動作認識のための時空間特徴量と特徴統合手法の提案

野口 顕嗣† 下田 保志† 柳井 啓司†

† 電気通信大学大学院 電気通信学研究科 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,shimoda-y,yanai}@mm.cs.uec.ac.jp

あらまし 映像中の動作を認識することは，様々なアプリケーションに応用することが可能で，多くの分野で扱われてきた研究課題であるが，Youtube 動画のような制限のない環境における認識を行った研究は少ない．そこで本研究では，新しい時空間特徴を提案し更に Multiple Kernel Learning(MKL) に基づく特徴統合による動作認識フレームワークを提案，Web 動画における動作認識を行った．提案した時空間特徴は，点の周辺パターンとその点の微少時間の動きを特徴化したものである．統合に利用した特徴は，時空間特徴，視覚特徴，動き特徴である．実験には KTH データセットと，二種類の Youtube データセットの計 3 種類のデータセットにおける分類を行った．結果として KTH データセットで最新手法に匹敵する分類率の 94.0%，Youtube データで最新手法を大幅に上回る分類率である 80.4%という結果が得られた．

キーワード 時空間特徴，動作認識，KTH，Youtube data，特徴統合，MKL

Action recognition employing a new spatio-temporal feature and a feature fusion method

Akitsugu NOGUCHI†, Yasushi SHIMODA†, and Keiji YANAI†

† Department of Computer Science, The University of Electro-Communications
Chōfugaoka 1–5–1, Chōfu-shi, Tokyo, 182–8585 Japan
E-mail: †{noguchi-a,shimoda-y,yanai}@mm.cs.uec.ac.jp

Abstract Since action recognition is one of the key techniques to organize a large number of videos automatically, action recognition has been studied by many researchers. Although most of the existing works have focused on videos taken in controlled environment so far, action recognition for uncontrolled videos such as Web videos has become important recently. In this paper, we propose a novel action recognition method by combining features based on Multiple Kernel Learning (MKL), which enables robust action recognition for both controlled and uncontrolled videos. In the proposed method, we combine appearance, motion and spatio-temporal features. As a spatio-temporal feature, we propose a novel feature which consists of local appearance and local motion. In the experiments, we evaluate our method using KTH dataset, and Youtube dataset. As a result, we obtain 94.0% as a classification rate for in KTH dataset which is almost equivalent to state-of-art, and 80.4% for Youtube dataset which outperforms state-of-the-art greatly.

Key words spatio-temporal feature, action recognition, KTH, Youtube data, combining featrues, MKL

1. はじめに

近年 Web 上の動画の数は爆発的に増えてきている．しかし一方でユーザが見たいシーンを探すことが非常に困難な問題となっている．現状の動画検索システムはタグなどによるテキストベーストな手法が用いられているが，これはユーザの主観によってつけられるもので，タグのみで動画を特定することは非常に難しい．そのため動画の内容を解析するコンテントベーストな検索手法が今後求められてくると考えられる．そのなかでも，動画の内容を解析し，分類を行うことは非常に意義のあることである．しかし現状このような Web 動画の分類を行った研究は少ない．

そこで本研究では，新しい時空間特徴の提案を行い，その特徴を利用することで，Web から収集された動画を，映像中で行われている動作に着目し分類を行う．分類には公開データセットにおける動作認識と独自に収集した大量 Web 動画における分類を行う．大量 Web 動画分類では教師信号ありの分類に加え，教師信号なしクラスタリングの二種類の分類を行う．また，それぞれの分類において Multiple Kernel Learning(MKL) に基づく特徴統合による精度の向上についても検討を行った．

動作認識 本研究では KTH データセット，Liu らが構築したデータセット [6]，本研究で独自に構築したデータセットの 3 種類のデータセットを用いて，動作認識を行う．本論文では独自に収集したデータセットを Our

Youtube データセット，Liu らが構築したデータセットを Wild Youtube データセットと呼ぶ．

教師信号あり分類　walking という単語で収集されたショットには walking を含むショットも存在するが，含まないショットも大量に存在する．そこで本実験では，そのショットが walking を含むかどうかランキング付けを行い，ランキングの上位には walking を含むショットが，ランキングの下位には walking とは関連のない動作を順位付けることを目的としている．

教師信号なしクラスタリング　soccer という単語で収集された動画には様々なシーンがある．そこで本実験では教師なしクラスタリングを行うことで，soccer を「試合のシーン」,「インタビューシーン」などのシーンに自動で分類することを目的としている．

2. 関 連 研 究

近年，動画の解析のために時空間特徴が注目を集めている．時空間特徴とは，動画から抽出される特徴の一つで動き情報と視覚情報を同時に表現することが可能な特徴である．

まず時空間特徴の抽出の考え方として，画像認識の三次元拡張の手法が挙げられる．Kobayashi らは自己相関特徴を三次元に拡張し，サーベイランスの分野に特化した cubic higher-order local auto-correlation(CHLAC) を提案した [1]．また Laptev らはハリス点検出を三次元に拡張した検出手法を提案している [2]．

次に主要な手法として，cuboid と呼ばれる立方体を抽出し，それを特徴化する手法がある．Dollar らは空間軸にガウシアンフィルター，時間軸にガボールフィルタを適応することでこの cuboid を抽出する手法を提案した [3]．また cuboid の記述子として，Laptev や Dollar らは Histgram of Orient Gradient(HoG) や Histogram of Orient Flow(HoF) で表現することを提案している．一方で Kläser らは，記述子として三次元的な HoG を利用することを提案している [4]．

しかしこのような cuboid 主体の手法は計算コストが高くなる傾向があり，本論文で行うような大量データを扱うのに向いていない．更に 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，特徴的な点と，その点の微少時間の動きを特徴化する新たな時空間特徴を提案する．この特徴は計算時間が少なく，かつ正確な分類が期待できる．

Web 動画に対する動作認識を行った研究は少ない．Cinbis らは Web 上から動作を自動学習する手法を提案し，Youtube データを動作認識に利用している [5]．この研究では，まず query word をもとにして画像を収集し，その画像から特徴を抽出することで動作モデルを構築し，実際の映像において認識を行っている．しかしこの手法の学習データは Web 上から収集された静止画像であり，動作の記述も全て動画像ではなく静止画像ベースで行われている．本研究では画像の視覚特徴のみではなく，動き特徴も考慮して動作を分類することを考える．

最も本研究の手法と類似している手法として，Liu らの研究が挙げられる [6]．この手法は特徴量として [3] で提案された時空間特徴を，視覚特徴として SIFT 記述子 [7] を利用し，Adaboost に基づき統合する手法を提案している．またこの研究では，Page Rank に基づく重要な特徴の選択を行っている．本研究では，動作を認識するために新たな時空間特徴を提案し，利用するが，この手法では既存の特徴をどのように扱うかに重点が置かれている．

3. 提案手法概要

本節では，時空間特徴抽出手法，特徴統合による Web 動画分類手法の概要について述べる．

3.1 時空間特徴抽出手法

時空間特徴には [8] を拡張した手法を利用する．まずフレーム画像から，SURF [9] に基づく視覚特徴を抽出する．SURF とは，回転，スケール変化に頑健な局所特徴である．次に，抽出された SURF の座標に対して，Lucas-Kanade アルゴリズム [10] に基づいて動きを計算する．ここで動きのなかった点は時空間特徴として適さないので，これらの点は排除し，動きのあった点のみを取り扱うこととする．その後，決定された時空間特徴点に関して Delaunay 三角分割法を適応する．以降特徴は三角形の頂点を構成する三点のペアで表現される．更に時空間特徴点の微少区間の動きを特徴化することにより動き特徴を抽出する．最後に抽出された視覚特徴と動き特徴を重みをつけて結合することで時空間特徴を抽出する．特徴抽出の手順は，手順 2 にて述べられている．

ここで抽出された特徴は Dollar らの手法 [3] と同様に bag of statio-temporal features(BoSTF) 表現によって認識に利用される．

3.2 Web 動画分類

手順 1 は Web 動画のショット分類の概要を示している．まず特定のキーワードで Youtube から動画を収集する．次に集められた動画を色特徴に基づきショット分割する．その後，それぞれのショットから時空間特徴，動き特徴，視覚特徴を抽出する．最後に分類を行うが，本研究では SVM や MKL による教師あり分類の他に，pLSA や k-means による教師なしクラスタリングも試みる．

手順 1 Web 動画分類の流れ

step1　：動画収集
step2　：色特徴に基づくショット分割
step3：特徴抽出
　▷step3-1：視覚特徴，動き特徴，時空間特徴
step4　：教師信号あり分類
　▷step4-1：ランダムに学習データを選択
　▷step4-2：SVM，MKL の出力値によるランキング
step5　：教師信号なしクラスタリング
　▷step3-2：k-means，pLSA によるクラスタリング

教師信号ありのショットランキング付けでは，例えばランダムに選ばれた running 学習セットで SVM のような分類器を学習する．その分類器によって収集した Web 動画ショットの分類を行う．その結果，ランキングの上位には running を含むショットが，ランキング下位には running を含まないショットがランキングされることが期待される．本研究では提案した特徴単体での分類と，MKL に基づく特徴統合による分類を行う．統合に利用した特徴は，時空間特徴，動き特徴，視覚特徴の三種類の特徴である．

次に教師なしの分類では，クラスタリングによるシーンの自動分類を行う．これは例えば soccer という単語で集められたショットには様々なサッカーのシーンが含まれると考えれる．そこで pLSA や k-means のようなクラスタリング手法に基づき，クラスタリングすることで「ドリブルシーン」,「シュートシーン」,「インタビューシーン」のようなシーンに分類されることが期待される．

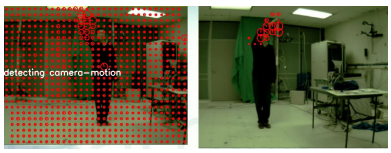


図 1　カメラモーション検出例 (左), カメラモーションが検出されない例 (右)

4.　提案手詳細

4.1　時空間特徴抽出

本研究では，Web 動画分類に適した新しい時空間特徴抽出手法を提案する．多くの動作認識の研究は，カメラモーションに対する対応がない．しかし Web 動画を分類するためには「カメラモーション」をどのように扱うかは非常に重要な問題となってくる．そこで本研究では，特徴を抽出する前にカメラモーションの検出を行う．

また既存の抽出手法は計算コストが高く，本研究のような大量のデータから抽出することに向いていない．本研究では，その問題を解決するために，点の周辺パターンと，その点の動きで動画を特徴化する手法を提案する．この手法は既存手法に比べ計算コストが低く，大量のデータから特徴を抽出することが可能である．

手順 2 に提案する特徴抽出手法の流れを示す．

手順 2 時空間特徴抽出手法の流れ

step1	：カメラモーション検出
step2	：時空間特徴における視覚特徴抽出
▷step2-1	：SURF 抽出
▷step2-2	：時空間特徴点の決定
▷step2-3	：Delaunay 三角分割
step3	：時空間特徴における動き特徴抽出
▷step3-1	：Lucas-Kanade 法による動き抽出
▷step3-2	：SURF の diminant rotation による方向の正規化
step4	：視覚特徴ベクトルと動き特徴ベクトルの結合

本手法は大きく 4 つに分けることが出来る．まず，カメラモーション検出部 (step 1)，二つ目に視覚特徴抽出部 (step 2)，三つ目が動き特徴抽出部 (step 3) ，最後に特徴統合部 (step 4) である．以下ではそれぞれのステップ毎に詳しく述べていく．

(Step 1) カメラモーション検出部　動画は撮影者の意図によって，ズームやパンなどのカメラモーションを含むことがある．しかし Web 動画においてのカメラモーションは手振れなどの，撮影者の意図しないカメラモーションが多く含まれている．また Web 動画は解像度が低い．これらのことから収集した動画の正確なカメラモーションを求めることは非常に難しい問題である．

Liu らの手法ではカメラモーションを検出した場合そのフレームを破棄することで対応をしている [6]．本研究でも，これを参考にし，カメラモーションを検出した場合，その特徴を破棄する．しかしカメラモーションを全て破棄していては,「動画全てにカメラモーションを含んでいるような動画から特徴が抽出されない」,「カメラモーション中に存在する重要な特徴を抽出できない」,などの問題点もある．

検出手法として，図 1 のようにグリッド上に Lucas-Kanade アルゴリズムに基づいて，動き情報を計算する．その中で動きのあった領域が一定以上だった場合カメラモーションの検出とする．

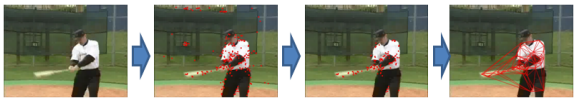


図 2　視覚特徴抽出の様子

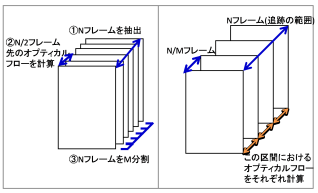


図 3　全体の動き特徴抽出概要 (左)，局所追跡の概要 (右)

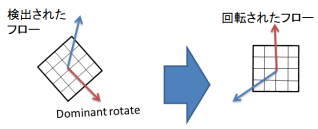


図 4　オプティカルフローの回転

(Step 2) 時空間特徴における視覚特徴抽出部　図 2 は視覚特徴の抽出の様子を示している．各ステップ毎に見ていくと，(1) フレーム画像から SURF を抽出する．本研究では動画としての特徴を抽出したいので，動きがない点は時空間特徴として適していない．(2) よってそれぞれの特徴点の動きを計算し，動きがなかった点を削除する (時空間特徴の決定)．(3) 最後に残った点について Delaunay 三角分割を行う．この時空間特徴を三角分割することは本手法独自のものである．これを行うことで，その点のみではなく隣接する特徴も考慮した特徴を構築することが可能となる．

視覚特徴には三角形の頂点を構成する三点の SURF 記述子を使用する．それぞれの三角形の頂点の点を SURF 記述子のスケール毎に整列させる．ただし同一スケールの特徴が存在した場合は，x 座標の最も小さい特徴点から見て右周りに特徴を配置する．SURF の次元数は 64 次元なので，結果として視覚特徴は $64 \times 3 = 192$ 次元で表現される．

(Step 3) 時空間特徴における動き特徴抽出部　動き情報として，本手法では三角形の頂点を構成する，それぞれの点の動きと，三角形の面積の変動を特徴化する．点の動きは視覚特徴の時と同様で SURF のスケールによって整列させる．

図 3 は動き特徴の抽出の概要を示している．まず図 3(左) が時空間特徴点の決定の様子を示している．最初に SURF を抽出したフレームから N フレーム先までのフレームをとり，以降この区間から動き特徴を抽出する．次に，抽出された SURF の点に対して，N/2 フレームの動き情報を計算する．この動き情報に基づいて時空間特徴は決定される (手順 2 の step 2-2)．

次に動き特徴の抽出に関して説明する，図 3(右) がその様子を示している．選択された N フレームを，M 分割する．そして分割したそれぞれの区間で Lucas-Kanade アルゴリズムによって，特徴点のオプティカルフローを計算する．ただし，インターバル区間 i の特徴点の座標 L_i は，前の区間 $i-1$ で推定された，動き情報によって求められる．この分割数 M が N の値に近いほど詳細な追跡が可能になり，逆に M が 1 に近くなると簡略な動き特徴が抽出される．

各区間の動き情報は x^+, x^-, y^+, y^-, 及び動きなし，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．実験において特徴を抽出する区間 N=5，分割数 M=5 に設定しているので，それぞれの点における次元数は $(M-1) \times 5$ で 20 次元で，三角形の面積変化は 5 次

図 5　動き特徴が有効な動作の例 (左)，視覚特徴が有効な動作の例 (右)

元で表現される．よって動きの次元数は $20 \times 3 + 5 = 65$ 次元となる．

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な特徴になってしまう．同じ「歩く」という動作においても，動作の方向によって異なる特徴が抽出されてしまう．そこで本研究では視覚特徴で得られる dominant rotation を利用し，オプティカルフローを回転させる．特徴は三点で一組になっているが，ここではそれぞれの dominant rotation を利用して，動き情報の回転を行う．その様子を図 4 に記載する．

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 rotation を θ とした場合，回転されたフローの座標 (x, y) は式 1 で定義されるものとなる．

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

最後に二つの特徴を単純に結合することによって時空間特徴を構築する．

4.2　MKL による特徴統合

動作によって重要となってくる特徴は変わってくる．図 5(左) ような “jogging” と “running” を視覚情報だけで判断することは難しいが，動き特徴を利用することで分類が可能となってくる．逆に図 5(右) は boxing と clapping という二つの動作を示しているが，これは動き特徴だけで判断しても，共に腕を左右に振っているので特徴的に類似したものになってしまう．しかし視覚特徴を用いることで，簡単に分類が出来る．

よって本研究では Multiple Kernel Learning(MKL) によって特徴の重みを自動で推定する手法を利用することで，これらのショットの分類を行った．統合には視覚，動き，時空間特徴の 3 種類の特徴を利用した．

4.2.1　Multiple Kernel Learning

本研究では動作認識を行うために Multiple Kernel Learning(MKL) を利用する．これは複数のサブカーネルを線形結合することで，新たな最適なカーネルを求める手法で，統合カーネルは以下の式で求められる．

$$K_{combined}(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{K} \beta_j k_j(\mathbf{x}, \mathbf{x}')$$

$$\text{with} \;\; \beta_j \geqq 0, \;\; \sum_{j=1}^{K} \beta_j = 1. \quad (2)$$

各サブカーネルの重み β_j の設定によって，統合された新たなカーネルの出力は変化する．そのためこの最適な重みをどのように求めるかが問題となり，この問題は MKL 問題と呼ばれている [11]．この MKL 問題は全ての β_j の組合わせを実際に試すことで解くことが出来る．しかし，特徴やカーネルの数が増えるにつれ，β_j の組合わせも爆発的に増えてしまう．実時間内にこの問題を解決するために，近年 MKL 問題を凸面最適化問題として解く手法が提案されている [12]．Sonnenburg らは単一カーネルの SVM 学習を反復することによって最適なカーネルの重み β_j を求める手法を提案している [12]．

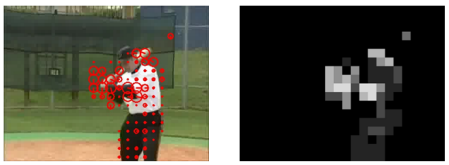


図 6　抽出された動き特徴

4.2.2　特 徴 抽 出

以下では統合に利用する動き特徴，視覚特徴，及び特徴の表現手法について述べる．

動き特徴　時空間特徴は，動き情報も内包しているが，それだけでは局所的な点の動きしか表現できない．よって本研究ではフレーム全体から動き特徴を抽出する．この動き特徴は，全体的な動きを表現出来るので，時空間特徴点の局所点の動き情報とは，異なる識別能力が期待できる．

動き特徴として，本研究ではグリッド点におけるオプティカルフローを利用する．それぞれの検出されたフローは，8 方向，7 段階の大きさからなるヒストグラムに投票されていく．図 6 は実際に抽出された動き特徴を示している．

視覚特徴　視覚特徴には 6 方向，4 周期のガボール特徴を使用する．ただし，フレーム画像を 20×20 の局所領域に分け，それぞれの領域から特徴を抽出し，それぞれのパッチを Bag-of-Keypoints 表現で認識に利用する．

特徴表現　視覚，動き特徴を時空間特徴と統合することで動作認識を行う．しかしフレーム一枚で認識しようとした場合，選ばれたフレームによって結果が大きく変わってしまうことがある．特に動き特徴の場合この傾向が顕著に現れる．そして，最も重要なキーフレームを選択することは非常に難しい問題である．よって本研究では，全てのフレームから特徴を抽出し，その特徴をベクトル量子化することで，認識に利用する．

前述の通り抽出された 257 次元の時空間特徴は BoSTF で表現される．視覚特徴は全てのフレームから抽出される，24 次元のパッチを，動き特徴はフレーム毎に抽出される 56 次元の特徴をベクトル量子化することで特徴を表現する．時空間特徴，視覚特徴におけるコードブックサイズは 5000 に，動き特徴は 3000 に設定した．

動き特徴はフレーム毎に抽出される，57 次元の特徴を用いて bag of frames(BoFr) で表現する．これはフレームから抽出された動き特徴をベクトル量子化して，その特徴の出現頻度で動画を表現する手法である．ただしこの時のコードブックサイズは 3000 に設定した．

カメラモーションが検出された場合，時空間特徴と動き特徴では特徴が破棄されるが，視覚特徴は抽出が行われる．ショット全体がカメラモーションを含んでいるショットでは，時空間特徴や動き特徴が検出されない場合がある．この場合，時空間特徴，動き特徴のベクトルは 0 ベクトルで表される．

5.　実　　験

評価実験として，動作認識，大量 Web 動画ショット分類を行った．以下ではその結果について述べていく．

図 7　Our Youtube データセット

図 8　Wild Youtube データセット

5.1　データセット

動作認識　動作認識には標準データセットである KTH データセット，本研究で構築した Our Youtube データセット，Liu らが構築した Wild Youtube データセットの三種類を使用した．

KTH データセットには「カメラモーションを含まない」「動作をしている人は一人だけ」など様々な制約が存在する．しかし Web 動画のような実世界の映像はそのような制約は存在せず，そのような制約のない動画を分類することはより難しい問題となってくる．

そこで本研究では Youtube から収集したショットから独自の Youtube データセットを構築した (図 7)．このデータセットは “batting”, “running”, “walking”, “shoot”, “jumping”, “eating” の動作から成り，カメラモーション，視点変化，背景ノイズなどを含む，より制約の少ないデータセットとなっている．

次に Web 動画分類における他の手法と比較するために，Liu らが構築した，データセット (図 8) における動作分類を試みた．このデータセットは 11 の動作 (“basketball shooting”, “volleyball spiking”, “trampoline jumping”, “soccer juggling”, “horse riding”, “cycling”, “diving”, “swinging”, “golf swinging”, “tennins swinging”, “walking_with_dog”) を含むデータセットで，Our Youtube データセットと同様に様々なノイズを含む挑戦的なデータセットである．本論文ではこのデータセットを Wild Youtube データセットと呼ぶ．

大量 Web 動画分類　教師信号ありランキングのデータセットとして，独自に収集した batting, running, walking, shoot, jumping, eating の 6 つの動作を利用した．表 1(上) に詳細なデータを記した．合計 984 の動画をショット分割して，得られた 37,179 のショットを本実験では使用した．そのデータの中から教師あり学習では，表 1 右のショット数を学習データとして使用している．

教師なし Web 動画クラスタリングでは Youtube から soccer，dancing のキーワードで収集されたものを利用する．データの詳細については表 1(下) に記載する．

表 1　大量 Web 動画ショット分類のためのデータ

動作	動画数	ショット数	平均時間 [秒]	合計時間 [時間]	学習データ	
					positive	negative
batting	174	8,980	5.9	14.6	31	75
running	170	7,342	6.6	13.4	28	66
walking	174	6,567	7.4	13.4	23	63
shoot	164	7,718	5.3	11.3	14	75
eating	142	3,442	7.7	7.3	22	64
jumping	160	3,130	6.6	5.8	27	40
dancing	185	8,235	6.1	13.9	-	-
soccer	178	10,430	4.5	12.9	-	-
合計	1,247	55,844	6.3	92.6	145	383

表 5　Wild Youtube dataset における分類率の比較

	visual	motion	VMR	MKL V+M	MKL V+M+VMR	Liu [6]
b_shooting	0.46	0.32	0.44	0.53	0.61	0.53
cycling	0.82	0.73	0.59	0.85	0.88	0.73
diving	0.88	0.72	0.82	0.92	0.93	0.81
g_swinging	0.78	0.57	0.72	0.82	0.87	0.86
h_riding	0.85	0.82	0.74	0.89	0.87	0.72
s_juggling	0.14	0.56	0.54	0.52	0.61	0.54
swinging	0.68	0.55	0.63	0.60	0.68	0.57
t_swinging	0.79	0.44	0.53	0.89	0.88	0.80
t_jumping	0.66	0.62	0.67	0.55	0.73	0.79
v_spiking	0.89	0.47	0.77	0.88	0.91	0.73
walking	0.66	0.43	0.53	0.83	0.86	0.75
average	0.691	0.565	0.634	0.735	0.804	0.712

5.2　実験結果

5.2.1　動作認識

KTH データセット　KTH データセットは動作認識の研究において最も広く利用されているデータセットである．本研究でもこのデータセットを使用して，leave one out で 1-vs-all によるマルチクラス分類を行うことで特徴統合の有用性を示し，最新手法との比較を行う．

表 6(左) は KTH データセットにおける MKL による分類の混合行列を示している．この結果から “running” と “jogging” を区別することが難しいことが分かる．表 2 は KTH データにおける動作ごとの分類率と最新手法との比較を示している VMR，visual，motion はそれぞれ，提案時空間特徴，視覚特徴，動き特徴を示している．MKL は，時空間特徴量を用いない場合の V+M と，3 つとも用いる V+M+VMR の 2 通りの結果を示してある．特徴を統合しない場合，最も良かったものは動き特徴の 92.7%で，VMR を用いない統合の結果は 93.5%であったが，特徴を統合することで 94.5%まで精度が向上し，提案特徴量である VMR の有効性が示されている．また KTH データセットにおいては視覚特徴より動き特徴による分類がより正確であった．

特徴統合の重みは図 9(左) に示されている．KTH データセットは視覚的変化に乏しいので，視覚特徴に重みはほとんど割り当てられていない．一方で動き特徴は “running”, “jogging”, “waving” で重みの値が高くなっている．実際 “running” と “jogging” を区別するためには動き特徴が重要であることを示している．

最新手法との KTH データセットの分類性能の比較について，表 3 に 4 つの結果を示す．本提案手法が 94.5%，Dollar らの手法 [3] で 81.2%，Liu らの手法 [6] で 91.8%，Kim ら [14] は 95.33%，Gilbert ら [13] は 96.7%の分類率であった．これらのことから，提案した分類手法はほぼ最新手法に匹敵することが分かる．

動作ごとに見ていくと，全ての研究で “running” と “jogging” を分類することは難しい問題であることが分かる．特に “running” の分類においては分類率は極端に下ってしまっている研究が多い．本研究は “walking” の精度が他より特に高く，他の動作においても，同精度の結果を残すことが出来ている．

Youtube データセット　実際の映像は KTH の映像のように単純ではない．そこで本論文では Youtube から集められたデータを利用して，より実用的な映像におけ

表 2　KTH データセットの分類率

	Point [8]	visual	motion	VMR	MKL V+M	MKL V+M+VMR
walking	0.94	0.47	0.93	1.00	0.94	0.99
jogging	0.76	0.09	0.93	0.86	0.87	0.92
running	0.81	0.41	0.87	0.83	0.92	0.87
boxing	0.91	0.76	0.96	0.90	0.96	0.96
waving	0.90	0.98	0.92	0.98	0.98	0.98
clapping	0.86	0.65	0.95	0.93	0.94	0.96
average	0.863	0.487	0.927	0.917	0.935	0.945

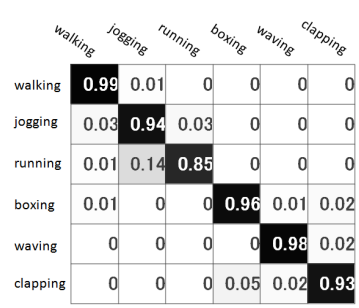
表 3　最新手法の KTH

Dóllar et al. [3]	81.2%
Liu et al. [6]	91.8%
Kim et al. [14]	95.33%
Gilbert et al. [13]	96.7%
ours	94.5%

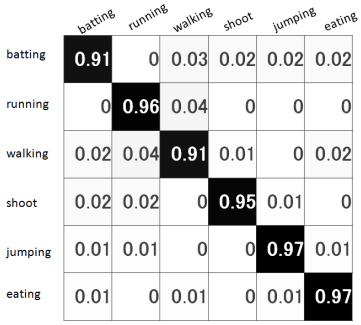
表 4　Our Youtube dataset の分類率

	visual	motion	VMR	MKL V+M	MKL V+M+VMR
batting	0.85	0.55	0.69	0.85	0.91
running	0.84	0.6	0.89	0.86	0.96
walking	0.85	0.41	0.68	0.88	0.91
shoot	0.94	0.77	0.82	0.93	0.95
jumping	0.85	0.69	0.85	0.93	0.97
eating	0.96	0.64	0.79	0.95	0.97
average	0.882	0.610	0.778	0.900	0.945

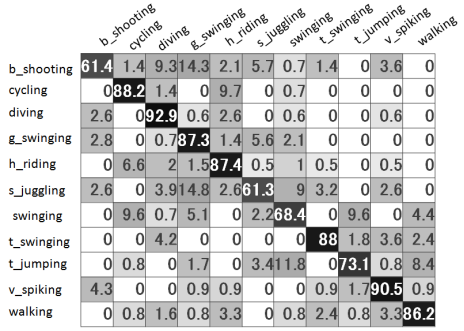
表 6　それぞれのデータセットにおける MKL による混合行列 (左)KTH データセット，(中)Our Youtube データセット，(右)Wild Youtube データセット



KTH データセット (94.0%)



Our Youtube データセット (94.5%)



Wild Youtube データセット (80.4%)

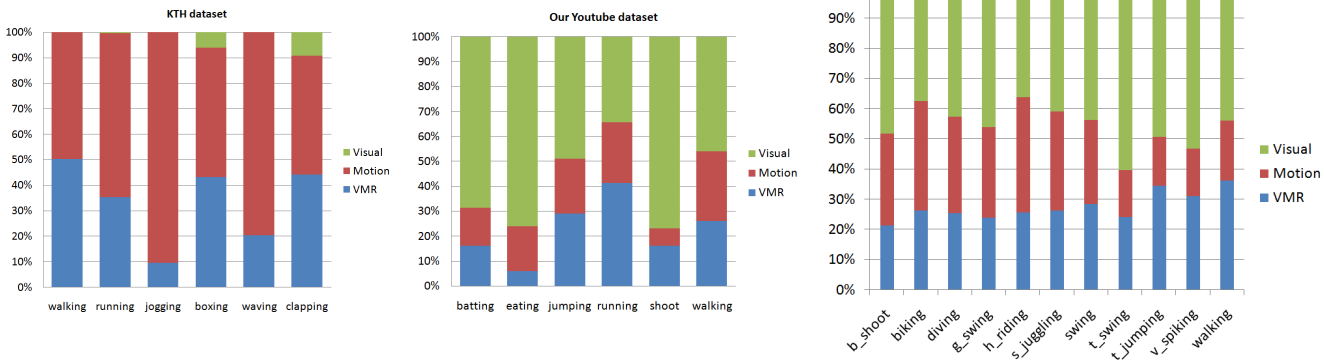


図 9　それぞれのデータセットにおける特徴の重み (左)KTH データセット (中)Our Youtube データセット (右)Wild Youtube データセット

る動作分類を行う．

本研究では 2 種類の Youtube データセットで動作分類を行った．一つ目が独自に構築した Our Youtube dataset．もう一つが Liu らが構築した Wild Youtube dataset である．

a)　Our Youtube dataset

表 6(中) は 5-fold cross validation で 1-vs-all によるマルチクラス分類したときの MKL の混合行列を示している．表 4 は動作ごとの分類結果を示している．Youtube のデータセットにおいては KTH データセットとは逆に視覚特徴が動き特徴の分類率を大きく上回っている．最も良かった視覚特徴で 88.1%で，時空間特徴を用いない統合の結果は 90.0%であったが，あったが，MKL で特徴を統合することで 94.5% まで向上した．統合する際の特徴の重みは図 9(中) に示されているが，“batting”, “shoot”, “walking” などのその動作が発生する場面が限定される動作においては，視覚特徴の重みが比較的高くなる傾向があった．

Wild Youtube dataset　次に Wild Youtube データセットにおいて動作分類を行い，他手法との比較を行う．ただし分類は 5-fold corss validation で行った．

表 6(右) に MKL による分類の混合行列を記載した．また表 5 に各特徴毎の分類率と Liu らの手法 [6] との比較について記載した．全体的に basket_shooting の結果が良くないことが分かる．これはこの動作が多くのカメラモーションを含んでいるためと考えられる．本手法ではカメラモーションがある場合，動き，時空間特徴は抽出されないので動画全てがカメラモーションの場合はそれらの特徴を抽出することが出来ないという問題がある．

このデータセットにおいても特徴を統合することにより，表 5 に示すように，時空間特徴のみでは 63.4%，視覚特徴で 69.2%の分類率であったが，80.4%まで向上した．その際の重みは図 9(右) に示すようになった．全体的に視覚特徴が大きな重みを得ていることがわかる．特に basket_shoot はそのカメラモーションにより，動き特徴による分類が困難になるため，視覚特徴に大きな重みがつけられている．

Liu らの手法 [6] との比較であるが，結果として本手法の分類率が 80.4%であったのに対し，Liu らの手法は 71.1%と Liu らの手法を大きく上回っている．これらのことから MKL による特徴統合は Web 動画分類において有効であることが分かる．

5.2.2　大量 Web 動画ショット分類

教師信号ありの分類　ここではランダムに選択された動画をアノテーションし，学習データとする．その後テストデータを SVM や MKL の 2 クラス分類によってラ

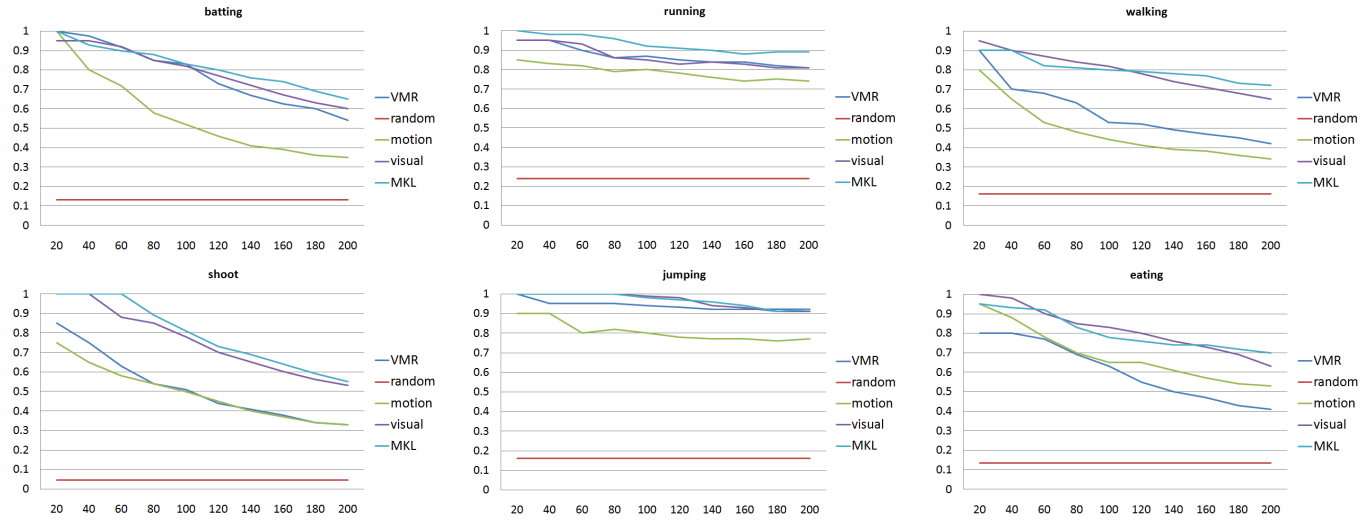


図 10　Web 動画ランキング付けの結果

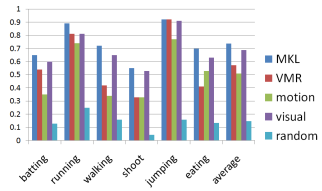


図 11　ランキング上位 200 位までの適合率

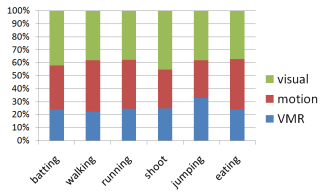


図 12　ランキング特徴毎の重み

図 13　各クラスタに付与されたラベル:上が dancing に付与されたラベル，下が soccer に付与されたラベル

ンキング付けをする．図 10 はその結果を示している．

この図の横軸はランキング N 位を，縦軸は適合率を示している．例えば，VMR(時空間特徴) による batting 分類ではランキングの上位 20 位までの適合率が 1.0，上位 40 位まで見ると適合率が 0.975 であることを示している．

実験として，提案した時空間特徴，視覚特徴，動き特徴，ランダムサンプリング及び MKL による特徴統合の 5 つの手法の比較を行っている．ランダムサンプリングは 200 のショットをランダムに選択し，その適合率を計算したものである．

MKL 以外の単体特徴での実験結果は，視覚特徴が最も良い結果で，上位 20 位までの適合率の平均は 0.98 (図なし)，上位 200 位までの適合率の平均は 0.69 (図 11) であった．これらのことから Web 動画における動作認識では，視覚特徴がより有効であることが分かる．

図 11 は各動作のランキング上位 200 位までの適合率を示している．すべての動作において MKL による特徴統合の値が最も良くなっている．しかし jumping のように単一の特徴でも良い分類が可能な動作に関しては，特徴統合による大きな変化はなかった．また図 12 はそのときの各特徴の重みを示している．

教師信号なしクラスタリング　まず，pLSA や k-means クラスタリングでクラスタリングした後，最も要素の多かった 10 のクラスタにラベルをつける．つけられたラベルに関しては図 13 に示す通り，dancing では「踊っている (dancing)」,「歌っている (sing)」,「映画やドラマの映像 (acting)」,「その他 (others)」分ける．また soccer では「遠いアングルで撮影された試合動画 (play-far)」,「近いアングルで撮影された試合動画 (play-near)」,「インタビューなど人の会話シーン (talking)」,「その他 (others)」のラベルを各クラスタに付与した．

図 14 は各クラスタリング結果の要素数の大きかった 1 位から 10 位までのクラスタのショットに，このラベルを付与したときの各ラベルの割合を示している．例えば k-means で dancing をクラスタリングした結果，最も要素数の大きかったクラスタ 1 は，dancing ショットを 65.0%含み，sing ショットを 6.0%含んでいることを意味している．ただし k-means, pLSA 共にクラスタ数は 200 に設定した．

同一のクラスタのショットは類似していることが望まれるので，本実験では，このラベルの割合が偏っているほど優れた結果であるといえる．

それぞれの動作毎に見ていくと，pLSA ではどのクラスタにおいても 4 つのラベルが多く混在してしまっている．一方で，k-means クラスタリングでは，クラスタ 5, 8, 10 を除いて，dancing や sing のような特定のラベルの割合が高い場合，他のラベルの割合は下がる傾向にあり，より正しい分類が出来ている．ここで k-means クラスタリングで，5, 8, 10 のクラスタの結果が悪かった理由として，k-means はその性質上，ノイズで構成されるクラスタが構築されることがある．実際これらのクラスタでは others が他のクラスタより高くなっている．よってこれらのクラスはノイズが多く集められたクラスタであったため，正しい分類が出来なかったと考えられる．

次に soccer のクラスタリング結果であるが，pLSA では play-far と play-near を正しく分類することが出来ず，全てのクラスタでこの二つのラベルが混在している．また，talking ラベルがどのクラスタにも均一に存在してしまっている．一方で k-means クラスタリングでは pLSA 分類に比べ，各ラベルの偏りが大きくなっており，正しく分類が出来ていることが分かる．

今回使用したデータは大量のノイズを含んでおり，い

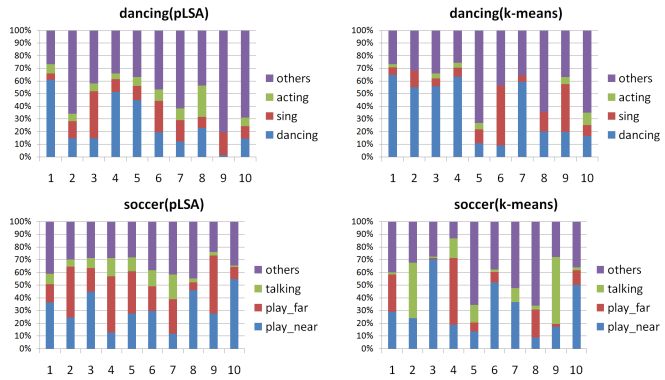


図 14　教師なしクラスタリングの結果

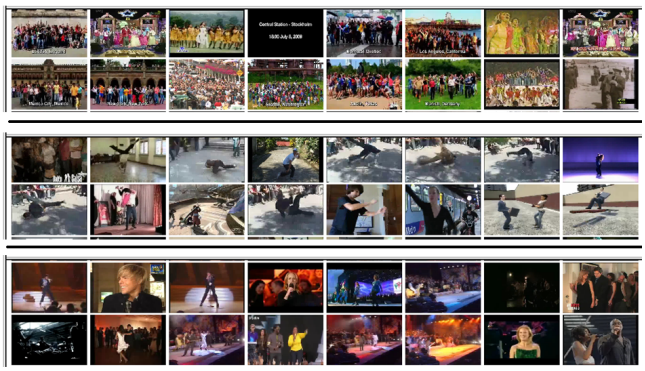


図 15　k-means クラスタリングによる dancing の分類結果の例

ずれの分類手法にしても，その影響を受けてしまっている．特に pLSA ではそれが顕著に現れてしまい，精度が大幅に下ってしまった．

図 15 は k-means クラスタリングによる dancing のクラスタリング結果の一例を示している．

6. おわりに

本研究では Web 動画分類のための時空間特徴抽出手法について提案し，さらに特徴統合を利用した Web 動画分類を行った．提案した時空間特徴は視覚的な局所特徴と，その点の動きを特徴化することで抽出を行った．

実験として Web から大量に収集したショット分類を行った．分類には教師信号ありと，教師信号なしの 2 種類の分類を試みた．教師信号あり分類実験では，ランキングの上位 20 位までについて 91.7%，上位 200 位までについて約 57.2%の適合率であった．また教師信号なしのクラスタリングでは，pLSA と k-means クラスタリングによる分類を行った．k-means クラスタリングは pLSA より正確に分類できた．しかしデータセットのノイズの影響を大きく受けてしまい，正確な分類は非常に困難であった．

時空間特徴の改良点として，現状ではカメラモーションを検出した場合，その特徴を排除することで対応してきた．しかし，それでは有用な特徴を抽出出来ない場合がある．よってカメラモーションの補正を入れることで対象の動きから，特徴を抽出するシステムを構築していくことが必要である．

また提案手法では一つのショットから大量の特徴が抽出されてしまっている．より良い分類を行うために，有用な特徴を選択して抽出するようにしなければならない．

Youtube のような動画には複数の人間が動作を行っている．「歩いている人」の隣で「走っている人」がいるときもある．しかし現在のシステムでは，このように同時に起こる，異なる動作を検出することができない．このような動作を検出するために，人検出とトラッキングに基づく，抽出手法を構築していくことも，精度を向上させるために有効なことである．

また動画像からの特徴だけでなく，ほとんどの Web 動画に付けられているタグなどのメタデータとの融合も教師なし分類の精度向上ためには有効な方法である．

文　献

[1] T. Kobayashi and N. Otsu. A three-way auto-correlation based approach to human identification by gait. In *Proc. of IEEE Workshop on Visual Surveillance*, pp. 185–192, 2006.

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

[3] 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.

[4] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, pp. 995–1004, sep 2008.

[5] R. I. Cinbins, R. Cinbins, and S. Sclaroff. Learning action from the web. In *Proc.of IEEE International Conference on Computer Vision*, pp. 995–1002, 2009.

[6] J. Liu, J. Luo, and M. Shah. Recognizing realistic action from videos. In *Proc.of IEEE Computer Vision and Pattern Recognition*, pp. 1–8, 2009.

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

[8] A. Noguchi and K. Yanai. Extracting spatio-temporal local features considering consecutuveness of motions. In *Proc. of Asian Conference on Computer Vision(ACCV)*, 2009.

[9] B. Herbert, E. Andreas, T. Tinne, and G. Luc. Surf: Speeded up robust features. *Computer Vision and Image Understanding*, pp. 346–359, 2008.

[10] 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.

[11] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, Vol. 5, pp. 27–72, 2004.

[12] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, Vol. 7, pp. 1531–1565, 2006.

[13] A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *Proc.of IEEE International Conference on Computer Vision*, pp. 925–931, 2009.

[14] Kim, T., Wong, S., Cipolla, R.: Tensor canonical correlation analysis for action classification. In: Proc. of IEEE Computer Vision and Pattern Recognition. (2009)