H-035

コンシューマビデオのジャンル分類方式に関する検討

Genre Classification Method for Consumer Generated Video

菅野　勝† Masaru Sugano　山田　徹‡ Toru Yamada　酒澤茂之† Shigeyuki Sakazawa　半谷精一郎‡ Seiichiro Hangai

1. まえがき

近年，ハンディカメラや携帯電話カメラなどが普及し，個人が容易にビデオコンテンツ（以下；コンシューマビデオ）を制作できる環境が整っている．コンシューマビデオは今後も増加することが予想され，これらを効率良くハンドリングする技術への需要が高まると考えられる．このような状況を鑑み，筆者らはコンシューマビデオの自動ジャンル分類に着目している．ジャンル分類は，例えばビデオライブラリにおいて効率的な検索を実現したり，ジャンルを事前知識として用いる自動要約などに適用したりすることができる．これまで，テレビや映画などのプロフェッショナルビデオに対するジャンル分類技術は多数提案されているが[1][2]，一般にコンシューマビデオはプロフェッショナルビデオのように制作品質が高くないことが多いため，既存手法の適用は困難であると考えられる．例えば，コンシューマビデオには意図しないカメラの手ぶれが含まれたり，編集が殆どされていないため構造の推定が難しいという制約などがある．筆者らの調査によれば，コンシューマビデオのジャンル分類はこれまでに取り組まれていない．

本稿では，コンシューマビデオの典型的なジャンルを定義した上で，各ジャンルに特徴的であり，且つ制作品質にロバストな特徴を抽出，評価することによって，ジャンル分類を実現する手法を提案する．筆者らによって撮影されたビデオや動画投稿サイトから取得したビデオを用いた実験により，その性能を評価したので報告する．

2．従来方式

前章に述べたように，これまでプロフェッショナルビデオに対する自動ジャンル分類が提案されている．例えば文献[1]では，ビデオの時間的特徴（平均ショット長，カメラワーク等）と空間的特徴（顔を含むフレーム割合，色エントロピー等）を用いて，テレビ番組を映画・CM・音楽・ニュース・スポーツに分類し，映画とスポーツについては更にそのサブジャンル（アクション・コメディや野球・サッカー等）を推定する．文献[2]ではスポーツ番組を競技ごとに分類する方式を提案している．カメラワークに着目し，その変化や連続する長さを特徴量として用いている．

コンシューマビデオでは，編集作業が行われることが一般的ではないため，ショット分割やショット単位での特徴抽出を適用することができない．また，カメラワークが安定していなかったり，手ぶれが含まれたりするため，プロフェッショナルビデオを前提としたカメラワーク推定だけでは信頼性が低いと考えられる．このため，コンシューマビデオに適したロバストな特徴を抽出する必要がある．

† 株式会社 KDDI 研究所, KDDI R&D Laboratories, Inc.
‡ 東京理科大学, Tokyo University of Science

3. コンシューマビデオのジャンル分類

本稿で提案する手法では，コンシューマビデオから低次の特徴を抽出し，それらの特徴と予め定義されたジャンルとの相関を機械学習によって学習する．そして，入力された未知のビデオコンテンツに対してジャンル分類を行う．以下の節では，まず本稿で対象とするジャンルを定義し，次にコンシューマビデオから抽出する特徴について述べる．本手法で用いる特徴は全て，MPEG 形式のオーディオビデオデータから復号なしで抽出することができる．

3.1 ジャンルの定義と傾向

本手法では，複数の動画投稿サイトで実際に扱われているジャンル（カテゴリ）を参考にし，表 1 に示すジャンルを定義する．動画投稿サイトでは，プロフェッショナルビデオに類するジャンルも含まれるため，これらを除いた上で，家族・友人とのイベントや旅行記など，ハンディカメラで撮影する機会が多いと考えられるジャンルを選択した．

表 1　ジャンルの定義

ジャンル	定義
旅行	旅行や観光で，風景や街の景観などを主に撮影した映像
スポーツ	運動会やスポーツ競技の様子
子供・ペット	子供やペットを撮影
パーティ	結婚式などのイベント
ステージ	学芸会や発表会などのイベント

以下に，各ジャンルの傾向を述べる．「旅行」では，風景や建物などの静止した被写体が多く撮影される．また，屋外で撮影されることが多いため環境雑音が含まれることが多い．「スポーツ」では，主に動的な被写体が撮影される．一般的には広い場所で競技が行われ，被写体が動き回ることが多いためカメラワークも頻繁になる．また，応援の叫び声や歓声などが伴うことも特徴的である．「子供・ペット」では，他のジャンルと比較して音量が小さいほか，被写体は静的であることが多く，このためカメラワークも少なくなると考えられる．「パーティ」では，イベントの主役（例えば新郎新婦）を主に撮影するが，その動きは活発ではないためカメラはあまり頻繁に動かない．また，このジャンルのイベントは薄暗い屋内で行われることが多く，カメラのフラッシュが多数発生する．「ステージ」では，客席からイベントを撮影することが多いため，カメラが静止している時間が長くなる．ステージ上の装飾や衣装などの色は，自然の風景などに比べて鮮明であることが多い．

3.2 分類に使用される特徴

3.1 節に述べた各ジャンルの傾向を考慮すると，本手法において抽出すべき特徴は，カメラの動き，被写体の動き，音声，音量，明度，彩度，フラッシュライトであると見なせる．以下に各特徴の抽出方法を概説する．

3.2.1 カメラの動き

以下により，ノイズを考慮したカメラのパンと固定の割合を検出する．

① 文献[3]の方式によりマクロブロックごとの動きベクトル（ローカル動きベクトル）からフレーム全体としての動きベクトル（グローバル動きベクトル）を算出
② 手ぶれを考慮し，グローバル動きベクトルの大きさが閾値以上ならパン，微小な動きはカメラ固定と判定
③ ゆっくりとカメラを動かしている場合は固定と判定されてしまうため，グローバル動きベクトルが連続して同じ方向を向いている場合はパンと判定

3.2.2 被写体の動き

ローカル動きベクトルの分散度合いを利用し，被写体の動きの有無（動物体，静止）を判断する．カメラを手で持ちながら撮影する場合，その動きが完全に静止することはほとんど無く，画面全体に動きベクトルが分布する．その中に動物体が存在する場合，動物体領域の動きベクトルは大きさや方向が周辺領域の動きベクトルとは異なるものになると考えられる．また，カメラを三脚などに固定して撮影した場合は，動物体以外の領域の動きベクトルが（0,0）となることを利用する．具体的には，方向の分散度合いを利用するために予め動きベクトルを 8 方向にサンプリングし，動き方向に関するヒストグラムを作成する．次に，以下の条件を満足すれば動物体が存在すると判定する．

① （0,0）の動きベクトル数が最も多い，且つ（0,0）以外の動きベクトル数＞閾値
② 最頻方向と第 2 の最頻方向が隣り合わない
③ 最頻方向の動きベクトル絶対値平均＜最頻方向以外の方向の動きベクトル絶対値平均

3.2.3 音声，音量，明度・彩度，フラッシュ

文献[4]の手法を用いてオーディオを無音・音声・音楽・雑音の 4 つのクラスに分類し，各クラスの割合を音声の特徴量とする．音量として MPEG オーディオのスケールファクタの平均値と標準偏差を算出する．また，ビデオフレームから抽出した輝度と色差から，明度（明るさ）と彩度（色の濃さ）を算出する．未編集のビデオでは，これらはコンテンツの時間経過によって大きく変化するものではないため，ビデオにおける先頭・中間・末尾のフレームで算出し，平均したものを特徴量とする．また，フラッシュのシーンでは映像が一瞬だけ明るくなるため，ビデオフレーム間の輝度差分にピークが存在すればフラッシュと見なす．

4. 検証実験

4.1 実験条件

各ジャンルの定義に沿うビデオを，筆者ら自身で撮影したり動画投稿サイトからダウンロードしたりすることで，合計 221 本を収集した．各ジャンルの内訳は，旅行:31，スポーツ:58，子供・ペット:50，パーティ:36，ステージ:46 である．全てのビデオファイルは事前に MPEG 形式に変換しておく．次に全ファイルをランダムに 3 つのグループに分割し，2 グループを学習用，1 グループを検証用とする．これらの組み合わせを変えて 3 パターンの実験を行い，実験 1，実験 2，実験 3 とする．また，機械学習としては WEKA ツール[5]を用いた．WEKA には多数の分類器が実装されており，特定の学習法を組み合わせることもできる．

4.2 結果と考察

WEKA における幾つかの分類器を適用したときの上位 3 つの結果を表 2 に示す．表中の数値は F 値（再現率と適合率の相乗平均）を表す．実験の結果，集団学習法である Random Forest 法[6]（以下，RF）が，決定木法である J48 や REPTree よりも高い精度を達成した．決定木法において生成された木を確認したところ，J48 では REPTree に比べてかなり複雑な木が作成され，J48 の方が高い精度が得られていることが分かった．より多くの特徴量が用いられた複雑な決定木を適用した方が高い精度であることは，本手法で選択した特徴が有効に作用していることを意味する．

表 2 異なる分類器による分類結果

分類器	実験 1	実験 2	実験 3	平均
RF	0.74	0.76	0.70	0.73
J48	0.68	0.68	0.62	0.66
REPTree	0.64	0.59	0.62	0.62

次に，学習法を組み合わせた精度を比較した．Bagging [7]に上記 3 つの分類器を組み合わせた結果を表 3 に示す．分類器の違いで比較すると，表 2 の結果と同様に全ての実験で RF の分類精度が高く，更に単独で行うよりも精度が向上している．尚，紙面の都合上割愛するが，この組み合わせによるジャンル毎の分類精度は 0.68～0.86 となった．誤分類の多くは，「旅行」が「スポーツ」に，「パーティ」が「ステージ」に分類される場合であり，前者は被写体の動きと音量，後者はカメラの動きに影響を受けている．精度向上に向け，今後新たな特徴を追加する必要がある．

表 3 学習法の組み合わせによる分類結果

集団学習法	分類器	実験 1	実験 2	実験 3	平均
Bagging	RF	0.78	0.75	0.77	**0.77**
	J48	0.75	0.74	0.75	0.75
	REPTree	0.78	0.68	0.69	0.72

5. まとめ

コンシューマビデオの自動ジャンル分類方式を提案した．実際の動画投稿サイトを参考にして定義したジャンルに対し，制作品質の高くないコンシューマビデオにロバストな特徴を抽出することによって，F 値 0.77 を達成した．

参考文献

[1] X. Yuan, et al., "Automatic video genre categorization using hierarchical SVM," IEEE ICIP, pp.2905-2908, 2006.
[2] 服部しのぶほか, "映像特徴に基づく自動映像分類システムの提案", 情処研報, Vol.2002, No.25, 2002.
[3] Roy Wang, "Fast Camera Motion Analysis in MPEG domain," IEEE ICIP, Vol.3, pp.691-694, 1999.
[4] 中島康之ほか, "MPEG 符号化データからのオーディオインデキシング," 信学論, Vol.J83-D2, No.5, 2000.
[5] I. H. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques," 2nd edition, Morgan Kaufmann Publishers, 2005.
[6] Leo Breiman, "Random Forests," Machine Learning, Vol. 45, No. 1, pp.5-32, 2001.
[7] Leo Breiman, "Bagging Predictors," Machine Learning, Vol. 24, No. 2, pp123-140, 1996.