社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

動き特徴とBoF特徴を組み合わせた動画像中の一般物体認識

中村　彰吾†　出口　大輔†　高橋　友和††　井手　一郎†　村瀬　洋†

† 名古屋大学 大学院情報科学研究科　〒 464-8601 愛知県名古屋市千種区不老町
†† 岐阜聖徳学園大学 経済情報学部　〒 500-8288 岐阜県岐阜市中鶉 1 丁目 38 番地
E-mail: †snakamura@murase.m.is.nagoya-u.ac.jp, †{ddeguchi,ide,murase}@is.nagoya-u.ac.jp,
††ttakahashi@gifu.shotoku.ac.jp

あらまし　Web 上の大量の動画像の分類・検索ができるようにするために，動画像に映っている物体を認識し，自動でタグ付けを行う技術が必要となってきている．従来，静止画像に対する一般物体認識は盛んに行われてきたが，動画像に対する一般物体認識はほとんど行われていない．動画像に対する一般物体認識では，動画像中に含まれる様々なフレームから得られる特徴を効果的に利用することが重要となる．本報告では，動画像の各フレームの BoF（Bag of Features）特徴と隣接フレーム間の動き特徴を統合し利用することで，動画像に映っている一般物体を認識する手法を提案する．実験により，動き特徴の統合利用による有効性を示した．

キーワード　一般物体認識，動画像分類，オプティカルフロー，Bag of Features

Video-based Generic Object Recognition by Combining Motion Features and BoF

Shogo NAKAMURA†, Daisuke DEGUCHI†, Tomokazu TAKAHASHI††,

Ichiro IDE†, and Hiroshi MURASE†

† Graduate School of Information Science, Nagoya University
Furo-cho, Chigusa-ku, Nagoya-shi, Aichi, 464-8601 Japan
†† Faculty of Economics and Information, Gifu Shotoku Gakuen University
Nakauzura 1–38, Gifu-shi, Gifu, 500-8288 Japan
E-mail: †snakamura@murase.m.is.nagoya-u.ac.jp, †{ddeguchi,ide,murase}@is.nagoya-u.ac.jp,
††ttakahashi@gifu.shotoku.ac.jp

Abstract　It has been needed to recognize objects in videos and attach tags automatically so as to categorize and search a large amount of videos on the Web. Recently, generic object recognition has been studied for still images actively, but not almost for videos. As for the generic object recognition in a video, it is important to make use of the features from various frames involved in the video efficiently. In this paper, we propose a method of recognizing generic objects in videos by combining BoF of each frames and motion features of consecutive frames. Experimental results showed the effectiveness of integrated use of motion features.

Key words　generic object recognition, video classification, optical flow, bag of features

1. はじめに

近年，Web 上には大量の動画像が存在し，それらを分類・検索する技術が求められている．そのための情報の一つとして，動画像に現れる物体が挙げられる．物体は動画像の内容に大きく関係するので，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．図 1 に，Web 上に存在する動画像の例を示す．

一般的に，これらの分類・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分類・検索が正しく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ことで，適切なタグを自動で付与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を用いる必要がある．このような実世

図 1　Web 上の動画像

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，困難な問題であると位置づけられている．

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，Bag of Features（BoF）[2] や constellation model（星座モデル）[3] を用いたものが挙げられる．これらは，画像のアピアランスを表現するために局所的な視覚特徴を用いた手法である．これに対して，動画像中の一般物体認識においては各フレームの視覚特徴に加えて，動き特徴を用いることが有効であると考えられる．野口らは，動画像中の動きのある部分の時空間特徴とフレーム全体の視覚特徴，動き特徴を統合することで Web 上の動画像に対する動作認識を行った [4]．この研究は動作認識を目的としたものであり，動画像中の一般物体認識を行う本研究とは目的が異なる．また，動画像中の一般物体認識を行った研究として，Noor らは，動画像における物体アピアランスの変化の連続性を利用した一般物体認識を行った [5]．しかし，この研究は剛体を対象とした視点の変化によるアピアランスの変化のみを扱うものであるため，本研究とは目的が異なる．

本報告では，動画像中の各フレームから抽出した動き特徴と BoF 特徴を組み合わせることで，動画像中の一般物体認識を行う手法を提案する．具体的には，動き特徴の抽出に各隣接フレーム間のオプティカルフロー，BoF 特徴の抽出に各フレームの BoF を用いる．

以降，2 節で提案手法について述べる．そして，3 節で実験の方法を述べ，その結果に対し考察を行う．最後に，4 節で本報告をむすぶ．

2. 提 案 手 法

2.1 提案手法概要

図 2 に，提案手法の流れを示す．まず入力動画像の各フレームにおいて，動き特徴，BoF 特徴を抽出する．動き特徴の抽出には各隣接フレーム間のオプティカルフロー，BoF 特徴の抽出には各フレームの BoF を用いる．そして，これらの特徴を用いて各フレームのカテゴリを認識する．最後に，全フレームの認識結果を統合することにより，入力動画像のカテゴリを決定する．以降で，各段階について詳しく述べる．

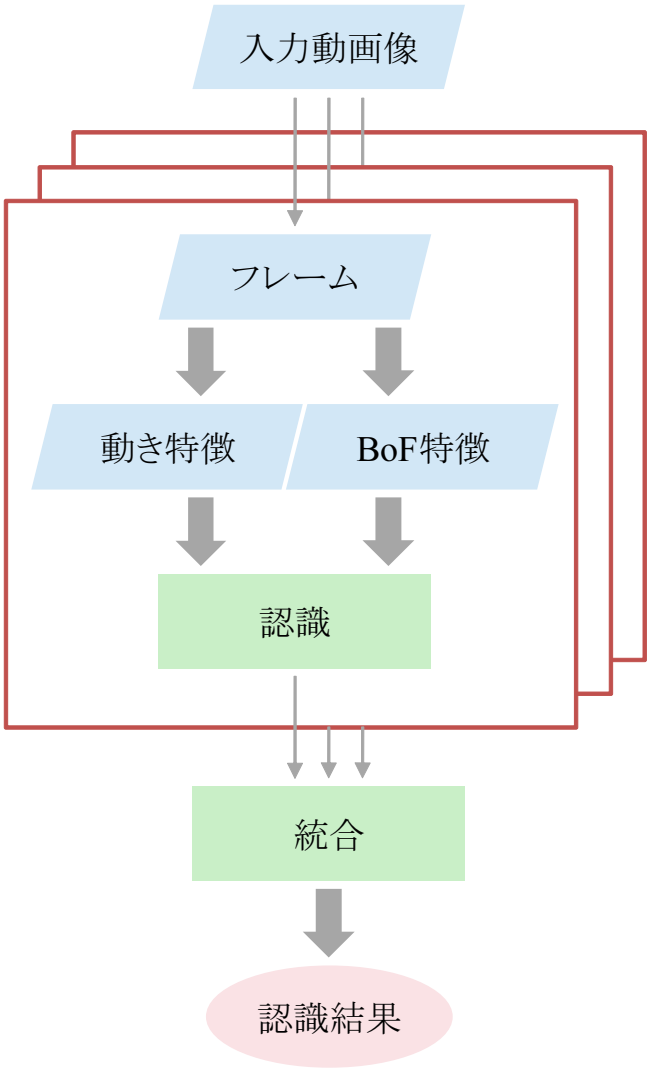


図 2　提案手法の流れ

2.2 動 き 特 徴

動き特徴として，隣接フレーム間のオプティカルフローから計算されるヒストグラムを用いる．本手法では，オプティカルフローの抽出にブロックマッチング法を利用する．ブロックマッチング法では，画像を一定サイズのブロックに分割し，一方の画像の各ブロックに対して他方の画像の近傍領域をテンプレートマッチングすることにより，ブロック毎にオプティカルフローを算出する．

提案手法では，まず動画像中の各フレームに対してオプティカルフローを計算する．ここで，カメラモーションの影響を抑え，認識したい物体の動き特徴のみを抽出する目的で，各フレームのオプティカルフローから全ブロックで求めたオプティカルフローの平均値を引く．すなわち，ブロック b のオプティカルフロー $\mathbf{v}'_b$ は，元のオプティカルフロー $\mathbf{v}_b$ を用いて次式のように計算する．

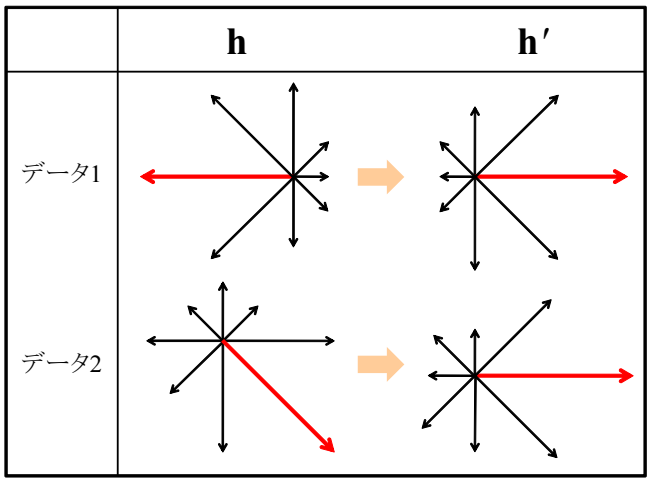


図 3　ヒストグラムの変換例（$M = 8$ のとき）

$$\mathbf{v}'_b = \mathbf{v}_b - \frac{1}{B}\sum_{i=1}^{B}\mathbf{v}_i \tag{1}$$

ただし，B はブロック数である．

次に，オプティカルフローの方向 θ を M 個のビンで分割したヒストグラム $\mathbf{h} = (h_{\theta_1}, \ldots, h_{\theta_M})$ を作成する．このヒストグラムを物体の動きの方向に不変な特徴とするため，最も頻度が高いビンが基準となるように回転する．この処理は，例えば左に走る自動車と右下に走る自動車では動きの方向が異なるため，同一のカテゴリであるにもかかわらずヒストグラムの形状が大きく異なってしまうという問題を防ぐために行う．最終的なヒストグラム $\mathbf{h}'$ の各ビン h'_{θ_m} は次式のようになる．

$$h'_{\theta_m} = \begin{cases} h_{\theta_m - \alpha} & (\alpha \leqq \theta_m < 2\pi) \\ h_{\theta_m + 2\pi - \alpha} & (0 \leqq \theta_m < \alpha) \end{cases} \tag{2}$$

$$\alpha = \arg\max_m h_{\theta_m} \tag{3}$$

図 3 に，ヒストグラム $\mathbf{h}$ から最終的なヒストグラム $\mathbf{h}'$ への変換例を示す．

2.3　BoF 特徴

BoF（Bag of Features）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る．

学習段階では，まず各学習用画像から複数の局所特徴を抽出する．そして，全学習用画像の全局所特徴を 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 とは，局所特徴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表現される．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．

本研究では，局所特徴に SIFT（Scale-Invariant Feature Transform）[6] を用いる．SIFT は，画像の回転・スケール変化・照明変化等に頑健な局所特徴であることが知られており，抽出した特徴点の周辺領域を 128 次元の特徴ベクトルで記述するものである．

2.4　フレームの認識

動き特徴と BoF 特徴それぞれで学習した識別器を用いて，フレームの認識を行う．識別器には，カーネル SVM（Support Vector Machine）[7] を用いる．カーネル 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．

フレーム f について，動き特徴のカテゴリ c への所属確率を $p_{f,c}$，BoF 特徴のカテゴリ c への所属確率を $q_{f,c}$，BoF 特徴の重みを w とすると，フレームの認識結果をカテゴリ c への所属確率 $r_{f,c}$ として以下のように計算する．

$$r_{f,c} = (1 - w)p_{f,c} + wq_{f,c} \tag{4}$$

2.5　統合処理

全フレームの認識結果を用いて，入力動画像の認識を行う．入力動画像のフレーム数を F とすると，認識結果カテゴリ $\hat{c}$ は，次式によって決定される．

$$\hat{c} = \arg\max_c r_c \tag{5}$$

$$r_c = \frac{1}{F}\sum_{f=1}^{F} r_{f,c} \tag{6}$$

3.　実　　験

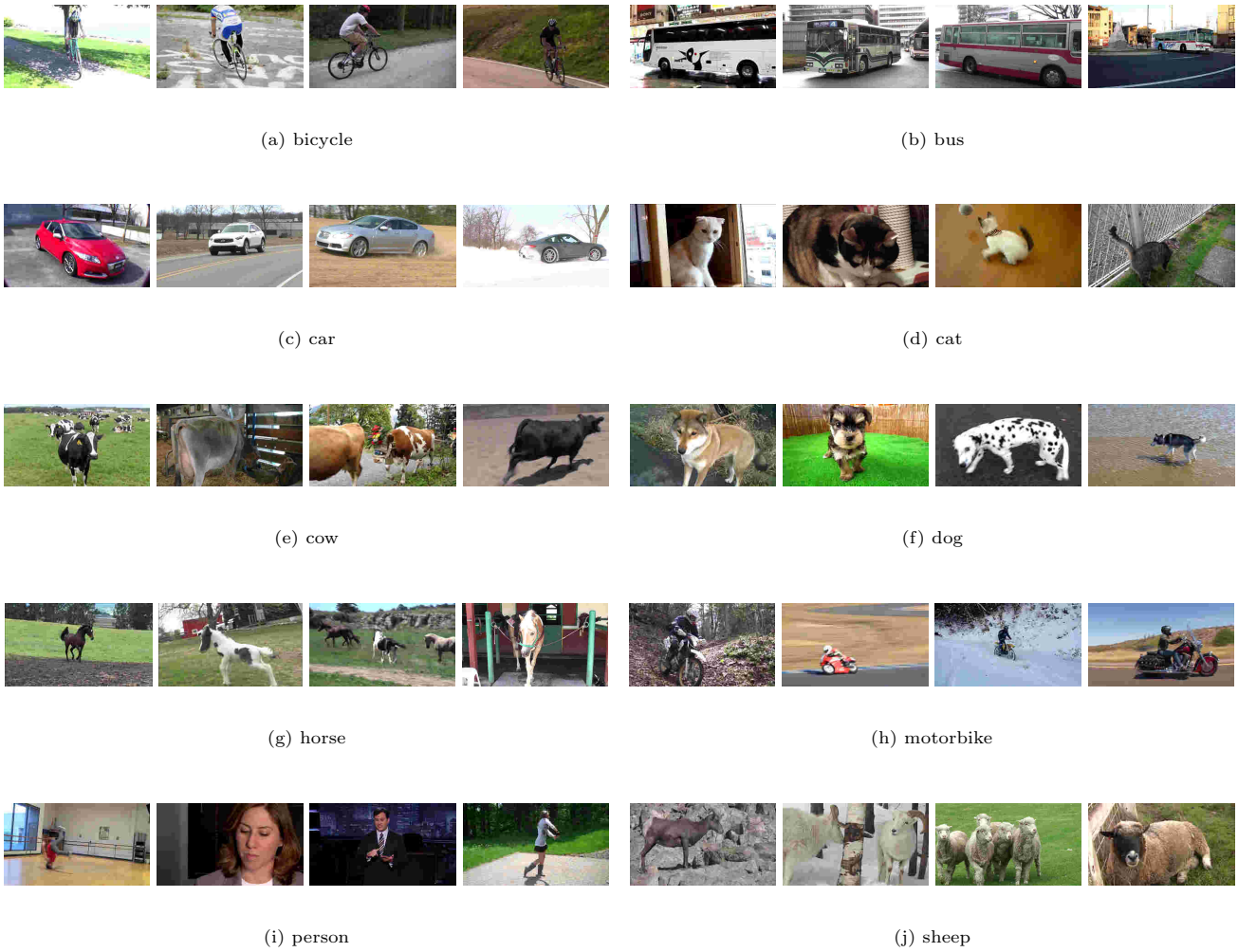
3.1　実験方法

データセットは，YouTube [8] にて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 [9] で使用されたデータセットと同一の 10 カテゴリとした．すなわち，bicycle, bus, car, cat, cow, dog, horse, motorbike, person, sheep である．図 4 にデータセットの一部を示す．データセット中の動画像はすべてのフレームに対象物体が映るように切り出した．ただし，それらの中には対象物体以外の物体が同時に映っている場合もあった．例えば，bicycle, motorbike については，すべて人が乗っているものであった．そのため，人と自転車が同時に映っているものを bicycle，人とバイクが同時に映っているものを motorbike と定義した．

データセットの動画像数は 226 であり，カテゴリ毎に 20〜26 であった．また，動画像のフレームサイズは 1280×720〜1920×1080 [pixels]，フレーム数は 60〜239 であった．

評価には，2-fold cross validation による第 n 位までの累積認識率を用いた．第 n 位までの累積認識率は，全入力動画像のうち第 n 位までに正しいカテゴリに認識された動画像の割合である．つまり，入力動画像のカテゴリを c とすると，式 (6) によって計算された r_c が n 番目までに大きければ認識成功となる．

提案手法の有効性を示すため，本実験では提案手法と BoF 特徴のみを用いた手法との比較を行った．BoF 特徴のみを用いた手法は，提案手法における式 (4) が $w = 1$ の場合と同等である．



(a) bicycle　(b) bus

(c) car　(d) cat

(e) cow　(f) dog

(g) horse　(h) motorbike

(i) person　(j) sheep

図 4　実験で使用した動画像の例

パラメータとしては，オプティカルフローの方向ヒストグラムのビン数 M の値を 36，BoF における visual word の数を 100，動き特徴と BoF 特徴の統合処理における BoF の重み w の値を予備実験により 0.2 とした．また，BoF における SIFT 特徴点は各フレームでスケールの大きいものから 100 個を利用し，SVM のカーネル関数としては RBF カーネルを用いた．なお，SIFT の実装には SIFT++ [10]，SVM の実装には LIBSVM [11] を利用した．

3.2　実 験 結 果

表 1 に，提案手法および BoF 特徴のみを用いた手法の第 n 位までの累積認識率を示す．

認識率はいずれの場合も BoF 特徴のみを用いて認識するよりも，動き特徴と BoF 特徴を組み合わせて認識したほうが高い値となっていることがわかる．このことから，隣接フレーム間の動き特徴を利用した提案手法の有効性を確認できた．

また，参考として動画像ではなく単一フレームの BoF 特徴のみで認識する手法の評価も行った．ただし，この場合は動画像で認識する場合とは異なり，フレーム毎に認識結果が出力されることになる．認識するフレームは，動画像の中央のフレームとした．この場合の認識率は，第 1 位で 24.3%，第 2 位で 37.6%，第 3 位で 50.4% となった．よって，単一フレームのみを用いるよりも，複数フレームを利用して認識する方が認識率が高くなることがわかる．このことから，複数フレームを用いた認識の有効性を確認できた．

表 1　各手法の第 n 位までの累積認識率 [%]

手法 ＼ n	1	2	3
提案手法（動き特徴+BoF 特徴）	32.3	54.0	65.5
BoF 特徴のみ	25.7	42.5	59.7

3.3　考　　察

動き特徴におけるカメラモーションの軽減，ヒストグラムの回転の有効性を確認するため，それらを行う場合と行わない場合で認識率を比較した．ただし，動き特徴のみを用いた手法，すなわち提案手法における式 (4) が $w = 0$ の場合で評価した．その結果，認識率は両方行わない場合の 29.2% から，カメラモーションの軽減で 29.6%，ヒストグラムの回転で 30.1%，両方行うことで 32.3% に向上した．このことから，どちらの手法も有効であることを確認した．

図 5 に，BoF 特徴の重み w を変化させたときの第 n 位まで

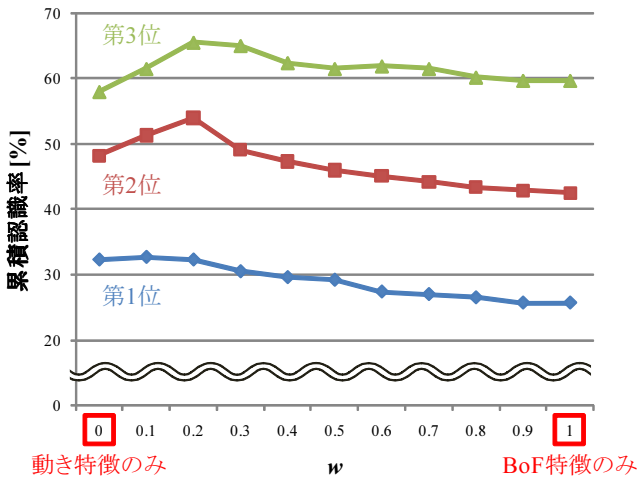


図 5　重み w を変化させたときの第 n 位までの累積認識率 [%]

の累積認識率を示す．最大の認識率は，第 1 位で $w = 0.1$ のとき 32.7%，第 2 位と第 3 位で $w = 0.2$ のときそれぞれ 54.0% と 65.5% であった．この結果から，BoF 特徴に比べ動き特徴が動画像の認識に有意な影響を与えることがわかった．アピアランスよりも動きが重要であるということは大変興味深い．特に，第 1 位累積認識率では，BoF 特徴のみの認識率は動き特徴のみの認識率より 6.6% も低く，認識率の最大値が動き特徴のみの認識率とほとんど変わらない．

表 2 に，$w = 0$（動き特徴のみ），$w = 0.2$（動き特徴+BoF 特徴），$w = 1$（BoF 特徴のみ）での Confusion Matrix を示す．表中の行は動画像の正しいカテゴリであり，列は提案手法により認識されたカテゴリである．この表から，w が変化すると認識結果が大きく変わることがわかる．すべての場合において car, dog に認識される動画像が多く，特に car は再現率も非常に高かった．逆に，horse に認識された動画像は非常に少なく，再現率もその分低かった．これは，動き特徴と BoF 特徴の両者において，car, dog のバリエーションが大きく，horse ではこれが小さいからであると考えられる．よって，BoF 特徴と動き特徴の抽出手法の改良が必要である．

また，$w = 0$（表 3(a)）と $w = 1$（表 3(c)）では認識しやすいカテゴリが異なることも読み取れる．動き特徴では主に cow, dog, person，BoF 特徴では主に bicycle, bus, motorbike である．このことから，動き特徴は動物のような複雑な動きをする物体，BoF 特徴は乗り物のような形状・動きのバリエーションが小さい物体の認識に強いと考えられる．さらに，各特徴を組み合わせた場合（表 3(b)），bus, car, horse, motorbike, sheep で単一の特徴で認識するよりも精度が向上している．しかし，全体の認識率としては動き特徴のみの場合と等しかった．そのため，特徴の統合手法を改良する必要がある．

4. むすび

本報告では，動画像中の動き特徴と BoF 特徴を組み合わせることで一般物体認識を行う手法を提案した．動き特徴は，各隣接フレーム間のオプティカルフローのヒストグラムを用いて抽出した．ただし，オプティカルフローはカメラモーションを軽減し，また最終的なヒストグラムは動きの主方向を揃えたものとした．BoF 特徴は，各フレームから抽出した SIFT 特徴点を用いた BoF によって抽出した．統合処理は，それぞれの特徴の各カテゴリへの所属確率を重み付けて足し合わせることにより行った．実験では，YouTube において収集した 226 個の動画像を使用し，BoF 特徴のみを用いた手法と提案手法を比較した．その結果，提案手法でより高い認識率が得られたことから，動き特徴の統合利用による有効性を示した．

今後の課題としては，BoF 特徴と動き特徴の抽出手法の改良，特徴の統合手法の改良が挙げられる．

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．

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] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual categorization with bags of keypoints,” Proc. ECCV2004 Workshop on Statistical Learning in Computer Vision, pp.1–22, May 2004.
[3] R. Fergus, P. Perona, and A. Zisserman, “Object class recognition by unsupervised scale-invariant learning,” Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol.2, pp.264-271, 2003.
[4] 野口顕嗣, 下田保志, 柳井啓司, “動作認識のための時空間特徴量と特徴統合手法の提案,” 画像の認識・理解シンポジウム（MIRU）2010, OS15-4, IS3-75, July 2010.
[5] H. Noor, S. Noor, S.H. Mirza, “Using video for multiview object categorization in security systems,” SPIE Defense, Security and Sensing Symposium, Paper 7696A-19, April 2010.
[6] D.G. Lowe, “Object recognition from local scale-invariant features,” Proc. 7th IEEE Int. Conf. on Computer Vision, pp.1150-1157, September 1999.
[7] V. Vapnik, “Statistical learning theory,” Wiley-Interscience Publication, 1998.
[8] YouTube - Broadcast Yourself -, http://www.youtube.com/
[9] The PASCAL Visual Object Classes Challenge 2006, http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2006/
[10] A. Vedaldi, “SIFT++,” http://www.vlfeat.org/~vedaldi/code/siftpp.html
[11] C.C. Chang and C.J. Lin, “LIBSVM,” http://www.csie.ntu.edu.tw/~cjlin/libsvm/

表 2　Confusion Matrix

正解＼結果	bicycle	bus	car	cat	cow	dog	horse	motorbike	person	sheep	合計
bicycle	3	2	4	2	1	6	0	1	1	2	22
bus	0	3	8	1	6	0	0	1	2	0	21
car	0	1	22	1	1	0	0	0	0	0	25
cat	0	1	5	5	4	3	0	0	2	3	23
cow	1	1	2	4	12	2	0	1	1	2	26
dog	0	0	3	1	1	16	0	0	1	1	23
horse	0	1	2	4	8	5	0	0	1	3	24
motorbike	1	3	1	2	4	6	0	2	3	0	22
person	1	1	1	4	3	1	0	0	6	3	20
sheep	0	0	3	1	8	2	0	0	2	4	20
合計	6	13	51	25	48	41	0	5	19	18	226

(a) $w = 0$（動き特徴のみ，認識率 32.3% ）

正解＼結果	bicycle	bus	car	cat	cow	dog	horse	motorbike	person	sheep	合計
bicycle	7	0	5	1	1	2	1	5	0	0	22
bus	0	6	8	0	0	0	0	6	1	0	21
car	0	0	24	0	0	0	0	1	0	0	25
cat	0	1	5	2	3	7	0	0	4	1	23
cow	0	2	2	1	7	6	0	4	2	2	26
dog	3	0	2	5	2	9	0	1	1	0	23
horse	1	0	3	4	2	5	2	4	2	1	24
motorbike	3	1	1	2	1	3	1	7	3	0	22
person	3	0	2	4	1	4	0	0	5	1	20
sheep	0	0	4	1	2	5	0	1	3	4	20
合計	17	10	56	20	19	41	4	29	21	9	226

(b) $w = 0.2$（提案手法，動き特徴＋ BoF 特徴，認識率 32.3% ）

正解＼結果	bicycle	bus	car	cat	cow	dog	horse	motorbike	person	sheep	合計
bicycle	10	0	7	0	0	1	1	3	0	0	22
bus	4	6	4	0	0	0	0	7	0	0	21
car	1	3	19	0	0	1	0	0	1	0	25
cat	0	0	3	3	2	13	0	0	2	0	23
cow	1	3	1	0	4	9	1	4	2	1	26
dog	4	0	2	4	4	7	1	1	0	0	23
horse	4	2	7	0	0	2	1	6	1	1	24
motorbike	2	1	8	0	1	1	1	5	3	0	22
person	2	0	4	4	0	6	0	2	1	1	20
sheep	4	0	2	1	1	7	1	1	1	2	20
合計	32	15	57	12	12	47	6	29	11	5	226

(c) $w = 1$（BoF 特徴のみ，認識率 25.7% ）