Web 動画からの教師なし特定動作ショット抽出における Web 画像の利用

DoHang Nga† 柳井 啓司†

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

あらまし 本研究は、タグ付きの Web 動画からの特定動作の対応ビデオショットの自動抽出フレームワークに Web 画像の導入を提案する。特定動作に対し、まずタグ共起によるビデオランキングを行い、動作と最も関連する 200 本の Web ビデオを収集する。収集したビデオをショット分割してからすべてのショットを bag-of-spatio-temporal-features として表現し、VisualRank メソッドを適用してランキングする。提案アプローチはショットランキングのステップに Web 画像の利用を導入することによって、より多くの対応ショットが得られる。Web 画像は直接 Web から収集したものであるが、人間動作の場合、Poselet 手法 [1] を利用して人間が検出された画像のみを利用するとする。前に提案したフレームワーク [2] によって精度が 20%以下の 28 人間動作カテゴリおよび 15%以下の 8 非人間動作カテゴリに Web 画像を用いた提案フレームワークを適用する。実験では、平均で人間動作カテゴリは 6%、非人間動作は 16% 改善できるという結果が得られた。

キーワード タグ付き Web 動画, Web 画像, Poselet, 教師なしの動作学習, VisualRank 手法

1. はじめに

動作認識分野においては KTH [3] や Weizmann [4] などのような小さく、制限ありの標準データセットにおける認識精度がほぼ完璧になってきたため、制限なしの大規模なデータセットが必要だと考えられている。最近構成された最も大きなデータセットは 51 種類の動作カテゴリ [5] から構成されているが、依然として、動作認識のための学習データベースの構築は、静止画と異なり、非常に時間と手間のかかる作業であると認識されている。その理由は、データセットを構築するためには、映画や Web ビデオのような動画データから特定動作の関連部分を手動で検索し、ラベル付けする必要があるためである。もし、Web 上の動画ソースのような簡単に取得可能なソース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れば、動作データベースの構築は容易になるであろう。なお、ここでビデオショットとは一つのシーンと対応する連続フレームセットとする。

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手法を提案した。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、タグリストは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によって関連ビデオを選択できず、最終的に対応ショットがあまり得られない場合があることが実験から分かった。

そこで、本研究では、以前のフレームワークを改良し、ショットランキングステップに動作の Web 静止画像を利用することを提案する。この手法は、動作の Web 画像とよくマッチングしたビデオショットは関連ショットの可能性が高いので、よりランキングを高くする、という考えに基づいている。最近の研究 [6]〜[9] によって、1 枚だけの画像からなる静止画像に対する動作認識が可能であることが示されている。我々は与えられた動作キーワードと関連した画像を Web から自動的に収集して、その画像とビデオショットの類似度を特徴マッチングで評価する。本研究では、新たに Web 画像を利用することになるが、Web 画像の収集もキーワードのみで行うため、キーワードのみで対応ショットを収集するというフレームワークの自動性はそのまま保持されることになる。

提案手法では、まず、タグ共起スコアの上位 200 ビデオを YouTube からダウンロードしてから、そのビデオをショットに分割し、ショットを bag-of-spatio-temporal-features として表現する 。同時に、Bing API を用いて、数百の動作キーワードの Bing 画像を収集し、その画像に Poselets [1] を適用して人間検出を行なう。次に、ビデオショットおよび人間が検出された画像から SURF 特

徴 [10] を抽出し、特徴マッチングで各ショットと画像集の類似度を計算する。ここで人間検出による画像選択は人間動作のみに適用し、非人間動作の場合は直接 Bing 画像を利用することに注意してほしい。最後に、グラフに基づくランキングメソッド VisualRank [11] を適用してビデオショットの視覚特徴と画像との類似度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。

以前のアプローチ [2] で精度が 20%以下の 28 人間動作カテゴリおよび精度が 15%以下の 8 非人間動作カテゴリに対し提案手法を適用した。我々の改良によって平均で精度が人間動作は 6%、非人間動作は 16% 向上できた。特に、精度が 10%であった以下の 8 人間動作カテゴリは 5.7%から 21.6%に改善できた。というのは、ノイズタグが圧倒的多数で、動作との関連性の少ない動画が多く収集されてしまった場合でも、提案手法によってそのビデオからかなりの対応ショットが抽出できるということである。

本論文では、第 2 章でで関連研究を述べる。第 3 章で提案アプローチの大まかな流れを表す。第 4 章では手法の詳細を説明する。第 5 章は実験結果について説明し、最後に本論文の結論を述べる。

2. 関 連 研 究

ここでは制限無しのビデオデータセットにおける動作認識の研究について説明する。

最近、YouTube ビデオセット [12], [13] と Kodak ビデオセット [14] のような制限の少ないデータセット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、その研究のほとんど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自動的に検出することが目的である。

教師なしの動作学習法として、Niebles らの研究 [15] 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。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、[16] のビデオデータセットに対し動作認識を行った [17]。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。一方で、本研究は Web 動画も利用し、動作を表現するために時空間特徴を使う。さらに、彼らの手法は人間動作だけ対応しているのに対して、我々は人間動作に限らず非人間動作を含むすべての動作を扱うのが目標にしている。

3. 提案手法の概要

本研究で提案されるアプローチは Web 動画を用いた特定動作のビデオショットデータベースの自動構築フレームワーク [2] に基づいて開発される。前のフレームワークでは、タグ付きの Web ビデオからの特定動作の対応ビデオショットの自動抽出手法が提案された。その手法によって、タグと動画内容のセマンティックギャップが極めて大きい動作カテゴリに対し対応ショットがあまり抽出できないことがわかった。提案される Web 画像の導入によってノイズタグが圧倒的の場合でもより多くの関連ショットが得られる。

3.1 Web 動画を用いた特定動作のショットデータベースの自動構築フレームワーク

前に提案したフレームワークでは、“walking” や “surfing+wave” のような動作キーワードが与えられた時、次の 3 ステップによって動作の対応ショットをタグ付きの Web 動画から自動的に抽出される：(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。

最初に、1000 YouTube ビデオに対しタグ共起統計によるビデオランキングを行い、ランク上位の 200 ビデオをダウンロードする。ここでは、YouTube API を用い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のでタグビデオランキングには 1000 ビデオのダウンロードが必要なく、実際にダウンロードするのはランキング後の上位 200 ビデオである。次に動画ショット分割およびショットの視覚特徴の抽出が行われる。視覚特徴として Noguchi らの提案時空間特徴 [18] が利用される。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 [11] 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。

多くアクションカテゴリに対しこのアプローチは高い精度が得られたが、非関連のタグが多く付けられているカテゴリの場合、ビデオランキングステップで得られるビデオのほとんどは動作を含まないことになって、最終に動作の対応ショットがあまり取得できない。

3.2 提案改良：Web 画像の導入

本研究は、Web 画像をビデオショットランキングステップに導入する。この改良によってノイズタグが圧倒的の場合でもより多くの関連ショットが得られるようになる。提案フレームワークの流れは (1) タグに基づくビデオ選択；ショット分割；ショットの時空間特徴の抽出；

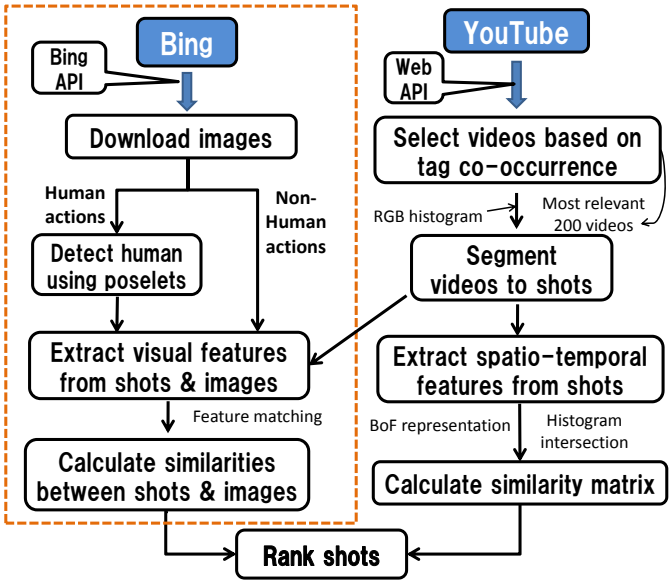

図 1　提案フレームワークの大まかな流れ。赤いボックスに含まれる左の部分は我々の改良

bag-of-spation-temporal-features 表現化；ショットの類似度行列の計算、(2)Web 画像収集、（人間動作の場合）Poselets [1] による人間検出を用いた画像選択 (3) ショットと画像の SURF 特徴の抽出；ショットと画像の特徴マッチングによる類似度の計算、(4) ショット同士の類似度およびショット-画像の類似度に基づくショットランキングとなっている (図 1)。

それぞれの動作に対し、まず、前のアプローチに従って、タグ共起によるビデオ選択によって 200 ビデオをダウンロード、RGB 特徴ヒストグラムを用いたショット分割、すべてのショットに対し時空間特徴抽出および bag-of-features 化を行ない、時空間特徴ヒストグラムによるショット同士の類似度行列を計算する。次のステップからは本研究の提案の改良点であり、図 1 の左に含まれる赤ドットボックスで示す。動作の関連画像の自動選択およびその画像による対応ショット選択は次で説明する。まずは Microsoft によって提供される Bing API という Web 画像検索 API を用いて数百の動作の画像をダウンロードする。次は Poselets [1] を適用してその画像に対し人間検出を行なう。人間が検出された画像のみは次のステップに進む。選択画像の適切な数は実験セクションで論議される。次のステップにおいては、選択した画像とショットから視覚特徴を抽出し、特徴マッチングによってショットと画像の類似度を計算する。これで視覚特徴として SURF [10] を利用する。最後に最初のステップで得られたショット同士の類似度行列およびショット-画像の類似度による補正ベクトルを用いて VisualRank 手法 [11] でショットランキングを行なう。

4. 使 用 手 法

ここでは、Poselets に基づいた人間検出による Web 画像選択；特徴マッチングによるショット-画像の類似度計算；および画像によるショットランキングについて詳細に説明する。

図 2　Poselets による人間検出後の最初の 6 画像

5. Web 画像選択

Web 画像検索エンジンにおいてユーザはクエリアクションキーワードを与えると何千の画像が得られるが、検索結果の上位でもユーザの返してほしいものではない場合がある。キーワードの意味のバリエーションと共に動作の多様性によって関連画像の検索は困難になる、特に人間動作の場合である。よって、検索結果の上位のままを使わずに、ここで検索結果から関連画像の選択を行なうとする。一方で、提案フレームワークの自動性を保持するため、Web 画像選択ステップも自動で実現するのは目標である。次の二つの仮説を前提とする。（1）検索結果画像には少なくともクエリアクションの対応画像が存在する（2）人間動作画像には人間やボディパートが含まれる上記の仮説に基づいて、Web 画像に Poselets による人間検出 [1] を適用して、人間が検出された画像のみを選択する。Poselets は 3D 人間ポーズアノテーションで学習された有効なボディパート検出器だと検証された。我々は Poselets の著者の公式に提供された Poselets 検出ツール 1 を利用する。パラメータはデフォルトパラメータとして設定する。図 2 は Poselets に基づいた人間検出を適用して選択した画像の例を示している。

ここで、画像は何枚選択すればよいという質問が出てくる。対応ショット抽出の改良に利用する画像の数を N とすれば、N は大きければ大きいほど多くの対応ショットが得られる？実際に N に複数の値：10、20、30、50 を与えて実験を行い、N による提案手法のパフォーマンスの変化の検討を行なう。その結果は実験セクションで表す。

5.1 ショット-画像の類似度計算

与えられたビデオショットと画像セットの類似度を評価するため、まず、すべての画像およびショットの 5 連続フレームごとの 1 フレームから SURF 特徴を抽出する。次は、特徴ベクトルのユークリッド距離の計算によってそれぞれのフレームと画像のマッチングポイントを数える。ショット S_i の特徴抽出されたフレームの数は M、画

像セット $\mathcal{I}$ は N 画像があるとすれば、S_i と $\mathcal{I}$ の類似度 $\mathrm{SI}(S_i)$ は以下の式のように計算する。

$$\mathrm{SI}(S_i) = \sum_{k=1}^{N} \max_{j=1} \mathrm{SI}(F_j|I_k), \quad (1)$$

$$\text{where } S_i(F|I_k) = \frac{2 * \mathrm{MatchPoint}(F_j, I_k)}{(\mathrm{Point}(F_j) + \mathrm{Point}(I_k))}, \quad (2)$$

$\mathrm{MatchPoint}(F_j, I_k)$、$\mathrm{Point}(F_j)$ と $\mathrm{Point}(I_k)$ はそれぞれフレーム F_j と画像 I_k のマッチポイントの数、F_j の特徴の数と I_k の特徴の数を示している。

5.2 画像を導入したショットランキング

ショットランキングには前のアプローチ [2], [19] と同様に、VisualRank 手法 [11] を適用する。ただし、[2] ではタグ共起スコアの大きいビデオからのショットのほうは上位にランキングされる可能性が高いことに対して、今回提案フレームワークにおいては選択した Web 画像セットとの類似度が高いショットにバイアスをかけるとする。ノイズタグが多い場合、タグに基づくショットランキングは有効ではないと考えているからである。

式 (3) は VisualRank 計算の公式を示している。

$$r = \alpha S r + (1 - \alpha) p \quad (0 \leqq \alpha \leqq 1) \quad (3)$$

ここでは S が正規化された類似度行列、p は補正ベクトル、r はランキングベクトルを指すものである。S は時空間特徴ヒストグラムによるショットの類似度行列として計算される [2]。α は 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 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。 [2] では、補正ベクトルは次のように定義された。

$$p_i^{(1)} = \begin{cases} 1/k & (i \leqq k) \\ 0 & (i > k) \end{cases} \quad (4)$$

$$p_i^{(2)} = \begin{cases} Sc(V_i)/C & (i \leqq k) \\ 0 & (i > k) \end{cases} \quad (5)$$

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

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 (4) 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。式 (5) 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。両方の補正ベクトルの定義し方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いうアイデアに基づく。タグ情報による信頼性が高い場合は、そのように補正ベクトルを設定すれば関連ショットを上位にランキングできるが、タグはあまりノイズが含まれたらタグスコアも有用ではないと考えられる。

本研究では、次のように補正値を定義し、動作画像とよく類似するショットを強調すると提案する。

$$p_i^{(3)} = \frac{\exp(\gamma \mathrm{SI}(S_i))}{\sum_{j=1}^{n} \exp(\gamma\ hrmSI(S_j))} \quad (7)$$

ここで γ は定数である。実験では、γ は $\log 3$ と設定する。

6. 実　　験

[2] によって精度が 20%以下の 28 人間動作カテゴリおよび 15%以下の 8 非人間動作カテゴリに Web 画像を用いた提案フレームワークを適用する。また、[2] のデータセットおよび評価法を利用する。よってここで精度はランク上位 100 ショットの適合率 (Precision@100) を示している。Web 画像の数 N は 10, 20, 30 と 50 の値を試す。N=0 は [2] と対応する。図 3 と図 4 はそれぞれ人間動作と非人間動作の結果を表す。各結果はすべてのカテゴリの平均とする。全カテゴリの結果は表 1(人間動作) と表 2(非人間動作) で示す。

結果からわかるように、人間動作と非人間動作の両方は N=20 または N=30 画像を利用したほうは最も高い精度が得られる。このとき、パフォーマンスは平均で人間動作は 6%、非人間動作は 16%向上できた。さらに、表 1 と表 2 をみてわかるように、10 人間動作カテゴリ：“bake+bread”, “jog”, “squat”, “swim+breaststroke”, “serve+volleyball”, “smile”, “cook+rice”, “grill+fish”, “swim+butterfly”, “tie+tie” および 6 非人間動作カテゴリ：“falling+leaves”, “snow+falling”, “typhoon”, “airplane+flying”, “earthquake”, “waterfall”. に対し提案手法は特に有効である。

ただし、“slap+face” や “wash+clothes” のように Web 画像の導入により精度が低下した場合もある。それを理解するために、次の 2 つの理由が考えられる。(1) 人間検出に基づいた画像選択によって関連画像があまり得られない (2) ショット・画像集の類似度の計算法は有効ではない一番目の場合は “slap+face” のようなカテゴリを含む。図 5 は “slap+face” の画像選択の最初結果を表す。ほとんどの選択された画像は動作と対応しないことがわかった。よってこれらの画像のショットランキングに適用によって対応ショットがあまり得られないことになる。二番目の場合は “wash+clothes” のようなカテゴリと対応する。この場合、画像選択ステップで関連画像がかなり選択でき (図 6)、ショットランキングでこれらの画像を利用してもより多くの対応ショットを抽出できること

になっていない。その理由として、動作の多様性のためビデオショットで現される動作 (図 6) と画像で現される動作 (図 7) は大きく異なるので画像との類似度に基づくショットランキングは効率が悪いからだと考えられる。“wash+clothes” の場合、選択した画像はほとんど「室外で手を使って服を洗濯する」についてであることに対し、ほとんどのダウンロードしたビデオは「室内で洗濯機を使って服を洗濯する」シーンを含む。

画像数によるショットランキングのパフォーマンスの検討実験の結果 (図 3 と図 4) からわかるように、画像数が 10 から 20 多くなると共に精度は上がってくるが、画像数が 30 から 50 増えると精度は低下する傾向がある。したがって、利用画像数が大きければ大きいほど対応ショットが多く抽出できるとは限らない。我々の理解では、ここで学習ソースとして Web 画像を使用するが、一般的には、Web 画像の検索結果の上位のほうはクエリキーワードと関連が深いので上位画像のみ利用したほうがよいと考えられる。また、この理解を検証するために、選択した画像に対する関連した画像の割合で画像選択の精度を評価してその結果は図 8(人間動作) と図 9（非人間動作）で示す。この結果によって N=10,20,30 の際、関連画像が多く選択できるが、N=50 の場合は関連画像の割合が低いことがわかった。

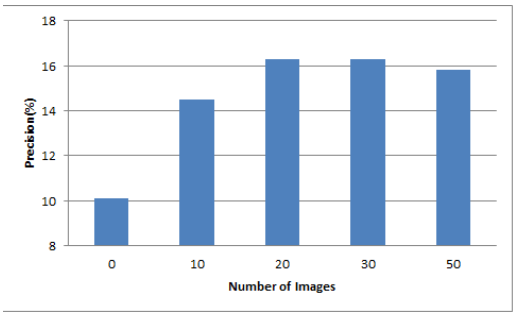

図 3　28 人間動作カテゴリの 100 ランク上位ショットの平均精度

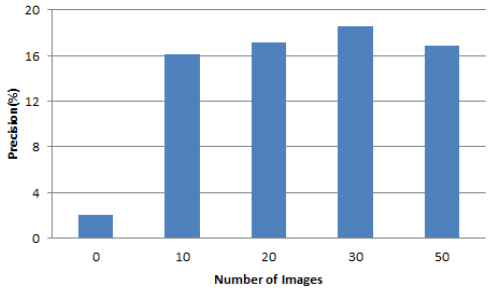

図 4　8 非人間動作カテゴリの 100 ランク上位ショットの平均精度

7. 結　　論

本研究は Web 動画からの特定動作と対応した動画ショットの自動抽出システムに Web 画像を導入すると提案した。人間動作の場合、Poselets [1] の適用により人間を検出し、人間やボディパートが検出された画像を利用する。非人間動作の場合、単に Web 画像検索結果の上位画像を利用する。実験結果からわかるように、選択した画像をショットランキングに適用することによってより多くの対応ショットが得られる。

今後の課題としては Web 画像選択の改良や結果分析による有効なキーワードの選択などだと考えられる。よりキーワードと関連する画像は収集可能であれば “slap+face” 動作のような場合の結果を改善できると思っている。また、提案システムの抽出結果により、“wash+clothes” のよう

表 1　選択した画像の数による 28 人間動作カテゴリの結果の変化

動作	[2]	N=10	N=20	N=30	N=50
slap+face	20	16	14	13	17
read+book	19	21	22	23	24
squat	19	34	38	32	35
row+dumbbell	16	23	23	24	22
wash+clothes	15	9	10	10	9
wash+dishes	15	23	21	25	24
comb+hair	14	12	15	12	23
drink+coffee	14	8	10	9	16
swim+breaststroke	13	23	27	31	24
cry	12	4	6	5	4
eat+sushi	12	13	13	11	10
serve+tennis	11	14	18	15	14
tie+tie	11	18	17	23	30
boil+egg	9	4	8	6	6
head+ball	9	5	9	7	6
swim+backstroke	9	10	12	14	12
take+medicine	8	5	8	7	6
serve+volleyball	7	20	24	31	23
swim+butterfly	7	29	33	31	36
bake+bread	6	18	19	18	14
cook+rice	6	15	16	15	13
grill+fish	5	21	23	26	17
jog	5	15	19	21	20
pick+apple	5	8	10	9	10
slice+apple	5	2	4	2	6
bowl+ball	4	18	17	15	5
smile	4	16	17	18	15
kiss	2	2	4	3	2
平均	10.1	14.5	16.3	16.3	15.8

表 2　選択した画像の数による 8 非人間動作カテゴリの結果の変化

動作	[2]	N=10	N=20	N=30	N=50
explosion	0	4	5	5	1
falling+leaves	3	12	14	16	9
snow+falling	0	18	21	22	24
typhoon	4	21	25	29	24
airplane+flying	2	29	30	32	27
earthquake	7	26	24	25	23
heavy+rain	0	4	3	3	4
waterfall	0	15	15	17	15
平均	2	16.1	17.1	18.6	15.9

図 5　Poselets による “slap+face” 動作の対応 Web 画像の選択結果の上位 18 画像。結果のほとんどは “slap+face” と関連しないことがわかる。

図 6　Poselets による “wash+clothes” 動作の対応 Web 画像の選択結果の上位 18 画像。結果のほとんどは “室外で手を使って服を洗濯する” シーンである。

図 7　提案システムによる “wash+clothes” 動作の対応ショットの自動抽出の結果例。すべては “室外で手を使って服を洗濯する” 動作を表すことがわかる。

な動作カテゴリは検索の際、キーワードに道具などを加えて、例えば “wash+clothes+by+hand” や “wash+clothes+in+washing+machine” などとすればより効果が高い検索結果が得られると考えられる。

文　献

[1] L. Bourdev and J. Malik: “Poselets: Body part detectors trained using 3d human pose annotations”, ICCV (2009).

[2] D. H. Nga and K. Yanai: “Automatic construction of an action video shot database using web videos”, ICCV (2011).

[3] C. Schuldt, I. Laptev and B. Caputo: “Recognizing human actions: A local SVM approach”, ICPR, pp. 32–36 (2004).

[4] M. Blank, L. Gorelick, E. Shechtman, M. Irani and R. Basri: “Actions as space-time shapes”, ICCV (2005).

[5] H. A. Jhuang, H. A. Garrote, E. A. Poggio, T. A. Serre and T. : “Hmdb: A large video database for human motion recognition”, ICCV (2011).

[6] K. Schindler and L. van Gool: “Action snippets: How many frames does human action recognition require?”, CVPR (2008).

[7] W. Yang, Y. Wang and G. Mori: “Recognizing human actions from still images with latent poses”, CVPR

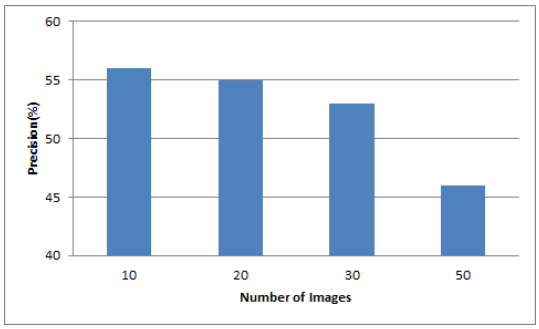

図 8　28 人間動作の関連画像の選択結果の上位 N の平均精度

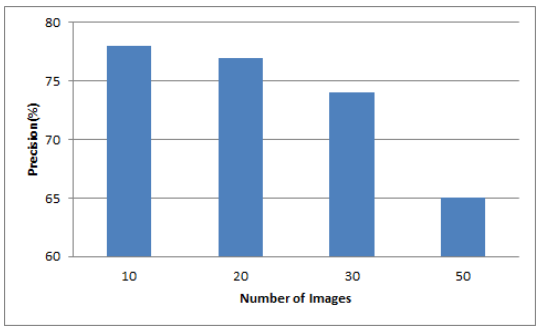

図 9　8 非人間動作の関連画像の選択結果の上位 N の平均精度

(2010).

[8] C. Thurau and V. Hlavac: “Pose primitive based human action recognition in videos or still images”, CVPR (2008).

[9] D. Weinland and E. Boyer: “Action recognition using exemplar-based embedding”, CVPR (2008).

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

[11] Y. Jing and S. Baluja: “Visualrank: Applying pagerank to large-scale image search”, PAMI, **30**, 11, pp. 1870–1890 (2008).

[12] Z. Wang, M. Zhao, Y. Song, S. Kumar and B. Li: “YouTubeCat: Learning to categorize wild web videos”, CVPR (2010).

[13] G. Toderici, H. Aradhye, M. Pasca, L. Sbaiz and J. Yagnik: “Finding meaning on YouTube: Tag recommendation and category discovery”, CVPR, pp. 3447–3454 (2010).

[14] L. Duan, D. Xu, I. W. Tsang and J. Luo: “Visual event recognition in videos by learning from web data”, CVPR (2010).

[15] J. Niebles, H. Wang and L. Fei-Fei: “Unsupervised learning of human action categories using spatial-temporal words”, BMVC (2006).

[16] J. Niebles, B. Han, A. Ferencz and L. Fei-Fei: “Extracting moving people from internet videos”, ECCV, pp. 527–540 (2008).

[17] N. I. Cinbins, R. G. Cinbins and S. Sclaroff: “Learning actions from the web”, ICCV, pp. 995–1002 (2009).

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

[19] D. Nga, 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 (MIRU2011) (2011).