TRECVID Semantic Indexing Task と Multimedia Event Detection Task への取り組み

○樋爪和也†, 柳井啓司‡
Kazuya HIZUME† and Keiji YANAI‡

†: 電気通信大学 大学院情報理工学研究科 総合情報学専攻, hizume-k@mm.cs.uec.ac.jp
‡: 電気通信大学 大学院情報理工学研究科 総合情報学専攻, yanai@cs.uec.ac.jp

概要: 本稿では米国国立標準技術局(NIST)主催の動画像における競争型ワークショップである TRECVID および Semantic indexing task、Mutimeda event detection task の概要について述べ、それらに対し我々が行った手法について解説する。200 コア以上の並列計算を用いて動画から抽出した特徴量を BoF 表現に変換し、Multiple Kernel Learning-SVM で学習・分類を行った。

大規模動画像認識、Bag-of-Features、Multiple Kernel Learning

1. はじめに

近年、動画像が持つ情報にパターン認識の技術を適用したコンテンツベースの映像検索手法が盛んに研究されている。そうした研究を促進するためにTRECVIDと呼ばれる国際的な評価ワークショップが存在し、世界中の多くの映像検索技術に関する研究チームが参加している。本稿では、このTRECVID で行われているタスクの中から我々がここ数年参加している 2 つのタスクに注目し、その概要と我々の手法について述べる。与えられた学習およびテストデータは膨大であり、それらを処理するために 200 コア以上の並列計算で延べ 6000 時間以上の大規模な動画像処理を行う必要があった。本稿ではこうした大規模映像データ処理についても述べる。

2. TRECVID

TRECVID（TREC Video Retrieval Evaluation）[1]とは、米国国立標準技術研究所（NIST）と Disruptive Technology Office（DTO）が主催する映像検索技術に関連する研究を促進する競争型ワークショップである。毎年、各国の研究グループが参加し、TRECVID が設定した共通のタスクおよび評価基準に対して各々の手法で研究に取り組む。共通のタスクに対して結果の優劣を競争させ、結果の比較検討および成果の共有を行うことで、映像検索技術の研究を促進させることが目的である。

図 1 SIN タスクのカテゴリ例

2012 年度は 6 つのタスクが設定されており、その中から我々が例年参加している Semantic indexing（SIN）と Multimedia event detection

(MED)の 2 つのタスクについて説明する。

2.1. Semantic indexing タスク

シーンごとに分割されたショットデータの中から特定の概念を含んだショットを検出するタスクである。対象となる概念は大まかに以下の 3 種類である。

- 物体、人(例 Airplane、Bus、Male_Person)
- 動き (例 Dancing、Ski、Walking)
- 風景 (例 Beach、Indoor、Mountain)

図 1 に示す通り、短く分割されたショット中では対象となる概念がはっきりと写っている。ただしデータに対するラベル付けは参加者が共同で行っているため、対象となる概念がショット中にどの程度出現するかは不定である。認識対象となるカテゴリ数とデータは年々増加しており、TRECVID2012 でのカテゴリ数は 346、学習データは計 600 時間、テストデータは計 200 時間である。データは全て MPEG-4/H.264 形式のインターネットアーカイブ映像であり、1 ショットあたり 10 秒から 3.5 分に制限されている。また、各々が用意した外部データを学習データとして扱うことも可能である。

評価形式はテストデータをランキング付けし、カテゴリごとに上位 2000 ショットを対象としてランキングを考慮した平均適合率である infAP[2]が計算される。提出方法として、すべてのカテゴリを認識する full、指定された 50 カテゴリのみの light がある。ただし評価結果が返却されるのは一部のカテゴリのみである。2012 年からは新たに Beach + Mountain のような 2 種類の認識対象を組み合わせる pair が追加された。Pair では 10 種類の組み合わせが課題として与えられる.

2.2. Multimedia event detection タスク

Birthday party、Parade、Rock climbing のような複雑なイベント(図2)を対象に、テストデータが各イベントを含むかどうかを判定する。ここでのイベントとは SIN タスクのように特定の動作ではなく、Repairing an appliance のような複数の動作から成るものや、Winning a race without a vehicle のような状態を指す。また SIN タスクとは異なり、動画はショット分割されていない。TRECVID2012 でのイベント数は 20、締め切り直前に検出イベントが与えられる追加のアドホックイベント数が 5 である。学習データは Linguistic Data Consortium が収集したユーザが撮影した映像であり、MPEG-4/H.264 形式で 1350 時間以上、テストデータは 4000 時間以上である。

評価形式には Normalized Detection Cost(NDC)が用いられている。まず各チームがイベントの有無を判定するための Threshold を独自に設定し、各テストデータのスコアを決定する。この結果から PMD、PFA が計算され、NDC は以下の式によって求められる。NDC は値が小さいほど良い結果であると言える。また Cost と P_{Target} の値は主催側で決定される。

$$\mathrm{NDC} = \frac{Cost_{MD} * P_{MD} * P_{Target} + Cost_{FA} * P_{FA} * (1 - P_{Target})}{MINIMUM(COST_{MD} * P_{Target}, Cost_{FA} * (1 - P_{Target}))}$$

P_{MD} = (Threshold を上回らなかった正解データ数)/(正解データ数)

P_{FA} = (Threshold を上回った非正解データ数)/(非正解データ数)

図 2 MED タスクのカテゴリ例

3. 関連研究

TRECVID2011 では SIN タスクにおいて東京工業大学と Canon [3]の合同チームが日本から参加チームとしては初の 1 位となった。彼らはショット中の全フレームから SIFT、HoG を、ショット全体から

MFCC 音響特徴といった特徴量を抽出し、これらに混合ガウス分布（GMMs）を適用している。このとき、木構造 GMMs を構築することでシステムの高速化を図っており、さらに正規化した平均ベクトルを連結した GMM Supervector を作成する。分類には SVM を使用し、Mean infAP=0.173 を記録した。

MED タスクではアメリカの BBN VISER[4]チームが 1 位となった。彼らは四つの局所特徴、三つの色特徴、二つの動き特徴、三つの音響特徴を抽出し、Audio-Visual Bi-Modal Words として視覚特徴と音響特徴の相関モデルを作成した。さらに物体検出や音声認識、テキスト OCR なども採用しており、最終的に 18 種類のシステムを Weighted Average Fusion により連結することで最高位を記録している。

4. SIN タスク認識手法

SIN タスクで扱う動画は既にショット分割が行われており、各概念はショット中において多くの時間で出現している。そのため、ショット全体から特徴量を抽出し、ショット単位で認識を行うことが可能である。本研究では画像単位で得られる局所特徴と動画全体から得られる動き特徴を使用し、パターン認識における基本的なアルゴリズムを用いて動画像認識を行う。

TRECVID2012 の両タスクで使用したシステム全体の流れは同じであるため、まず SIN タスクで使用した手法について述べる。図 3 に処理の流れを示す。図 3 の点線枠で囲まれた部分は MED タスクで行う追加処理であり、SIN タスクでは行わない。

4.1 特徴抽出

4.1.1 画像特徴

フレーム画像から抽出する特徴量として、SURF[5]と色特徴を使用する。

SURF(Speeded-Up Robust Feature)は、特徴点とその周辺の局所特徴を記述した特徴量である。画像の照明変化、スケール変化、回転に対して頑健であり、128 次元の特徴ベクトルで表現される。

色特徴は画像の RGB 画素値を使用する。SURF、および RGB 色特徴は全フレームより抽出する。

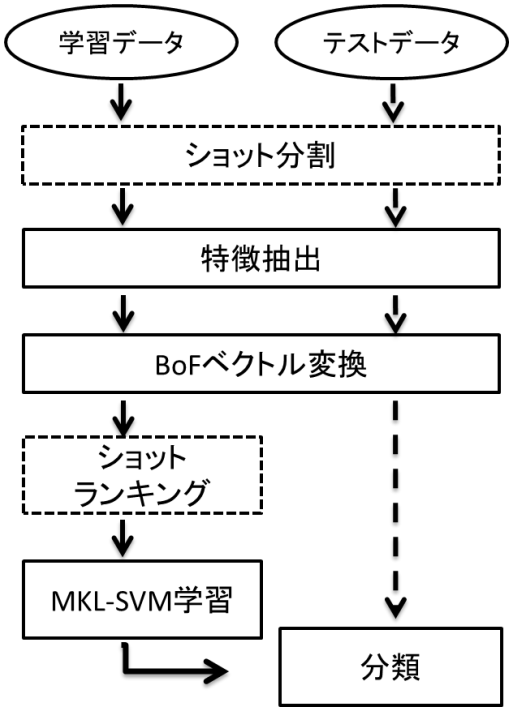

図 3 処理の流れ、点線枠は MED のシステムのみ

4.1.2 時空間特徴

動画像全体から抽出する時空間特徴として、野口が提案した時空間特徴抽出手法[6]を用いる。

まず、動画に含まれるズームやパンなどのカメラモーションの検出を行う。Web 動画には手振れのような意図しないカメラモーションが多く含まれており、さらに解像度が低い。これを解決するために、カメラモーションを検出した場合、その特徴を破棄する。検出手法には Lucas-Kanade アルゴリズムを用いる。まず、動画に含まれるズームやパンなどのカメラモーションの検出を行う。Web 動画には手振れのような意図しないカメラモーションが多く含まれており、さらに解像度が低い。これを解決するために、カメラモーションを検出した場合、その特徴を破棄する。検出手法には Lucas-Kanade アルゴリズム 9) を用いる。

次にフレーム画像から SURF を抽出する。この

SURF の点の内、時空間特徴として適しているのはフレーム間で動きのある点のみであるため、動きのなかった点は削除する。そして残った点について Delaunay 三角分割を行う。Delaunay 三角分割とは空間内の点を連結して三角形のグループを作成することで、点の特徴だけでなくその

周辺の特徴も考慮する手法である。この 3 点の SURF 記述子を視覚特徴とするため、64×3 = 192 次元で表現される。

動き特徴は、まず SURF を抽出したフレームから N フレーム先までのフレームを取得する。上記の SURF 点の動き情報はこの内の N=2 フレームから計算される。取得した N フレームを M 分割し、その区間内で Lucas-Kanade アルゴリズムによって特徴点のオプティカルフローを計算する。各区間の動き特徴はオプティカルフローの x,y 成分の正方向および負方向に動きなしを加えた 5 次元で表現される。設定は区間 N = 5、分割数 M = 5 としており、三角形の面積変化は 5 次元で表現されるため、動き特徴は 20×3 + 5 = 65 次元となる。また、動き特徴を回転に対して有効にするため、SURF 記述子で得た回転角を利用してオプティカルフローの回転も行う。

最後に視覚特徴、動き特徴を単純に結合することによって 192 + 65 = 257 次元の一つの時空間特徴とする。図 4 は処理の一連の流れを図式化したものである。

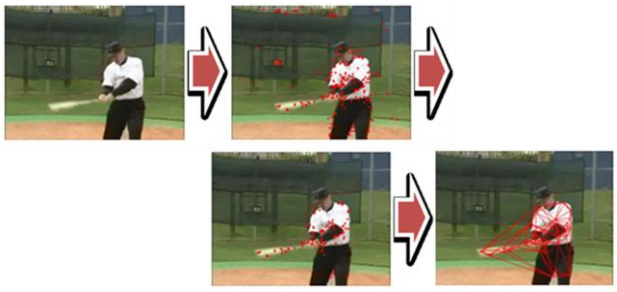

図 4 動き特徴における視覚特徴抽出の様子

4.2 Bag-of-Features

抽出した特徴量から Bag-of-Features 表現を用いて局所特徴の出現頻度の特徴ベクトルを作成する。フレーム画像から抽出した SURF および色特徴にはマルチフレーム処理を適用し、全フレームから抽出した特徴量を一つの BoF ベクトルに変換する。BoF ベクトルの生成には soft assignment[7]を利用した。これは BoF におけるコードワードへの割り当ての際、複数のコードワードに割り当てを行う。割り当てる値はコードワードとの距離の逆数を L1 正規化して使用した。

4.2.1 空間ピラミッド表現

抽出された画像特徴は空間ピラミッド表現[8]を用いて、領域毎の特徴ベクトルを作成した。空間ピラミッド表現では領域を階層的にグリッドで分割し、それぞれのグリッドに対して BoF ベクトルを作成する事で、局所特徴の空間情報も考慮した特徴ベクトルを得ることができる（図5）。時空間特徴は元々マルチフレームから抽出する特徴なので、ショット単位での BoF ベクトルを生成する。

本研究ではピラミッドレベルを 1 および 2 として、画像全体および 2×2 の領域分割を行ったヒストグラムを作成、連結した。これに伴い、コードワード数は画像特徴では 1000、時空間特徴では 5000 とした。よって最終的な BoF ベクトルの次元数はすべての特徴量において 5000 次元である。

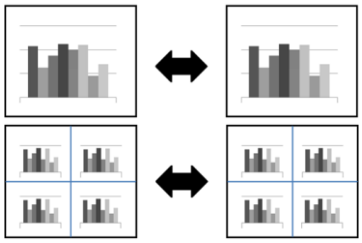

図 5 空間ピラミッドマッチング例

4.3 分類器

本研究では、分類器として Support Vector Machine（SVM）を使用する。SVM は 2 値分類問題を解くために考えられた学習アルゴリズムであり、認識性能の優れた学習モデルのひとつである。SVM は線形識別器であるが、カーネル関数を用

いることで非線形への拡張が可能である。本研究では RBF-χ^2 カーネルを用いる。

$$K(x,y) = \exp\left(-\frac{1}{2\sigma^2}\sum_i \frac{\|x_i - y_i\|^2}{x_i + y_i}\right)$$

4.3.1 MKL-SVM

Multiple Kernel Learning（MKL）は複数の SVM カーネル（サブカーネル）を式 1 のように統合することにより、新しい最適なカーネルを求める手法である。各サブカーネルをそれぞれの特徴と対応させることによって、MKL は特徴選択や特徴統合に用いることができる。

$$K(x, x') = \sum_{j=1}^{K} \beta_j k_j(x, x') \quad \beta_j \geq 0, \quad \sum_{j=1}^{K} \beta_j = 1 \quad (1)$$

SIN タスクでは、MKL-SVM の出力値をソートすることで上位 2000 のテストショットを決定する。

5. MED タスク認識手法

MED タスクのデータは対象となるイベントが動画中の全体または一部に含まれており、さらに複数の動作から成るイベントはいくつかの構成要素が含まれる。そのため、SIN タスクと同様に動画全体を使って認識しようとするとノイズとなる部分が多く含まれたり、複数の動作を一度に表現したりしてしまう。そこで、まずフレーム画像間のカラーヒストグラム差分を使って各動画をショット分割する（図 3 点線枠）。このとき各ショットの長さは最大 3000 フレームとした。元の動画は 30fps であるため、これは 100 秒分に相当する。そして SIN タスクと同様に各ショットから特徴を抽出し、BoF ベクトルを生成する。ただし、MED タスクで使用する特徴量は SURF、時空間特徴のみである。これは認識する対象がほぼ動きを中心としたものであるため、SIN タスクで使用した色特徴は除外した。

ここで、イベントを含まないショットが多く存在している。学習の際にこれらのデータをポジティブショットとして与えてしまうとノイズとなるため、ショットの選定を行う。選定には VisualRank 手法を使用し、教師なしショットランキングを行う。VisualRank については 5.1 で詳細を述べる。

MKL-SVM の学習、分類も SIN タスクと同様である。スコアはショット分割前の元の動画単位で決定されるため、元の動画に含まれるショットのうち、上位 3 つの出力値の平均をその動画のスコアとした。また、Threshold はポジティブ学習データの 2-fold cross validation で得たスコアの平均と設定した。

5.1 VisualRank

対象となるイベントを含んだショットを探すため、Web ページランキング手法の PageRank を画像に応用した VisualRank 手法[9]を適用する教師なしショットランキングを行う。VisualRank 手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計算にヒストグラムインタセクションを用いる。X は BoF ベクトルであり、本研究では時空間特徴の BoF ベクトルを使用する。

$$s(X_i, X_j) = \sum_{l=1}^{|X|} \min(x_{i,l}, x_{j,l}) \qquad (2)$$

類似度行列が得られたら、VisualRank 計算式を使ってショットのランキング付けを行う。

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

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。α は p の影響を補正するパラメータである。一般的に α は 0.8 以上の値が設定され、本研究では 0.85 と設定した。また、補正ベクトルは均一に与える。

MED タスクのシステムでは得られたランキングベクトルの上位 500 のショットをポジティブショットとして扱い、学習データに使用する。

6. 実験環境

表 1 に両タスクで使用したラベル付きデータ数と、

テストデータ数を示す。なお、SIN タスクはカテゴリによってラベル付き学習データが非常に多いため、学習の都合上、クラスごとに最大で10000ショットに制限した。そのため、配布されたラベル付きデータ数は記載した数字よりも多い。

特徴抽出や学習、分類などの計算には 80 台の計算用マシンを使用した。各マシンは種類が様々ではあるが 4 コアの CPU、8GB のメモリを搭載している。1 台あたり 3 コアを使用し、240 コアを使用した並列分散処理によって全体的な処理の高速化を図った。これにより、全行程の計算に SIN タスクでは約 4 日、MED タスクでは約 6 日かかった。

表 1 データ数

	学習データ	テストデータ
SIN タスク ショット数	344960	145634
MED タスク 動画数	4225	98118
MED タスク 分割後ショット数	48792	733764

7. 実験結果

本研究では実験として、TRECVID2012でのSINタスク、MEDタスクで得られた実験結果を記載する。

図 6 は SIN タスクの我々と他チームの最高値、中央値との比較である。今年は 346 カテゴリ中、図中の50のカテゴリが対象となった。右端は全カテゴリの平均値である。一部のカテゴリで中央値に追いついているものの、全体としては下回る結果となった。理由の一つとして、使用した特徴量の少なさがあげられる。過去に我々は顔検出や音響特徴を使用したが思うような結果を得られず、今年度は時間的な制約もあり比較的基本的な特徴量しか使用しなかった。他チームと比較すると、全体的に認識率の高いカテゴリについては我々の結果もある程度上昇しているため、現在の特徴量だけでもある程度の有効性は伺える。しかし Airplane や Baby、Kitchen や Glasses など、動きよりも画像特徴が重視されるようなカテゴリで大きく下回っている。そのため、有効な画像特徴の追加が精度向上につながると考えられる。

次に MED タスクの結果を図 7 に示す。20 クラスの Pre イベントと 5 クラスのアドホックイベントはそれぞれ個別に評価されているため、グラフを別にしている。我々のチームの結果は黒枠で囲んだ部分である。NDC は小さいほど良い結果であるため、我々のチームは残念ながら下位の方の結果に終わった。これは他チームの P_{MD} の値に比べ、P_{FA} の値が非常に大きいことが原因である。NDC の計算式において正解データを検出できなかったエラー率 P_{FA} に対するコスト $Cost_{FA}$ は非正解データを認識できなかったエラー率のコスト $Cost_{MD}$ に対して、非常に大きく設定されている。これにより、P_{MD} は他チームと大きく差があるようには見られないが、P_{FA} の差によって NDC の差が広がってしまった。これには SIN タスクでも使用した認識システムの精度も影響しているが、ショットのスコアから元の動画のスコアを決定する方法も再検討する必要がある。 イベントショットの検出のための threshold の値については、P_{FA} を下げようと値を高くしすぎれば P_{MD} が上昇してしまうトレードオフの関係のため、threshold は cross validation で設定するしかない。本研究では元の動画のスコア決定に上位 3 ショットの値を使用しているが、元の動画のショット分割数や各ショットの時間など、より元の動画を意識したスコアの決定方法が求められる。

また図 8 は 20 カテゴリに対して VisualRank 手法を適用した際の、選定画像の適合率を比較したグラフである。ベースラインとしてランダムに選んだ 100 ショット、実験で使用した VisualRank を適用し上位 500 ショットからランダムに選んだ 100 ショット、さらに VisualRank を適用した上位 100 ショットの適合率を求めた。ほぼすべてのカテゴリで VisualRank 適用前以上の適合率を得られた。しかしトップ 100 の場合は適用前よりも値が下がる結果

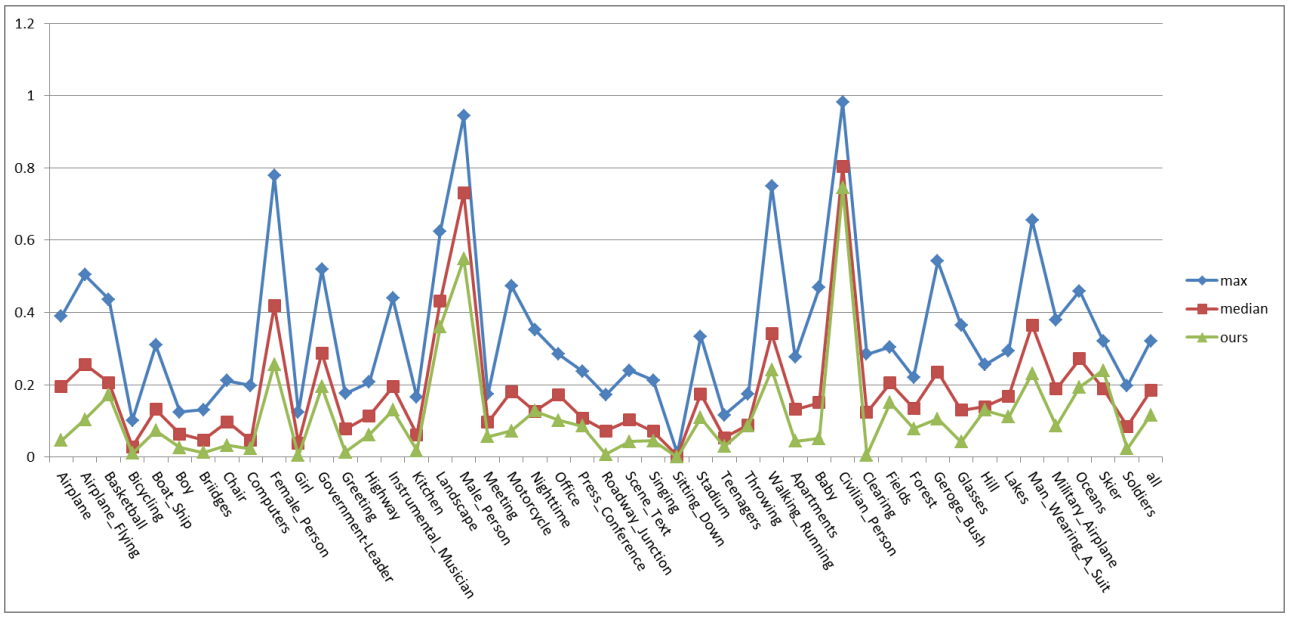

図 6　SIN タスク　他チームの最高値、中央値との比較

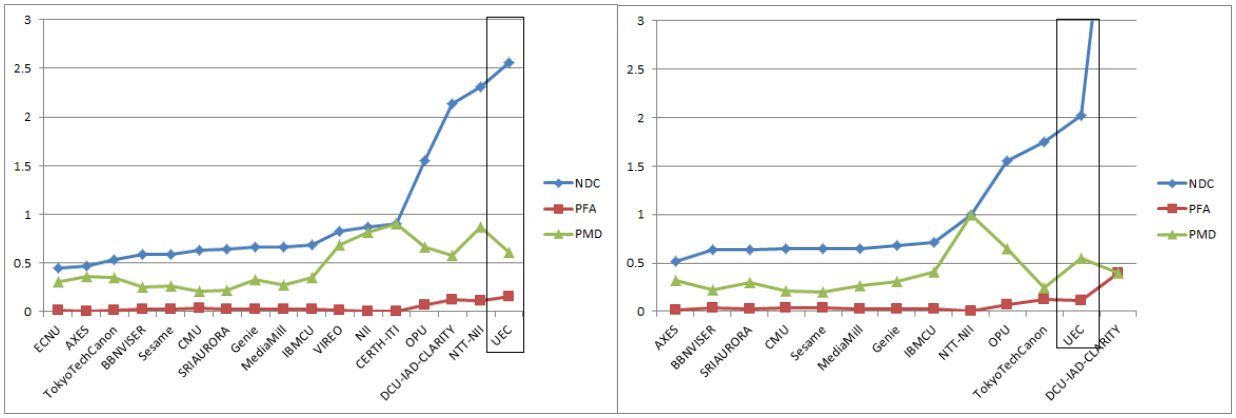

図 7　MED タスク　他チームとの比較

左：Pre イベント 20 クラス平均　右：アドホックイベント 5 クラス平均

となった。これは複雑で大きな動作の中間にあるショットが上位に入ってしまったためである。このようなショットはショットの中で動きが少なかったり、ショットの時間が非常に短かったりすることで十分な特徴量が得られず、さらに中間のショットなので比較的数が多く存在するため、上位にランクインしてしまったと考えられる。図 8、9、10 に VisualRank によって上位にソートされたショットのフレーム画像を示す。Birthday party や Parade はカテゴリに即した動画が上位に現れていると言える。しかし Making a sandwich のように人物が瞬間的に映り込んだようなショットが上位に含まれる場合もあり、認識に大きな影響が出ていると考えられる。よってこのような中間のショットを極力排除する方法を適用する必要がある。

8.　まとめ

本稿は動画像認識のワークショップである TRECVIDについて、およびSemantic Indexingタスク、Multimedia event detectionタスクへの取り組みについて述べた。SINタスクでは動画像に関連した特徴量やパターン認識における基本的なアルゴルズムを用いて、全体の中で中位程度の結果を得ることができた。MEDタスクは独自にショット分割したデータにVisualRank法を用い、SINタスクのシステムをMEDタスクに適用できるように改良した。

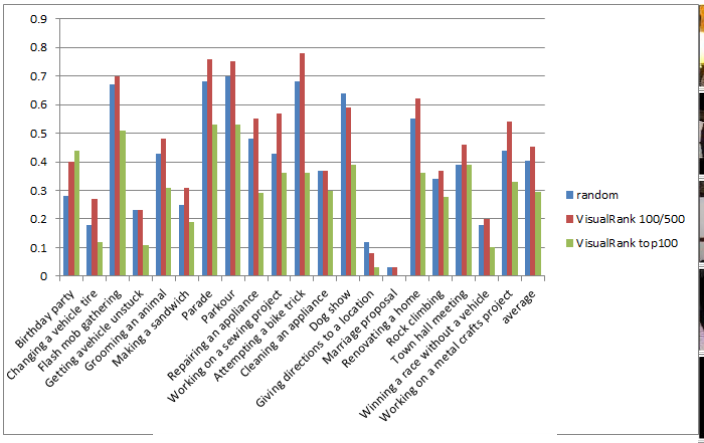

図 8　VisualRank 適用の比較

図 9　VisualRank 上位ショット

Birthday party

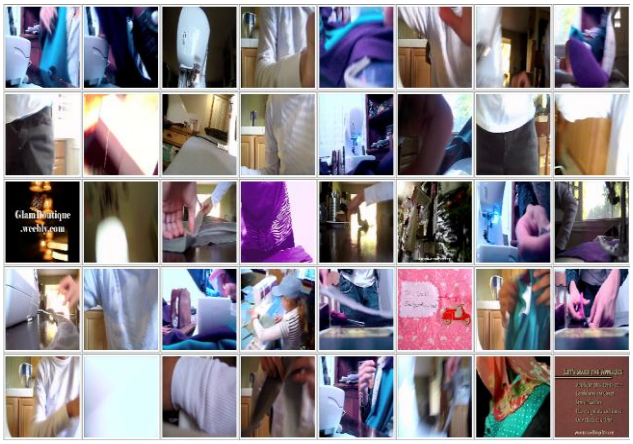

図 10　VisualRank 上位ショット

Working on a sewing project

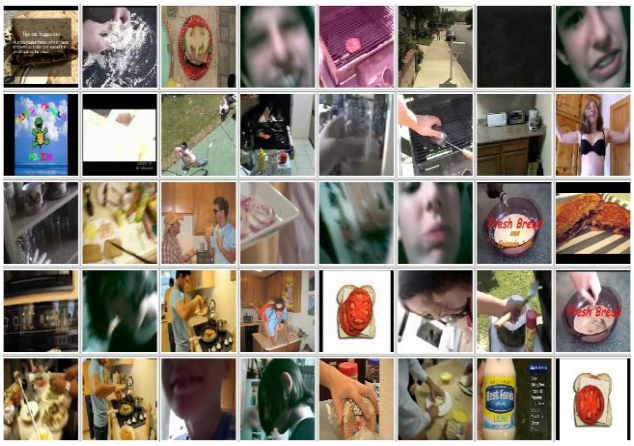

図 11　VisualRank 上位ショット

Making a sandwich

今後はシステム全体として追加すべき有効な画像特徴の再検討が必要である。Flisher ベクトルなど物体認識の分野で成功を収めている特徴量の導入を検討している. また、MED タスクにおいては追加の処理としてショット分割を行っているため、各ショットの元の動画に占める割合などを考慮し、最終的な動画のスコアの決定方法を探って行きたい。またポジティブショットを選定する際のアルゴリズムもさらに検証したい。

参考文献

[1] TRECVID Home Page. http://www-nlpir.nist.gov/projects/trecvid/.

[2] Yilmaz, E. and Kanoulas, E. and Aslam, J.A. A simple and efficient sampling method for estimating AP and NDCG. In Proc. SIGIR, pp.603-610, 2008.

[3] http://www-nlpir.nist.gov/projects/tvpubs/tv11.slides/tv11.tokyotechcanon.sin.slides.pdf

[4] http://www-nlpir.nist.gov/projects/tvpubs/tv11.papers/bbnviser.pdf

[5] H. Bay, T.Tuytelaars, and L.VanGool. SURF: Speeded up robust features. In Proc. ECCV, pp. 404-415, 2006.

[6] A. Noguchi and K. Yanai. A surf-based spatio-temporal feature for feature-fusion-based action recognition.. ECCV WS on Human Motion, 2010.

[7] . C. V. Gemert, J. mark Geusebroek, C. J. Veenman, and A. W. M. Smeulders. Kernel codebooks for scene categorization. In Proc. ECCV, pp. 696–709, 2008.

[8] S. Lazebnik, C. Schmid, and J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proc. CVPR, pp.2169-2178, 2006.

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