大量の Web 動画からの教師なし特定動作ショット抽出

DoHang Nga† 柳井 啓司†

† 電気通信大学大学院 情報理工学研究科 総合情報学専攻
E-mail: †{dohang,yanai}@mm.cs.uec.ac.jp

あらまし Web 上にはテキストタグ付きの動画が多く存在する．そこで，本研究では “walking”, “eating” などのような動作キーワードを与えるだけで，タグ付きの Web 動画から特定動作の対応ビデオショットを自動的に抽出する手法を提案する．提案手法の大まかな流れは，(1) タグ共起に基づくビデオランキング，(2) 動画分割と特徴抽出，(3) 視覚特徴とタグスコアによるビデオショット選択，となっている．まず，Web API を用いて，与えられるキーワードについての 1000Web 動画のビデオ ID とタグリストを収集する．事前に作成したタグ共起辞書を利用して各ビデオのタグ共起スコアを計算する．次に，タグ共起スコア上位の 200 ビデオをダウンロードして，そのビデオに対しショット分割を行う．それぞれのショットから時空間特徴，全体動き特徴，静的視覚特徴を抽出し，ショットを bag-of-features として表現する．最後に，VisualRank 手法を適用して特定動作の対応ショットを選択する．実験では，６種類の動作について上位 100 ショットは平均で 49.5%の精度が得られた．

キーワード タグ付き Web 動画，動作の教師なし学習，VisualRank 手法，タグ共起辞書

1. 導　　入

近年，Web 上では大量の動画が公開され，誰もが容易に視聴することができるようになっている．特に，YouTube や DailyMotion などの一般ユーザが動画を自由にアップロードできる動画共有サイトには毎秒，何百もの動画がアップロードされている．動画をアップロードした人は通常，他の人がキーワードサーチによってその動画を探せるように動画に “タグ” と呼ばれる，動画内容と関連するテキストキーワードを付ける．ただし，ビデオ全体にタグが付けられることが一般である．よって，どのタグはビデオのどの部分に対応するのが未知である．例えば，“eating” というタグが付いている動画は，レストランに入店するシーン，食べ物の注文シーン，会話シーンなどを含んでいるかもしれない．“eating” シーンにしか興味がない人は手動でスキップして目的のシーンを探さなければならない．これは面倒で時間がかかることだと考えられる．また，特定の動作シーンの認識はいくつかの研究が既に提案されているが，ほとんどの場合，教師データが必要で事前に用意された動作以外には対応ができないという問題点がある．

そこで，本研究は教師なしで，キーワードを与えることだけによってタグ付きの Web 動画からキーワードに対応した動画ショットを検出する新しい手法を提案する．キーワードとしては “eating” や “running” などの人間動作に関する単語を用いることとする．

もし特定動作の対応ビデオショットが自動的に取得可能となれば，Web 動画のような制限なしの動画を用いて人間動作認識のための学習データを生成することが容易になる．物体認識のため静止画像の学習データ収集とは異なり，特定動作のビデオショットの学習データを収集することは，対象が動画であるために一般には容易ではない．そのため実際には多くても 14 種類程度の限られた動作についてしか動作認識の実験が行われていない [1]. 教師信号なしで Web 動画から特定動作に対応したショットの自動抽出が成功すれば，多種類の動作に対応した認識システムが実現可能になることが期待できる．我々の最終目的は，動作認識の研究のための大規模ビデオショットデータベースの自動作成である．

提案手法では，まず，WebAPI によって得られたビデオの ID とタグリストを用いてタグ共起に基づいて 1000 ビデオをランキングする．次にタグ共起スコアの上位 200 のビデオをダウンロードし，そのビデオをショットに分割する．最後に，グラフに基づくランキングメソッド VisualRank [2] を適用してビデオショットの視覚特徴とタグ共起スコアに基づいて与えられたキーワードに対応したショットを上位にランキングする．

本研究の主な貢献としては，以下の３点が挙げられる．

(1) 特定動作に対応したビデオショットデータベースの完全教師なしの構築.
(2) 大量のタグ付きのビデオからのタグに基づく動画選択と時空間特徴を用いた視覚特徴に基づくショット選択の２つのステップの対応ショット選択手法の提案.
(3) 100 種類の動作についての 100,000 の YouTube ビデオのメタデータ分析と 20,000 の YouTube ビデオの時空間特徴分析による大規模な動作ショットデータベース構築実験.

本論文では，セクション 2 で関連研究を述べる．提案手法の概念はセクション 3 にて述べる．セクション 4 ではタグに基づくランキング，特徴表現と視覚特徴に基づくランキングの手順の詳細を説明する．セクション５は実験結果について説明する．最後に本論文の結論をセクション 6 にて述べる．

2. 関 連 研 究

関連研究として人間動作認識，Web 画像マイニング，タグランキングメソッドの研究について説明する．

動作認識：ここ数年，時空間特徴（Spatio-Temporal(ST) 特徴）とその bag-of-features (BoF) 表現が人間の動作認識のための特徴表現として注目を集めている．それは，この特徴を利用することにより動作認識問題は，特徴抽出のステップ以外は物体認識問題と同じように考えることができるためである．2〜3 年前は，ほとんどの動作認識研究は，KTH データセット [3] や Weizmann データセット [4] などのような制限のあるビデオデータを実験データとしていた．最近は KTH と Weizmann データセットの認識率はほぼ完璧になった（それぞれ 95.5% と 100%）ので [5]，いくつかの研究は"in the wild" YouTube データセット [6] や Hollywood アクションデータセット [7] などの制限のない動画セットを使うようになってきた．2010 年になって YouTube ビデオセット [8], [9] と Kodak ビデオセット [10] に対し動画分類を行う研究が多く発表されるようになった．ただし，その研究のすべては学習データが必要であり，動画を事前定義されたカテゴリーに分類することが目的である．それに対して，提案手法は学習データが必要なく，大量の Web ビデオから与えられたキーワードに対応したビデオショットを検出することが目的である．

本研究の目標は，人間動作に対応するショットデータベースの完全自動構築であるため，教師なしの学習法を利用する．教師信号なしの手法を使う研究としては Niebles らの研究 [11] がある．彼らは PLSA モデルを用いて KTH データセットと彼らの ice-skating データセットに対し動作分類を行った．彼らの提案手法は教師なしであるがカテゴリー数を事前に与える必要がある．Niebles らはさらに，制限なしの動画から動作シーケンスを検出する教師なしの手法を提案した [12]．Cinbis らは Web 画像検索エンジンから収集される画像を利用して動作モデルを自動学習するメソッドを提案し, [12] のビデオデータセットに対し動作認識を行った [13]．この研究は本研究と最も関連があるが，彼らは学習ソースとして Web 画像，特徴として静的特徴だけ使う．それに対して，本研究は Web 動画と時空間特徴を使う．Niebles らの研究と Cinbis らの研究は人間の存在する領域を検出するために HOG(Histogram of Oriented Gradient [14]) に基づく人間検出器を適用するので，動画から人間の全身が検出可能である必要があり，動作の種類が限られる．一方で，我々の提案手法はどの様な人間動作にも適用可能であり，さらに人間動作以外の "airplane-flying" や "tornado" などの非人間動作にも適用できる．もう 1 つ関連研究として Ballan らの研究がある [15]．彼らは Flickr からの Web 画像を学習サンプルとしてビデオショットにタグを付けるメソッドを提案した．

Web 画像マイニング：Web 画像を利用した画像データベースの自動建設の研究は以前よりよく行われている [16]〜[21]．ほとんどの研究はキーワードについての Web 画像検索エンジンによって収集される画像から物体認識手法を適用してキーワードの対応画像を選択する．時空間特徴の BoF 表現を使うことによって動画認識ドメインにこのアイデアを容易に適用できることが考えられ，よって，本研究は Web 画像の自動収集のビデオショットバージョンと見なすことができる．

タグランキング：本論文ではタグに基づく関連スコアを計算するためタグ分析を行う．タグが付けられることは Web 上の消費者生成メディア (Consumer-Generated Media, CGM) データの共有の特徴である．一般には Web 上の画像とビデオに複数のタグが付けられる．Yang らはタグの出現統計による画像に付与された各タグのタグ共起スコアの計算法を提案した [22]．この計算法は "Web2.0 Dictionary" と呼ばれる．我々はこの計算法を適用してタグ共起スコアの高い Web 動画を選択する．この手法はタグ分析だけを必要とし，視覚特徴の分析の必要がない．他に視覚特徴を利用しない関連研究として Dong らの研究がある [23]．彼らは関連スコア確率推定と random-walk に基づく改善化の結合によってタグの関連性を評価する．彼らの提案手法も我々の提案手法と同様に 2 段階手法であるが，タグ情報しか使わない．それに対して，本研究はタグだけではなく 2 段階目に視覚特徴も利用する．

3. 提案手法の概要

本論文では，動作キーワードを入力するだけで，タグ付きの Web ビデオからキーワードに対応する特定動作の対応ビデオショットを自動的に抽出する新手法を提案する．提案手法の大まかな流れは，(1) タグ共起による動画選択，(2) 動画分割と特徴抽出，(3) 視覚特徴とタグスコアによるビデオショット選択，となっている (図 1)．

最初に，動画を Web から実際にダウンロードする前に，与えられたキーワードに対する各動画のタグ共起スコアを計算する．これによって，指定キーワードがタグとして付けられる Web ビデオは大量に存在するが，その中でも関連性が高いビデオだけを選択してダウンロードすることが可能となる．ここでは，Web 動画共有サイトによって提供される WebAPI を利用して，指定キーワードをタグに含むビデオのビデオ ID とタグリストを取得して，動画自体のメタデータのみを用いてタグ共起スコアに基づくビデオランキングを行う．

次に，図 1 に示すようにビデオショットランキングの前に視覚特徴抽出が行われる．本研究は視覚特徴として Noguchi らの提案時空間特徴 [24]，全体的動き特徴，ガボール視覚特徴と，これらの統合を利用する．

3 番目のステップにおいて，グラフに基づくランキングメソッド VisualRank [2] を適用してビデオショットをランキングする．類似度行列として視覚特徴による類似度行列，補正ベクトルとしてタグ共起スコアによるバイアスベクトルを設定する．元々VisualRank は画像集合において代表的な画像を自動的に上位にランキングする手法で，これをビデオショットに適用することによって，最終的に与えられたキーワードに対応する動画ショットが上位にランキングされることが期待される．なお，ここで注意すべき点は，第 1 ステップではビデオ全体のビデ

オランキングを行ったのに対して，第 3 ステップでは分割したショットのビデオショットランキングを行うことである．

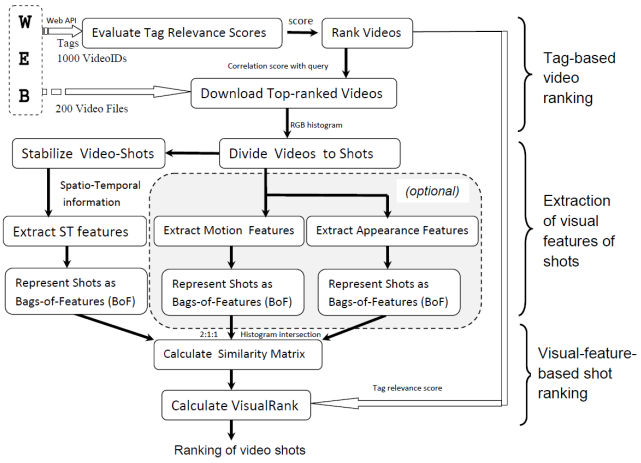

図 1　提案手法の概要

4. 提案手法の詳細

ここでは，タグに基づくビデオランキング，ビデオショットからの特徴の抽出法，視覚特徴によるビデオショットランキングの詳細を説明する．また，最後に動き補正についても説明する．

4.1 タグに基づくビデオランキング

動画共有サイトが提供する WebAPI を利用することによって，与えられたキーワードに対応する Web ビデオが簡単に得られる．YouTube の場合，指定キーワードをタグとして含んでいるビデオデータベース中のビデオを検索可能な YouTube WebAPI を提供している．ただし，タグはビデオをアップロードした人の主観によって付けられるので，タグとビデオとの関連が弱い場合がある．また，タグがビデオの一部のみと関連する場合もある．このステップの目的としては，タグ分析のみで，指定クエリキーワードとの関連が深いビデオだけを選択することである．

最初に，YouTube の WebAPI にキーワードを送ることによってビデオ ID とタグリストのセットが得られる．タグの共起出現を利用しビデオのキーワードとの関連性を評価する．評価法として Yang らが提案した “Web 2.0 Dictionary” を適用する [22]．Web2.0 辞書とは Web からの大量のビデオタグに基づくタグ共起の統計を表すものである．

指定キーワード t をタグに含んでいるビデオの数を $N(t)$，t 以外のワードを t_i，$F(t,t_i)$ を t と t_i の両方のタグを含むビデオの数を示す，と仮定したとき，親語 t とその子語 t_i の関連度は**式** (1) によって計算される．

$$w(t,t_i) = \frac{F(t,t_i)}{N(t)} \tag{1}$$

さらに，$\mathcal{T}$ をビデオ V が含む t 以外のタグの集合とすると，次のように V を $\mathcal{T}$，$P(t_i|t)$ を $w(t,t_i)$ と置替えればビデオ V のワード t との関連性が推定できる．

$$\begin{aligned} P(V|t) &= P(\mathcal{T}|t) \\ &= \prod_{t_i \in \mathcal{T}} P(t_i|t) \\ &= \prod_{t_i \in \mathcal{T}} w(t,t_i) \end{aligned} \tag{2}$$

以上が [22] による画像・ビデオの与えられたキーワードとの関連値の計算法である．これはクエリタグ以外のタグがクエリタグのサポーターであり，クエリワードとの関連が強いサポータータグが多く付けられるほど画像・ビデオがクエリワードとより深く関連するというアイデアに基づいている．

ただし，**式** (2) によりクエリタグとビデオのほかのタグの関連値を掛けることでタグが増えるとともに値が小さくなる．これを避けるため計算に用いる共起出現ワードの最大数を m と制限し，平均ログ尤度を適用して関連値を定義するとする．**式** (2) は次のように改善する．

$$\begin{aligned} S(V|t) &= \frac{1}{n} \sum_{t_i \in \mathcal{T}'} \log_2 w(t,t_i) \\ &= \frac{1}{n} \left(\sum_{t_i \in \mathcal{T}'} \log_2 F(t,t_i) - n \log_2 N(t) \right) \\ &= \frac{1}{n} \sum_{t_i \in \mathcal{T}'} \log_2 F(t,t_i) - log_2 N(t) \end{aligned} \tag{3}$$

$$Sc_t(V) = \frac{1}{n} \sum_{t_i \in \mathcal{T}'} \log_2 F(t,t_i) \tag{4}$$

ここで，$\mathcal{T}'$ は $w(t,t_i)$ に関して上位 m 語の t_i を含み，n $(n \leqq m)$ は $|\mathcal{T}'|$ を示す．特定動作キーワードについてのビデオセットのどのビデオに対しても**式** (3) の第 2 項は不変なので省略して**式** (4) のように関連値 $Sc_t(V)$ を定義する．実験では m は 10 と設定し，キーワードの検索結果の 1000 ビデオから関連値 $Sc_t(V)$ の上位の 200 ビデオを選択する．このタグに基づく動画選択のステップは，関連性が高いと考えられるビデオだけを計算コストが高いステップ 2（特徴抽出）で処理することにするために必要である．ここで注意すべき 2 つのことがある．1 つ目は “drink coffee” のような 2 以上の単語からなる統合ワードの場合，結合ワードも 1 つのワードとし，$N(t)$ は t の各単語の全部が付けられるビデオの数とし，$w(t,t_i)$ は t と t_i のすべての単語を持つビデオの数として考えることである．2 つ目は共起タグのない，タグが検索キーワード一つのみのビデオは共起スコアの計算が不可能であるため利用しないとすることである．

タグの共起確率 $w(t,t_i)$ はビデオデータペース全体について事前に求めておく必要があるので，実験ではシードワード (seed word) として “ride bicycle” と “launch shuttle” のような動詞と名詞の 150 セットを準備した．各シードワードについて 1000 ビデオのタグリストを収集する．結果として集められたタグの中に 12,471 タグが 5 回以上出現した．この 12,471 タグワードのそれぞれに対し，さらに 1000 ビデオのタグを収集し，**式** (1) を用いて共起確率 $w(t,t_i)$ を事前に求める．この $w(t,t_i)$ の値こそ

が “Web 2.0 Dictionary” である．

4.2 視覚特徴に基づくショットランキング

次に，タグ共起に基づくビデオランキングメソッドによるランクの上位 200 本の Web 動画を実際にダウンロードし，その後，隣接フレームの間のカラーヒストグラムの距離による閾値の設定による簡易的な手法によってビデオをショットに分割する．

ショットから特徴を抽出する前に，特徴抽出を行うショットを選択する．200 ビデオからのショットの総数は 10000 を超え，総時間は 15 時を越える場合があるため，ショット数が多い場合は計算量をある一定以下に制限する必要がある．そこで実験では各ビデオの利用ショットの上限数を制限する．また，各動作に対し最大 2000 ショットを利用することとする．ビデオはランクが高いほどより多いショットが選択されること，可能な限り多くのビデオから様々なショットを選択すること，の両方のバランスを保つために次のようにヒューリスティックを用いてショットを選択する．

$$N_{upper}(V_i) = c \times Sc(V_i) + f(N(V_i)) \tag{5}$$

$$\text{where} \quad f(x) = \begin{cases} 20 & (20 \leqq x) \\ 20 + (x-20)/4 & (20 < x < 100) \\ 40 & (x \geqq 100) \end{cases}$$

ここで，$N_{upper}(V_i)$ と $N(V_i)$ はそれぞれビデオ i-th の利用ショット上限数とショット総数を示す．$Sc(V_i)$ はビデオ i-th のタグ共起スコアを指す．c は “Web 2.0 Dictionary” のサイズによる定数である．実験では，c を 10 と設定した．200 ビデオからタグ共起スコアの順にビデオを選択し，選択されるビデオに対し利用ショット上限数を決めてショットを選択する．

ショット選択の後は時空間特徴，全体動き特徴と静的視覚特徴の抽出である．特徴抽出の詳細は次のサブセクションにて説明する．

視覚特徴に基づくショットランキングメソッドとして，よく知られている Web ページランキングメソッド PageRank を画像に応用した VisualRank 法 [2] を利用する．VisualRank 法では，画像の類似度行列を用いて反復計算によって各画像のランク値が求められる．本研究では，ショットの類似度を計算するには，ヒストグラムインタセクションを用いる．2 つのヒストグラム化されたショットの類似度は次によって求められる．

$$s(H_i, H_j) = \sum_{l=1}^{|H|} min(h_{i,l}, h_{i,l}) \tag{6}$$

ここで，H_i，$h_{i,l}$，$|H|$ はそれぞれショット i-th の BoF ベクトル，その l-th 要素，BoF ベクトルの次元数を示す．

また統合特徴を用いる場合，3 種類の特徴を線形統合によって統合し，類似度行列は次のように計算する．

$$S_{combined} = w_{ST} \times S_{ST} + w_{mot} \times S_{mot} + w_{app} \times S_{app} \tag{7}$$

ここで，S_{ST}，S_{mot} と S_{app} はそれぞれ時空間特徴，動き特徴，視覚特徴による類似度行列を指すものであり，w_{ST}，w_{mot}, w_{app} はそれぞれに対応する重みである．

教師ありの認識システムの場合，MKL や AdaBoost などの機械学習手法を用いて最適な重みを推定することができるが，本研究は，完全に教師なしの動作ショットデータベースの構築を目的とするので，学習データを利用しない．ここで，最適の重みの推定の代わりに，次のようにヒューリスティックに 3 つの重みを設定する．

$$w_{ST} = \frac{1}{2},\ w_{mot} = \frac{1}{4},\ w_{app} = \frac{1}{4} \tag{8}$$

時空間特徴の重みは最も高く設定したのは [24] によって時空間特徴は他の 2 つの特徴より優れた結果が得られることが示されているからである．

VisualRank 計算によって，類似画像が多い，より代表的な画像が上位にランク付けられる．**式** (9) は VisualRank 計算の公式を示している．

$$\boldsymbol{r} = \alpha S \boldsymbol{r} + (1-\alpha)\boldsymbol{p} \quad (0 \leqq \alpha \leqq 1) \tag{9}$$

ここで S は列が正規化された類似度行列，$\boldsymbol{p}$ は補正ベクトル，$\boldsymbol{r}$ はランキングベクトルを指すものである．α は $\boldsymbol{p}$ の影響を制御する補正パラメータである．一般には α は 0.8 以上の値が設定される．補正ベクトル $\boldsymbol{p}$ を不均一なベクトルとして与える場合，補正値が高いほど対応イメージのランクスコアは高くなる傾向がある．本研究では，動作のタグ共起スコアの高い動画ショットに大きな補正値を与えて強調するとする．次のように補正ベクトルは 2 つの種類を定義する．

$$p_i^{(1)} = \begin{cases} 1/k & (i \leqq k) \\ 0 & (i > k) \end{cases} \tag{10}$$

$$p_i^{(2)} = \begin{cases} Sc(V_i)/C & (i \leqq k) \\ 0 & (i > k) \end{cases} \tag{11}$$

$$\text{where } C = \sum_{j=1}^{k} Sc(V_j)$$

ここで，$Sc(j)$ はショット j のビデオのタグ共起スコアを示す．**式** (10) では，タグ共起スコアのトップ k ショットに同一バイアス値を与える．一方で，**式** (11) では，トップ k ショットの各ショットは補正値が対応動画のタグ共起スコアに比例する．実際には**式** (10) は画像検索エンジンのトップ k ショットだけにバイアスをかける場合の Jing らが提案した計算式と類似している [2].

4.3 視 覚 特 徴

本研究では，動画からの視覚特徴量として主に時空間特徴量を利用する．また，複数種類の特徴を統合して用いる場合に，時空間特徴量に加えて，全体動き特徴と静的視覚特徴も利用する．さらに抽出された特徴を用いて 1 ショットを 1 つの BoF ベクトルとしてヒストグラム表現する．

4.3.1 時空間特徴抽出

本研究では，時空間特徴を抽出するための手法として Noguchi らの提案した手法を利用する [24]．まず最初の

フレームから SURF [25] に基づく動的な視覚特徴を抽出する．次に，抽出された SURF 点（候補点）に対して，Lucas-Kanade アルゴリズム [26] に基づいて動きフローを計算する．動きのあった点のみを時空間特徴として取り扱う．その特徴点は Delaunay 三角分割法を適応して三角形の頂点を構成する三点のペアで表現する．ここで，特徴点の微小区間の動きを特徴化することにより局所的な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付けて結合することで時空間特徴抽出ができる．図 2 は時空間特徴抽出のプロセスの例を示している．Noguchi らのメソッドは速い検出器に SURF と

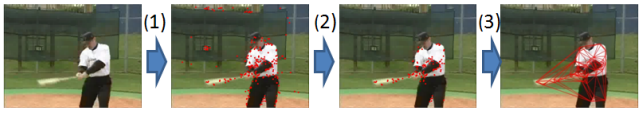

図 2　時空間特徴抽出手順 (1) SURF 点検出,(2) 時空間特徴点検出, (3)Delaunay 三角作成 ([24] から引用)

Lucas-Kanade を適用するので cuboid に基づくメソッドより比較的に実行時間が速い．5 フレームごとに特徴が抽出され，特徴ベクトルの次元数は 256 である．

4.3.2　全体的動き特徴抽出

時空間特徴は局所的な動きであるが，ここではそれとは別に画像の全体的な動きを全体的動き特徴として抽出する．全体的動き特徴の抽出には Lucas-Kanade アルゴリズム [26] に基づくグリッド点におけるオプティカルフローを利用する．オプティカルフローとは連続画像の中の物体の動きをベクトルで表現するものである．Lucas-Kanade 法を用いることにより，高速で精密なオプティカルフロー推定ができる．またこの方法は同一物体の局所領域内ではオプティカルフローはほぼ一様になると仮定する空間的局所最適化法の一つである．8 ピクセルごとのグリッドポイントから動き特徴を抽出する．各グリッドから抽出される動き特徴は 7 方向，8 モーションマグニチュードのヒストグラムに投票する．よって各特徴は 56 次元数のベクトルとして表現する．

4.3.3　静的視覚特徴抽出

一方，静的な視覚特徴として時空間特徴で用いられている SURF 特徴と異なる特徴量として，ガボール特徴を利用する．ガボール特徴は，画像から局所的な濃淡情報の周期と方向を表した特徴量である．カーネルの形を固定し，それを周期を変えて伸び縮みさせたり，回転させて方向を変えたりして，様々な周期や方向のカーネルフィルタを作成する．本研究では，6 方向，4 周期の 24 カーネルフィルタを利用する．ガボール特徴は局所的な情報を見るので，画像の照明変動の影響を受けにくいという利点がある．ガボールフィルタの適用の前にフレームを 20×20 ブロックに分割する．各ブロックに対し 24 フィルタを用いてガボール特徴を抽出する．結局は各フレームから 400 の 24 次元のガボール特徴ベクトルが得られる．

4.4　ショットの特徴表現

本研究では，抽出した特徴を用いて，各ショットに対し特徴によるベクトル量子化を行う．全体的動き特徴と静的視覚特徴の場合は 4 フレーム，時空間特徴の場合は 5 フレームごとに特徴を抽出し，特徴によるベクトル量子化によって各ショットを BoF として表現する．元の BoF は一枚のフレームの局所特徴の分布を表すが，本研究での BoF は複数のフレームを含むショットの特徴の分布を表す．この BoF は bag-of-frames(BoFr) と呼ぶ．

実験では，時空間特徴，動き特徴，視覚特徴のコードブックのサイズは, [24] と同様にそれぞれ 5000，3000，5000 と設定する．

4.5　動き補正

Web 上の動画は全体的にカメラモーションを含んでいて，背景の動きも強いことが多いので，時空間特徴抽出する前に，その動きを減らすことを行い，その有効性を検討する．動き補正の目的としては連続したフレームに対し，全体のフローから背景のフローを抜き出し，前景のフロー (主に人間の動き) を残すことである．動き補正は Cinbis らの研究 [27] と同様に RANSAC によるホモグラフィー推定を用いて背景フローを推定するすることによって行う．

Noguchi らの時空間特徴抽出手法では，特徴点の決定には背景フローも含む全体オプティカルフローを利用する．全体フローが強い候補点は動きがあった点 (特徴点) とみなす．そこで，本研究で動き補正を行う場合は，全体フローの代わりに前景フローだけを使用する．前景フローは全体オプティカルフローから背景フローを引いたものとする．ここで，全体フローは Lucas-Kanade アルゴリズムによるオプティカルフローとして求める．背景フローを計算するために，まずは，キーフレームとその連続したフレームの間の対応関係を表すホモグラフィーを RANSAC によって求める．対応点としてはその 2 つのフレームの SURF 点と利用する．推定されるホモグラフィーを用いて，背景フローを計算することができる．候補特徴点における前景フローは**式** (12) によって計算する．

$$m_f(x,y) = \|m_o(x,y) - m_g(x,y)\|_2^2 \qquad (12)$$

ここで，$m_f(x,y)$，$m_o(x,y)$，$m_g(x,y)$ はそれぞれ (x,y) における前景フロー，全体フロー，背景フローを指すものである．

Cinbis らの研究では Web 動画の動作認識において動き補正法を行ったが [27]，彼らは動き補正のない結果との比較を行っていないので動作認識タスクに対するこの動き補正法の有効性は検証されていない．実験ではその有効性を検討するため，動き補正ありの実験と動き補正なしの実験の両方を行う．

5.　実験結果

提案手法の有効性を検証するため Noguchi ら [24] のデータセットに対し様々な設定で実験を行った．このデータセットは 6 種類の人間動作キーワード：“batting”，“eating ramen”，“jumping trampoline”，“running marathon”，“shooting fooball” と “walking street” についての YouTube

図 3　Examples of six kinds of the actions.

ビデオからなる．動作のイメージの例は図 3 で表す．さらに，6 種類での様々な条件での比較実験に加えて，非人間動作のビデオショットの自動抽出と 100 種類の人間動作のデータベースの構築実験も行った．

ランキング結果を評価するには一般的には平均適合率がよく使われているが，ここではランク上位 100 位での適合率を用いる．理由としては，本研究の目的は動作認識の研究のための動作ビデオショットデータベースの自動作成であり，一般的に用いられる KTH データセット [3] や “in-the-wild” YouTube データセット [6] のような動作データセットは 1 種類の動作あたり約 100 ビデオショットを含んでいるため (注1)，ランク上位 100 位までに多くの正しいショットが含まれていることが重要だと考えるからである．

5.1　様々な設定による実験集

表 1 で表すように，条件を変化させて 10 種類の実験を行った．各実験の設定は以下の通りになる．

RND　Youtube のキーワード検索の結果の上位 200 ビデオをダウンロードしショット分割したのち，ランダムに 100 ショットを選択．

TAG　1000 ビデオからタグに基づくランキングによって 200 ビデオをダウンロード，ショット分割後，ランダムに 100 ショットを選択．

Exp.1　RND と同様に Youtube のキーワード検索上位 200 ビデオをダウンロード，ショット分割後，2000 ショットをランダムに選択，時空間特徴だけを利用し，VisualRank によるショットランキングを実行．ショットランキングで均一補正ベクトルを使用．

Exp.2　Exp.1 にタグに基づくビデオショット選択を追加．

Exp.3　(1) **式** (10)，(2) **式** (11) によって得られるバイアス補正ベクトルを利用．

Exp.4　Exp.3(1) にカメラモーション補正を追加．

Exp.5　Exp.3(1) で使われる時空間特徴の代わりに全体的動き特徴を使用．

(注1)：KTH データセットと “in-the-wild” データセットはそれぞれは 6 動作について 599 ショットと 11 動作について 1168 ショットを含む．

表 1　実験設定と実験結果

Exp. no	タグに基づくランキング	補正ベクトル	動き補正	視覚特徴	**上位 100 の平均適合率**
RND	randomly-selected 100 shots				14.2 %
TAG	√	-	-	-	23.5 %
1	-	-	-	時空間	33.7 %
2	√	-	-	時空間	41.0 %
3(1)	√	√(1)	-	時空間	**47.3 %**
3(2)	√	√(2)	-	時空間	44.8 %
4	√	√	√	時空間	39.8 %
5	√	√	-	動き	31.8 %
6	√	√	-	視覚	39.7 %
7	√	√	-	**fusion**	**49.5** %
8	√	√	√	**fusion**	41.2 %

表 2　6 動作の上位 100 の適合率 (%).

Exp. no	batting	eating	jumping	running	shoot	walking	AVG.
RND	12	13	17	23	4	16	14.2 %
TAG	32	14	45	23	20	7	23.5 %
1	56	37	69	20	5	23	33.7 %
2	77	30	75	23	17	24	41.0 %
3(1)	66	42	82	35	33	29	47.3 %
3(2)	73	28	90	38	27	30	44.8 %
4	38	31	78	45	12	**36**	39.8 %
5	69	10	67	12	32	20	31.8 %
6	35	**57**	64	**46**	7	29	39.7 %
7	69	39	**87**	30	**36**	**36**	**49.5 %**
8	61	37	74	38	16	21	41.2 %
9 (0.75)	68	32	80	29	26	29	44.0 %
9 (0.80)	**82**	34	83	29	30	33	48.5 %
9 (0.85), 7	69	39	87	30	36	36	**49.5 %**
9 (0.90)	77	31	79	26	28	35	46.0 %
9‘(0.95)	74	31	81	28	25	32	45.2 %
MKL [24]	83	78	98	87	82	52	80.0 %

Exp.6　Exp.3(1) で使われる時空間特徴の代わりに静的視覚特徴を使用．

Exp.7　Exp.3(1) で使われる時空間特徴の代わりに統合特徴を使用．

Exp.8　Exp.7 にカメラモーション補正を追加．

以上で述べた 10 実験の結果は表 1 の最後の列で表す．結果からわかるようにランダムに 100 ショットを選択する RND は一番低い適合率になる．TAG はタグに基づくビデオランキングの導入によって RND を 9.3%改良した．Exp.1 は時空間特徴に基づショットランキングの適用によって TAG より優れた．また，ビデオランキングとショットランキングの両方を行うことによって Exp.2 は Exp.1 より 7.3%向上できた．

Exp.2 では VisualRank 計算で均一補正ベクトルを利用した．Exp.3 では 2 つの方法でバイアスをかけた．Exp.3(1) はタグ共起スコアのトップ 1000 ショットに均一の重みを与え，Exp.3(2) はタグ共起スコアと比例する重みを与える．実験 Exp.3(1) と Exp.3(2) の結果からバイアス補正ベクトルの設定法としてメソッド (2) よりメソッド (1) は良い結果が得られることがわかる．よって，Exp.3 以降はバイアス補正ベクトルの設定にはメソッド (1) を利用する．

実験 Exp.4 では，動き補正を行った．ただし，Exp.3 の結果との比較によって動き補正の行わないほうは結果が良いとわかった．使用した動き補正法は本研究の不規則

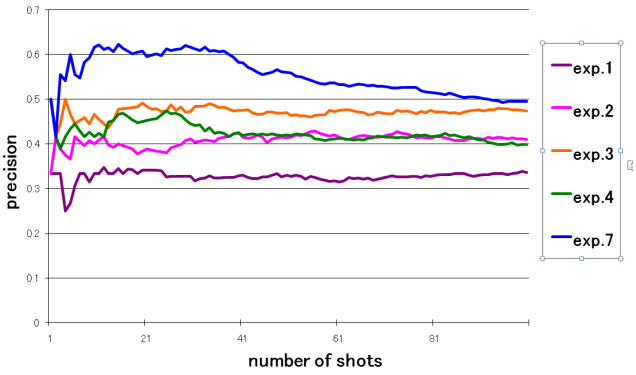

図 4　6 動作の n-th ショットの平均適合率

なカメラモーションを含む Web データセットに対し有効ではないと考えられる．

実験 Exp.5 と Exp.6 では，時空間特徴の代わりに全体的動き特徴と視覚特徴を使用し，それぞれ 31.9%と 39.7%の平均適合率が得られた．よって，47.3%が出た時空間特徴はその 2 つの特徴より優れた結果を示すことが言える．

実験 Exp.7 では，全体を通して一番良い結果の 49.5%が得られた．実験 Exp.7 はカメラモーション補正を除いて提案手法のすべてのステップを行った．実験 Exp.3(1) と比べて，統合特徴の利用により結果が 47.3%から 49.5%に少し改良できることがわかる．実験 8 では実験 7 に動き補正を追加したが Exp.4 と同じように結果が劣化した．

6 種類の各動作の詳細結果は表 2 にて表す．“walking” と “running” に対して Exp.4 は Exp.3(1) より良い結果が得られた．その理由としては，この 2 つの動作に対応する動画には水平方向のカメラモーションが多く含まれていて，こうした一定方向のカメラモーションが含まれる場合は利用した動き補正法は有効であると考えられる．

ページランクの補正ベクトルの最適な補正パラメータ (**式** (9) の α) を確認するために実験 Exp.9 を行った．Exp.9 では補正パラメータを 0.75 から 0.95 まで 0.05 ごとに変化させる．Exp.9 では α の設定以外は最も良い結果が得られた実験 Exp.7 と同様に設定する．Exp.9 の結果は表 2 の下部に示す．括弧には α の値を示す．Exp.9(0.85) は Exp.7 と等価である．Exp.9(0.85) の結果からわかるように，α を 0.85 と設定した実験は結果が一番良い．また，Noguchi らの教師あり学習による分類結果を，参考のため表 2 の最後の行に表す [24]．“batting” と” jumping” は Noguchi らの手法に近い精度が得られていることが分かる．なお，彼らは教師ありの手法で，一方で我々は完全に教師なしの手法である．

図 4 には 6 種類動作の上位 n 動画の平均の適合率を表している．このグラフからわかるように Exp.7 は一番高い適合率になり，特に，上位 10 位から 40 位までは適合率が 60%を超えている．

5.2　追 加 実 験

Noguchi らのデータセットについての実験だけではなく，“airplane flying” のような 6 種類の非人間動作と 100 種類の人間動作にも実験を行った．6 種類の実験では実験 Exp.7 の結果が最も良かったが，統合特徴は特徴抽出にコストが掛るため，単一の特徴でも Exp.7 と遜色のない結果を示した実験 Exp.3(1) と同じ設定をここでの実験ではすべて利用した．

表 3　人工物動作と自然現象の結果 (%)

aircraft +landing	tornado	blooming +flower	airplane +flying	earthquake	shuttle +launching	AVG.
30	22	19	14	2	2	14.8

表 4　追加実験の 100 動作の結果 (%)

soccer+dribble	100	swim+crawl	36	swim+breaststroke	9
fold+origami	96	cut+hair	35	climb+tree	9
crochet+hat	95	paint+wall	33	clean+floor	8
arrange+flower	94	lunge	32	tie+tie	8
paint+picture	88	hit+golfball	32	jump+rope	8
boxing	86	fieldhockey+dribble	32	swim+butterfly	7
comb+hair	83	shave+mustache	31	brush+teeth	7
parachute+jump	82	chat+friend	31	boil+egg	7
do+exercise	79	pick+lock	30	cook+rice	6
do+aerobics	78	play+guitar	28	iron+clothes	6
do+yoga	77	plant+flower	28	bake+bread	6
surf+wave	75	catch+fish	28	slap+face	5
serve+volleyball	75	serve+tennis	27	grill+fish	5
shoot+arrow	73	lift+weight	27	smile	4
fix+tire	67	row+dumbell	26	weep	2
blow-dry+hair	64	hang+wallpaper	26	run	2
basketball+dribble	64	jump+trampoline	24	kiss	2
ride+bicycle	62	sew+button	24	blow+candle	2
curl+bicep	58	roll+makizushi	24	cut+onion	1
shoot+ball	58	ride+horse	24	wash+face	0
bowl+ball	58	fry+tempura	23	read+book	0
tie+shoelace	57	row+boat	20	knit+sweater	0
laugh	50	massage+leg	20	watch+television	0
play+drum	49	play+piano	19	walk	0
ski	49	drive+car	17	slice+apple	0
harvest+rice	49	wash+dishes	15	plaster+wall	0
dive+sea	49	wash+clothes	15	pick+apple	0
twist+crunch	47	draw+eyebrows	15	peel+grape	0
dance+flamenco	45	sing	12	jog	0
dance+hiphop	43	squat	12	head+ball	0
dance+tango	41	raise+leg	12	drink+medicine	0
play+trumpet	41	cry	12	drink+coffee	0
skate	37	eat+sushi	11	count+money	0
AVG (1-33)	65.7	swim+backstroke	9	**AVG (68-100)**	3.2
		AVG (34-67)	23.2	**AVG (1-100)**	30.6

人工物動作と自然現象を含む 6 種類の非人間動作の結果は表 3 にて表す．結果は 6 種類の人間動作の結果ほど良くないことがわかった．その理由の 1 つとしては時空間特徴が非人間動作に適した特徴ではないからだと考えられる．非人間動作にも有効の特徴を探すことは今後の課題の 1 つになる．

さらに，大規模な動作ショットデータベースを構築する実験として，100 種類の人間動作のビデオショットを収集した．実験では，合計で 100,000 本の YouTube ビデオのビデオメタデータ分析と，20,000 本の YouTube ビデオの時空間特徴分析を 300 コアからなるクラスタ計算機を利用して行った．100 種類のキーワードは主に動詞と動作対象に相当する名詞を組み合わせて構築した．

表 4 に 100 種類のキーワードで自動選択した上位 100 ショットの適合率の結果を示す．100 種類の平均の適合率は 30.6%となった．そのうち，適合率の上位 33 位まで

のキーワードに関する結果は 65.7%, 34 位から 67 位までは 23.2%, 68 位以下は 3.2%であった．上位 33 位までには比較的多くのスポーツに関係する動作キーワードが含まれている．一方，日常的な動作や動詞のみのキーワードはよい結果が得られず 0%に近い結果が多かった．

この結果から適合率は動作の種類と動作キーワードの選択のし方に大きく依存することがわかった．100 種類中の一部の 23 種類のサムネイル画像と要約動画は本論文の付録に添付した．また，100 種類すべてのサムネイル画像および要約動画は，http://img.cs.uec.ac.jp/dohang/miru2011/ にて公開中である．

6. 結論 と 今後の課題

本研究では，動作に関係するキーワードを与えるだけで Web ビデオから指定キーワードに対応したビデオショットを自動抽出する新しい手法を提案した．実験結果として完全に教師なしで 6 種類の人間動作について上位 100 位において 49.5%の適合率が得られた．また Youtube からダウンロードをした 20,000 本の動画を解析して，100 種類のキーワードに対応する大規模ショット自動抽出実験を行った．

今後の課題としては，VisualRank の補正ベクトルの設定方法の詳細な検討と非人間動作にも有効である視覚特徴の検討が挙げられる．こうした検討を行うことで，提案手法は制限のない動作種類に適用可能となることを期待している．さらに，多種多様な動詞に対応するショットが自動収集可能になることで，視覚情報を用いた動詞概念の関係分析が可能となり，従来は主に名詞概念で主に行われていた視覚情報による概念の関係分析が動詞概念においても可能となる．

文　献

[1] J. Sun, X. Wu, S. Yan, L. F. Cheong, T. S. Chua and J. Li: “Hierarchical spatio-temporal context modeling for action recognition”, Proc. of IEEE Computer Vision and Pattern Recognition (2009).

[2] Y. Jing and S. Baluja: “Visualrank: Applying pagerank to large-scale image search”, IEEE Transactions on Pattern Analysis and Machine Intelligence, **30**, 11, pp. 1870–1890 (2008).

[3] C. Schuldt, I. Laptev and B. Caputo: “Recognizing human actions: A local SVM approach”, Proc. of IAPR International Conference on Pattern Recognition, pp. 32–36 (2004).

[4] M. Blank, L. Gorelick, E. Shechtman, M. Irani and R. Basri: “Actions as space-time shapes”, Proc. of IEEE International Conference on Computer Vision (2005).

[5] C. Liu, J. Yuen and A. Torralba: “Nonparametric scene parsing: Label transfer via dense scene alignment”, Proc. of IEEE Computer Vision and Pattern Recognition (2009).

[6] J. Liu, J. Luo and M. Shah: “Recognizing realistic action from videos”, Proc. of IEEE Computer Vision and Pattern Recognition (2009).

[7] M. Marszalek, I. Laptev and C. Schmid: “Actions in context”, Proc. of IEEE International Conference on Computer Vision (2009).

[8] Z. Wang, M. Zhao, Y. Song, S. Kumar and B. Li: “Youtubecat: Learning to categorize wild web videos”, Proc. of IEEE Computer Vision and Pattern Recognition (2010).

[9] G. Toderici, H. Aradhye, M. Pasca, L. Sbaiz and J. Yagnik: “Finding meaning on YouTube: Tag recommendation and category discovery”, Proc. of IEEE Computer Vision and Pattern Recognition, pp. 3447–3454 (2010).

[10] L. Duan, D. Xu, I. W. Tsang and J. Luo: “Visual event recognition in videos by learning from web data”, Proc. of IEEE Computer Vision and Pattern Recognition (2010).

[11] J. Niebles, H. Wang and L. Fei-Fei: “Unsupervised learning of human action categories using spatial-temporal words”, Proc. of British Machine Vision Conference (2006).

[12] J. Niebles, B. Han, A. Ferencz and L. Fei-Fei: “Extracting moving people from internet videos”, Proc. of European Conference on Computer Vision, pp. 527–540 (2008).

[13] N. I.-. Cinbins, R. G. Cinbins and S. Sclaroff: “Learning action from the web”, Proc. of IEEE International Conference on Computer Vision, pp. 995–1002 (2009).

[14] N. Dalal and B. Triggs: “Histograms of oriented gradients for human detection”, Proc. of IEEE Computer Vision and Pattern Recognition, Vol. 1, pp. 886–893 (2005).

[15] L. Ballan, M. Bertini, A. D. Bimbo, M. Meoni and G. Serra: “Tag suggestion and localization in user-generated videos based on social knowledge”, Proc. ACM Multimedia Workshop on Social Media, pp. 3–7 (2010).

[16] K. Yanai: “Generic image classification using visual knowledge on the web”, Proc. of ACM International Conference Multimedia, pp. 67–76 (2003).

[17] R. Fergus, P. Perona and A. Zisserman: “A visual category filter for google images”, Proc. of European Conference on Computer Vision, pp. 242–255 (2004).

[18] R. Fergus, L. Fei-Fei, P. Perona and A. Zisserman: “Learning object categories from google’s image search”, Proc. of IEEE International Conference on Computer Vision, pp. 1816–1823 (2005).

[19] K. Yanai and K. Barnard: “Probabilistic Web image gathering”, Proc. of ACM SIGMM International Workshop on Multimedia Information Retrieval, pp. 57–64 (2005).

[20] L. Li and L. Fei-Fei: “OPTIMOL: automatic Online Picture collecTion via Incremental MOdel Learning”, Proc. of IEEE Computer Vision and Pattern Recognition (2007).

[21] F. Schroff, A. Criminisi and A. Zisserman: “Harvesting image databases from the web”, Proc. of IEEE International Conference on Computer Vision (2007).

[22] Q. Yang, X. Chen and G. Wang: “Web 2.0 dictionary”, Proc. of ACM International Conference on Image and Video Retrieval, pp. 591–600 (2008).

[23] D. Liu, X. Hua, L. Yang, M. Wang and H. Zhang: “Tag ranking”, Proc. of the International World Wide Web Conference, pp. 351–360 (2009).

[24] A. Noguchi and K. Yanai: “A surf-based spatio-temporal feature for feature-fusion-based action recognition”, Proc. of ECCV WS on Human Motion: Understanding, Modeling, Capture and Animation (2010).

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

[26] B. Lucas and T. Kanade: “An iterative image registration technique with an application to stereo vision”, Proc. of International Joint Conference on Artificial Intelligence, pp. 674–679 (1981).

[27] N. I. Cinbis and S. Sclaroff: “Object, scene and actions: Combining multiple features for human action recognition”, Proc. of European Conference on Computer Vision, pp. 494–507 (2010).