動きの連続性を考慮した動画からの局所的な時空間特徴の抽出

野口　顕嗣†　　柳井　啓司†

† 電気通信大学大学院 情報工学専攻　〒 182–8585 東京都調布市調布ヶ丘 1–5–1
E-mail: †noguchi-a@mm.cs.uec.ac.jp, ††yanai@cs.uec.ac.jp

あらまし　本論文では時空間特徴に基づく映像からの特徴量の抽出手法について提案する．従来の局所的な時空間特徴は cuboid と呼ばれる立方体を抽出して，それを特徴化する手法が一般的に用いられている．しかし cuboid 全体を特徴化することは非常に計算コストが高い．そこで本研究では特徴的な点のみを抽出し，その局所的な追跡を行うことで，特徴ベクトルを求める．本研究の手法は視覚特徴抽出部と動き特徴抽出部の二つに分かれる．視覚特徴抽出部では SURF 検出器に基づき特徴の候補点を求める．次に求められた候補点に関して動き特徴を求めていく．その際に一定の局所的な時間において分割を行いオプティカルフローを計算することで動きの連続性を考慮した特徴量を構築していく．特徴量を変化に関して頑健にするために．本研究では視覚特徴の回転に合わせてフローを回転させる事で，回転に対して頑健な特徴を構築する．実験として人間の簡単な 6 つの動作を分類することで評価を行い，結果として 85%という精度が得られた．またこの特徴の応用例として Web 動画のショット分類を行うことで特徴量の有効性を示した．

キーワード　映像認識，動作認識，時空間特徴，Web 動画分類

Extracting Spatio-temporal Local Features Considering Consecutiveness of Motions

Akitsugu NOGUCHI† and Keiji YANAI†

† The University of Electro-Communications　1-5-1 Chofugaoka,　Chofu-shi Tokyo,　182-8585
E-mail: †noguchi-a@mm.cs.uec.ac.jp, ††yanai@cs.uec.ac.jp

Abstract　Recently spatio-temporal local features are proposed as image features to recognize events or human actions in videos. In this paper, we propose a novel local spatio-temporal feature which is applicable to large amounts of video data. Our method consists of two parts: extracting visual features and extracting motion features. First, we select candidate points based on the SURF detector, which is a very fast detector. Next, we calculate motion features at each points with local temporal units divided in order to consider consecutiveness of motions. Since our proposed feature is intended to be robust to rotation, we rotate optical flow vectors to the dominant direction of extracted SURF features. In the experiments, we evaluate the proposed spatio-temporal local feature with the common dataset containing six kinds of simple human actions. As the result, the accuracy achieves 85%, which is almost equivalent to state-of-the-art. In addition, we make experiments to classify large amounts of Web video clips downloaded from Youtube.

Key words　video recognition, action recognition, spatio-temporal local feature

1. はじめに

今日では Web 上，及び個人が所有している動画の数が爆発的に増えてきている．それにつれて映像を解析するコンテントベーストなアプリケーションは重要なポジションになりつつある．例えば映像の重要な部分だけを表示するような映像要約や，内容的に似通った動画を探す類似動画検索のようなシステムは，ユーザの見たい動画を効率的に探す手助けとなり得る．

本論文では，映像から効果的な特徴量を抽出する手法を述べる．過去の研究においては映像からの特徴抽出の手法は大きく二つに分類される．一つ目はビデオ全体を特徴化する手法，二つ目が短い時間から局所的な特徴を抽出する手法である．本研究では後者の局所的な特徴を抽出する手法を提案する．

局所的な時空間特徴は，動画の中から cuboid と呼ばれる立方体を抽出する手法が一般的に用いられている．しかしこの手法では cuboid のサイズをどのようにするか，その cuboid をどのように特徴化するかという問題がある．[1], [2] では cuboid から HoG や HoF を抽出する

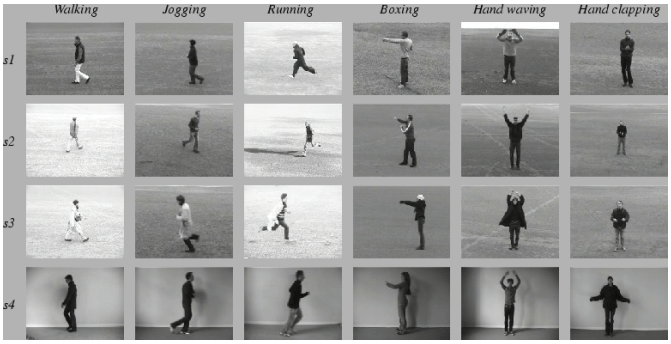

図 1　KTH データセット

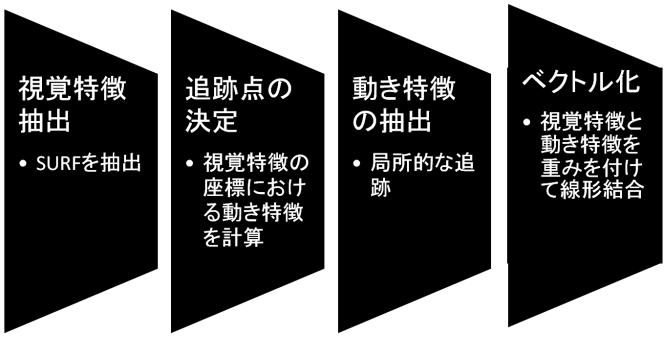

図 2　提案手法の概要

手法を提案している．しかし cuobid の内部からそれらの特徴を抽出することは計算コストが非常に高い．また立方体全体を特徴化するので周りのノイズに対して敏感になる事が考えられる．

そこで本研究では，時空間的に特徴的な点のみを抽出し，その点を追跡することで単純で高速，効果的な，新規的な特徴量を構築する．

局所的な特徴は，オクルージョンなどのノイズに対して頑健という特徴がある．そのため画像認識においては主流となっている．本研究でもそれに従い映像から局所的な時空間特徴を抽出する．ここで抽出される特徴はオクルージョン，照明変化，回転に対して頑健であることが望まれる．例えば，同一の「歩く」という動作において「右から歩く」と「左から歩く」から同一の特徴量が抽出されなければならない．

これを行うために，SURF 記述子 [3] で求められた回転を利用し，求められたフローを視覚特徴に合わせて，回転を行う．実験によって，この回転は精度向上に貢献することが分かった．

実験として図 1 に示すような動作認識の標準データセットである KTH を利用した．このデータセットには walking，running，jogging, boxing，hand waving, hand clapping という動作がある．ただし一つのビデオにつき一つの動作以外は存在せず，動作をする人間も一人だけである．結果として 85%という結果が得られた．

最後にこの特徴量の有効性を示すためにクラスタリングによる Web 動画のショット分類を行った．用いたデータは Youtube から収集された 100 本のサッカー動画である．

以降では 2 節で関連研究の紹介を行う. そして 3 節にて提案手法の説明を行い，4 節で実験について述べる. 最後に 5 節でまとめと今後の課題を挙げる.

2. 関 連 研 究

多くの研究で人間の動きの認識を行うため特徴抽出を試みている．それらは大きく分けて二つに分けることが出来る．一つ目が人間の体の主要な部位を追跡するという手法である [4]〜[6]．しかしこれら手法は正しく人間の部位を特定できることと，正しい追跡が行われることが前提となっている．

もう一つの手法が局所的な Cuboid と呼ばれる立方体を動画から抽出し，それを特徴化する手法である．Dollor らは空間にガウシアンカーネルを空間軸に一次元ガボールフィルタを適応し Cuboid を探し出す手法を提案した [1]. またこの研究ではこの局所的な特徴を visual word 化し分類を行っている．本研究でもこれにならい，抽出された特徴は visual word 化した後に行動分類を行う．

また Laptev らは STIP(spatio-temporal interest point) と呼ばれる特徴量を提案している [2]．これはハリスコーナ検出の 3 次元拡張である．この研究では抽出されてた Cuboid を HoG や HoF で表現することを提案している．

Alireza らは Cuboid から低レベルオプティカルフローを抽出し，ブースティングを行うことで，精度を向上させる手法を提案している [7].

しかし cuboid 全体を特徴化することは計算コストが非常に高い．また正確な cuboid のサイズを求めることは非常に困難な問題である．そこで本研究では，高速に画像の特徴点を求めることが出来る SURF 検出器と動き特徴を基に高速に時空間的特徴点を求める．その後その点を局所的に追跡することで特徴量を計算する新しい手法を提案する．

3. 提 案 手 法

図 2 は提案手法の概要を示している．まずフレーム画像に対して，SURF (Speeded-up Robust Feature) [3] を抽出する (第一フェイズ)．これは画像から局所的な特徴を抽出する手法である．ここで抽出された特徴の座標が追跡の候補点となる．次にその候補点の中から実際に追跡を行う点を決定していく．本論文では時空間的な特徴の抽出を目的としている．よって空間的に特徴的な点でも，以降動きがない点は抽出される特徴として適していない．そこで各候補点に関して動きを計算し動きがあった点を追跡点とする (第二フェイズ)．その追跡点において局所的な追跡を行うことで，動き情報を得る (第三フェイズ)．最後に視覚特徴と動き特徴を重みを付けて結合することで，特徴点をベクトル化する (第四フェイズ).

図 3　SURF の抽出

3.1　視覚特徴抽出

視覚特徴として SURF を抽出する．SURF とは照明変化，スケール変化，回転に対して頑健な特報量である．図 3 は実際に SURF で抽出された特徴を示した画像である．似たような特徴量として SIFT [8] がある．しかし SURF は SIFT に比べ処理が高速で，同等の精度がある．以下では SURF の概要について説明していく．

3.1.1　SURF 検出器

SURF は主に検出部と記述部の二つに分けることが出来る．ここではまず検出部の概要について説明していく．

SURF 検出器は積分画像を用いることで処理の高速化を図っている．座標 $\mathbf{X} = (x, y)^T$ における積分画像 $I_\Sigma(\mathbf{X})$ は式 1 で示させるようになる．

$$I_\Sigma(\mathbf{X}) = \sum_{i=0}^{i \leqq x} \sum_{j=0}^{j \leqq y} I(i, j) \tag{1}$$

ただし $I(x, y)$ は座標 (x, y) における輝度値を示している．要するに (x, y) における値は原点から (x, y) の輝度値の合計値で表される．あらかじめこの値を計算することで，これ以降のフィルタ処理が定数時間で行えるようになる．

SURF 検出器はヘッセ行列に基づき候補点を求めていく．画像 I 座標 $\mathbf{X} = (x, y)$ のヘッセ行列 $H(\mathbf{X},\sigma)$ は式 2 のようになる．

$$H(\mathbf{X},\sigma) = \begin{bmatrix} L_{xx}(\mathbf{X},\sigma) & L_{xy}(\mathbf{X},\sigma) \\ L_{xy}(\mathbf{X},\sigma) & L_{yy}(\mathbf{X},\sigma) \end{bmatrix} \tag{2}$$

ただし $L_{xx}(\mathbf{X},\sigma)$ は画像 I の座標 $\mathbf{X}$ におけるガウシアン二階導関数 $\frac{\partial^2}{\partial x^2} g(\sigma)$ で平滑化したものである．また $L_{yy}(\mathbf{X},\sigma)$, $L_{xy}(\mathbf{X},\sigma)$ についても同様である．

ただしここでガウシアンの二階導関数に関してブロブを行うことでレスポンスマップを作成する．ブロブというのは一定の範囲にある数をまとめることである．図 4 は L_{yy}, L_{xy} におけるガウシアン二階導関数とブロブを行い作成されたレスポンスマップを示している．このフィルタを適応することで得られる値を D_{xx}, D_{yy}, D_{xy} とする．そしてレスポンス関数はヘッセ行列式の近似で表される．

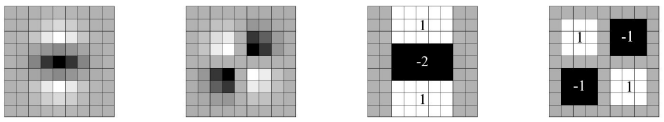

図 4　y 方向と xy 方向のガウシアン二階導関数 (左) とブロブによって作成されたレスポンスマップ (右)

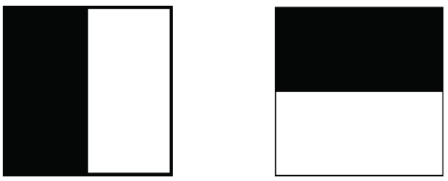

図 5　haar wavelet フィルタ．左が x 方向に関する傾き (dx)，右が y 方向に関する傾き (dy)

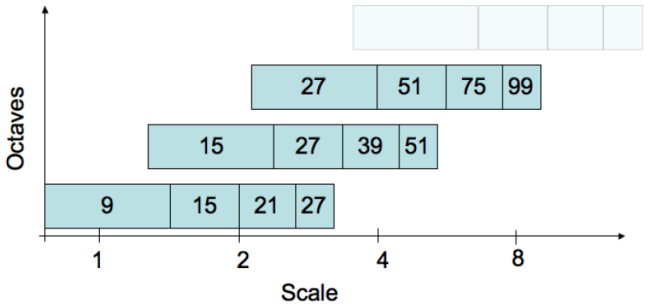

図 6　フィルタの拡大の概要

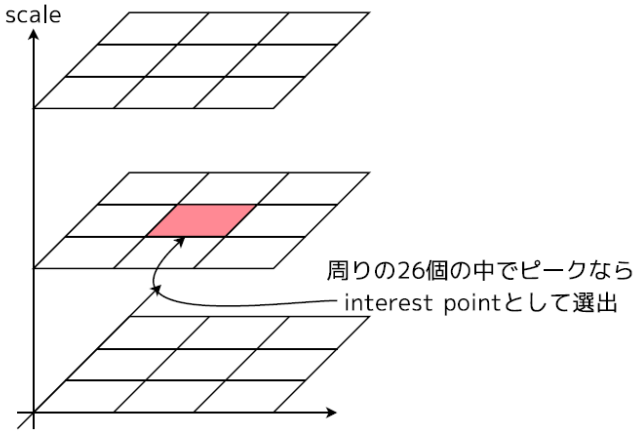

図 7　non-maximum suppression

$$det(H_{approx}) = D_{xx}Dyy - (0.9 \times D_{xy})^2 \tag{3}$$

最後にスケール描写であるが，SIFT などの手法では画像をダウンサンプリングすることでスケールを求めていた．しかし積分画像の導入で，フィルタ処理が高速化されているので画像サイズを変えずにフィルタの数を大きくしすることで，計算の高速化をはかる．図 6 はその概要を示している．ただし横軸がスケールを縦軸がオクターブの番号を，図の中の数字はフィルタサイズを示している．

まず第一オクターブ目はフィルタサイズが 9 から始ま

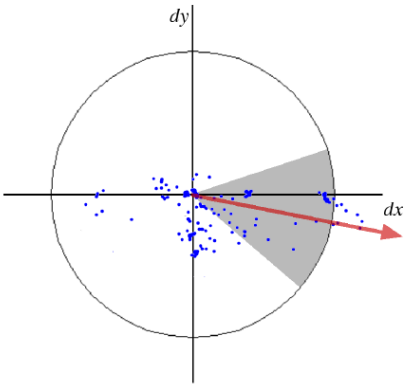

図 8　Dominant rotation の決定

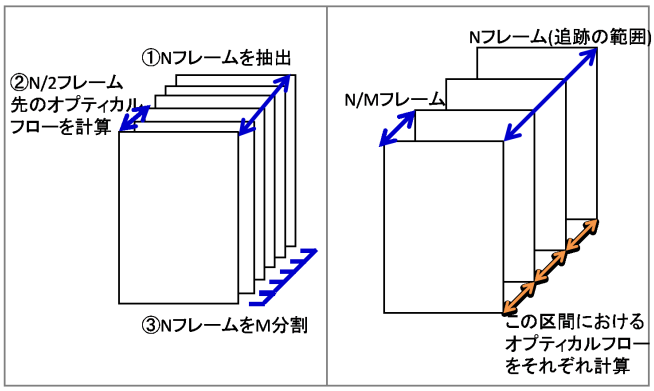

図 9　全体の動き特徴抽出概要 (左)，局所追跡の概要 (右)

る．その後サイズを 6 づつ上げて 27 まで繰り返す．次に第二オクターブはフィルタサイズ 15 から始まり，サイズを 6 × 2 づつ上げていき 51 まで繰り返す．第三オクターブではさらに上げていくサイズを倍にして同様の作業を繰り返していく．

次に図 7 に示すように隣接するスケールにおいて 3 × 3 × 3 近傍に対して non-maximum suppression を行い，ピークなら特徴点として選ばれる．

3.1.2　SURF 記述子

記述子は Haar Wavelet に基づき求められる．そのために図 5 に示すようなフィルタを用いて特徴点周辺の dx, dy を求めていく．その dx, dy を図 8 に示すようにプロットしていく．図 8 のように回転する窓を利用し，この窓から見えるベクトルを加算していき，最も長くなった時のベクトルを Dominant rotation とする．

その後特徴点の周辺を 4 × 4 に分割しそれぞれの区画において $\sum dx, \sum dy, \sum |dx|, \sum |dy|$ を計算する．よって各特徴点は $4 \times 4 \times 4 = 64$ 次元のベクトルになる．

3.2　動き特徴抽出

前節で求められた視覚特徴を基にして，その点を局所的な追跡を行うことで動き特徴を抽出する．ここではその手法について述べる．

3.2.1　追跡点の決定

図 9(左) に示すように SURF を抽出したフレームから

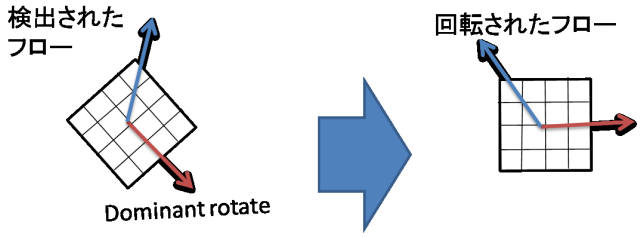

図 10　オプティカルフローの回転

N フレーム先までのフレームをとる．この N が追跡を行う時間に相当する．実験においては N=5 とする．その後，視覚特徴における特徴点を N/2 フレーム先のフレームに対して Lucus-Kanade アルゴリズム [9] を用いてオプティカルフローを計算する．特徴点の中からフローが検出されたものを追跡点として利用する．

3.2.2　局所的追跡

次に図 9(左) に示すように抽出された N フレームを M 分割する．ここで空間的な分割を行うことによって，動きの連続性を考慮した特徴を抽出することが出来るようになる．M の値が N に近いほど，より精密な追跡が可能になる．逆に M が 1 に近いほど，追跡が簡約化される．実験においては $N = 5, M = 5$ に設定してある．図 9(右) は局所追跡の概要を示している．追跡点に関して N フレームの局所的な追跡を行うが，その特徴量として，各分割区間におけるオプティカルフローの角度を利用する．

各区間のオプティカルフローは x^+, x^-, y^+, y^-， フロー無しの 5 次元で表現される．ただし x^+ はオプティカルフローの x 成分の正の方向を，x^- は x 成分の負の方向を示している．また各区間における 5 次元の特徴は合計が 1 になるように正規化を行う．よって本手法において動きの大きさは無視されることになる．また一点におけるオプティカルフローの計算なので，この特徴はフローの角度を意味することになる．動きの次元数は $(M-1) \times 5$ となる．

ただし，このままの計算では動き特徴は回転に関して敏感になってしまう．同じ「歩く」という動作においても歩いている方向によって異なる，動き特徴が抽出されることになる．

そこで本研究では，視覚特徴で得られた dominant rotate を使い，オプティカルフローを回転することを考える．図 10 が抽出された視覚特徴とオプティカルフローが実際に回転される様子を示している．

特徴点の座標を (x_1, y_1)，オプティカルフローが検出された座標を (x_2, y_2) とした，SURF の dominant rotate を θ とした場合，回転されたフローの座標 (x, y) は式 4 で定義されるものとなる．

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} cos\theta & -sin\theta & x_2 \\ sin\theta & cos\theta & y_2 \end{bmatrix} \begin{bmatrix} x_1 - x_2 \\ y_1 - y_2 \\ 1 \end{bmatrix} \quad (4)$$

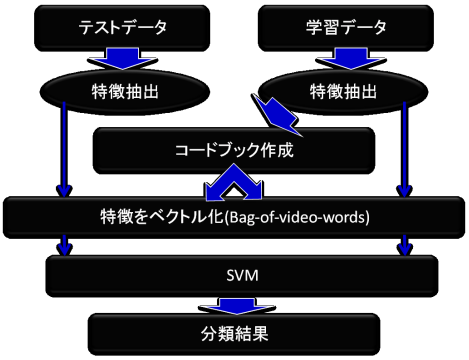

図 11　行動分類手法の概要

3.3　特 徴 結 合

抽出した視覚特徴を *visual*，動き特徴を *motion* とすると，本論文では 64 次元の *visual* に $5 \times (M-1)$ 次元の *motion* に重み *weight* を掛けたものを結合した $64+5 \times (M-1)$ 次元ベクトルを特徴ベクトルとして利用する．実験においては M=5 としたためベクトルの次元数は 84 次元となった．

4.　実　　験

本研究では人間の単純な動きを分類することで特徴の評価を行う．ここでは実験に使用したデータセット，実験結果について述べていく．

また最後にこの特徴の応用例として，Web 動画のショット分類を行うことで，特徴の有益性について考える．

4.1　行 動 分 類

これまで映像からの特徴の抽出について述べてきたが，本論文では実験として人間の単純な動きを分類する．ここではその分類手法について述べる．

Dollar [1] らは人間の動きの分類手法として，特徴を visual word 化して分類を行っている．本手法でもビデオから抽出された特徴を基にして bag-of-video-word を作成し，SVM によって分類を行う．

図 11 は分類手法の概要を表している．まずデータセットを学習データとテストデータに分ける．それぞれのデータセットから本論文の手法で特徴を抽出する．その後テストデータから抽出された全ての特徴に対して k-means 法を用いてクラスタリングを行うことでコードブックを作成する．

次にこのコードブックをもとに bag-of-video-word(BoVW) を構築していく．BoVW は bag-of-feature(BoF) [10] を動画に適応したものである．BoF は特徴の位置関係を無視して，特徴の出現頻度をベクトル化したものである．コードブックから特徴の出現頻度を求める．その特徴を用いて，support vector machiene(SVM) で学習，分類する．SVM のカーネルには RBF カーネルを使用した．

4.1.1　データセット

人間の動きのデータセットとしてここでは，KTH を使用する．分類される動作として walking，running，jogging，boxing，hand clapping，hand waving の 6 つのものがある．「カメラは動かない」,「一つの映像中には一つの動作しかない」ことが前提として存在している．

それぞれの動作において 25 名の人が，4 つの異なる背景でその動作を行っている．よって各動作について 100 のデータが存在する．実験ではそのデータを 5 人単位に区切ることで 5 fold cross validation で学習，テストを行い，それぞれの validation セットの合計値をもって評価を行う．

ただし KTH の動画の長さは平均で 20 秒ほどであり，本手法を用いることで，それぞれの動画の中から 4000 程度の特徴を抽出することが出来た．

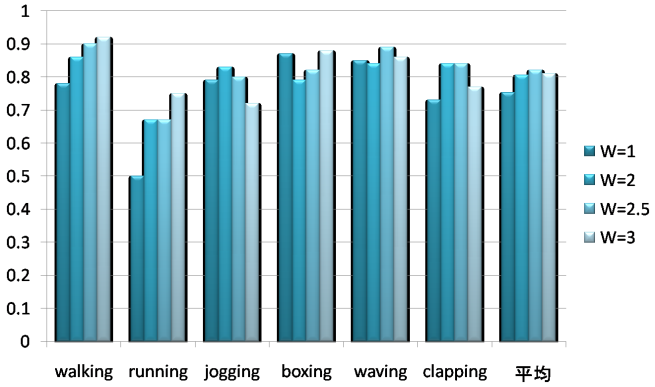

図 12　動作の重みにおける行動分類の結果

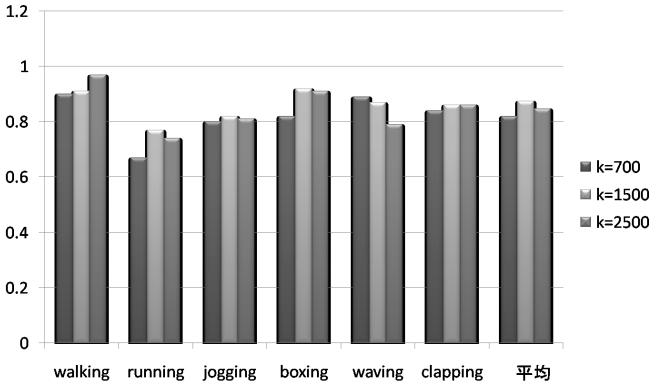

図 13　codebook サイズにおける行動分類の結果

4.1.2　評　　価

まず動き特徴の重み *weight* とコードブックサイズ k の最適化を行う．まず図 12 は重み *weight* を変更していったときの分類率を示している．ただしここでコードブックサイズ k は 700 とする．

次に図 13 ではコードブックサイズを変更していったときの分類率である．ただしこのときの *weight* の値は 2.5 とする．この実験の結果 $weight = 2.5, k = 1500$ のときに最適な分類率であることが分かった．以下の実験ではこの値を用いる．

本論文では特徴量を変えた 4 つのシステムについて評価を行っている．実験に用いた特徴は以下の四つである．

（1） 視覚特徴＋動き＋回転考慮 (VMR)
（2） 視覚特徴＋動き (VM)
（3） 視覚特徴 (V)
（4） 動き特徴 (M)

結果は図 14 に示す通りになった．視覚特徴と動き情報を組合わせたもの VMR と VM は特徴を単独で用いるよりも精度が格段に向上することが分かった．また動きの回転を考慮したシステムでは，若干であるが考慮しないシステムよりも改善が見られた．

動きのみの場合 running，jogging，walking，hand waving においてはある程度の精度を出すことが出来ているが，boxing，hand clapping においては精度が大幅に下っている．図 1 に示すように boxing，hand clapping は左右への動きがメインになっており，動き情報だけでは分類が出来なかった．hand waving の場合では，これも左右の方向がメインではあるが，斜め方向への動きも含まれるのでこの場合ある程度分類できていると考えられる．

逆に視覚特徴のみでは良く動く部分が似通っている walking，running，jogging の分類率が極端に悪くなってしまっていることが分かる，

表 1 から表 4 はそれぞれの手法による，混合行列を示している．セルの値が大きくなるほど色が濃くなる．どの手法でも walking で高い精度となっている．一方で running と jogging の分類が困難であることが分かる．これはこの二つの動作が非常に似ていることが原因で，手動で分類した場合でも完全な分類を行うことが難しい動作である．

それぞれの手法について見ていく．まず表 1 では視覚特徴，動き特徴，回転を考慮した結果である．いずれの行動でも高い精度であることが分かる．

次に表 2，表 3 であるが視覚特徴のみでは walking，running， jogging の分類が困難であることに対して，動き特徴では boxing， waving， clapping の分類が困難であることを示している．

表 4 では視覚特徴と動き特徴を組合わせた結果であるが，ある程度高い精度であることが分かる．しかし表 1 と比較して waving 以外は精度が少し下っていることが分かる．全体的な結果を見た場合でも回転を考慮した場合，85.5%，回転を考慮しない場合，83.3%であった．このことからも回転を考慮した動き特徴は精度を向上させることが分かる．

最後に提案手法を最新の手法と比較する. 比較に用いた手法は，Dollar らの手法 [1]，Alireza らの手法 [7]，Laptev らの手法 [2] である．図 15 は比較結果を示している. 本手法における分類率は 85%, Dollar らの手法で 82.3%，Alireza らの手法で 91.5%，Laptev らの手法では 91.8%であった．どの動作においても最新の手法には一歩及ばないことが分かる．

次に図 16 に提案手法と [1] の計算時間の比較を行った結果を示す．ただし [1] の手法は，我々が独自に実装したものを利用した．これは KTH のデータセット 600 の

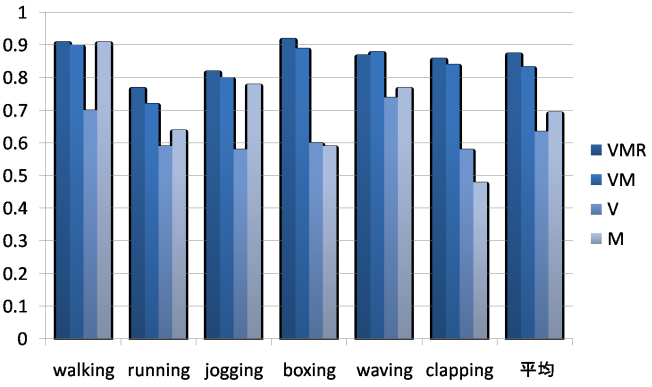

図 14　手法毎の行動分類の結果

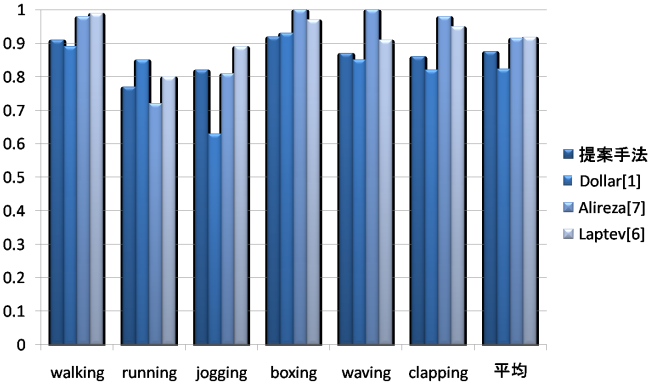

図 15　行動分類における最新手法との比較

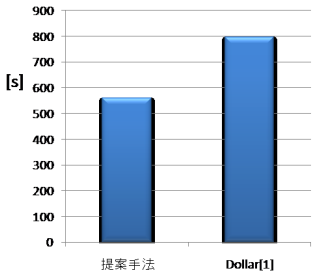

図 16　特徴抽出における計算時間の比較

動画から特徴点を検出するのにかかった時間を示したものである．実際に時間を比較した場合 2/3 の時間で特徴点の探索が出来ることが分かる．精度は最新手法に比べ，若干劣るが計算コストの削減に成功していることが分かる．

4.2　Web 動画のショット分類

本論文で提案した特徴を応用して，クラスタリングによって Web 動画のショット分類を行う. これはユーザが見たい動画を効率的に探すための，手助けとなり得る. 図 17 はその手法を示している.

まず収集された Web 動画をそれぞれショット分割する. これは隣接するフレームの色特徴を比較することで行われる. 色特徴としては HSV 色空間を使用した. また二つのフレームの距離として，χ^2 距離を使用した (第一

表 1　視覚特徴＋動き特徴＋回転あり

	walking	running	jogging	boxing	waving	clapping
walking	0.91	0.01	0.06	0.01	0.01	0
running	0.02	0.77	0.21	0	0	0
jogging	0.03	0.15	0.82	0	0	0
boxing	0	0	0	0.92	0.02	0.06
waving	0	0	0	0.05	0.87	0.08
clapping	0	0	0	0.07	0.06	0.86

表 2　視覚特徴のみ

	walking	running	jogging	boxing	waving	clapping
walking	0.7	0.13	0.16	0.01	0	0
running	0.1	0.59	0.21	0	0	0
jogging	0.12	0.29	0.58	0	0	0.01
boxing	0.13	0.13	0.1	0.6	0.03	0.01
waving	0.03	0.09	0.01	0.05	0.74	0.08
clapping	0.04	0.05	0.02	0.06	0.25	0.58

表 3　動き特徴のみ

	walking	running	jogging	boxing	waving	clapping
walking	0.91	0	0.06	0.03	0	0
running	0	0.64	0.3	0	0.02	0.04
jogging	0.04	0.13	0.78	0.02	0.03	0
boxing	0.01	0	0	0.59	0.32	0.08
waving	0	0	0.01	0.17	0.77	0.05
clapping	0	0	0	0.18	0.33	0.48

表 4　視覚特徴＋動き特徴＋回転無し

	walking	running	jogging	boxing	waving	clapping
walking	0.9	0.01	0.07	0.01	0	0
running	0.01	0.72	0.27	0	0	0
jogging	0.01	0.18	0.8	0.01	0	0
boxing	0	0	0	0.89	0	0.11
waving	0	0	0	0.06	0.88	0.06
clapping	0	0	0	0.13	0.02	0.84

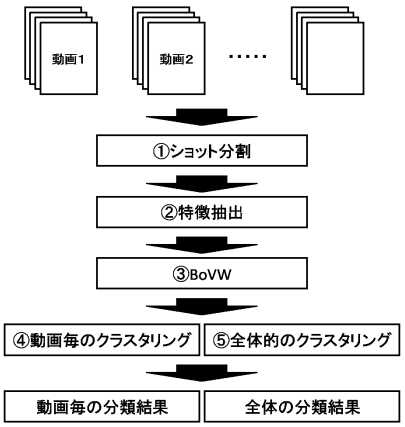

図 17　Web 動画分類手法

フェイズ).

次に各ショットから本論文で提案した特徴を抽出する (第二フェイズ). その後行動分類の手法と同様に bag-of-video-words を用いる (第三フェイズ). 分類対象であるが, 本論文では動画毎に対する分類 (第四フェイズ) と, 収集された動画全体における分類 (第五フェイズ) を k-means クラスタリングによって行うこととする. ただし $k = 50$ とした.

4.2.1　収 集 動 画

ここでは Youtube から収集したスポーツの Web 動画を使用する. ここではサッカーのキーワードで収集した 100 の動画を使用する.

4.2.2　評　　価

図 18 はあるサッカー映像のショットのクラスタの一例である. 上は比較的遠くから撮影されたショット, 真ん中が近くで撮影されたショット, 下が人をズームアップして撮影されたショットと, 効率的なクラスタリングが出来ていることが分かる.

図 19 にすべての動画におけるショットの分類結果の一部を記載した. 上が遠くからのアングルで撮影されたものが集まったクラスタ, 真ん中が比較的選手を大きく撮影したクラスタとなっていることが分かる. 一方で一番下のクラスタのように様々なものが混ざりあって構成されているクラスタも存在した. 提案手法を用いたショット分類は概ね上手く動作しており, この特徴量の応用性を示している.

しかし本実験において各ショットから平均で約 1 万個, 多いもので 20 万個の特徴が抽出された. 特にカメラモーションがある動画に関しては, 抽出される視覚特徴は, そのまま追跡点となってしまうので, 特徴数は多くなっていく傾向があることが分かった. それに従い計算コストも高くなった. 実用的なシステムにするためには抽出される特徴数は出来るだけ減らし, 計算コストを低くする必要がある. その問題をどのように扱うかは今後の課題となってくる.

またカメラモーションがある動画では, 動き情報が正しく反映されない可能性がある. よってカメラモーションの方向と早さを特定するシステムが必要となってくる.

5.　おわりに

本論文では新規的な時空間特徴の抽出法について提案した. この手法は視覚特徴抽出部と動き特徴抽出部に分かれている. 視覚特徴抽出部では SURF に基づいて特徴点を検出, ベクトル化する. その候補点に対して動き特徴抽出部では動き特徴を計算する. この際に時間軸において分割を行うことで動きの連続性を考慮した特徴量を構築する. また回転に関して頑健にするために検出された動き情報は SURF の dominant rotation に合わせて回転される.

評価として KTH データセットを用いて 6 種類の人間の簡単な動作について分類実験を行った. 結果として 85%という高い分類率であった. また回転を考慮した動き特徴は動画分類の精度向上に貢献することが分かった. また最新手法と比較をすると精度では一歩及ばないが, 計算コストを削減出来ていることが確認出来た.

今後の課題として大きく二つの方向が考えられる. 一つ目がこの特徴量の精度の向上である. そのためには特徴量の追加, 特徴記述の改良, 視覚特徴と動き特徴の結合手法の改良, カメラモーションの検出などが挙げられる.

二つ目は, 本論文で行ったような, この特徴を利用した, システムの開発である. 例を挙げると, 類似動画検

図 18　一つの動画のクラスタリング結果: 遠くからの (上)，比較的近くのショット (中)，人をクローズアップしたショット (下)

図 19　全ての動画ショットのクラスタリング結果: 遠くから撮影されたショット (上)，近くからのショット (中)，様々なものが混ざったクラスタ (下)

索，動画要約，自動サーベイランスシステムなどが考えられる．

文　　献

[1] P. Dollar, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proc. of Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65–72, 2005.

[2] I. Laptev and T. Lindeberg. Local descriptors for spatio-temporal recognition. In *Proc.of IEEE International Conference on Computer Vision*, 2003.

[3] B. Herbert, E. Andreas, T. Tinne, and G. Luc. Surf: Speeded up robust features. In *CVIU*, pp. 346–359, 2008.

[4] C. Fanti and P. Perona. Hybrid models for human motion recognition. In *Proc.of IEEE Computer Vision and Pattern Recognition*, 2005.

[5] C. Rao, A. Yilmaz, and M. Shah. View-invariant representation and recognition of actions. *Int J Comput Vision*, Vol. 50(2), pp. 203–226, 2002.

[6] Y. Yacoob and M.J. Black. Parameterized modeling and recognition of activities. *Comput Vis Image Und*, Vol. 72(2), pp. 203–226, 2002.

[7] F. Alireza and M. Greg. Action recognition by learning mid-level feature. In *Proc.of IEEE Computer Vision and Pattern Recognition*, 2008.

[8] D. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, pp. 91–110, 2004.

[9] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. of International Joint Conference on Artificial Intelligence*, pp. 674–679, 1981.

[10] G.Csurka, C.Bray, C.Dance, and L. Fan. Visual categorization with bags of keypoints. In *Proc. of ECCV Workshop on Statistical Learning in Computer Vision*, pp. 1–22, 2004.

[11] S.Konrad and G.Luc. Action snippets: How many frames does human action recognition require? In *Proc.of IEEE Computer Vision and Pattern Recognition*, 2008.