時空間 CoHOG 特徴を用いた一般物体認識による動画像検索

○ 中村 彰吾†，出口 大輔†，高橋 友和‡，井手 一郎†，村瀬 洋†

○ Shogo NAKAMURA†, Daisuke DEGUCHI†, Tomokazu TAKAHASHI‡,
Ichiro IDE†, and Hiroshi MURASE†

†: 名古屋大学大学院情報科学研究科，snakamura@murase.m.is.nagoya-u.ac.jp
{ddeguchi,ide,murase}@is.nagoya-u.ac.jp
‡: 岐阜聖徳学園大学経済情報学部，ttakahashi@gifu.shotoku.ac.jp

Web 上の大量の動画像を効率よく検索するための重要な要素技術の一つとして，動画像に映っている物体を認識する技術が挙げられる．動画像に含まれる物体は様々であるため，本研究では近年静止画像を対象として盛んに研究されている一般物体認識のアプローチをとる．動画像を対象とした一般物体認識では，動画像中の様々なフレームから得られる形状特徴と動き特徴の双方を効果的に利用することが重要となる．本報告では，形状の記述能力が高い CoHOG（Co-occurrence Histograms of Oriented Gradients）特徴を時間方向に拡張した時空間 CoHOG 特徴を提案する．時空間 CoHOG 特徴は，動画像中の局所領域での勾配方向の共起ヒストグラムである．実験では，Web 上から収集した 10 カテゴリ，計 1,000 本の動画像を用い，時空間 CoHOG 特徴を用いる手法と CoHOG 特徴を用いる手法で動画像の検索精度を比較することにより，時空間 CoHOG 特徴の有効性を確認した．
＜キーワード＞ 一般物体認識，動画像検索，時空間 CoHOG 特徴，BoF 表現

1. はじめに

近年，Web 上には大量の動画像が存在し，それらを効率よく検索する技術が求められている．検索するためのクエリの一つとして，動画像中に現れる物体が挙げられる．動画像中の物体を認識することができれば，ユーザは自分の見たい動画像を容易に探し出すことができる．例えば，自動車を検索クエリとすることで，図 1 のような動画像を検索できる．一般的に，これらの検索は動画像に付随したタグ（テキスト）を用いて行われる．しかしながら，このようなタグはユーザが主観的に付けるものであるため，表記ゆれなどが原因で正しく検索できない場合があり，そもそもタグが付随しないものも多い．そのため，動画像中の物体を計算機で認識する技術が必要となる．

Web 上の動画像に含まれる物体は様々であるため，特定の物体に依存しない認識手法が必要である．実世界に含まれる物体を計算機が一般的な名称で認識する手法は一般物体認識と呼ばれ，近年盛んに研究が行われている[1]．一般物体認識はカテゴリ内で見た目のバリエーションが大きい物体を扱うため，高い認識率を得ることが難しく，解決すべき課題が多い．

従来，静止画像を対象とした一般物体認識手法に関する研究が盛んに行われている．代表的な手法として，SIFT（Scale-Invariant Feature Transform）特徴[2] を用いたものが挙げられる．SIFT 特徴は，画像の回転や照明変化に頑健な局所特徴であり，抽出した特徴点の周辺領域のエッジやテクスチャなどの形状特徴を記述するものである．これに対して，動画像を対象とした一般物体認識においては形状特徴に加えて，動き特徴を用いることが有効である．

一方，複数の特徴量の共起をとることで対象の記述能力を向上させる研究が行われている．それらの研究の一つとして，Watanabe らは輝度勾配の共起を利用した CoHOG（Co-occurrence Histograms of Oriented Gradients）特徴を提案している[3]．画像空間中の様々な位置関係で勾配の共起をとることで複雑な形状が表現可能となる．この論文では，歩行

図 1　Web 上の動画像

者検出を対象とした実験により CoHOG 特徴の有効性を示している．

本発表では，CoHOG 特徴の共起を時間方向に拡張した時空間 CoHOG 特徴を提案する．共起を時間方向に拡張することで，形状だけでなく動きに対しても高い記述力が得られると考えられる．

以降，2 節で時空間 CoHOG 特徴とそれを用いた動画像の検索手法について述べる．そして，3 節で実験の方法を述べ，実験結果について考察する．最後に，4 節で本報告をむすぶ．

2.　時空間 CoHOG 特徴を用いた検索手法

2.1　手法の概要

図 2 に，手法の流れを示す．まず，入力動画像から単位区間を抽出し，そこから時空間 CoHOG 特徴を抽出する．単位区間は，連続した複数フレームから成るスコアの算出を行う単位である．そして，単位区間から抽出された特徴を BoF（Bag of Features）[4] 表現し，カテゴリ毎に用意した 2 クラス SVM（Support Vector Machine）[5] を用いて対象カテゴリに属する信頼度（スコア）を算出する．最後に，全単位区間のスコアから，入力動画像の検索スコアを決定する．この検索スコアを利用して，各カテゴリに対して入力動画像群をランキングする．以降で，各段階について詳しく述べる．

2.2　時空間 CoHOG 特徴の抽出

単位区間から一定間隔で時空間局所領域（ブロック）をサンプリングし，ブロック毎に時空間 CoHOG 特徴を抽出する．図 3 に，ブロックのサンプリングの様子を示す．以下で時空間 CoHOG 特徴量の計算方法について述べる．

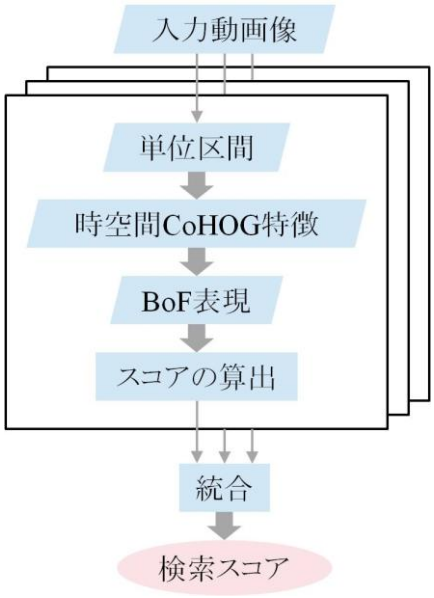

図 2　検索手法の流れ

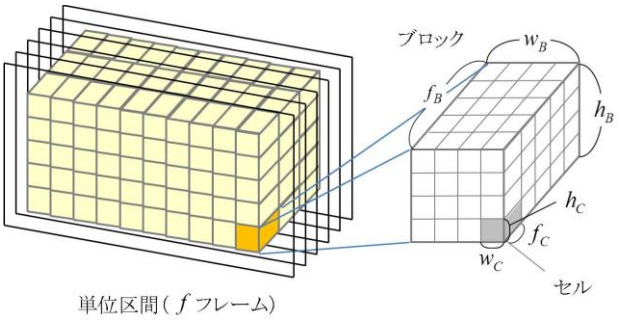

図 3　単位区間からのブロックのサンプリング

まず，ブロック内の全画素位置で空間方向の輝度勾配 $\mathbf{g}(x,y,t)$ を計算する．$I(x,y,t)$ を画素の輝度として，$\mathbf{g}(x,y,t)$ の各要素を次のように求める．

$$\begin{cases} g_x(x,y,t) = I(x+1,y,t) - I(x-1,y,t) \\ g_y(x,y,t) = I(x,y+1,t) - I(x,y-1,t) \end{cases} \quad (1)$$

次に，ブロックを複数の領域（セル）に分割し，各セルで平均勾配 $\bar{\mathbf{g}}(x,y,t)$ を計算する．そして，各平均勾配 $\bar{\mathbf{g}}(x,y,t)$ を 0°，30°，…，330° の 12 方向に量子化する．

ブロック内の各セルで計算した勾配方向を利用して，勾配方向の共起ヒストグラムを作成する．これが時空間 CoHOG 特徴である．以下で，図 4 に沿って作成の流れを説明する．

(a)　まず，基準セルとの位置関係が異なる多数のセルのペアを形成する．例えば，時空間で 2 つ隣のセルまでの位置関係を用いる場合，ペアの数は対称性による冗長性を排除することによって 63 種類となる（ただし，基準セ

ル自身のペアも含む）．

(b) 各ペアの位置関係にあるセルの組をブロック内から列挙し，ヒストグラムの対応するビンに投票していく．

(c) 各ペアに対するヒストグラムを連結することで最終的なヒストグラムとする．ただし，基準セル同士のペアはそのセル自身の勾配方向のみでヒストグラムを作成する．また，勾配強度が 0 に等しいセルは利用しない．

時空間 CoHOG 特徴量は12×12×62+12＝8,940次元となり，非常に次元が大きくなる．そのため，PCA（Principal Component Analysis）により累積寄与率が 0.9 となる次元に圧縮する．

2.3 時空間 CoHOG 特徴の BoF 表現

スコア算出を行う前処理として，単位区間の全ブロックで作成した時空間 CoHOG 特徴を BoF 表現する．BoF 表現は画像を局所特徴量のヒストグラムで表現する手法であり，一般物体認識の分野で広く用いられている．この手法は学習段階と認識段階に分かれており，一般に以下の手順で行う．

学習段階では，まず各学習用画像から複数の局所特徴量を抽出する．そして，全学習用画像の全局所特徴量を k-means クラスタリングすることにより，visual word を生成する．visual word は，局所特徴量を表す特徴ベクトルをベクトル量子化したものである．各局所特徴量を最も類似する visual word として表現することにより，各学習用画像は visual word の出現頻度のヒストグラムで記述される．認識段階では，まず学習段階と同様に，各入力用画像から局所特徴量を抽出する．そして，学習段階で生成された visual word を用いて，各入力用画像を visual word の出現頻度のヒストグラムで表現する．

2.4 SVM の学習

学習動画像の単位区間で作成した時空間 CoHOG 特徴の BoF 表現を用いて，カーネル SVM を学習する．なお，SVM はカテゴリ毎に用意しておく．SVM は高い認識性能を持っていることが知られ，様々な画像認識問題に応用されている．本研究では，カーネル関数として χ^2 カーネルを用いる．χ^2 カーネルは以下の式で表される．

$$k(\mathbf{x},\mathbf{y}) = \exp\left(-\gamma \sum_i \frac{(x_i - y_i)^2}{x_i + y_i}\right) \tag{2}$$

χ^2 カーネルは画像分類において最も性能が良いカーネルであると報告されている[6]．γ には，全学習ベクトルの組み合わせにおける平均 χ^2 距離の逆数を設定する．

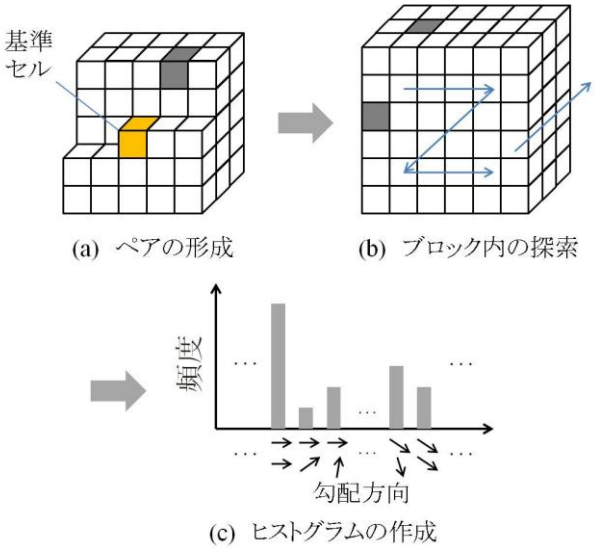

図 4 時空間 CoHOG 特徴の抽出

2.5 スコアの算出

事前に学習しておいた SVM を用いて，単位区間に対して対象カテゴリに属する信頼度（スコア）を算出する．SVM の出力結果として，分離超平面からの符号付き距離が得られる．この値を単位区間の対象カテゴリに対するスコアとする．

2.6 スコアの統合

単位区間 t のスコア c_t を用いて，入力動画像の検索スコアを算出する．入力動画像の単位区間数を T とすると，検索スコア $\hat{c}$ は次式によって決定する．

$$\hat{c} = \frac{1}{T}\sum_{t=1}^{T} c_t \tag{3}$$

本研究では，$\hat{c}$ をそのまま利用して，各カテゴリに対して入力動画像群をランキングする．

3. 実験

3.1 実験データセット

データセットは，YouTube[7] から動画像を収集して構築した．動画像中の一般物体のカテゴリは，PASCAL Visual Object Classes Challenge 2006[8] で使用されたデータセットと同一の 10 カテゴリとした．具体的には，bicycle, bus, car, cat, cow, dog, horse, motorbike, person, sheep である．図 5 にデ

図 5　実験で使用した動画像の例

ータセットの一部を示す．データセット中の動画像は対象物体が映っている部分を切り出した．ただし，オクルージョンや画面外へのはみ出しで物体の一部が映っていないフレームが存在した．また，異なるカテゴリの物体が同時に映っている動画像も収集した．例えば，bicycle，motorbike が含まれる動画像については，すべて person も同時に映っているものであった．そのため，これらの動画像を person の識別器に学習させる際は，すべてポジティブデータとして扱った．

データセットの動画像数は 1,000 本であり，各カテゴリ 100 本（person は bicycle, motorbike と同時に映っている動画像も含めると 300 本）とした．また，動画像のフレームサイズは 1,280×720 ～1,920×1,080 [pixels]，フレーム数は 45～270[フレーム]であった．

3.2　実験方法

提案する時空間 CoHOG 特徴を用いた手法と，時間情報を利用しない CoHOG 特徴を比較した．パラメータとしては，単位区間の長さを $f=30$[フレーム]，ブロックの空間方向のサイズを $w_B=h_B=80$[pixels]，セルの空間方向のサイズを $w_C=h_C=5$[pixels]，BoF における visual word の数を 500 とした．時空間 CoHOG 特徴を用いた手法では，ブロックの時間方向のサイズを $f_B=5$[フレーム]，セルの時間方向のサイズを $f_C=1$[フレーム]とした． CoHOG 特徴を用いた手法は，ブロック・セルの時間方向の大きさを $f_B=f_C=1$[フレーム]とした．このことにより，勾配方向の共起は空間方向のみでとることになる．また，ブロックは時空間 CoHOG 特徴で抽出するブロックの中央フレームから抽出した．

3.3　評価方法

評価には，2-fold cross validation による Mean AP（Average Precision）を用いた．Mean AP はカテゴリ毎に算出した AP の平均値である．AP は検索結果を評価する際に用いられる評価基準であり，以下のように求める．まず，入力動画像群を対象カテゴリに属する信頼度が高い順に並び替える．そして，ランキング上位から順に調べていき，正解であればその時点での適合率を計算する．すべての正解データが現れたら，次式で AP を計算する．

$$\mathrm{AP}=\frac{1}{V}\sum_{v=1}^{V}\mathrm{Pr}(v) \tag{4}$$

$\mathrm{Pr}(v)$ は対象カテゴリに属する動画像 v が現れたときの適合率，V は対象カテゴリに属する動画像数である．本研究の場合，person では $V=150$，それ以

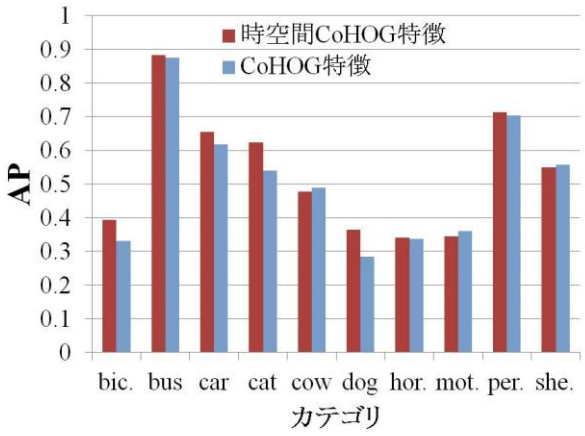

図 6　各特徴を用いた手法のカテゴリ毎の AP

外のカテゴリでは $V=50$ となる．AP は Precision-Recall 曲線と座標軸に囲まれた領域の面積に相当する．

3.4　実験結果

時空間 CoHOG 特徴を用いた手法と CoHOG 特徴を用いた手法の Mean AP はそれぞれ 0.534, 0.509 となった．このことから，CoHOG 特徴に対する時空間 CoHOG 特徴の優位性を確認した．

3.5　考察

図 6 に，各特徴を用いた手法のカテゴリ毎の AP を示す．この図から，時空間 CoHOG 特徴を用いた手法は bicycle, cat, dog の検索精度が特に良いことがわかる．

表 1 に，dog のランキング上位 10 位までの結果を示す．この表は，dog のランキングにおいて検索スコアが上位 10 位以内であった動画像の実際のカテゴリを表している．両者とも horse が誤って検索されているが，CoHOG 特徴ではさらに cat や cow も誤って検索されてしまっていることがわかる．

Cow, motorbike, sheep については CoHOG 特徴にわずかに及ばなかった．その中でも，CoHOG 特徴と比較して最も検索精度の悪かった cow のランキングを表 2 に示す．CoHOG 特徴に比べて horse や sheep の動画像が多く検索されてしまっていることがわかる．

ランキングが上位となる動画像の傾向を見てみると，時空間 CoHOG 特徴は CoHOG 特徴と比較して動きの大きい動画像がやや多いことがわかった．このことから，時空間 CoHOG 特徴は動きの小さい動画像に弱いと考えられる．現段階では勾配強度を考慮しておらず，特に 0 となる勾配は利用していない．

表 1　dog のランキング

(a) 時空間 CoHOG 特徴 (AP=0.400)

順位	カテゴリ
1	dog
2	dog
3	dog
4	horse
5	dog
6	dog
7	horse
8	dog
9	horse
10	dog

(b) CoHOG 特徴 (AP=0.301)

順位	カテゴリ
1	horse
2	dog
3	dog
4	horse
5	dog
6	horse
7	cat
8	cat
9	dog
10	cow

表 2　cow のランキング

(a) 時空間 CoHOG 特徴 (AP=0.435)

順位	カテゴリ
1	cow
2	cow
3	sheep
4	cow
5	horse
6	cow
7	sheep
8	cow
9	horse
10	sheep

(b) CoHOG 特徴 (AP=0.509)

順位	カテゴリ
1	cow
2	cow
3	cow
4	horse
5	cow
6	person
7	sheep
8	cow
9	cow
10	cow

そのため，より高い検索精度を得るためには勾配強度の導入が必要であると考えられる．

4.　むすび

本報告では，動画像を対象とした一般物体認識のための時空間 CoHOG 特徴を提案した．時空間 CoHOG 特徴は CoHOG 特徴における共起を時間方向に拡張したものである．単位区間から一定間隔で時空間 CoHOG 特徴を抽出し，PCA により次元を圧縮し，BoF 表現することで単位区間に映る物体の形状特徴と動き特徴を記述した．そして，カーネル SVM によるスコアの算出を単位区間毎に行い，それらの結果を平均することにより，入力動画像の検

索スコアを決定した．実験では，YouTube において収集した 1,000 本の動画像を使用し，時空間 CoHOG 特徴と時間情報を利用しない CoHOG 特徴を比較した．実験の結果，時空間 CoHOG 特徴を用いた手法がより高い検索精度となることを確認した．

今後の課題としては，勾配強度の導入や，一般物体認識で利用されている他の特徴との比較が挙げられる．

謝辞

日頃より熱心に御討論頂く名古屋大学村瀬研究室諸氏に深く感謝する．本研究の一部は，科学研究費補助金による．また，本研究では画像処理に MIST ライブラリ（http://mist.murase.m.is.nagoya-u.ac.jp/）を使用した．

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. SIG 16 (CVIM 19), pp. 1–24, November 2007.

[2] D.G. Lowe, “Object recognition from local scale-invariant features,” Proc. 7th IEEE Int. Conf. on Computer Vision, pp. 1150–1157, September 1999.

[3] T. Watanabe, S. Ito, and K. Yokoi, “Co-occurrence histograms of oriented gradients for pedestrian detection,” Proc. 3rd Pacific-Rim Symposium on Image and Video Technology, Lecture Notes in Computer Science, vol. 5414, pp. 37–47, January 2009.

[4] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual categorization with bags of keypoints,” Proc. ECCV 2004 Workshop on Statistical Learning in Computer Vision, pp. 1–22, May 2004.

[5] V. Vapnik, “Statistical learning theory,” Wiley-Interscience Publication, 1998.

[6] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, “Local features and kernels for classification of texture and object categories: A comprehensive study,” International Journal of Computer Vision, vol. 73, no. 2, pp. 213–238, June 2007.

[7] YouTube —Broadcast Yourself—, http://www.youtube.com/.

[8] The PASCAL Visual Object Classes Challenge 2006, http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2006/.