on Applications of Computer Vision (WACV), pp. 37-44, January, 2005.

[22] Alaa E. Abdel-Hakim and Aly A. Farag, “CSIFT: A SIFT descriptor with color invariant characteristics”, Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1978-1983, 2006.

[23] W. Cheung and G. Hamarneh, “N-dimensional scale invariant feature transform for matching medical images,” Proc. of IEEE International Symposium on Biomedical Imaging (ISBI), pp. 720-723, 2007.

[24] 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.

[25] N. Dalal, B. Triggs and C. Schmid, “Human detection using oriented histograms of flow and appearance”, Proc. of IEEE European Conference on Computer Vision (ECCV), vol. 2, pp. 428-441, May, 2006.

[26] 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.

[27] F. Suard, A. Broggi, “Pedestrian detection using infrared images and histograms of oriented fradients”, Proc. of IEEE Intelligent Vehicles Symposium (IV), pp. 206-212, Jun, 2006.

[28] 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.

[29] Q. Zhu, S. Avidan, M. Yeh, K. Cheng, “Fast human detection using a cascade of histograms of oriented gradients”, Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2, pp. 1491-1498, Jun, 2006.

[30] 小林拓也, 日高章理, 栗田多喜夫, “Histograms of oriented gradients を用いた対象識別での特徴選択”, 信学技報, Vol. 106, pp. 119-124, Mar, 2007.

[31] `http://www.cs.ubc.ca/~lowe/keypoints/`

[32] `http://vision.ucla.edu/~vedaldi/code/siftpp/siftpp.html`

[33] `http://vision.ucla.edu/~vedaldi/code/sift/sift.html`

[34] `http://www.cs.cmu.edu/~yke/pcasift/`

[35] N. Sinha, J. M. Frahm, M. Pollefeys, Y. Genc, “GPU-based video feature tracking and matching”, Proc. of Workshop on Edge Computing Using New Commodity Architectures, 2006.