H-019

一般物体認識技術を利用したテレビ番組からの人物検索手法

Finding Specific Person from TV Program Video using General Object Recognition Technique

河合 吉彦 †　**望月 貴裕** †　**住吉 英樹** †
Yoshihiko Kawai　Takahiro Mochizuki　Hideki Sumiyoshi

1 はじめに

大量の蓄積映像を有効に活用するためには，映像に何が映っているのか，といった意味内容に基づく検索が重要である．特にテレビ番組映像の検索においては，特定の人物が映っているシーンを検索したいという要望が多くある．しかし，テレビ番組における顔画像は，顔向きや表情，照明条件などの変動に加えて，画質や解像度にも制限があるため，セキュリティ分野を対象とした人物識別技術をそのまま適用することは難しい．そこで本稿では，これらの変動にロバストな一般物体認識のアプローチ [1] を利用して，テレビ番組映像からの特定人物の検索を試みる．提案手法では，顔領域画像から勾配特徴 [2] や色，テクスチャ特徴 [3] を求め，それらを機械学習で識別する．実験では、実際に放送されたドラマ番組映像を利用して有効性を検証する．

2 提案手法

図 1 に提案手法の概要を示す．まず，入力映像をカット単位に自動分割 [4] した後，各カットからキーフレーム画像を抽出する．本稿では，ショットの先頭フレームをキーフレーム画像とする単純な方法を利用した．次に，キーフレーム画像から顔領域を検出し，一定のサイズに正規化する．続いて，検出した顔領域から画像特徴ベクトルを算出する．学習フェーズにおいては，画像特徴ベクトルを用いて，検索対象とする人物ごとにモデルを学習する．学習にはサポートベクターマシンを利用する．また検索フェーズにおいては，学習したモデルを用いて目的の人物の画像であるか否かを識別し，スコア順にソートして検索結果とする．

画像特徴ベクトルについては，領域ごとに求めた geometric phrase pooling（GPP）特徴 [2] と，色やテクスチャ特徴とを，spatial pyramid matching（SPM）[5] を用いて統合することで算出する．図 2 に画像特徴ベクトルの算出手順を示す．以下，算出手順を詳述する．

2.1 GPP 特徴の算出

まず始めに，入力画像とそのエッジ検出画像から，一定の画素間隔で格子状に特徴点をサンプリングし，各特徴点の周りの小領域から勾配ヒストグラムを算出する．算出した特徴量の集合を $\mathcal{M}$ で表す．

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

ここで，$\mathbf{d}_m$ と $\mathbf{I}_m$ は，それぞれ m 番目の特徴点における特徴量と座標を表す．

† NHK 放送技術研究所

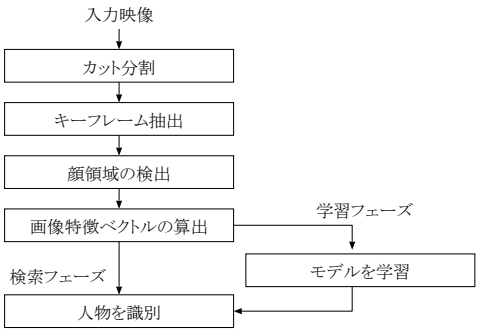

図 1　提案手法の概要

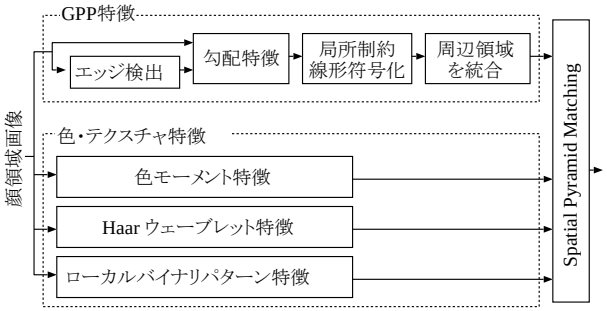

図 2　画像特徴ベクトルの算出手順

次に，特徴量 $\mathbf{d}_m$ を局所制約線形符号化 [6] 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．B はコードブックのサイズを表す．続いて，座標 $\mathbf{I}_m$ の K 近傍の特徴点に対応する特徴ベクトル $\mathbf{v}_{m,k}$ $(k = 1...K)$ を max pooling で統合し，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の最大値演算を表す．また，s_k は，$\mathbf{I}_m$ からの距離に基づく重みを表す．

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

最後に，領域ごとに $\mathbf{w}_m$ を集計し，その領域に対応する GPP 特徴とする．画像全体をひとつの領域とした場合の算出式を以下に示す．

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

w_m は重みを表し，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．これにより，エッジが密集した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の重みは小さくなる．

表 1　実験結果

俳優	正例数	平均適合率		
		100 件	300 件	500 件
A	11	0.766	0.689	0.617
B	15	0.648	0.515	0.433
C	37	0.996	0.965	0.929
D	28	0.615	0.476	0.443
E	13	0.854	0.695	0.624
F	17	0.874	0.758	0.685
G	8	0.696	0.573	0.531
H	58	0.988	0.973	0.951
I	18	0.802	0.688	0.614
J	59	1.000	0.967	0.922
K	16	0.731	0.645	0.612
L	41	0.961	0.853	0.798
M	265	1.000	1.000	0.999
N	38	0.945	0.826	0.770
O	165	0.995	0.992	0.988
P	6	0.576	0.461	0.401
Q	39	0.956	0.891	0.826
平均		0.847	0.763	0.714

2.2　色・テクスチャ特徴の算出

色・テクスチャ特徴としては, 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 [7] の 3 種類を算出する．色モーメント特徴については，入力画像を HSV 色空間および L*a*b*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 を算出する．Haar ウェーブレット特徴については，画像領域に対して Haar ウェーブレット変換を 3 段階適用し，各サブバンド領域の画素値の分散を算出する．LBP 特徴については，領域内の全画素から LBP を算出し，その頻度ヒストグラムを求める．

以上の処理よって求めた GPP 特徴と, 色・テクスチャ特徴を，SPM で分割した矩形領域ごとに求め，すべてを連結することで顔領域全体の特徴ベクトルとする．

3　実験

提案手法の有効性を検証するため，実際のテレビ番組映像を用いた評価実験を実施した．実験には，2013 年に放送された「連続テレビ小説 あまちゃん」の全 156 話分（1 話あたり 15 分）を使用した．実験に用いた映像の解像度は 432×240 ピクセルであり，検出した顔画像は 64×64 画素に正規化して利用した．実験映像から検出されたカットの総数は 41,269 カットであり，そのうちの 1/10 を学習データ，残りをテストデータとして利用した．出演者の中から登場回数がある程度多い 17 名を選択し，それぞれについてモデルを学習した．学習データは人手で作成した．検出精度の評価には，検索結果の上位 n 件の平均適合率 AP を利用した．

$$AP(n) = \frac{\sum_{i=1}^{n} d_i \cdot P(i)}{\sum_{i=1}^{n} d_i}, \quad P(n) = \frac{\sum_{i=1}^{n} d_i}{n} \qquad (5)$$

ここで，d_i は，第 i の検索結果が正解なら 1，不正解なら 0 を表すものとする．

3.1　実験結果

実験結果を表 1 に示す．実験の結果，俳優 17 名に対する平均適合率の平均（MAP）は，上位 100 件が 84.7%，上位 300 件が 76.3%，上位 500 件が 71.4% となった．実験映像には表情や顔向き，照明などの変動が多く含まれていたが，良好な結果を得ることができた．

俳優ごとの結果については，正例（その俳優が映る学習用の画像）が多いほど精度が高くなる傾向が見られた．特に正例数の多かった俳優 M，俳優 O については，上位 500 件の平均適合率がほぼ 100% という非常に高い結果となった．また，正例数が 50 件以上あった俳優 H と俳優 J についても，上位 500 件の平均適合率が 90% 以上という高い精度が得られた．一方，正例数の少なかった俳優 B や俳優 D，俳優 P などについては精度が 50% 以下に低下する結果となった．少ない学習データからは，顔の特徴を十分に捉えることができなかったものと考える．しかし，同様に正例数が少なかった俳優 A や俳優 I については 60% 以上の精度が得られており，俳優や学習データの内容によって，精度にばらつきがでることが分かった．今後は，より多くの番組映像，俳優を対象とした実験を実施し，提案手法の特性を詳細に分析したい．あわせて，従来手法との比較実験を実施し，提案手法の有効性を定量的に評価したい．

4　まとめ

本稿では，一般物体認識技術を利用したテレビ番組からの特定人物の検索手法を提案した．提案手法では，番組映像の各カットから検出した顔領域に対して特徴ベクトルを求め，機械学習で識別することによって特定人物を検索した．実際に放送されたドラマ番組を用いた実験では，俳優 17 名に対する検索結果の平均適合率の平均が，上位 100 件で 84.7%，上位 500 件で 71.4% という良好な結果が得られた．今後は，既存技術との比較実験を実施し，提案手法の有効性を検証したい．また，より多くの番組映像を用いた実験を実施し，さらなる高精度化に向けた改良を進めたい．

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] L. Xie, *et al.* “Spatial pooling of heterogeneous features for image applications,” Proc. ACM Multimdia, pp. 539–548, 2012.

[3] Y. Kawai, *et al.* “NHK STRL at TRECVID 2013: Semantic Indexing”, Proc. TRECVID Workshop, 2013.

[4] 河合ら “逐次的な特徴算出によるディゾルブ，フェードを含むショット境界の高速検出手法”, 信学論, vol. J91-D, no. 10, pp. 2529-2539, 2008.

[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] J. Wang, *et al.* “Locality-constrained linear coding for image classification,” Proc. IEEE CVPR, pp. 3360–3367, 2010.

[7] 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.