I-012

時空間特徴を用いた YouTube 上のビデオ内の暴力シーン検出
Violence Detection in YouTube Video Using Sparse Spatio-Temporal Features

内藤 貴† 王 彧† 加藤 ジェーン† 間瀬健二†
Takashi Naito Yu Wang Jien Kato Kenji Mase

1.はじめに

暴力的な表現は成長過程の子供に悪い影響を与える。1999 年 4 月にアメリカのコロラド州コロンバイン高校で 17 歳の少年二人が銃乱射事件を起こし 12 名の高校生と 1 名の教師の合計 13 名を殺害した。また 2012 年 12 月にアメリカのコネチカット州サンディフック小学校で 20 歳の少年が銃乱射事件を起こし児童 20 名を含む 26 名の当時小学校にいた人間を殺害した。これらの事件で犯人の少年らに暴力ビデオや FPS（一人称視点のオンラインゲーム）の影響を受けていたという見方がある。このように暴力的な映像が与える影響は深刻である。そのため近年では、テレビや映画、テレビゲームでは年齢制限や暴力シーンの削除などで規制が厳しくなってきている。しかし、インターネットの普及により、誰でも映像を投稿できる YouTube やニコニコ動画などの動画サイトで、数多くの映像コンテンツの視聴が可能になった。このような動画サイトでは、規制が投稿する個人の良心によるのでテレビや映画などに対して比較的緩い。サイトの管理者による規制もあるが、一日に膨大な量の映像コンテンツが投稿されるため手動でチェックすることは困難である。そこで、これらインターネット上の動画から暴力を自動で検出する方法が求められる。

行動認識の分野における暴力シーンの検出の研究は、数多く存在する。Clarin ら[1] は、殴る、蹴るなどの動きの抽出に加え、色ヒストグラムから皮膚や血液の色を特徴とするフレームを抽出し暴力シーンの検出を行った。Acar ら[2] は、映像内の感情の高揚した音を特徴として暴力シーンの検出を行った。これらの暴力シーンの検出は、映画やテレビなどの映像を対象にしている。一方で、YouTube などのインターネット上の動画には、「低解像度」、「映像内のノイズ」、「音声のズレ」を含む動画も大量にある。そのため、解像度とノイズは Clarin ら[1] の色ヒストグラムを、音声のズレは Acar ら[2] の音声表現のヒストグラムを正確に作成できない可能性がある。そのため彼らの手法をインターネット上の動画に対して適用することは難しい。そこで、音や色に影響されない方法で、映像内の動作から暴力シーンを検出する必要がある。

1.1 関連研究

色や音以外に映像から抽出できる特徴を用いた研究は存在する。Dollar[3] らは、ビデオ内のフレーム順から輝度の勾配を抽出し、行動を検出した。そして、特徴量として取り出すため、通常の二次元上のフレームワークを、時間方向を加えた三次元に拡張する時空間特徴の追跡を行った。

また、柳井[4] らは、Dollar[3] らの時空間特徴を用いて動画を表現し、動き特徴に対し特徴量を取り出した。そして、画像集合の代表的な画像を選出するためのランキング手法である VisualRank を適用した。この手法により、映像をランキング付けすることで行動認識を行った。上記二つの研究は、KTH データセット[5] を用いて、6 種類の行動に対し実験が行われた。しかし、実験で用いられた KTH データセットは、対象を一人としてカメラを固定した状態で撮影された映像からなるデータセットである。インターネット上の動画には「対象が複数」、「カメラが動く」といった映像も多く存在する。本研究ではそれらの映像も対象とする。

†名古屋大学大学院 情報科学研究科

1.2 目的

本研究では、自動で暴力シーンを削除、規制することや、手動によるレビューコストを削減するために、インターネット上の映像から暴力シーンを検出することを目的とする。本研究では、映像を扱いやすくするため以下のようにいくつかの映像単位を定義する。

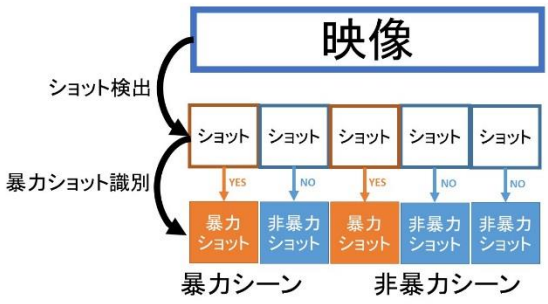

図 1.1 映像の単位

映像単位の相互の関係のイメージを図 1.1 に示す。連続的に撮影された映像区間のことをショットと呼ぶ。ショットは映像の最小単位であり文章の単語にあたるものである。映像を時間順に見たとき一つのショットから次のショットに切り替わる変わり目をカットと呼ぶ。シーンはショットの意味的内容を表す。本研究ではカットによって検出されたショットに対して、暴力を含むショットをまとめて暴力シーンと呼ぶ。また、暴力を含まないショットをまとめて非暴力シーンと呼ぶ。本研究の提案手法は以下の二つのステップからなる。

1. 本研究で用いる映像から自動でショットを検出

2. 検出したショットから時空間特徴を用いた暴力ショットを検出

映像内からのショット検出を手動で行うと、カットする基準に個人差が生まれる。そのため、カットを同一手法で客観的に算出する必要がある。そこで本研究では、使用するすべてのショットの検出基準を統一するため、映像内からショットの自動検出を行う。本研究では、暴力シーン特有の動き特徴を映像中から抽出するために時空間特徴を用

いる。これにより、「ノイズや音声のズレが存在するインターネット上の動画」に対して暴力シーンの検出を可能にする。

2.映像のショット分割

本章では映像からカット点を算出し、ショットを検出する手法について述べる。提案手法によって背景の切り替わりごとに映像からショットを検出することを可能にする。本研究では、入力として与えられた YouTube 上の動画に対し、ショット検出が行われ 0.5 秒から 2 秒のショットに切り分けられる。検出点を表す映像内の特徴として、フレーム間の色ヒストグラム距離と Edge Change Ratio を用いた。算出された 2 種類の特徴点から極大値を取る共通の時刻をカットの時刻とする。

2.1 フレーム間の色ヒストグラム距離

色ヒストグラムは画像中の各色の要素数をヒストグラムで表したものである。本研究では、カラー画像の赤成分、緑成分、青成分を 8 段階で表現し、各画素に対し、8 × 8 × 8 = 512 通りの 3 次元セルの内のどれかを当てはめる。連続した二つのフレームの色ヒストグラム A, B をそれぞれ計算する。フレーム内のすべての画素に当てはめ、一つのフレームからヒストグラムを作成する。そして、AB 間のユークリッド距離を色ヒストグラム距離とする。この特徴はフレーム間の色の相違の大きさを表す。

2.2 Edge Change Ratio

Edge Change Ratio はフレーム間のエッジの出現の割合を特徴とし、輝度の急激な変化を特徴として表す。まず、フレーム内の画素間の輝度の差からエッジ点を求める。そして、すべての画素数に対する連続した二つのフレーム間の出現と消滅するエッジ点の総数の割合を表し特徴とする。前のフレームのエッジ画素の総数を s_i、後ろのフレームのエッジ画素の総数を s_i+1 とし、追加されたエッジ画素の数をρ_{in}、削除されたエッジ画素の総数をρ_{out} とすると、フレーム i の Edge Change Ratio(ECR) は次式に表す。

$$ECR_i = \max\left(\frac{\rho_{in}}{s_i}, \frac{\rho_{out}}{s_{i+1}}\right)$$

2.3 カット点の算出

第 2.1 節と第 2.2 節において、二つの特徴量をフレームごとに算出する。両方の特徴量が極大値を取る共通の時刻をカットの時刻とする。算出されたカットを用いて、映像のショット分割が行われる。

3.時空間特徴を用いた暴力ショットの検出

本章では検出された各ショットから暴力ショットを検出する手法について述べる。提案手法は以下のようになる。まず、Dense Trajectories[6] を用いて時空間特徴点を検出する。Dense Trajectories は Dense Sampling を用いてフレームごとに輝度の勾配から特徴点を取り出し、時間方向に追跡することで特徴量を算出し、また求められた特徴点から特徴量を算出する手法である。本研究では特徴量として HOG 特徴量を用いる。算出した HOG 特徴量から、Bag-of-Features 法を用いて特徴ベクトルを作成する。最後に、SVM 用いて識別を行い、暴力と非暴力のショットを認識することで、暴力ショットの検出を行う。

3.1 Dense Trajectories を用いた特徴点の抽出

まず、すべてのフレームに対して Dense Sampling を行う。Dense Sampling は与えられたショットに対し、フレーム内の一定の画素ごとに密集したサンプルを特徴点として求める。一定の画素毎に特徴点を取り出す。特徴点は輝度の勾配によって抽出される。サンプリングされた点の例は図 3.1 に示す。図 3.1 中の赤い点で示してある部分が抽出した特徴点である。

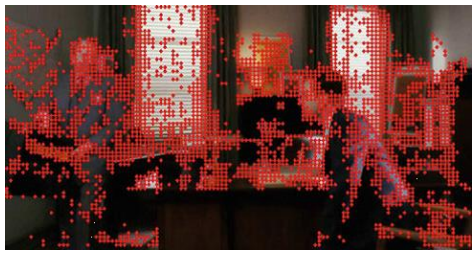

図 3.1 閾値を用いて、密集した特徴点を可視化した結果

3.2 HOG 特徴量の算出

本研究ではノイズの影響を受けにくい局所特徴量である HOG 特徴量を用いる。HOG 特徴量は、画像から輝度の勾配強度と勾配方向を数ピクセルからなる局所領域セルごとに計算し、勾配方向ごとヒストグラムを作成する。そして数セルからなる三次元の局所領域パッチを正規化することで得られる特徴である。本研究では、第 3.1 節で検出された特徴点から 3 次元パッチを抽出する。3 次元パッチのサイズは 2 × 2 × 3 ピクセルとした。ヒストグラムによる分割する勾配方向は 8 方向とし、特徴量は 96 次元とした。検出するパッチのイメージを図 3.2 に示す。パッチから HOG 特徴量を求める。

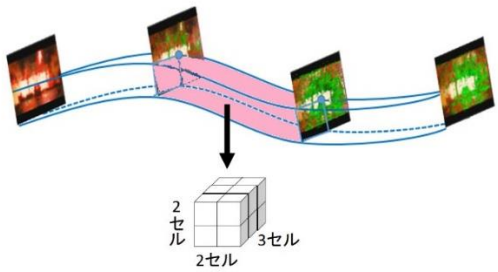

図 3.2 HOG 特徴のイメージ

3.3 Bag-of-Features を用いた特徴ベクトルの作成

本研究では、第 3.2 節から求めた大量の局所特徴量のデータから、Bag-of-Features 法を用いてショットごとの特徴ベクトルを求める。Bag-of-Features 法の流れについて示す。

この手法ではまず、大量のデータから得られる局所特徴量の集合を用いて visual words を求める。visual words は局所特徴量の集合を k 個のクラスタへクラスタリングすることによって求まる各クラスタの中心となるベクトルのことを表す。本研究では、クラスタリングの手法として k-means クラスタリングを用いた。図 3.4 に k = 4 の場合の、

k-means クラスタリングを用いた visual words 算出について示す。局所特徴量を四つのクラスタに分割し、　クラスタごとにクラスタ中心を求める。図 3.4 のオレンジの点がクラスタ中心である。

次に、特徴ベクトルを求めたい新たな映像から取得できる局所特徴量の集合に対して、各局所特徴量と最も距離の近い visual words を探索する。ここで距離はユークリッド距離とする。判定された visual words の出現頻度をヒストグラムで表す。このヒストグラムを特徴ベクトルとして扱う。図 3.5 にヒストグラム算出について示す。図 3.5 の赤い点が新しく加わった特徴量である。各クラスタの特徴量の数から投票を行い、ヒストグラムを作成する。本研究では、クラスタ数 k は 400 とした。

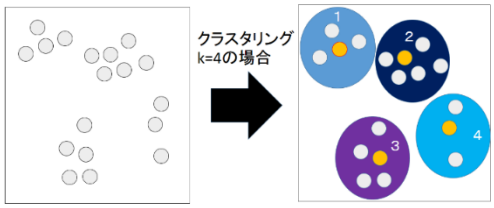

図 3.4 visual words の算出

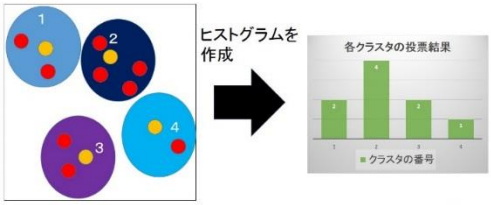

図 3.5 ヒストグラムの算出

4.実験

本章では提案手法の評価実験について述べる。本研究では学習とテスト用のデータセットは YouTube 上の動画を使用した。第 2 章で提案した手法で検出したすべてのショットに手動でタグ付けを行い、第 3 章で提案した手法でショット内から、暴力ショットを検出した。本研究では、タグ付けされたショットを用いて、交差検定により検証を行った。また、交差検定の出力結果に混同行列を用いて暴力ショットの検出率を評価した。

4.1 データセット

本研究の評価実験には、YouTube 上の動画から作成したデータセットを用いた。2013 年 12 月付の Amazon の洋画アクションランキングから、同じタイトルのものを除く上位 40 の映画の予告動画を収集した。ショット検出後、再生時間の短い 0.5 秒以下のショットを除いたすべての検出したシーンに対し手動でタグ付けを行った。本研究では以下の 6 種類の内容を一つでも含んでいるショットを暴力ショットとした。図 4.1 に暴力ショットの例を示す。収集した 43 本の映画予告から、248 個の暴力ショットと 1032 個の非暴力ショットを準備した。

図 4.1 各暴力のクラス

4.2 実験設定

本実験では、提案手法を用いて予告動画から暴力ショットの検出を行う。そして、その精度を検証することで提案手法の有用性の評価を行う。本研究では、提案手法によって作成した特徴ベクトルから以下のような交差検定によって検証を行う。

まず、テストデータと学習データに分ける。テストデータは収集した 43 本の映画予告の内の一本から検出したショットである。学習データは残りの 42 本の映画予告の暴力ショットと非暴力ショットを用いた。学習データ数の偏りをなくすため、学習データに用いる暴力ショットと同じ数になるように、非暴力ショットの選択をランダムに行った。また、次の二つの条件ごとに検証実験を行う。

1. 暴力ショットをまとめて検出
2. 暴力ショットを暴力の種類ごとに検出

検出した結果は混同行列を用いて評価する。

4.3 結果

第 4.2 章で述べた手法から、実験ごとの精度を示す。また、検出率と正解率は次のように定義する。

$$\text{Precision} = \frac{\text{正しく検出した暴力のショット数}}{\text{暴力のショット数}}$$

$$\text{Recall} = \frac{\begin{pmatrix}\text{正しく検出した暴力ショット数} + \\ \text{正しく検出した非暴力ショット数}\end{pmatrix}}{\text{暴力ショット数} + \text{非暴力ショット数}}$$

A)　暴力ショットをまとめて検出した結果

実験結果は表 4.1 に示す。表 4.1 の各数字はショット数を表す。正しく暴力ショットを検出できた Precision が

63.7%、正しく全体のショット検出した Recall が 70.2% を示した。また、非暴力のショットから正しく非暴力のショットを検出した割合が 740=(292 + 740) = 71.7% となり、暴力ショットの検出率を上回った。

表 4.1 暴力ショットをまとめて検出

		識別結果	
		暴力	非暴力
正解ラベル	暴力	158	90
	非暴力	292	740

B)　暴力ショットをタグごとに検出した結果

実験結果は表 4.2 に示す。表 4.2 の各数字は暴力ショット検出率とショット全体の正解率を示す。本研究では、時空間特徴を用いたが、爆発以外のジャンルでは平均より低い精度を示した。殴打やカーチェイス、武器を使った暴力の抽出は不向きだと考えられる。また、銃撃と流血のある暴力ショットはデータセットが足りなかったため、検出自体が難しかった。

表 4.2 暴力ショットを暴力の種類ごとに検出

	Precision (%)	Recall (%)
殴打	36.62	75.39
爆発	68.37	33.75
銃	63.53	51.33
流血	50.00	70.23
カーチェイス	0.00	32.11
武器	3.45	48.52

4.4 考察

本研究の提案手法ではインターネット上の予告動画から全暴力ショット中の 63.7% を検出した。しかし、全ショット 1280 個の内 90 個のショットで暴力ショットを非暴力と認識し、292 個のショットで非暴力ショットを暴力ショットと認識した。非暴力を暴力と認識したショットでは、映画予告特有の字幕が流れたショット、暴力がなくても激しい動作を行うショット、CG などの演出が混じったショットのような時空間特徴点の抽出箇所が多いショットに誤認識する傾向があった。本研究では日本語字幕など映像の下に映る字幕が入った映像は実験で使用していないが、出演者のクレジットタイトルなど映像中に映るものに対して特別な処置をしていない。映画予告の演出で動きをもつ字幕も存在し、誤認識に繋がったと考えられる。解決策としてショット検出のカットの基準を上げるなどして細分化する必要がある。また、暴力がなくても激しい動作を行うショットや CG などの演出が混じったショットは、その他の特徴と関連を持たせることによって判断する必要がある。時空間特徴の周りに、血や皮膚などの色特徴が存在するかなど考えられる。

暴力を非暴力と認識したシーンでは、銃撃や流血、カーチェイスの認識率が低い傾向があった。銃撃や流血が含まれるショットでは、非常に局所的な動きの変化をするため、爆発のような画面全体に広がるような大域的な動きの変化の含まれるショットと同じショットに存在するとき検出することは難しい。また、テストデータの数が少ないことも原因として考えられる。これから対策として種類別に特徴量を増やすことを検討する。銃撃を検出する場合は、火薬の色も検出するように色特徴量を組み合わせることを考える。また暴力ショットのデータ数も増やす必要がある。

5. おわりに

暴力シーンは成長過程の子供に悪影響を与え、インターネット上では暴力シーンを含む動画は誰でも見ることができるため規制が必要である。しかし、インターネット上の動画は非常に数が多いため規制が難しい。インターネット上の動画には、背景ノイズや音声のズレを含む動画も多く存在するため有効に使える特徴量が限られる。そこで本研究では、インターネット上の動画から時空間特徴を用いて暴力シーンを検出する手法を提案する。時空間特徴は映像内から局所的に動き特徴を抽出するため、本研究では有用である。得られた特徴量から暴力ショットの検出を行った。

YouTube 上の動画を利用したデータセットから 63.7% の暴力ショットを検出した。二値分類でのチャンスレートである 50%を超えたため本手法は有用であるといえる。

今後の課題として、アクション映画の予告のみをデータセットとして用いた。しかし、暴力的なシーンを含むジャンルは他にも SF、戦争、ホラーなど数多く存在する。これらのジャンルからも検出を行い、傾向を確かめる必要がある。また、タグ別の暴力の数、特に流血とカーチェイスが少ないので、全体的にショット数も増やすことを検討したい。またインターネット上の動画の特徴から、本研究では色特徴やその他の特徴は用いなかったが、局所的な領域を動き特徴として取り出す時空間特徴には限界があることが分かった。そのため、今後は適用できる場合は時空間特徴と組み合わせることを検討する。爆発の検出の際には、色ヒストグラムも特徴量として加え検証を行う。

参考文献

[1] C Clarin, J Dionisio, M Echavez, and PC Naval. Dove: Detection of movie violence using motion intensity analysis on skin and blood. PCSC, 6:150156, 2005.

[2] Esra Acar and Sahin Albayrak. Dai lab at mediaeval 2012 aect task: The detection of violent scenes using aective features. In MediaEval, 2012.

[3] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on, pages 6572. IEEE, 2005.

[4] DO HANG NGA and 柳井啓司. 時空間特徴量を用いた youtube 動画からの特定動作ショットの自動抽出 (テーマセッション, 映像処理と trecvid). 電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解, 110(414):159164, 2011.

[5] I. Laptev. Recognition of human actions. http://www.nada.kth.se/cvap/actions/.

[6] [6] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action Recognition by Dense Trajectories. In IEEE Conference on Computer Vision & Pattern Recognition, pages 31693176, Colorado Springs, United States, June2011.