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.

時空間特徴量を用いたYoutube動画からの特定動作ショットの自動抽出

DOHANG NGA†　　柳井　啓司†

† 電気通信大学 情報工学科　〒 182–8585 東京都調布市調布ヶ丘 1–5–1
E-mail: †{dohang,yanai}@mm.cs.uec.ac.jp

あらまし　本研究では、教師なし手法によって単語を入力するだけで、様々な動詞に対応するショットを Web ビデオから自動抽出することを目的とする。まず、Web 上での一般ユーザが動画像を自由にアップロードできる動画共有サイトである Youtube から特定動作に対応する動画に付与されるタグの共起情報を用いて、動作キーワードについての Web2.0 辞書を作成する。次に、この辞書に基づくキーワードに対する相関値の高い順に動画を大量に収集する。次は、収集動画に対して色情報に基づいてショット分割する。次に、ショットから時空間特徴、動き特徴、視覚特徴を抽出する。そして、各ショットを bag-of-features として表現する。最後に、画像集合の代表的な画像を選出するためのランキング手法である VisualRank を適用して、特定動作に対応するショットを自動抽出する。時空間特徴は野口らによって提案された時空間特徴を利用する。さらに、時空間特徴に全体的な動きを表現する動き特徴とガボールフィルタによる視覚特徴を統合し、ショットをランキングすることも行い、その有効性を示す。また、VisualRank の計算では、動作のタグデータベースから構築された Web2.0 辞書に基づく相関値が高い動画からのショットを強調するとする。実験では、6 種類の動詞についてランキングをおこない、上位 10 ショットに 53.3%,、上位 100 ショットに 48.7%の精度でそれぞれ動詞に対応したショットをランキングすることができた。
キーワード　時空間特徴、Web 動画、自動学習、VisualRank、Web2.0 辞書

Mining Specific Actions from Youtube Video with Spatio-Temporal Features

DO HANG NGA† and Keiji YANAI†

† The University of Electro-Communication,　Choufugaoka 1–5–1, Choufu, Tokyo, 182–8585 Japan
E-mail: †{dohang,yanai}@mm.cs.uec.ac.jp

Abstract　In this paper, we present a new method of automatically extracting from tagged Web videos the video-shots correspond to specific actions with just inputing the action keywords such as walking, eating etc.
Key words　Spatio-Temporal Feature, Web video, unsupervised learning

1. はじめに

1.1 背　景

近年、インターネットの普及と共に Web 上の動画の数は爆発的に増大している。Web 上での YouTube やニコニコ動画など一般ユーザが動画像を自由にアップロードできる動画共有サイトという動画ソースは膨大で、容易に収集し、利用できる。よって、動作学習の研究では、Web 上の動画をデータソースとして使う研究は増えてきている。

ただし、Web 上の動画の利用だと、さまざまな難点が存在する。通常、Web 上の動画にはユーザによってタグと呼ばれるキーワードが付与されているが、タグは主観的であり、必ずしも動画全体の内容と関係しているとは限らない。例えば, ”eating” というタグを持った動画であって、最初から最後まで”eating” している訳ではなく、店に入るシーンや注文して待っているシーン、場合によっては調理しているシーンなども含まれていることがある。こうした場合、eating に対応したシーンを自動的に検出出来れば、eating 以外のシーンを取り除いて、eating シーンのみを閲覧することが可能となる。そこで、本研究では、タグの中でも特に動詞に対応するタグに直接対応するシーンの自動抽出を行う。

また、Web 上の動画の多くは,「解像度が低い」、「カメラは固定ではない」、「動作は複数」、「動作を行う人が一人に限らない」、「背景ノイズは強い」のような制限のない動画である。現在行われている動画像認識手法のほとんどが制限のある映像下における動作の認識を行っているので、Web 動画像に対してそのまま適用することは不可能である。

さらに、ほとんどの研究における動作の学習は教師信号あり学習法で、そのために多くても 15 種類程度の限られた動作についてしか動作認識が実現していない [1]。教師信号なしで特定動作に対応するショットの自動抽出が成功すれば Web 動画像に対する新しい動作認識システムを実現可能になることが期待

できる。

1.2 目　的

本研究では、6 種類の動詞の動画についての時空間特徴、動き特徴、静的な視覚特徴を使用して、その動詞に対応するショットを Web2.0 辞書を参考にして VisualRank 手法によって自動的に抽出し、考察を行う。図 1 に示すように、ショットのランキング付けを行う。この図では、Web からの検索結果の大量にある動画ショットから”running marathon”という動作と対応したショットを自動的に抽出している。本研究では、この例のようにランキング上位には”running marathon”を含むような動画ショット、下位には”running marathon”とは関係のない動画ショットをランキングすることを目的としている。

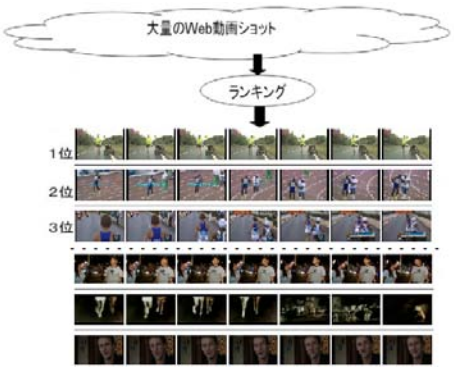

図 1　提案手法の概念

1.3 関連研究

最近、動作認識、動作学習は広く研究されるようになったが、YouTube などからの動画のような制限のない動画に実験する研究は多くない。研究のほとんどが「カメラは動かない」、「動作している人間が一人」、「高い解像度」、「不変の背景」のような制限のある映像下において、学習と認識を行っている。Niebles ら [2] はこのような制限のある動画における人間動作の自動学習手法を提案した。この研究は Dollar ら [3] により提案された時空間特徴を用いて動画を表現して、pLSA モデルによる教師なしの学習を行う。したがって、この研究は時空間特徴の利用と教師なしの手法の提案という点では、本研究と似ているが、KTH データセット [4] のような制限のある動画セットでのみ実験している。

野口らは Web 動画マイニングのための新しい時空間特徴の抽出手法を提案した。この手法は処理が単純で、計算時間が少なく、動きの回転に関しても頑健であるので、大量の Web データセットに適する [5]。野口らによる教師あり分類は 91% の分類精度も達成した。

もうひとつの関連研究としては Liu らの研究が挙げられる [6]。この研究も YouTube からの動画をデータソースとする。彼らは 11 種類の人間動作を含む Wild YouTube というデータセットを作成した。彼らの提案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 記述子を利用し、Adaboost に基づき統合する手法を提案している。この研究では、PageRank に基づく重要な特徴の選択を行っている。本研究では、Liu らの研究とは異なり、PageRank をショットに対して適用することによって、特定の動詞について代表的なショットの自動抽出を行なう。

Liu らのデータセットを利用した研究として、Cinbis らの人間動作認識研究 [7] がある。この研究は、動作認識で、背景と関連物体の情報も重要であることを論じて、人間、背景と物体からの複数の特徴を Multiple Instance Learning (MIL) によって統合する人間動作認識システムを提案する。本研究も統合特徴を利用するが、彼らの研究と違って、完全に教師なしでの動作学習を目的とするので、MIL などは適用して、特徴統合を行うとしない。

もうひとつの関連研究としては Satkin らの研究である [8]。この研究は動作認識に対し最も識別性が高い部分を動画から推定するフレームワークを提案する。トリミングされる動画以外の動画を用いて、multi-class SVM 分類器を学習してから、それらを利用して最適なトリミングを選択する。さらに、その部分を用いて分類器をまた学習することで動作認識システムを改善する。本研究は、動画のどの部分も使用して、動作の自動学習を行う。

本研究と最も関連がある研究としては Cinbis らの動作学習研究である [9]。この研究も Web データを利用して人間動作の自動学習を行う。ただし、彼らは動作についての静画像を利用することで動画像における動作のアノテーションの精度を改良することを目的とする。それに対して、本研究は動画像に付けられるタグを利用することでショットランキングの精度を改善する。また、彼らの手法では、動作の記述は動き特徴を使わず、全て静止画像ベースで行われている。本研究では画像の視覚情報のみではなく，動き情報も利用して動画を表現する。

3. 提案手法の概念

本研究では、時空間特徴を用いた新しい動画学習手法を提案する。図 2 は提案手法の流れを示している。

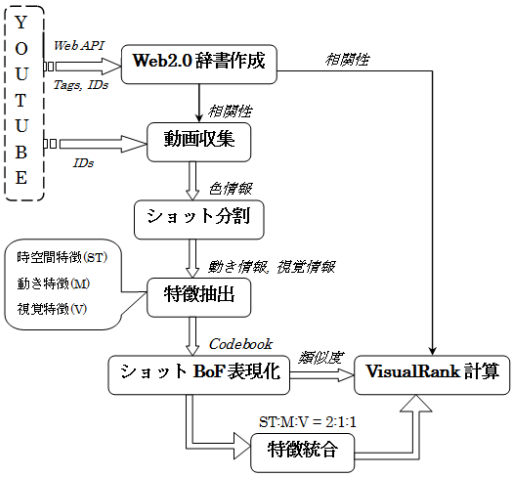

図 2　提案手法の流れ

まず、Web 動画に付けられるタグを用いて、動作についての Web2.0 辞書 [10] を作成する。次に動画を Web2.0 辞書に基づく相関値の高い順に収集する。次は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、静的視覚特徴を抽出する。次に、抽出された特徴を用いて、ショットをヒストグラム化する。最後に、特徴統合を行って、ショットを VisualRank 手法によってランキングする。ランキングについてであるが、Web2.0 辞書による相関値が高い動画のショットを強調するとする。

4. 提案手法の詳細

4.1 Web2.0 辞書の作成

本研究では、動画につけられるタグから Web2.0 辞書を構築する。この辞書に基づいて、動画収集と VisualRank 計算での動画の利用順番が決められる。

ここで、Web2.0 辞書の概念を紹介する。Web2.0 辞書の定義は [10] によって提案される。この辞書は、同時出現の関係を相関により定量化することですべての言葉を定義する辞書である。この辞書では、順序付きの語のリストを親語と呼ぶ。その説明は子語と呼ばれるキーワードのリストである。それぞれの子語には親語と子語の相関を定量化する浮動小数点値が割り当てられる。相関値は 0(子語のリストが存在しない) から 1(親語自身) までの間の値である。子語は相関値減少順にリストされる。

本研究は、Web API を利用して、各キーワードを入力した結果の上位の 1000 動画のタグと動画 ID を収集する。そして、集められたタグを利用して、そのキーワードについての Web2.0 辞書を構築する。Web2.0 辞書を適用して、動画は動作キーワード (eating、running など) との相関値の高い順に利用する。動画のそのキーワードとの相関値は動画に付けられるタグの相関値の合計として計算する。

4.2 動画収集

Youtube API を用いて動詞キーワードに対応する動画を 1000 動画分検索し、その動画 ID と計算できた相関値に基づいて、相関値が高い動画を優先に収集する。1000 動画の中、上位の 200 動画を利用する。このように、Web2.0 辞書による動詞キーワードと動画の関連度スコアを利用することによって動作との関連性が低い動画を減らすことが実現可能になる。

4.3 ショット分割

集められた動画をショットに分割する。ショットとは場面の切り替わるまでのシーンのことである。

動画の隣接するフレームの色情報を比較することでショット分割を行う。色情報としては、RGB 色空間の各軸を 4 分割することで、本来なら 256×256 次元の画像を $4 \times 4 \times 4$ 次元で表現した特徴を使用した。フレーム間の距離としてはヒストグラムインターセクションを利用する。

4.4 特徴抽出

4.4.1 時空間特徴

時空間特徴は Noguchi ら [5] によって提案された時空間特徴抽出手法を用いて抽出する。この手法の流れは次で示す。

時空間特徴抽出手法の流れ ([11] から引用)
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

Noguchi らの提案時空間特徴抽出手法では、まず、フレームから、SURF 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 アルゴリズムに基づいて動きを計算する。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。

4.4.2 動き特徴

動き特徴抽出には Lucas-Kanade アルゴリズム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。

4.4.3　静的視覚特徴

視覚特徴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。

4.5 ショットの特徴表現

本研究では、抽出した特徴を用いて、各ショットの特徴表現を行う。それぞれのショットは 1 つの bag-of-features(BoF) として表現する。

BoF では画像を局所特徴の集合と見なす。そして特徴の位置情報を無視して、特徴の出現頻度を利用し、分類を行う手法である。ここでは 1 つの動画ショット毎に BoF によって特徴ベクトルを生成する。その手順を以下にまとめる。

手順 4.1 BoF 表現の手順
step1：特徴をヒストグラム化
step2：ベクトル量子化を行い，codebook を作成
step3：codebook を基に，特徴の出現頻度を計算

a）　特徴のヒストグラム化 (step 1)

抽出された特徴をベクトルとして表現する。時空間特徴の次元数は 257 とした。動き特徴の場合は、それぞれの検出されたフローを 8 方向、7 段階の大きさからなるヒストグラムに投票する。よって動き特徴は $8 \times 7 = 56$ 次元数のベクトルとして表現する。視覚特徴の場合は、フレームを 2×2 に分割し各領域から 24 次元ベクトルを抽出する。領域ごとのベクトルを結合することによって一つの特徴とする。よって視覚特徴の次元数は各領域のベクトル 24 次元 ×4 で表される。

b）　codebook の作成 (step 2)

抽出された特徴をベクトル量子化することで codebook を作成する。ベクトル量子化には k-means クラスタリングを利用する。計算されたクラスタ中心を visual words と呼び、その集合は codebook と呼ばれる。時空間特徴，視覚特徴におけるコードブックサイズは 5000、動き特徴におけるコードブックサイズはに 3000 に設定した。

c）　動画の特徴表現 (step 3)

抽出された特徴は codebook の中で最も近い visual words に投票される。これにより動画を visual words の出現頻度のヒストグラムで表現することが出来る。ただし抽出される時空間特徴の数は動画によって異なるので、codebook によって投票された特徴数を全特徴数で割ることで正規化を行う。

4.6 特徴統合

時空間特徴は、視覚情報と動き情報を内包しているが、それは動的な視覚情報と局所的な動き情報である。ここで、静的視覚特徴とフレーム全体からのグローバル的な動き特徴も抽出して、それを時空間特徴に結合する。特徴結合により、より高いランキング精度が期待できる。

特徴統合では、各特徴の重みの決定は最も重要であるが、本研究は完全に教師なしで動作ショットを抽出することを目標とするので、多くの研究と違って、MKL などの教師ありの学習法や cross-validation を適用して、各特徴の最適な重みを計算

するとは困難である。

ここで、動き情報と視覚情報の両方を表現できる時空間特徴は認識精度が一番高いであると証明されたので [5]、時空間特徴の重みは最も高いと設定する。実験では、時空間特徴、動き特徴、視覚特徴の重みの比例は 2:1:1 とする。

4.7 VisualRank 手法によるショットランキング

特定動作に対応したショットを自動抽出するため、VisualRank というランキング手法を適用する [12]。VisualRank 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式 (1) によって求められる。

$$s(H_1, H_2) = \sum_{i=1}^{|H|} min(H_{1i}, H_{2i}) \quad (1)$$

ここで、H はショットのヒストグラム、$|H|$ はヒストグラムの次元数を示す。

統合特徴を用いる場合、類似度行列は式 (2) によって計算する。

$$S_{combined} = w_{st} \times S^*_{st} + w_m \times S^*_m + w_v \times S^*_v \quad (2)$$

$$where \;\; w_{st} = \tfrac{1}{2}, w_m = \tfrac{1}{4}, w_v = \tfrac{1}{4}$$

ここで、st、m、v はそれぞれ時空間特徴、動き特徴、視覚特徴を指す記号であり、S^* は対応する正規化した類似度行列、w は対応重みである。

VisualRank 計算によって、より重要な画像が上位にランク付けられる。ただし、更新ステップごとに補正をかけるために、補正ベクトルが用いられ。本研究では、補正ベクトルとしては、不均一なベクトルを与える。動作のタグデータベースから構築された Web2.0 辞書に基づく相関値が高いと評価される動画ショットを強調するとする。VisualRank の計算は式 (3) によって行う。

$$VR = VR \times dS^* + (1-d)q \quad (3)$$

$$where \;\; q_j = \begin{cases} \frac{1}{m}, & j < m \\ 0, & j \geqq m \end{cases}$$

ここで、m は強調するショットの数である。実験で $m = 1000$ と設定した。VisualRank 計算の入力ショット数は動詞によって違うが、実験では Web2.0 のスコアの高い順に 1800 ショットを利用したので、このうち Web2.0 のスコアの上位の動画の 1000 ショットに重みを付与することとした。なお、この補正ベクトルの設定方法は多様な方法が考えられるため、設定方法の詳細な検討は今後の課題とする。類似度行列の大きさはショット数の 2 乗となるため、大量のデータセットを処理するとき、そのまますべてのデータを用いると、類似度行列の計算は不可能になる。Web2.0 スコアの上位のみに実際に処理する動画を限定することで、VisualRank の計算量は制御可能となる。

また、補正の強度はパラメータ d によって調整する。多くの場合、$d \geqq 0.8$ を満たす値が設定される。実験では $d = 0.85$ と設定した。

4. 実　験

4.1 データセット

実験には野口らが構築した Youtube データセットを使用した。

野口らのデータセットは “batting”, “running marathon”, “walking street”, “shoot football”, “jumping trampoline”, “eating ramen” の 6 動作からなる (図 3)。表 1 はこのデータセットの詳細を示している。ここで、利用ショット数は実際に VisualRank 計算に利用したショット数を指すものである。VisualRank 計算に利用したショット数は 2000 を上限とした。ショットが多すぎる動画に対し、式 (4) によってショットの上限数を設定した。その理由としては、動画が Web2.0 スコアの上位であるが、不適応で計算に悪い影響を与える場合を避けるからである。

図 3　野口らの Youtube データセット

$$MS(V_i) = S(V_i) \times AS \times 3 + P \quad (4)$$

$$where \;\; P = \begin{cases} NS/2, & 20 < NS < 50 \\ NS/3, & 50 \leqq NS < 90 \\ 40, & NS \geqq 90 \end{cases}$$

ここで、MS はビデオの利用ショットの上限数、S は Web2.0 スコア、AS は平均スコア、NS はビデオのショット数を示している。ショット数が 20 以下の場合はすべて利用するとする。P は動画のショット数によって調整する。利用ショットの上限数はその動画のスコアに比例する。また、平均スコアにも比例する。平均スコアはあまり低かったら Web2.0 の信頼性も低いので、ショットの上限数を減らす必要がある。

また、表 1 には [5] のランク結果の上位 100 の適合率も表す。[5] は教師信号ありの手法を使っている。

表 1　利用データセット

動作	動画数	ショット数	利用ショット数	[5] のランク結果
batting	174	6391	2000	83%
eating	170	1795	1795	78%
jumping	174	2005	1650	98%
running	164	4213	2000	87%
shoot	142	4698	2000	82%
walking	160	3841	1950	52%
合計	1,247	22943	11395	平均: 80%

4.2 評価法

結果を評価するには、ランキング後のランク上位の 10 ショット、30 ショット、50 ショット、100 ショットとランク下位の 100 ショットについての適合率を利用する。

4.3 実験結果

本論文では動作自動学習フレームワークについて提案をした。ここでは野口らの YouTube データセットに対する実験結果を表す。

Web2.0 辞書の適用の有効性を確認するには、Web2.0 辞書適用なしで、ランダムに 2000 ショットを選択し、VisualRank の計算で均一な補正ベクトルを与える実験も行った。その実験の結果は表 2 で示している。ただし、使用特徴としては時空間特徴のみである。表 3 に Web2.0 辞書適用ありの結果を示す。結

表 2　Web2.0 辞書適用なしの結果の適合率

動作	上位 100	下位 100
batting	68%	8%
eating	42%	19%
jumping	76%	19%
running	18%	17%
shoot	6%	5%
walking	7%	8%
平均	36.17%	12.67%

表 3　Web2.0 辞書適用ありの結果の適合率
時空間特徴のみ

動作	上位 10	上位 30	上位 50	上位 100	下位 100
batting	100%	86.7%	72%	68%	2%
eating	50%	33.3%	40%	47%	0%
jumping	80%	83.3%	82%	82%	5%
running	30%	33.3%	30%	34%	5%
shoot	20%	26.7%	32%	33%	1%
walking	20%	23.3%	20%	21%	1%
平均	50%	47.8%	46%	47.5%	2.2%

表 4　Web2.0 辞書適用ありの結果の適合率
統合特徴使用

動作	上位 10	上位 30	上位 50	上位 100	下位 100
batting	70%	83.3%	76%	66%	3%
eating	50%	60%	48%	41%	1%
jumping	100%	90%	86%	85%	9%
running	30%	40%	42%	38%	8%
shoot	50%	53.3%	38%	29%	1%
walking	20%	50%	40%	38%	2%
平均	53.3%	62.8%	55%	48.7%	4%

果からわかるように、Web2.0 辞書の適用によって平均適合率がはるかに改善された。特に、“running”, “walking”, “shoot” の 3 種類の動作については、上位にランキングされる動作と対応したショット数は比較的に増加した。Web2.0 辞書適用なしの結果では、“batting” や “jumping” などは、上位の適合率は高いが (それぞれ 68% と 79%)、“shoot” と “walking” は非常に低い (それぞれ 6% と 7%) ことがわかる。その理由のひとつとしては、ランダムにショットを選択することによって、結果の安定性が失われてしまうからと考えられる。VisualRank の計算で均一な補正ベクトルを与えたので、選ばれたショットの中は、対応するショット数が多ければ、良い結果が得られるが、逆に、不対応なショット数が大部分を含めてしまうと当然結果が良くなくなる。それに対して、Web2.0 辞書によるスコアを使えることにより、適切なショットが選択される機会はどの動作にも多くなる。このように、動作による結果の非常に大きなギャップを避けることが出来る。

それに、下位の適合率については、Web2.0 辞書の使ったほうが比較的に良いことがわかった。下位には、できるだけ多く不対応なショットをランキングすることも目的である。Web2.0 辞書を適用した場合、下位に対応したショットは平均が 2.2%しか占めないので、動作認識にこの下位のショットはネガティブショットとして使うことは信頼性が高いと考えられる。

また、特徴を統合した結果は表 4 にて表される。結果からわかると、特徴統合により、平均としては上位 100 の適合率があまり改善されないが、上位 10 ランク、上位 30 ランク、上位 50 ランクの適合率が良くなっている。特に、上位 30 の結果では、平均適合率が、時空間特徴のみの場合、47.8% であるが、統合特徴の場合、62.8% である。

5. 結　論

本研究は、Web 動画を利用して、動作の自動学習フレームワークを提案した。結果として、時空間特徴のみ使った場合は、上位 100 ショットの平均適合率は 47.5% で、下位 100 ショットの平均適合率は 2.2% であった。また、特徴統合を行った場合は、上位 100 ショットの平均適合率は 48.7%、下位 100 ショットの平均適合率は 4% になった。

VisualRank 計算に Web2.0 辞書の適用の有効性が実証できた。Web2.0 辞書適用なしの結果と比べて、上位 100 の適合率が約 12% も改良できた。特徴統合により上位 10、上位 30、上位 50 により多く対応したショットをランキングすることができた。

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、大規模な実験が挙げられる。今回は Web2.0 スコアの上位 1000 ショットにバイアスを与えて、それ以外のショットのバイアスは 0 としたが、上位ショットの数の検討や、Web2.0 スコアの値や順位に応じたバイアス値の設定方法など、様々なバリエーションが考えらる。また、現在 100 種類の動詞についての実験を実施中であり、今後はさらに大規模な実験を予定している。多数の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. *CVPR 2009*, pp. 2004–2011, 2009.

[2] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, Vol. 79, pp. 299–318, 2008.

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pp. 65–72, 2005.

[4] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, Vol. 64, pp. 107–123, 2005.

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

[6] J. Liu, L. Jiebo, and M. Shah. Recognizing realistic actions from videos in the wild. In *CVPR 2009*, pp. 1996–2003, 2009.

[7] N. Ikizler-Cinbis and S. Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. In *ECCV 2010*, Vol. 6311, pp. 494–507. 2010.

[8] S. Satkin and M. Hebert. Modeling the temporal extent of actions. In *ECCV 2010*, Vol. 6311, pp. 536–548. 2010.

[9] N. Ikizler-Cinbis, R.G. Cinbis, and S. Sclaroff. Learning actions from the web. In *ICCV 2009*, pp. 995–1002, 2009.

[10] Q. Yang, X. Chen, and G. Wang. Web 2.0 dictionary. In *Proc. of ACM International Conference on Image and Video Retrieval*, pp. 591–600, 2008.

[11] 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 (MIRU2010), 2010.

[12] Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30, pp. 1877–1890, 2008.

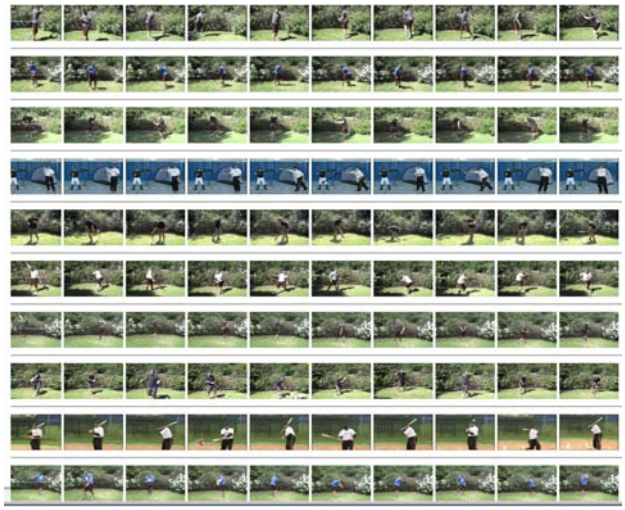

図 4　batting の上位 10 ランクショット

図 7　batting の下位 10 ランクショット

図 5　jumping の上位 10 ランクショット

図 8　jumping の下位 10 ランクショット

図 6　eating の上位 10 ランクショット

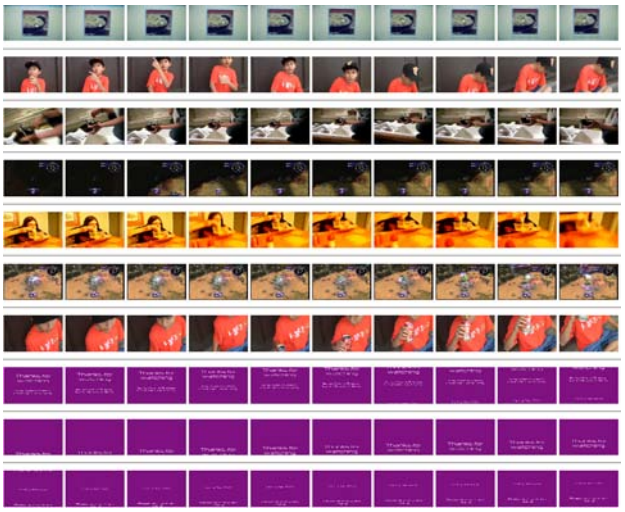

図 9　eating の下位 10 ランクショット