H-070

局所特徴の共起を考慮した映像意味内容の解析手法

A method for analyzing video semantic content based on co-occurrence of local features

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

1　まえがき

効率的な映像検索のためには，色やテクスチャなどの表層的な特徴だけでなく，そこに何が映っているのかといった意味内容に基づいた解析が重要である．一般物体認識は，映像中に出現するオブジェクトやイベントを識別する手法であり，学習データの変更によって様々なオブジェクトに対応できることを目的としている．従来手法としては，SIFT [1] や SURF [2] といった局所特徴の出現頻度に基づく手法 [3] がある．このアプローチは bag-of-features 法と呼ばれ，様々な先行研究において有効性が認められている．しかしながら，従来手法には，局所特徴を visual word に変換する際に情報が欠落するという問題や，特徴点間の関係をまったく考慮していないという問題が残されている．そこで，本稿ではエッジ情報に基づく局所特徴ベクトルの空間的な位置関係を考慮した新たな特徴量を提案し，一般物体認識に適用する．実験では，国際的な評価型ワークショップである TRECVID 2010 [4] のデータセットに対して提案手法を適用し，その有効性を検証する．

2　局所特徴の共起に基づく映像内容解析

図 1 に提案手法の概要を示す．まず，入力映像からショット境界を検出し，各ショットからキーフレーム画像を抽出する．次に，抽出されたキーフレーム画像に対してエッジ検出を適用し，各画素におけるエッジ方向とその強度を求める．その後，グリッドサンプリングによって画像から特徴点を検出した後，各特徴点の周辺の局所領域からエッジ方向ヒストグラムに基づく局所特徴ベクトルを算出する．このとき，画像スケールを様々に変化させてエッジ方向ヒストグラムを算出することにより，スケール変動や異なるスケール間の特徴点の関係も考慮できるようにする．その後，各特徴点の間の共起関係に基づいて共起特徴ベクトルを算出する．算出された共起特徴ベクトルは，画像のブロック領域ごとに平均化し，それらを連結することによってキーフレーム全体の特徴ベクトルとする．最後に，ラベル付きの学習データと，算出した特徴ベクトルを用いて，検出対象のオブジェクト種別ごとに識別器を学習する．以降では，特徴ベクトルの算出処理と学習処理について詳述する．

2.1　エッジ検出と特徴点検出

エッジ検出については，Sobel フィルタを利用して各画素におけるエッジ方向 θ とエッジ強度 m を求める．特徴点の検出については，従来手法では，SIFT [1] などの特徴点検出手法が利用されていたが，画像によって検出される特徴点数が大きく変動し，特に特徴点数が少ない場合に精度が低下するという問題があった．そこで本手法では，一定の画素間隔で特徴点を取得するグリッドサンプリングを採用することにより，どのような画像からも一定数の特徴点が得られるようにする．

2.2　局所特徴ベクトルの算出

まず，各特徴点について，その周辺画素からエッジ方向ヒストグラムを算出する．座標 (x, y) の特徴点を処理する場合を例に説明する．エッジ方向ヒストグラムは，分散 σ のガウス窓を利用した空間的重み付けを利用して求める．座標 (x, y) におけるエッジ方向 $\theta(x, y)$ を n 方向に量子化する場合，エッジ方向ヒストグラム $\mathbf{h}_\sigma = (h_1, h_2, ..., h_n)$ の各要素 h_i は次式で求める．

$$h_i = \sum_u \sum_v G(u, v, \sigma) \cdot m(x+u, y+u) \cdot \delta_i(\theta(x+u, y+v)) \tag{1}$$

ここで，$G(u, v, \sigma)$ は，座標 $(x+u, y+v)$ におけるエッジ強度 $m(x+u, y+v)$ に対する重みを表し，座標 (x, y) からの距離が近いほど大きな重みとなるような分散 σ のガウス窓を表す．また，$\delta_i(\theta)$ は量子化した θ が i 番目のビンに属する場合には 1，それ以外の場合には 0 を返す関数を表す．

式 (1) に示したエッジ方向ヒストグラムを，ガウス窓の分散 σ の値を様々に変化させて算出し，それらを連結することによって特徴点 (x, y) に対する局所特徴ベクトル $\mathbf{g}_{x,y}$ を求める．具体的には，局所特徴ベクトル $\mathbf{g}_{x,y}$ は，以下のように表される．

$$\mathbf{g}_{x,y} = (\mathbf{h}_{\sigma_1}, \mathbf{h}_{\sigma_2}, ..., \mathbf{h}_{\sigma_t}), \quad \sigma_i = \sigma_0 \cdot s^{i-1} \tag{2}$$

σ_0 はガウス窓の分散の初期値を表し，s は分散を変化させる割合を表す．本稿の実験では，$t = 3$, $s = \sqrt{2}$, $\sigma_0 = 1.6$ とした．ガウス窓の分散を様々に変化させることによって，様々なスケールに対応した特徴量を得ることができる．

2.3　共起特徴ベクトルの算出

各特徴点に対する局所特徴ベクトル $\mathbf{g}_{x,y}$ の空間的な共起に基づいて，共起特徴ベクトルを算出する．特徴点 (x, y) と特徴点 $(x+u, y+v)$ との関係に基づく共起特徴量 $\mathbf{U}_{x,y,u,v}$ を次式のように定義する．

$$\mathbf{U}_{x,y,u,v} = \mathbf{g}_{x,y} \cdot \mathbf{g}^t_{x+u,y+u} \tag{3}$$

算出される $\mathbf{U}$ は，行数と列数が $\mathbf{g}$ の次元数に等しい正方行列となるが，以降の処理では，これを 1 次元ベクトルの形式 $\mathbf{U}'$ に変換して利用する．

特徴点 (x, y) に対して，図 2 に示した 19 種類（自分自身を含む）の位置関係についてそれぞれ $\mathbf{U}'$ を算出し，それらをつなぎ合わせることで特徴点 (x, y) に対する共起特徴ベクトルとする．

† NHK 放送技術研究所

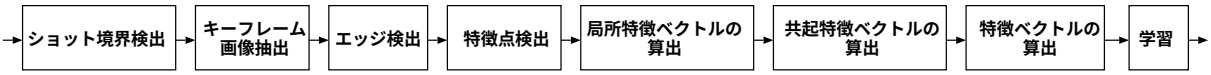

図 1　提案手法の概要

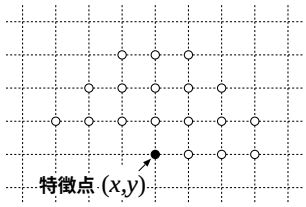

図 2　共起特徴ベクトルの算出

2.4　特徴ベクトルの算出

画像全体に対する特徴ベクトルの算出手順について説明する．まず，画像内でのオブジェクトの出現位置を考慮できるようにするため，キーフレーム画像を横縦 2×2，および 1×3 のブロック領域に分割し，各領域ごとに共起特徴ベクトルの平均ベクトルを求める．次に，算出された平均共起特徴ベクトルをつなぎ合わせることによって，画像全体の特徴ベクトルとする．

2.5　識別器の学習

ラベル付きの学習データと，各画像から算出した特徴ベクトルに基づいて，検出対象のオブジェクトごとに識別器を学習する．多くの従来手法では，識別手法としてサポートベクターマシンが利用されているが，本手法では学習処理や識別処理に要する計算時間を削減するため，ランダムフォレスト法 [5] を利用する．ランダムフォレストはアンサンブル学習の一種であり，多数の決定木の多数決によって入力データを分類するような識別アルゴリズムである．ランダムサンプルした一部の学習データと特徴量のみを利用して各々の決定木を構築するため，計算時間を大きく削減することができる．

3　実験

TRECVID 2010 [4] のデータセットを利用して，評価実験を実施した．TRECVID データセットは，インターネットから取得した約 400 時間分の短い映像クリップからなり，半分が学習データ，残りの半分がテストデータとなっている．本実験では，TRECVID 2010 の評価に利用された 30 種類のオブジェクトの中から 10 種類を選択し，検出精度の評価に用いた．評価尺度には，推定平均適合率 xinfAP [6] を利用した．また，機械学習のためのラベルデータ，およびショット境界データ，評価のための正解データ，xinfAP の算出ソフトウェアは，ウェブ [4] で公開されているものを利用した．

実験結果を表 1 に示す．比較手法として，SIFT を利用した bag-of-features 法 （表中，従来手法）[3] を用いた．実験の結果，従来手法の平均精度は 0.015，提案手法の平均精度は 0.019 となり，従来手法に比べて 0.004 だけ精度が向上した．オブジェクトごとの精度を比較すると，「Dark-skinned_People」や「Demonstration_Or_Protest」において大きく精度が向上した．構図やオブジェクトの形状などの特徴が，共起性を考慮した本特徴量に良好に反映されたものと考えられる．一方で，「Female-Human-Face-Closeup」や「Flowers」は従来手法よりも精度が低下した．これらのオブジェクトにおいては，共起性の考慮によって，正例に対する過剰な適応（過学習）が発生した可能性がある．オブジェクトの種別によって有効性に差が生じており，今後も詳細な解析が必要と考える．

表 1　実験結果

オブジェクト名	xinfAP [6]	
	従来手法	提案手法
Airplane_Flying	0.016	0.021
Bus	0.001	0.004
Dark-skinned_People	0.020	0.044
Demonstration_Or_Protest	0.020	0.047
Female-Human-Face-Closeup	0.052	0.049
Flowers	0.031	0.013
Singing	0.006	0.005
Sitting_Down	0.001	0.000
Telephones	0.005	0.008
Throwing	0.001	0.002
平均	0.015	0.019

4　あとがき

本稿では，映像からの一般物体認識を目的として，特徴点の周辺における局所特徴ベクトルの共起に基づいた新たな画像特徴量を提案した．特徴点の位置関係や異なるスケール間での特徴点の関係を考慮することにより，オブジェクトの形状や特徴をより正確に捉えられるようにした．TRECVID データセットに対する実験では，従来手法と比較して，推定平均適合率が向上することが確認できた．今後は，検出対象とするオブジェクトの種別を拡大し，手法の有効性を詳細に検証したい．

参考文献

[1] D.G. Lowe, “Object recognition from local scale invariant features,” Proc. IEEE ICCV, pp. 1150–1157, 1999.
[2] H. Bay, T. Tuytelaars and L.V. Gool, “SURF: Speeded up robust features,” Proc. ECCV, vol. 3951, pp. 404–417, 2006.
[3] G. Csurka, C. Bray, C. Dance and L. Fan, “Visual categorization with bags of keypoints,” Proc. ECCV Workshop on Statistical Learning in Computer Vision, pp. 59–74, 2004.
[4] http://trecvid.nist.gov/
[5] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[6] E. Yilmaz, E. Kanoulas and J.A. Aslam, “A simple and efficient sampling method for estimating AP and NDCG,” Proc. ACM SIGIR, no. 8, pp. 603–610, 2008.