Vol.2012-CVIM-181 No.6
2012/3/15

Web 動画・画像を用いた特定動作ショットの自動収集

Do Hang Nga[†1]　樋 爪 和 也[†1]　柳 井 啓 司[†1]

Web 上にはテキストタグ付きの動画が大量に存在している．しかしながら，タグが付与されているのは通常は動画全体に対してであり，動画のどの部分が最も重要なシーンであるかはテキストタグだけでは知ることはできない．そこで，我々は以前，“walking”,“eating” などのような動作キーワードを与えるだけで，大量のタグ付きの Web 動画から特定動作の対応ビデオショットを教師データなしで自動的に抽出する手法を提案した．しかしながら，動作キーワードによっては精度がほぼ０％になってしまうなど，必ずしもどのような動作に関してもうまくいくという訳ではなかった．そこで，本発表では，Web 画像検索エンジンから得られた静止画像を利用することによって，教師なし動画ショット検出の精度を向上させる方法について述べる．具体的には PageRank に基づく教師なしショットランキングを行う際に，Web 画像検索エンジンから得られた上位画像にさらにフィルタリング処理を施した画像を疑似教師データとして利用することによって，精度が向上することを示す．

1. はじめに

Web 上には膨大な動画が存在し，日々その量が急速に増加している．特に，一般のユーザが誰でも気軽に動画像を公開できる Youtube などの動画共有サイトでは，極めて多くの動画が登録されている．実際，Youtube は，2012 年 1 月現在で毎秒 1 時間分の動画像がアップロードされていると公表されている．その動画像の多くには，アップロードしたユーザが付けた検索のための単語である「タグ」が付与されている．タグは，一般に，動画の内容と関連したキーワードであり，動画を試聴するユーザが動画を検索する際の大きな手がかりとなる．しかしながら，タグが付与されているのは通常は動画全体に対してであり，動画のどの部分がタグに対応するのは未知である．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．

†1 電気通信大学 大学院 情報理工学研究科 総合情報学専攻

人手でタグを付与する場合，静止画の場合は一目見れば付与すべきタグは容易に想像ができるが，動画の場合は動画像を試聴する必要があり，特に特定の場面だけにタグを付与するといったような時間軸を考慮したタグの付与は極めて時間のかかる作業である．そのため，そうした時間情報を持ったタグが付与されている Web 上の動画像は極めてまれである．

そこで，我々はキーワードを与えるだけで学習データなしで，大量のタグ付きの 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研究している．これまでに，タグに基づくランキングと，動画の特徴量に基づくランキングを組み合わせた手法を提案し，100 種類のキーワードに関する実験によって手法の有効性を示している[1),2)]．この提案手法では，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し，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てショット分割してから，最後にグラフに基づくランキングメソッド VisualRank[3)] 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．

従来より，動画から与えられたキーワードもしくはタグに対応する動画中のショットを検索する研究はビデオ検索の基本的なタスクとして長年研究が取り組まれている．代表的な例としては，国際的なビデオ検索ワークショップである TRECVID が挙げられる．TRECVID では 1 万本以上の Web 動画について，与えられた 300 種類以上の単語について対応するショットを検索する Semantic Indexing タスクが設定されている．しかしながら，こうした従来の手法では，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．実際，TRECVID も教師データを用いることが前提となっており，世界中の参加者が一ヶ月以上の期間をかけて，与えられた学習動画データに対して，検索対象となる 300 種類以上の単語のラベル付けを行っており，膨大な労力が学習データセットの構築のために費やされている．

一方，我々の研究では，主に “eating” や “running” などの人間動作に関するキーワードに対応する動画ショットを，教師データなしに自動的に検出する方法の実現を目標としている．教師データを用いないことによって，事前に学習データを用意することが不要になり，どのようなキーワードに対しても対応が可能となる．

もし人間動作のキーワードに対応する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 50 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[4]．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高精度で可能となれば，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．

本発表では，1)，2) での提案手法の改良手法を提案し，実験結果を報告する．本研究における改良点は，Web 動画に加えて Web 動画を Web 画像検索エンジンを用いて収集し，フィルタリング，再ランキングをした後，その上位の画像をビデオショットランキングにおける擬似的な学習データとして利用する点である．

2. 関 連 研 究

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[5] 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[6]．

本研究に最も近い研究は Cinbis らの研究である[7]．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，6) のビデオデータセットに対し動作認識を行った[7]．この研究は本研究と類似しているが，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．一方，本研究では，Web 画像と静的特徴に加えて，Web 動画の類似性も考慮するため時空間特徴も使う．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[8]) 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，我々の提案手法は人間のポーズを認識するための Poselet 特徴を人間検出器として用いるために，より幅広い人間動作に適用可能であるという特徴がある．

3. 手法の概要

最初に本発表の元となる手法である 1)，2) について概要を述べる．この手法では，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．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 の赤枠で囲った部分以外が対応する部分となる．

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[9]，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[3] 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．

本発表で行った拡張は，図 1 の赤枠で囲った部分である．Web 画像の収集，Poselets[10] による人物画像のフィルタリング，VisualRank[3] による再ランキングからなる．これらの処理の結果の上位の 10 枚〜20 枚程度をビデオランキングする際の擬似的な学習データとして利用した．具体的には，ビデオショットランキングで VisualRank を用いる際のバイアスベクトルの設定を，Web 画像から自動選択した 10〜20 枚の画像とビデオショットの視覚的な類似度に基づいて行うことによって，ビデオショットランキングの精度向上を図っている．

4. 手法の詳細

ここでは，1)，2) で提案したタグに基づくビデオランキング，視覚特徴によるビデオショットランキングの詳細，および本発表で新たに試みた Web 画像の利用について説明する．

4.1 従 来 手 法

4.1.1 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

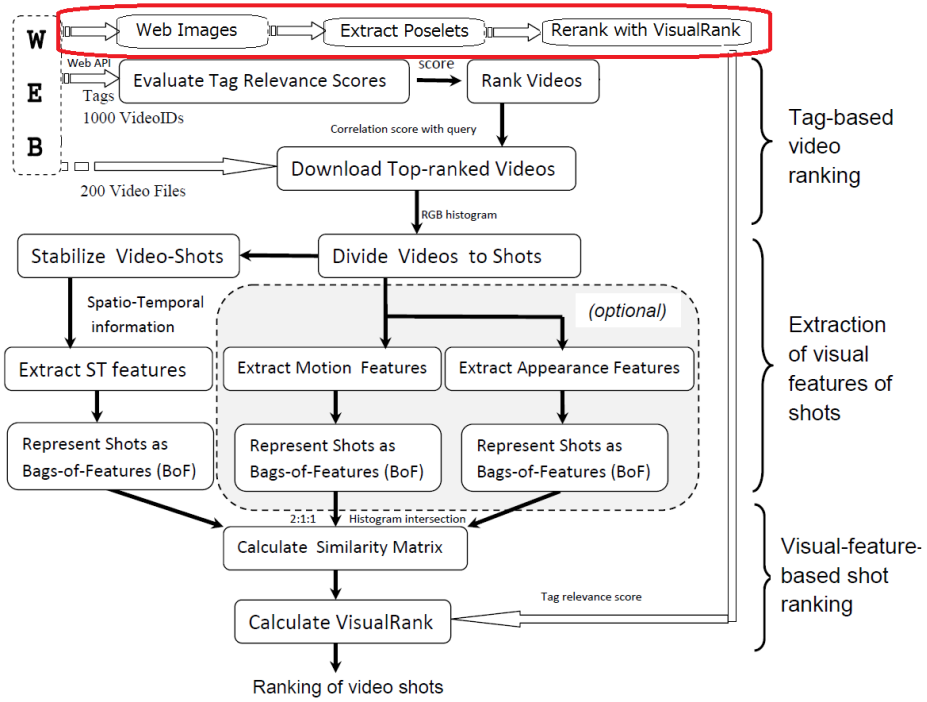

図 1 提案手法の概要

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[11]．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[11] 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 $(n \leq m)$ 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．

4.1.2 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．

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

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

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視覚特徴量のである．動画ショットからの視覚特徴量としては，Noguchi らの提案した時空間特徴を利用する[9]．各ショットから時空間特徴量を抽出し，事前に求めておいたコードブックを用いて 5000 次元の Bag-of-Features 表現に変換して，ショット毎の視覚特徴ベクトルとして用いる．

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[3] 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．

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 (7) は VisualRank 計算の公式を示している．

$$\boldsymbol{r} = \alpha S \boldsymbol{r} + (1-\alpha)\boldsymbol{p} \quad (0 \leq \alpha \leq 1) \tag{7}$$

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．

$$p_i^{(1)} = \begin{cases} 1/k & (i \leq k) \\ 0 & (i > k) \end{cases} \tag{8}$$

$$p_i^{(2)} = \begin{cases} Sc(V_i)/C & (i \leq k) \\ 0 & (i > k) \end{cases} \tag{9}$$

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

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 (8) 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 (9) 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 (8) 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[3]．

4.2 提案する改良手法

本発表で提案する改良点は，式 (8) および式 (9) に代わる新たな補正ベクトルの求め方である．動作キーワードに対応する Web 画像を収集し，ノイズ除去を行った後，10〜20 枚程度の画像を擬似的な学習画像として選択し，それらの学習画像と各ショットのキーフレーム画像との類似度を求め，類似度に基づいて補正ベクトルを設定する．

具体的には，以下の手順で擬似学習画像を Web から自動収集する．

(1) 動作キーワードに対応する画像を Web 画像検索エンジンの WebAPI を用いて 300 枚程度収集．
(2) 人物検出器である Poselets detector[10] を用いて人物が写っている画像のみを選択．
(3) VisualRank[3] を画像に適用して，上位 10〜20 枚の画像を擬似的な学習画像として選出．類似度行列の計算には VisualRank[3] と同様に局所特徴点のマッチング個数を用いる．局所特徴量としては SURF を用いる．

なお，本研究では「人間の動作キーワード」に対応するショットの抽出を目的としているので，原則的に人間がショットに写っている方が抽出するショットとしては望ましいため，擬似学習画像の選択に Poselets を用いることとする．実際には，動画ショットに対して直接 Poselets を適用することも可能であるが，処理時間の問題と Poselets detector の recall の低さから多くのショットが除外されてしまう可能性があるため，今回の発表では擬似学習画像のフィルタリングのみに利用する．

次に，選ばれた擬似学習画像と各ショットのキーフレーム画像との類似度を算出する．類似度は，VisualRank 計算時と同様に SURF 特徴量のマッチングの個数によって行った．ショット i の特徴点のマッチング個数を C_i とすると，補正ベクトル (バイアスベクトル) は次の式 (10) で求めることができる．なお，実験では $k = 100$ とした．

$$p_i^{(3)} = \begin{cases} C_i/C_{all} & (i \leq k) \\ 0 & (i > k) \end{cases} \tag{10}$$

$$\text{where } C_{all} = \sum_{j=1}^{k} C_j$$

5. 実験結果

実験は 1)，2) で実験した 100 種類のうち，上位 100 ショットの適合率の結果があまり思わしくない動作キーワードについて行う．まず，参考までに，今まで得られている 100 種類の動作キーワードに対する結果を表 1 に示す．

まず，予備実験として，“brush+teeth”, “iron+clothes”, “jog”, “jump+rope”, “read+book” の 5 種類について，擬似学習画像を手動で 10 枚選択し，提案手法による補正ベクトルの設定法の有効性を検証した．なお，評価は 1)，2) と同様に上位 100 ショットの適合率で行う．結果を表 5 に示す．5 種類中 4 種類で適合率が向上していることが分かる．“read+book” に関しては逆に精度が低下してしまった．

次に，擬似学習画像の自動抽出による結果であるが，原稿提出時点では実験中であるため，研究会の当日にご報告する予定である．

6. まとめ

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に対する改良案を示し，実験を行った．手動画像選択による予備実験の結果では，十分であるとは言えないが精度の向上が確認できた．

今後の課題としては，さらなる VisualRank の補正ベクトルの設定方法の詳細な検討と，“airplane-flying” や”car-runnig” のような人間動作以外にも有効である手法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

表 1 100 動作の上位 100 ショットの適合率 (%)

soccer+dribble	100
fold+origami	96
crochet+hat	95
arrange+flower	94
paint+picture	88
boxing	86
comb+hair	83
parachute+jump	82
do+exercise	79
do+aerobics	78
do+yoga	77
surf+wave	75
serve+volleyball	75
shoot+arrow	73
fix+tire	67
blow-dry+hair	64
basketball+dribble	64
ride+bicycle	62
curl+bicep	58
shoot+ball	58
bowl+ball	58
tie+shoelace	57
laugh	50
play+drum	49
ski	49
harvest+rice	49
dive+sea	49
twist+crunch	47
dance+flamenco	45
dance+hiphop	43
dance+tango	41
play+trumpet	41
skate	37
AVG (1-33)	65.7

swim+crawl	36
cut+hair	35
paint+wall	33
lunge	32
hit+golfball	32
fieldhockey+dribble	32
shave+mustache	31
chat+friend	31
pick+lock	30
play+guitar	28
plant+flower	28
catch+fish	28
serve+tennis	27
lift+weight	27
row+dumbell	26
hang+wallpaper	26
jump+trampoline	24
sew+button	24
roll+makizushi	24
ride+horse	24
fry+tempura	23
row+boat	20
massage+leg	20
play+piano	19
drive+car	17
wash+dishes	15
wash+clothes	15
draw+eyebrows	15
sing	12
squat	12
raise+leg	12
cry	12
eat+sushi	11
swim+backstroke	9
AVG (34-67)	23.2

swim+breaststroke	9
climb+tree	9
clean+floor	8
tie+tie	8
jump+rope	8
swim+butterfly	7
brush+teeth	7
boil+egg	7
cook+rice	6
iron+clothes	6
bake+bread	6
slap+face	5
grill+fish	5
smile	4
weep	2
run	2
kiss	2
blow+candle	2
cut+onion	1
wash+face	0
read+book	0
knit+sweater	0
watch+television	0
walk	0
slice+apple	0
plaster+wall	0
pick+apple	0
peel+grape	0
jog	0
head+ball	0
drink+medicine	0
drink+coffee	0
count+money	0
AVG (68-100)	3.2
AVG (1-100)	30.6

表 2 手動選択画像により設定した補正ベクトルを用いた場合の上位 100 位の適合率 (%).

動作キーワード	従来結果	画像利用あり
brush+teeth	28	34
iron+clothes	46	51
jog	5	11
jump+rope	22	28
read+book	19	16

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．

参 考 文 献

1) Nga, D.H. and Yanai, K.: Automatic Construction of an Action Video Shot Database using Web Videos, *Proc. of IEEE International Conference on Computer Vision* (2011).
2) Nga, D.，柳井啓司：大量の Web 動画からの教師なし特定動作ショット抽出，画像の認識・理解シンポジウム (MIRU) (2011).
3) Jing, Y. and Baluja, S.: VisualRank: Applying PageRank to Large-Scale Image Search, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.30, No.11, pp.1870–1890 (2008).
4) Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. and Serre, T.: HMDB: A Large Video Database for Human Motion Recognition, *Proc. of IEEE International Conference on Computer Vision* (2011).
5) Niebles, J., Wang, H. and Fei-Fei, L.: Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words, *Proc. of British Machine Vision Conference* (2006).
6) Niebles, J., Han, B., Ferencz, A. and Fei-Fei, L.: Extracting moving people from internet videos, *Proc. of European Conference on Computer Vision*, pp.527–540 (2008).
7) Cinbins, N. I.-., Cinbins, R.G. and Sclaroff, S.: Learning Action From the Web, *Proc. of IEEE International Conference on Computer Vision*, pp.995–1002 (2009).
8) Dalal, N. and Triggs, B.: Histograms of oriented gradients for human detection, *Proc. of IEEE Computer Vision and Pattern Recognition*, Vol.1, pp.886–893 (2005).
9) Noguchi, A. and Yanai, K.: 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).
10) Bourdev, L. and Malik, J.: Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations, *Proc. of IEEE International Conference on Computer Vision* (2009).
11) Yang, Q., Chen, X. and Wang, G.: Web 2.0 Dictionary, *Proc. of ACM International Conference on Image and Video Retrieval*, pp.591–600 (2008).