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.

視覚特徴およびタグ共起を用いた大規模 Web ビデオショットランキング

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　本研究はビデオショットの間の視覚的関連およびビデオとそのタグの共起関係を利用した新しいランキング手法を提案する。提案手法は特定動作の関連 Web ビデオショットの自動抽出システムに適用する。提案手法ではノイズが多く含まれたタグ情報はビデオのコンテンツ情報によって洗練し、より効率的に使うことができる。また、ビデオショットの視覚特徴と同時にそのビデオの洗練されたタグ特徴も考慮することによって、視覚特徴のみを利用することよりランク上位で多くの関連ありのビデオショットが得られる。大規模実験でベースラインとの比較を行い、提案手法の有効性を検証できた。

キーワード　Web データ、大規模、タグ共起、教師なし、共ランキング

Large-scale Web Video Shot Ranking Exploiting Visual Features and Tag Co-occurance

Nga HANG DO† 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 propose a novel ranking method which aims to detect automatically relevant Web video shots of specific actions using visual links between the video shos as well as textual links between the videos and their tags. Our method adopts content based features to reduce tag noise and exploits the refined tags to improve VisualRank method which employs only visual features. We conduct large-scale experiments and show the effectiveness of our proposed method over the baseline.

Key words　Web data, large-scale, tag co-occurance, unsupervised learning, co-ranking

1. 導　　入

最近、動作認識分野では、“in-the-wild” YouTube データセット [1] や HMDB51 [2]、Hollywood アクションデータセット [3] などのような大規模で、制限なしのデータベースが広く利用されるようになって来ている。その理由としては以前からの小規模で、制限有りのデータベース (KTH [4]、Weizmann [5]) における認識精度はほぼ 100%となっているからである [1]。ほとんどの新しいデータベースはインターネットから収集されたものである。YouTube API などの Web API を利用することによって、動画共有サイトから容易に動画を大量に取得することができる。さらに、Web 動画のほんとどは CGM (Consumer Generated Media) なので制限が無く、以前より難しいデータベースの構築には適切だと考えられる。

多くの動画共有サイトはキーワードに基づいたサーチエンジンを提供する。このサーチエンジンは “タグ” というビデオに付けられるキーワードを活用してユーザによって挙げられたサーチキーともっとも関連があるようなビデオを返す。ただし、タグはユーザによって主観的に与えられるものなので信頼性が低いという問題があり、また、サーチキーにも多様性が含むという場合もある。よって、検索結果には非関連の動画があることは少なくない。また、CGM ビデオは元々動作認識研究のために作られるものではないので特定動作についてのビデオではなく様々な動作が含まれていることが一般的である。したがって、検索結果のビデオは特定動作と対応する部分がない場合もあり、またはあってもそのビデオの一部だけである場合もある。Web 動画を用いて動作のビデオショットデータベースを生成するのに研究者たちは特定動作の関連部分を手動で検索しないといけないため時間及び手間が非常にかかるということである。ここでビデオショットとは動作の一部であり、一つのシーンと対応する連続フレームセットを示す。もし、Web 上の動画のような簡単に取得可能な動画から特定動作の対応ビデオショット

を自動的に得られるようになったらアクションデータベースの構築は容易になると考えられる。

そこで、以前に我々は教師なしで，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 2-ランキングステップフレームワークを提案した [6]。キーワードとしては “brushing+teeth” や “jogging” などの人間動作および “airplane+flying” や “typhoon” などの非人間動作に関する単語を対象とする。ステップ 1 ではタグ共起に基づくビデオランキングを行い、キーワードとよく共起するタグが多く付けられるビデオを関連ビデオとして収集する。ステップ 2 ではそのビデオからのショットのすべてに対し視覚特徴によるショットランキングを行い、ランク上位のショットをキーワード動作の対応ショットとして得られる。この手法は多くの動作種類に有効であるが、タグリストに極めてノイズが多い場合、タグ情報によるビデオランキングステップは関連ビデオを選択できず、最後に対応ショットはあまり得られないということが実験より判明した。

そこで本研究では、以前に提案したシステムのように、タグと与えられたキーワードの関連性を評価するにはタグ共起頻度のみを利用するのではなく、対応動画との相互関係も利用することを考えた。例えば、もし、あるビデオショットは動作キーワードと関連することがわかったらそのビデオに付けられるタグもキーワードとの関連があると推定することができる。逆に、もし、あるタグがキーワードと関連深いであるならそのビデオも重要であるという可能性が高いと考えられる。

本研究は、VisualTextualRank というランキング手法を提案する。VisualTextualRank は VisualRank [7] と Liu らの Biparite-Graph [8] を拡張し、Web ビデオの視覚特徴とメタデータを同時に効率的に活用する手法である。実験結果からわかるように、VTR(VisualTextualRank) を適用したフレームワークは以前の VisualRank を適用したフレームワーク ([6], [9]) より Web ビデオから多くの関連ショットを検出することができた。

本論文の構成は次のようになる：セクション 2 ではいくつの関連研究を紹介し、提案手法はセクション 3 で詳しく説明し、セクション 4 では実験とその結果を示し、結論および討論はセクション 5 で述べる。

2. 関 連 研 究

最近、動作認識世界においては、動画を ST(Spatio-Temporal) 特徴の BoF(Bag-of-Features) として表現し、SVMs(SuperVectorMachines) を利用して動作を認識するという教師ありの手法はよく使われている。この手法の有効性は多くの研究によって検証されている。その研究の中は Jones らの研究 [10] が最近の研究である。Jones らの研究はさらに、ユーザのフィードバックを活用して認識結果を改良するのを提案する。

本研究の目標は、完全自動システムを目指して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(Probabilistic Latent Semantic Analysis) モデルを用いて KTH データセット [4] 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]。また、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のみ使う。それに対して、本研究は Web 動画と時空間特徴も利用する。Niebles らの研究と Cinbis らの研究は人間動作認識のみに目指すが、我々の提案手法はどの様な動作にも適用可能である。人間動作以外に “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 [14]．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。

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 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 [15]。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。視覚特徴を利用しない他の関連研究として Dong らの研究がある [16]。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。Liu ら [8] は Web ビデオとそのタグの相互関係を利用し Web ビデオトピック検出手法を提案する。本研究は、彼らの研究を元に着想に至ったもので、Web ビデオとタグの相互関係を利用する点は共通している。ただし、彼らはビデオのタグ情報のみを活用する。一方で、我々はタグだけではなく、ビデオの視覚特徴も利用する。

3. 提 案 手 法

本研究では、タグが付けられた Web 動画から自動的に特定動作の対応ビデオショットを抽出するシステム ([6], [9]) の改良として、VisualTextualRank を提案する。システムの全体はセクション 3.1 で述べられる。VisualTextualRank はセクション 3.2 にて詳細に説明される。

3.1 ビデオショット抽出システムのあらまし

図 1 は提案するビデオショット自動抽出システムの大まかな流れを示す。提案システムは次の 4 ステップからなる：

（1） 動画選択および動画とタグの関連スコアの計算

（2） ショット分割およびショット類似度の計算

（3） 画像選択およびショットと画像の類似度の計算 (オプション)

（4） ショットランキング：VisualTextualRank

最初に、第一ステップでは、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。ここでは，Web 動画共有サイトによっ

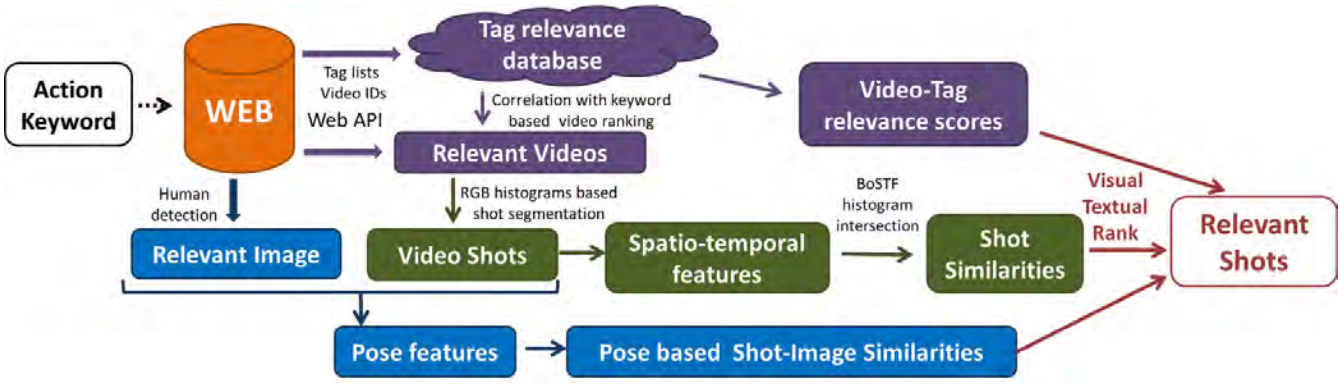

図 1　Web 動画から自動的に特定動作の関連ショットを抽出するシステムのあらまし。我々によって初めに提案されたシステム [6] に新しいショットランキング手法、VisualTextualRank、を導入する。VisualTextualRank はビデオとタグのテキスト類似度と共に、ビデオショットの間のコンテンツ類似度を利用する。

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する。タグ共起スコアに基づくビデオランキングを行って、ランク上位のビデオのみをダウンロードする。

次に、第 2 ステップでは、ダウンロードしたビデオに対しカラー情報によってショット分割を行い、各ショットから視覚特徴を抽出する。このステップでは視覚特徴として Noguchi らの提案した時空間特徴 [17] を利用する。ショットを時空間特徴の BoF として表現し、ショットの類似度行列を計算する。本論文で適用されるタグ共起スコア計算法、ショット分割法、視覚特徴抽出法および類似度行列計算法は我々の以前の論文 [6] で詳細に説明される。

3 番目のステップはオプションである。ここで、Web 画像および人間ポーズという視覚情報が導入される。そのアイデアとしては動作は環境と実現者によって見た目で異なる場合もあるが、動作を行なう基本的なポーズがあり、動作とポーズは対応関係がある。動作を映す画像にある人間のポーズの含むビデオショットは関連ショットである可能性が高いということである。このステップの導入に関しては我々の以前の研究 [9] を参考にしてほしい。

最後に、本研究で提案する VisualTextualRank を適用して、ビデオショットランキングを行い、上位で関連ショットを得る。

3.2　VisualTextualRank

VisualTextualRank の基本アイデアとしては重要な動画に付けられるタグは重要であり、重要なタグのビデオのショットも重要で、かつ視覚的に類似するというものである。このアイデアに基づいて、VTR では、各ランキングステップにおいて、ショットのランクスコアはショットの相互視覚類似度および対応タグとの関連性によってアップデートされる。またタグは、ビデオショットとの関連性によって洗練される。図 2 は VTR のアイデアを示す。

VTR は Liu らの研究 [8] のアイデアを継承して VisualRank [7] を拡張する。[8] では、ビデオとタグもその相互関係に基づいて特定トピックとの関連性の順で共ランキングされる。ただし、[8] ではビデオの全体が評価され、ビデオの視覚特徴がまったく考慮されていない。一方で、本研究はビデオの各シーン (ビデオショット) を評価し、コンテンツに基づく特徴も利用する。また、VisualRank [7] は画像の視覚連結のみを使用し、テキスト情報 (メタデータ) は活用しない。提案 VTR は Web ビデオの視覚特徴とテキスト特徴の両方を利用する。ビデオショットの間の視覚関連およびビデオとタグの相互関係の分析結果をタグランキングに適用することによって、単に出現頻度が多いタグは上位にランクすることはない。よって、ノイズタグの影響は小さくなり、ショットランキングにより多くのよいタグが利用される。

提案ランキング手法 VTR は次で示す式に従ってビデオショットとタグを反復的に共ランキングする：

$$RS_k = \alpha \times SM^* \times SC^* \times RT_k + (1-\alpha)\, p \qquad (1)$$

$$RT_{k+1} = (SC')^* \times RS_k \qquad (2)$$

RS と RT はそれぞれショットとタグのランクスコアを表すベクトルである。ショット数は n_s とタグ数は n_t とすれば、RS は $n_s \times 1$ 次元のベクトル、RT は $n_t \times 1$ 次元のベクトルになる。SM はショット類似度行列であり、$SM[i][j]$ はショット i とショット j の視覚類似スコアを表す。SM^* は SM の列正規化されたものを示し、$n_s \times n_s$ 行列である。SC はショット-タグ類似度行列であり、$SC[i][k]$ はショット i のビデオとタグ k のテキスト関連スコアを表す。SC^* は SC の列正規化されたものを示し、$n_s \times n_t$ 行列である。SC' は SC の転置行列であり、タグ-ショット類似度行列を示す。VTR では、ショットの ST 特徴のヒストグラムインターセクションを視覚類似度として利用される。ビデオとタグのテキスト関連スコアはビデオとアクションキーワードの関連スコアと同じように計算される。ここで注意して欲しいのは、テキスト特徴 (タグ共起) は視覚特徴より信頼性が低いと思われているのでビデオショットが先にランキングされてから、タグはアップデートされたビデオショットによって洗練されるということにした。

図 2　VisualTextualRank のアイデアの説明図。ここで示すアクションは “catch+fish” である。青リンクはビデオショットとタグのテキスト的関連、紫リンクはビデオショットの間の視覚的関連を示す。弱リンクは表示されない。最初は (図の左のほう) 赤 BB が付いているショットのみは重要ショットだとわかったが、そのショットは強く関連するタグに自分の投票を投げる。そして次のランキングステップでは、そのタグはまた強く関連するショットに自分の投票を投げる。最後は (図の右のほう) で示すような評価結果が得られる。このように VTR はショットとタグの両方のランクスコアを反復的にアップデートする。

RT は初期で均一ベクトルとして定義される。各ランキングステップにおいて、式 1 でビデオショットのランクスコアは視覚関連およびタグとの関連性によってアップデートされてから、式 2 でビデオショットはタグに自分のスコアを分配する。重要ショットは強く関連するタグに大きなスコアを分配するのでそのタグも重要になる。そして次のランキングステップでは、そのタグはまた自分の関連深いのショットのランクスコアを高くする。このように徐々に重要なショットとタグのスコアは上位に行くことになる。

VisualRank のように、我々も補正値 α と補正ベクトル p をショットランキングに導入する。α はランキング結果にあまり影響を与えない ([15]) ので [6] に従って α は 0.85 とする。p の設定は以下の 2 つの方法を試す。

$$p_i^{(1)} = [1/n] \tag{3}$$

$$p_i^{(2)} = \frac{\exp(\mathrm{SI}(S_i))}{\sum_{j=1}^{n} \exp(\mathrm{SI}(S_j))} \tag{4}$$

式 3 で示す均一補正ベクトルはオプションステップ (第 3 ステップ) を適用しないときに使う。一方で、式 4 で示すバイアス補正ベクトルは Web 画像を導入するときに使う。この場合は画像に映す人間ポーズの含むビデオショットはバイアスされる。画像を導入した補正ベクトル計算に関しては我々の査読中の論文(注1)で詳しく説明される。

4.　実験とその結果

まず、VisualTextualRank および Web 画像導入の有効性を検証するために前に提案したフレームワーク [6] によって関連ショットがうまく検出できていない 20 人間動作カテゴリを対象に次の二つの実験を行った。

(注1)：Do Hang Nga and Keiji Yanai. Automatic Extraction of Relevant Video Shots of Specific Actions Exploiting Web Data. Computer Vision and Image Understanding.

（1）　Exp.1: [6] のフレームワークに VisualTextualRank のみを適用

（2）　Exp.2: [6] のフレームワークに Web 画像+ VisualTextualRank を導入

Exp.1 は視覚特徴として時空間特徴のみを利用し、補正ベクトルは式 3 で示すものを使用する。Exp.2 はポーズ特徴も考慮し、式 4 の補正ベクトルを利用する。実験したカテゴリは精度が 30%以下の 45 動作種類からランダムに選択された。その中は、[6] によって 7 カテゴリは 20%〜30%、10 カテゴリは 10%〜20%、残りは 10%以下の精度が得られた。ここで [6] と同じように、精度とは上位ランク 100 ショットの関連ショットの割合を示すものである。実験 Exp.1 と Exp.2 の結果は表 1 で表す。

表 1 からわかるように、VTR はほとんどの実験したカテゴリの精度を改善できた。特に、“blow+candle”, “jump+rope”, “catch+fish”, “play+guitar”, “wash+dishes”, “drive+car”, “slap+face”, “squat”, “serve+tennis”, “tie+necktie” は精度が 15%以上改良できた。図 3 で VTR のいくつの成功例を表す。

また表 1 で示すように、人間ポーズ特徴導入の有効性は、カテゴリによって違うことがわかる。この特徴を利用することによって、“serve+tennis” や “row+dumbbell” などのカテゴリの結果は改良されたが、“blow+candle”、“eat+sushi” や “drive+car” などのカテゴリの精度は下がった。その理由としては、“serve+tennis” や “row+dumbbell” のように、ポーズ特徴は人の全体が写され、オクルージョンが少ない場合に有効であるが、“blow+candle” や “eat+sushi” の場合、上半身のみが写され、テーブルなどでよく抽出できない、ということである (図 4 を参考)。

また、面白い結果として、VTR は VisualRank と比べてランキング結果の多様性が改良されるということである。VisualRank は視覚特徴のみを使用するため、ランク上位では見た目で類似する画像のみが出てくる。[6] のように、ショットランキ

表 1　Results of 20 human action categories compared between baseline and our proposed method. All of these categories have precision lower than 30% by using baseline method. VTR means VTR computed with a uniform bias vector(3) while Pose-VTR refers to VTR which uses pose features and damping vector(4).

Action	[6]	Exp.1	Exp.2
blow+candle	29	44	35
climb+tree	24	24	24
eat+sushi	12	23	15
jump+rope	26	49	47
catch+fish	28	59	54
read+book	19	21	20
boil+egg	9	11	14
grill+fish	5	13	19
play+guitar	28	41	43
wash+clothes	15	29	31
wash+dishes	15	39	39
drive+car	28	40	34
slap+face	20	45	44
squat	19	35	36
serve+tennis	11	27	30
cook+rice	6	11	15
comb+hair	14	26	27
roll+makizushi	24	36	32
row+dumbbell	16	30	33
tie+necktie	11	28	27
Average	17.9	31.5	30.9

ングに VisualRank を適用することによって、上位で視覚類似のショットのみが見られる場合がある。一方で、VTR は視覚情報だけではなく、ビデオとタグのテキスト情報も利用するので関連ショットと強いテキストリンクを持つショットも上位に行く可能性がある (図 5)。

提案 VTR の有効性をもっと検証するために、以前に提案したフレームワーク [6] で 30%より高い精度が得られたカテゴリにも実験を行った。上の結果から画像適用なしのフレームワークのほうは平均精度が高いことがわかったのでこの実験では VTR のみを適用したフレームワーク (Exp.1) を利用する。この実験での 10 カテゴリは精度が 30%より高いカテゴリからランダムに選択されたものである。その結果は表 2 で表す。結果例は図 6 で示す。

表 2 で示す結果からわかるように、“harvest+rice” の以外はすべての実験したカテゴリの結果はかなり改良できた。特に、“play+trumpet”、“ski”、“dance+hiphop”、“swim+crawl” の 4 種類の精度はよく改善された。平均精度は約 8%改善できた。したがって、ほとんどのカテゴリに対し、VTR を適用することによって、より多くの関連ショットが得られたことがわかる。

5. 結　　論

本論文では、VisualRank を拡張した新しいランキングメソッド、VisualTextualRank、を提案した。VisualTextualRank はビデオショットの視覚リンクおよびビデオとタグのテキストリンクを利用してビデオショットとタグを同時にランキングする手法である。特定動作の関連ショットを自動抽出するフレームワークに VisualTextualRank を適用して VisualTextualRank の有効性を検証できた。

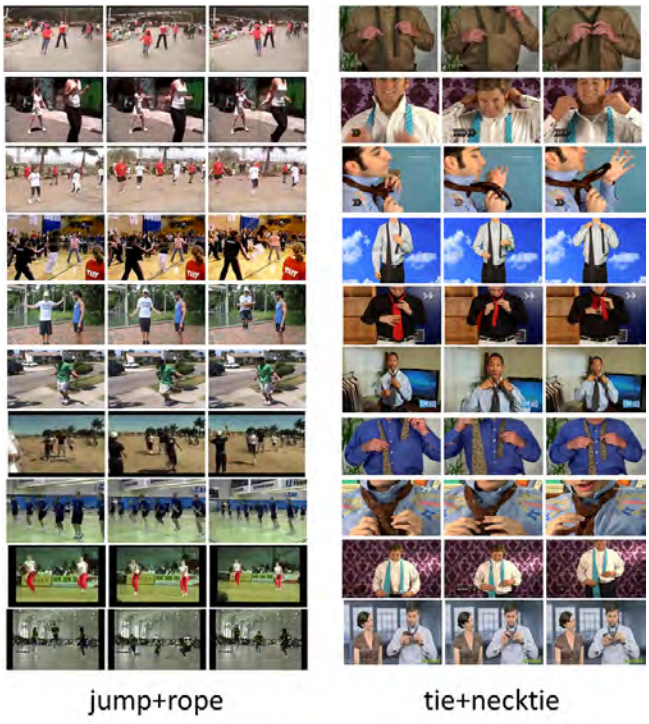

図 3　VTR を適用することによって検出できた 10 関連ショット。[6] のフレームワークは上のショットを検出できなかった。ここで示すカテゴリは [6] によって 30%以下の精度を得たものである。

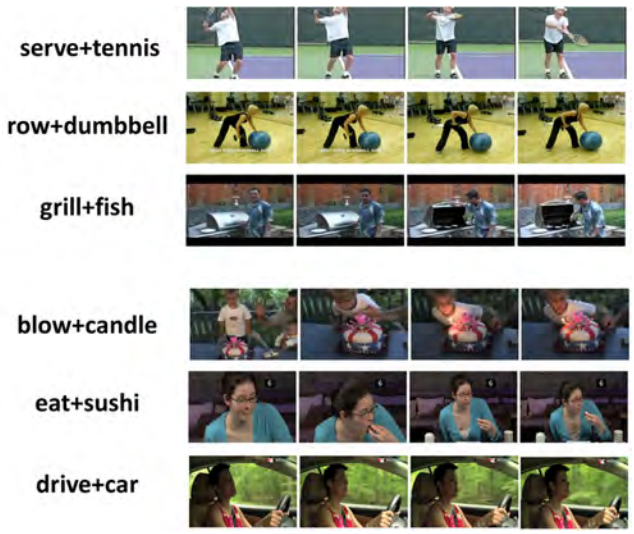

図 4　ポーズ特徴の有効性。上のカテゴリ:　“serve+tennis”、“row+dumbbell” (全身可視) と “grill+fish” (上半身がよく見える) はポーズ特徴が有効な場合である。下のカテゴリ: ‘blow+candle”、“eat+sushi”、“drive+car” (上半身のみ可視、オクルージョンあり) はポーズ特徴を適用することによって精度が下がる場合である。この場合、VTR のみ (Exp.1) の方はよい精度が得られる。

図 5　ベースライン手法 [6] と提案手法による結果の多様性の比較例。この図は “play+guitar” の上位 10 ショットを示す。ベースライン手法によって上位で videoID が “6P–1elQwRE” のビデオのショットのみが得られた。一方で、VTR は視覚情報だけではなくテキスト情報も利用するので異なるビデオからのショットを検出できる。

表 2　VTR の有効性の検証のための追加実験。実験したカテゴリは VTR 適用なしのフレームワーク [6] によって 30%より高い精度が得られたものである。

Action	[6]	VTR
harvest+rice	49	46
play+trumpet	41	59
ski	49	60
dance+hiphop	43	68
play+drum	40	45
shave+mustache	31	30
dance+flamenco	45	53
clean+floor	31	38
pick+lock	30	28
swim+crawl	36	49
Average	39.5	47.6

図 6　VTR によって検出された関連ショットのなかの 5 ショット。ここで実験したカテゴリは VTR 適用なしのフレームワーク [6] よりも提案手法によって 30%高い精度が得られたカテゴリ。

文　献

[1] J. Liu, J. Luo, and M. Shah. Recognizing realistic action from videos. In *CVPR*, 2009.

[2] H. A. Jhuang, H. A. Garrote, E. A. Poggio, T. A. Serre, and T. . Hmdb: A large video database for human motion recognition. In *ICCV*, 2011.

[3] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *ICCV*, 2009.

[4] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *ICPR*, pp. 32–36, 2004.

[5] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, 2005.

[6] D. H. Nga and K. Yanai. Automatic construction of an action video shot database using web videos. In *ICCV*, 2011.

[7] Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *PAMI*, Vol. 30, No. 11, pp. 1870–1890, 2008.

[8] L. Liu, L. Sun, Y. Rui, Y. Shi, and S. Yang. Web video topic discovery and tracking via bipartite graph reinforcement model. In *WWW*, pp. 1009–1018, 2008.

[9] D. H. Nga and K. Yanai. Automatic collection of web video shots corresponding to specific actions using web images. In *CVPR Workshop on Large-Scale Video Search and Mining*, 2012.

[10] S. Jones, L. Shao, J. Zhang, and Y. Liu. Relevance feedback for real world human action retrieval. *Pattern Recognition Letters*, 2012.

[11] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*, 2006.

[12] J. Niebles, B. Han, A. Ferencz, and L. Fei-Fei. Extracting moving people from internet videos. In *ECCV*, pp. 527–540, 2008.

[13] N. I. Cinbins, R. G. Cinbins, and S. Sclaroff. Learning actions from the web. In *ICCV*, pp. 995–1002, 2009.

[14] L. Ballan, M. Bertini, A. D. Bimbo, M. Meoni, and G. Serra. Tag suggestion and localization in user-generated videos based on social knowledge. In *Proc. ACM MM WS on Social Media*, pp. 3–7, 2010.

[15] Q. Yang, X. Chen, and G. Wang. Web 2.0 dictionary. In *CIVR*, pp. 591–600, 2008.

[16] D. Liu, X.S. Hua, L. Yang, M. Wang, and H.J. Zhang. Tag ranking. In *WWW*, pp. 351–360, 2009.

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