H-027

Geometric Phrase Poolingと大域特徴を利用した物体検出手法

An object detection method using geometric phrase pooling and global feature

河合 吉彦 †　**藤井 真人** †
Yoshihiko Kawai　Mahito Fujii

1 はじめに

大量の画像データを効率的にハンドリングするためには，画像に何が映っているのか，といった意味内容の解析が重要である．意味内容解析の研究においては，画像を SIFT [2] などの局所特徴の集合で表現する bag of feature（BoF）フレームワーク [1] の有効性が広く確認されている．Geometric phrase pooling (GPP) [3] は近年着目されている局所特徴のひとつであり，注目領域とその周辺領域における輝度勾配とエッジ勾配を考慮することによって，従来手法よりも頑健で表現能力の高い特徴ベクトルを算出する手法である．しかし GPP には，画像の大域的な色特徴やテクスチャ特徴が反映されていないという課題が残されていた．そこで本研究では，GPP に基づく局所特徴と，色やテクスチャなどの大域特徴を組み合わせて特徴ベクトルを算出することにより，画像中に出現する物体を精度よく検出する手法を提案する．実験では，Caltech-101 データセットに対して提案手法を適用し，有効性を検証する．

2 提案手法

図 1 に提案手法の概要を示す．提案手法では，特徴点の周辺領域から算出した GPP による局所特徴と，より広い領域から算出した大域特徴とを，spatial pyramid matching（SPM）[5] を用いて統合し，画像全体の特徴ベクトルを算出する．その後，算出した特徴ベクトルを，学習データに基づいて分類することにより，画像に映る物体を判定する．分類にはサポートベクターマシン（SVM）を利用する．以下，特徴量の算出の詳細を示す．

2.1 局所特徴の算出

まず始めに，入力画像と，入力画像に対するエッジ検出画像のそれぞれから SIFT 特徴 [2] を算出する．特徴点は，一定の画素間隔で格子状にサンプリング（dense sampling）する．算出した SIFT 特徴の集合 $\mathcal{M}$ は次のように表される．

$$\mathcal{M} = \{(\mathbf{d}_1, \mathbf{I}_1), .., (\mathbf{d}_M, \mathbf{I}_M)\} \tag{1}$$

ここで，$\mathbf{d}_m$ と $\mathbf{I}_m$ は，それぞれ m 番目の特徴点の特徴記述子（descriptor）と座標を表す．M は特徴点の総数を表す．文献 [3] では，エッジ検出に Compass オペレータを利用していたが，提案手法では予備実験の結果に基づき Sobel オペレータを利用した．

次に，特徴記述子 $\mathbf{d}_m$ を LLC [4] を利用して量子化し，B 次元の特徴ベクトル $\mathbf{v}_m$ に変換する．ここで，B はコードブックのサイズを表す．コードブックは，学習データから算出した特徴記述子を k 平均法などでクラスタリングすることで作成する．

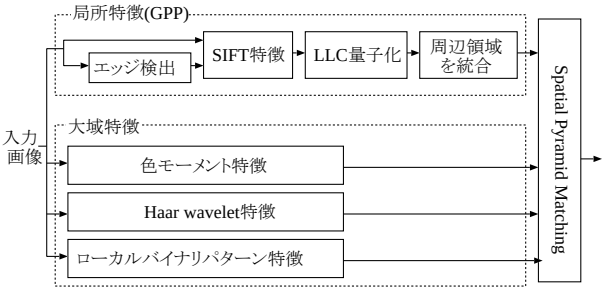

図 1　提案手法の概要

続いて，$\mathbf{I}_m$ の K 近傍の特徴点の特徴ベクトル $\mathbf{v}_{m,k}$ ($k = 1...K$) を max pooling で統合することにより，周辺領域を考慮した特徴ベクトル $\mathbf{w}_m$ を算出する．$\mathbf{w}_m$ の算出式を以下に示す．

$$\mathbf{w}_m = \max_{1 \le k \le K} \{\mathbf{v}_m + s_k \cdot \mathbf{v}_{m,k}\} \tag{2}$$

ここで max は，ベクトルの要素単位で最大値演算を実施する関数を表す．提案手法では $K = 20$ に設定した．また，式中の s_k は，$\mathbf{I}_m$ からの距離に基づく重みを表し，以下のように定義する．

$$s_k = \exp\{-\sigma_w \times ||\mathbf{I}_m - \mathbf{I}_{m,k}||_2\} \tag{3}$$

上式における σ_w は重みを調整するパラメータであり，提案手法では $\sigma_w = 0.01$ に設定した．

最後に，画像領域ごとに $\mathbf{w}_m$ を集計することで，その領域に対する特徴ベクトルを求める．画像全体をひとつの領域とした場合，特徴ベクトル $\mathbf{w}$ の算出式は以下のようになる．

$$\mathbf{w} = \max_{1 \le m \le M} \{w_m \cdot \mathbf{w}_m\} \tag{4}$$

ここで w_m は，特徴ベクトル $\mathbf{w}_m$ に対する重みを表し，特徴点の座標 $\mathbf{I}_m$ におけるエッジ強度に基づいて定義する．具体的には，エッジ検出画像にガウシアンフィルタを適用した結果を利用する．これにより，エッジが密集した（何かが映っている可能性が高い）領域にある特徴点の重みが大きくなり，エッジの少ない平坦な領域にある特徴点お重みは小さくなる．

以上の処理によって局所特徴を算出する．

2.2 大域特徴の算出

大域特徴としては，色モーメント特徴，Haar ウェーブレット特徴，ローカルバイナリパターン (LBP) 特徴 [6] の 3 種類を利用する．

色モーメント特徴は，色分布を反映した特徴量である．提案手法では，入力画像を HSV 色空間および Lab

† NHK 放送技術研究所

表 1　実験結果

手法	局所特徴	大域特徴	コードブックサイズ B	認識精度
GPP-512	○	×	512	0.723
GPPGLO-512	○	○	512	0.767
GPP-1024	○	×	1024	0.754
GPPGLO-1024	○	○	1024	0.777
GPP-2048	○	×	2048	0.765
GPPGLO-2048	○	○	2048	0.794

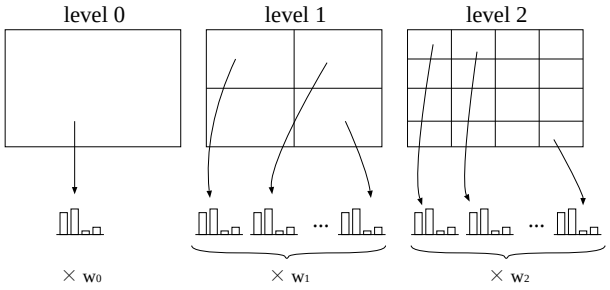

図 2　特徴ベクトルの算出

色空間に変換し，コンポーネント c ($c \in \{h, s, v, l, a, b\}$) ごとに，画素値の平均 μ_c，標準偏差 σ_c，歪度の立方根 s_c を算出し連結する．これを画像領域ごとに求める．

Haar ウェーブレット特徴は，画像のテクスチャを反映した特徴量である．まず，画像領域に対して Haar wavelet 変換を 3 段階適用する．次に，それぞれのサブバンド領域の画素値の分散を算出し，それらを連結して特徴量とする．

LBP 特徴は，注目画素に対する周辺画素の濃度の大小パターンを表したテクスチャ特徴量である．領域内のすべての画素から LBP を算出し，その頻度ヒストグラムを求めて特徴量とする．

2.3　特徴ベクトルの算出

最後に，局所特徴と大域特徴を組み合わせて画像全体の特徴ベクトルを算出する．画像内における空間的な位置情報を反映するため，3 段階の SPM [5] を利用して各領域ごとに局所特徴および大域特徴を集計し，それらをすべてを連結することで画像全体の特徴ベクトルとする．図 2 に SPM による特徴ベクトルの概要を示す．

3　実験

提案手法の有効性を検証するため，Caltech-101 データセット [7] を用いた評価実験を実施した．101 カテゴリからランダムに 30 画像ずつを選択して学習データとし，残りをテストデータとして利用した．局所特徴の算出におけるパラメータは文献 [3] と同様の設定とした．また比較のため，大域特徴を利用した場合と，利用しなかった場合について識別精度を評価した．加えて，局所特徴の量子化におけるコードブックサイズ B について，512, 1024, 2048 の 3 種類の設定で識別精度を比較した．

3.1　実験結果

実験結果を表 1 に示す．実験の結果，大域特徴と局所特徴の両方を利用した手法（GPPGLO-512, GPPGLO-1024, GPPGLO-2048）の方が，従来の局所特徴のみを使用した手法（GPP-512, GPP-1024, GPP-2048）と比較して，認識精度が 2% から 3% 程度向上することが確認された．局所特徴と大域特徴の両方を利用することで，より正確にオブジェクトの特徴を捉えることが可能となり，認識精度の向上につながったものと考えられる．もっとも精度が高くなったのは，GPPGLO-2048 の 79.4% であった．

コードブックサイズについては，同じ特徴量で比較した場合，サイズが 512 の手法（GPP-512, GPPGLO-512）が最も低い精度となり，サイズが 2048 の手法（GPP-2048, GPPGLO-2048）が最も高い精度となった．コードブックは，特徴ベクトルの量子化における基底として利用されるが，3 種類のサイズの中では，量子化誤差と汎化性能のバランスにおいて 2048 が最良であることが実験的に確認された．今後，他のデータセットにおいても同様の傾向が得られるか検証が必要である．また，2048 以上のサイズについても検証を進めたい．

4　まとめ

本稿では，近年着目されている geometric phrase pooling 局所特徴と，色やテクスチャなどのより広い領域を考慮した大域特徴を組み合わせることにより，画像に映る物体を精度よく検出する手法を提案した．Caltech-101 データセットを利用した評価実験では 79.4% という高い認識精度が得られ，従来手法に比べ約 3% の精度向上が確認された．今後は，他のデータセットに対する実験を実施するとともに，さらなる高精度化に向けて改良を進めたい．

参考文献

[1] G. Csurka, *et al.* “Visual categorization with bags of keypoints,” Proc. ECCV Workshop on Statistical Learning in Computer Vision, pp. 59–74, 2004.
[2] D.G. Lowe, “Object recognition from local scale invariant features,” Proc. IEEE ICCV, pp. 1150–1157, 1999.
[3] L. Xie, *et al.* “Spatial pooling of heterogeneous features for image applications,” Proc. ACM Multimdia, pp. 539–548, 2012.
[4] J. Wang, *et al.* “Locality-constrained linear coding for image classification,” Proc. IEEE CVPR, pp. 3360–3367, 2010.
[5] S. Lazebnik, *et al.*, “Beyond bags of features: spatial pyramid matching for recognizing natural scene categories,” Proc. IEEE CVPR, pp. 2169–2178, 2006.
[6] T.Ojala, *et al.* “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol.24, no.7, pp. 971–987, 2002.
[7] http://www.vision.caltech.edu/Image_Datasets/Caltech101/