2004－DBS－132　(5)
2004／1／15

映画の文法に基づく要約映像の生成

出口　嘉紀†　　吉高　淳夫†

映画の撮影や編集の際に制作者によって、特定の意味や意図を強調する目的で使用される映画の文法に基づき、映画の内容と文脈が理解しやすい要約映像を生成する手法を提案する。そのためには、内容が伝わるように編集上強調された区間に加え、それに従属する区間を抽出する必要がある。まず編集上強調された区間として、アクション区間、緊迫した区間、落ち着いた区間を抽出し、重要度の高い順に要約映像の候補とする。次に性質の異なる区間が連続している場合、それらの区間で従属関係が成り立つため、前後の区間においてその度合いを求め、強い従属関係のある区間を要約映像として採用する。したがって、強調された区間だけでなくそれに至る経緯も要約映像に含めることができる。

Movie Summarization based on Film Grammar

Yoshiki DEGUCHI†　　Atsuo YOSHITAKA†

In this paper, a framework of movie summarization based on film grammar is proposed. In order to summarize a movie, segments impressed by editing and segments depended on those segments should be detected. First, action segments, tension segments, and calm segments are detected as segments impressed by editing, and included in the summary in order of the important degree. Secondly, if segments which property is different from are adjacent, there is a dependent relationship between those segments. A dependent degree between those segments is calculated, and segments impressed by editing and tight dependent segments are included in the summary. Consequently, both segments impressed by editing and context are included in the summary.

1. はじめに

インターネット上での通信速度の増大により、映像配信やディジタル放送が一般的になりつつあり、HDD内蔵のビデオレコーダなどが普及してきていることから、ユーザは多くの映像をインターネットを通じて取得し、それらを蓄積し、視聴することが可能となってきている。そのためユーザは、多くの映像の中から観たい映像を選択する必要がある。短時間で映像の内容や雰囲気を理解することを目的とした手法の一つとして、映像を要約する手法が挙げられる。

†広島大学大学院工学研究科
Graduate School of Engineering, Hiroshima University

映像にはドラマ、映画、スポーツ、ニュース、音楽番組など様々なものが存在するが、特に映画は時間が長いため、短時間で内容が理解しやすい要約映像を生成することができれば、ユーザにとっては有用なものとなる場合がある。例えば、蓄積した映画をブラウジングする場合、映画評論家が過去に観た映画の紹介や批評を書く際、その映画の内容を思い出したい場合などが挙げられる。

映画を対象とした映像要約に関する研究として、[1]では、主要人物のクロースアップ、銃声や爆発、タイトルやテロップなどの特別なイベントを検出し、これらをつなぎ合わせることで映画の予告編を目的とした

要約映像を生成している。また[2]では、ドラマの心理的印象の高い区間に注目し、音楽の開始や終了、カットが頻出する箇所など心理的に重要な箇所を切り出した要約映像を生成している。[3]では、視聴者が視覚、聴覚に注意を向ける要素を元にして作成した User Attention Model に基づき、視聴者が注意を向けたと考えられる区間を要約映像に採用している。これらの手法で生成されている要約映像は、特定の特徴が検出された区間を単純につなぎ合わせているに過ぎないため、断片的な映像になってしまい、どのような出来事が起こったかを十分に知ることが出来ない上に、その出来事の前後関係が分からない要約映像となる。

[4]では、ショットを視覚的な類似度に基づきクラスタリングし、各クラスタから一番長いショットを要約映像として採用している。この研究では、視覚的に冗長なショットを除いただけであり、話の内容を伝える上で重要なショットの選択はされていない。また各クラスタから一番長いショットを要約映像として採用しているが、内容を伝える上で一番長いショットが適切であるかが不明である。

[5]では、画像、音の特徴から映画をショット、ストーリ・ユニット、シーンに構造化し、それぞれの単位における従属性を検出することによって、映画の文脈を考慮に入れた要約映像を生成している。この研究では、文脈を考慮しているが、従属関係にあるショットすべてを要約映像に採用しているため、要約映像に偏りがあり映画全体の話の内容を知ることは困難である。

本研究では、映画の文法[6]に基づき、内容が効果的に視聴者に伝わるように、編集上強調された区間としてアクション区間、緊迫した区間、落ち着いた区間と、それらの区間と従属関係にある区間を抽出する。そして制約時間を満たすように、重要度の高い順にそれらの区間内のショットを要約映像として採用することで、映画の内容と文脈が理解しやすい要約映像の生成手法を提案する。

本稿では、2 章に映画の文法と処理の流れについて述べ、3 章でショットの性質の定義を行う。4 章で要約映像の生成手順を述べ、5 章で実験と評価結果について述べ、6 章でまとめとする。

2. 処理内容

2.1 映画の文法

映画には、撮影や編集の際に制作者によって特定の意味や意図を強調する目的で使用される技法がある。それを映画の文法[6]という。映画の文法によると、以下のことが述べられている。

- **アクションシーン**
  短いショットが連続し、映像の動きが激しい
- **緊迫したシーン**
  ショットの長さが徐々に短くなる
- **落ち着いたシーン**
  長いショットが連続し、映像の動きが緩やか
- **効果的な内容伝達**
  原因と結果の関係にある映像区間を結合する

2.2 処理の流れ

映画の文法に基づき、話の内容を視聴者に効果的に伝えるために、編集上強調された区間として、アクション区間、緊迫した区間、落ち着いた区間を抽出する。その際、各ショットにおいて、ショットの長さ、画像の動きの激しさや緩やかさ、音楽（楽器音成分の継続時間）に基づき、ショットの性質として、アクション性、緊迫性、落ち着き性を定義する。そして性質を表す値が連続して高い値をとるショット群をそれぞれアクション区間、緊迫した区間、落ち着いた区間とする。これら 3 つの区間を抽出し、各性質を表す値の高い順に要約映像を生成する際の候補とすることにより、映画の中で編集上強調された区間を要約映像に加えることが可能となり、その要約映像は映画の内容が分かりやすいものとなる。ここで、ショットとは一台のカメラから撮影された連続するフレームの集合のことである。またカットとは、ショットの境界のことである。

主体の存在するショットを検出する。主体の存在するショットは、話の内容を視聴者に伝える上で重要なショットとなり、そのショットを中心に採用した要約映像は、それを考慮しないものに比べて、映画の内容を理解しやすくなる。画像の中で強調されているオブジェクトが主体である可能性が高いことから、ある一定以上の大きさで、同一色で輝度の変化が周囲と異なるオブジェクトが存在するショットを検出する。

さらにアクション区間、緊迫した区間、落ち着いた区間のいずれか 2 つの区間が隣接している場合、それらの区間には原因と結果を表す従属関係がある。そのため、それら 2 つの区間を含めた要約映像は、含めない映像に比べてより文脈の理解しやすいものとなる。抽出した区間内でアクション性を表す値、あるいは落ち着き性を表す値の平均値を求め、前後の区間においてその差を求めることにより、それらの区間での従属関係の度合いを求める。ここで従属関係の度合いを前後の区間の値の差としているのは、前後の性質の違いが大きいほど、視聴者に強い印象を与えて内容を効果的に伝えることができるからである。

最後に要約映像を生成する際、映画全体から満遍なく要約映像となる映像区間を選択し、話の内容を理解しやすくするため、映画を n(=20)等分する。そしてその分割された区間の中から、視聴者が指定した制約時間を満たすように、アクション性を表す値、緊迫性を表す値、落ち着き性を表す値のいずれかが高く、主体が存在するショットを優先して要約映像として採用し、それと強い従属関係のある区間内の主体の存在するショットも要約映像として採用することで、映画の内容と文脈をより理解しやすい要約映像を生成する。

3. ショットの性質の定義

3.1 アクション性

3.1.1 ショットの長さによるアクション性

アクション区間では、短いショットが連続するという特徴があるため、それを以下の条件で抽出し、アクション性を表す値を求める。

k 番目のショット s_k でのショットの長さを $SL(s_k)$[秒]とすると、s_k でのショットの長さによるアクション性を表す値 $SLV_A(s_k)$を以下のように定義する。これは、アクションを視聴者に効果的に伝えるためには、短いショットを用いることに基づき、あるショットの長さが短いと判定された場合、アクションを表しているショットとみなし、アクション性を 1 とする。ここで、ショットの長さによるアクション性を 2 値としているのは、ショットの長さが短ければ短いほど、アクション性が高くなることは映画の文法により示されていないためである。

$$SLV_A(s_k)=\begin{cases}1, & if\ SL(s_k)<Th_{shot}\\ 0, & otherwise\end{cases}$$

$$Th_{shot}=\frac{1}{2}\left(SL_{mean}+SL_{mode}\right)$$

ただし、Th_{shot}[frame]はショットの長さが短いことを表す閾値で、SL_{mean}[frame]はある映画全体のショットの長さの平均値である。SL_{mode}[frame]は、ショットの長さの最頻値を表す。ただし最頻値は、0.5 秒間隔でショットの累積頻度を求め、その度数が最大になる 0.5 秒間での中間値としている。

3.1.2 画像内の変化によるアクション性

時空間投影画像[7]（図 1）は、映像中のオブジェクトやカメラワークによって生じる動きを可視化した画像であるため、[7]ではカメラワークを検出する際に用いられている。本研究では、時空間投影画像中に、画像の動きの激しさに伴う特徴が現れることに着目し、その特徴を検出することによってアクション性を求める。映像の動きが激しい場合、時空間投影画像上では垂直方向のエッジが現れる（図２）。

ショット s_k での時空間投影画像における垂直方向のエッジの数を $E_v(s_k)$とすると、時空間投影画像によるアクション性を表す値 $VTIV_A(s_k)$を以下の式のように定義する。以下の式では、映像内の激しさを単位時間に現れるエッジの数として表している。これは、アクション区間で映像内の動きが激しいほど、時空間投影画像中に現れる垂直方向のエッジの数が多くなることに基づいている。

$$VTIV_A(s_k)=E_v(s_k)/SL(s_k)\times 30$$

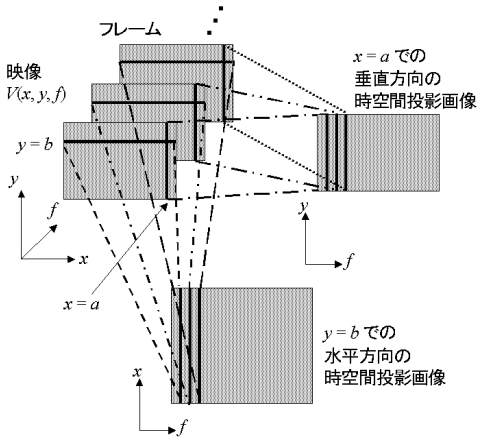

図 1　時空間投影画像

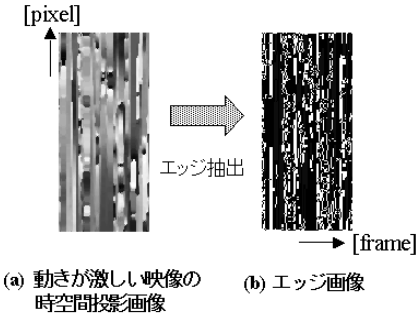

図 2 動きが激しい映像の時空間投影画像とエッジ画像

3.1.3 音楽によるアクション性

[8]では、サウンドスペクトログラム上に現れる時間軸に沿った周波数ピークを示す楽器音成分（図 3）を検出し、ある時間間隔における楽器音成分の数により音楽が流れていることを判定している。本研究では、音楽の特徴がその楽器音成分の継続時間に表れることに着目し、その時間によって音楽の性質を検出する。実験により、アクション区間で流れている音楽は、楽器音成分の継続時間が短い傾向があることを確認して

いる。また、音楽の中でベースに分類される楽器は音楽の基準になるため、ベースが担う周波数帯の楽器音成分に着目する。映画では、オーケストラで演奏された音楽が流れることが多いため、オーケストラでベースを担う楽器の周波数帯(30-300Hz)の楽器音成分の継続時間を指標とする。

ショット s_k での楽器音成分の長さを $IL(s_k)$ [秒]とし、楽器音成分の継続時間が短いことを判定する閾値を Th_{inst_A}[秒]とすると、音楽によるアクション性を表す値 $MV_A(s_k)$を以下のように定義する。

$$MV_A(s_k)=\begin{cases}1, & if\ \dfrac{1}{10}\displaystyle\sum_{i=-5}^{4} IL(s_{k+i}) < Th_{inst_A}\\ 0, & otherwise\end{cases}$$

ただし、Th_{inst_A} は実験により求めた値で 1.24[秒]とした。

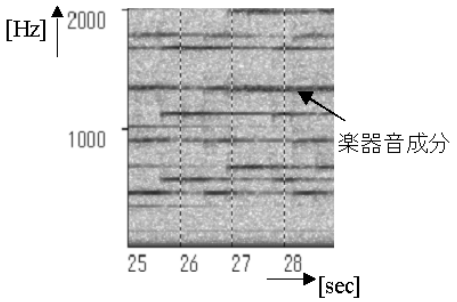

図 3　サウンドスペクトログラム

3.1.4　アクション性

以上で求めた各特徴によるアクション性を表す値に基づき、ショット s_k でのアクション性を表す値 $Action(s_k)$を以下の式のように表す。以上で求めた 3 つの値に基づき、ショット s_k でのアクション性を表す値を求めるが、ある要素のみが必ずアクション区間に表れるのではなく、各要素が満たされる可能性があるため、各要素の平均を求めアクション性を表す値としている。

$$Action(s_k)=\frac{1}{3}\{SLV_A(s_k)+VTIV_A(s_k)+MV_A(s_k)\}$$

3.2　緊迫性

緊迫した区間ではショットの長さが徐々に短くなるという特徴がある。その特徴に基づいて緊迫した区間を抽出する。緊迫した区間内でショットの平均時間が短いほど、緊迫性が高く感じられるため、それを緊迫性を表す値の指標とし、$Tension(s_k)$と表す。以下にその式を定義する。

$$Tension(s_k)=1-\frac{SL_{Tension}}{Th_{shot}}$$

$$SL_{Tension}=\frac{1}{n}\sum_{i=1}^{n} SL(s_{k+m_i})$$

$$if\ \ Th_{shot} > SL(s_{k+m_1}) > SL(s_{k+m_2}) > \cdots > SL(s_{k+m_n})\ \ and$$
$$-5 \le m_1 < m_2 < \cdots < m_n \le 4\ \ and\ \ n \ge 4$$

ただし、$SL_{Tension}$ は緊迫した区間内でのショットの長さの平均値、n は緊迫した区間内のショットの数、m_i は k 番目のショットからの変位を表す。

3.3　落ち着き性

3.3.1　ショットの長さによる落ち着き性

落ち着いた区間では、長いショットが連続するという特徴があるため、それを以下の条件で抽出し、落ち着き性を表す値を求める。

ショット s_k でのショットの長さによる落ち着き性を表す値 $SLV_C(s_k)$を以下の式のように定義する。これは、落ち着いた雰囲気を視聴者に効果的に伝えるためには、長いショットを用いることに基づき、あるショットの長さが長いと判定された場合、落ち着いた感じを表しているショットとみなし、落ち着き性を 1 とする。ここで、ショットの長さによる落ち着き性を 2 値としているのは、ショットの長さが長ければ長いほど、落ち着き性が高くなることは映画の文法により示されていないためである。

$$SLV_C(s_k)=\begin{cases}1, & if\ SL(s_k) > Th_{shot}\\ 0, & otherwise\end{cases}$$

3.3.2　画像内の動きによる落ち着き性

落ち着いた区間では、映像内でオブジェクトやカメラワークによる動きがあまり見られないため、時空間投影画像上には水平方向に沿ってエッジが存在する。そのエッジの平らさを検出することによって落ち着き性を定義する。この場合、平らさの尺度が落ち着き性を表す値とする。

ショット s_k での平らさの尺度を求めるには、時空間投影画像上でエッジとなる部分を追跡し、図 2(a)に示す値を図 2(b)に示す追跡順序に従って加算していく。そしてこの加算結果を追跡したピクセル数で除算することにより求めた値を平らさの尺度とする。これは、エッジが水平方向の直線となる場合、最大値 1 をとり、図 2(b)の追跡順序において 7, あるいは 9 の位置に繰り返しエッジとなる部分が存在する場合、最小値 0 を

とる。

図 2(a)に示す値の加算結果を $Sum(s_k)$、追跡ピクセル数を $N(s_k)$とすると、ショット s_kでの時空間投影画像による落ち着き性を表す値 $VTIV_C(s_k)$を以下のように定義する。

$$VTIV_C(s_k)=\frac{1}{2}\times\left(Sum(s_k)\Big/N(s_k)\quad+\quad 1\right)$$

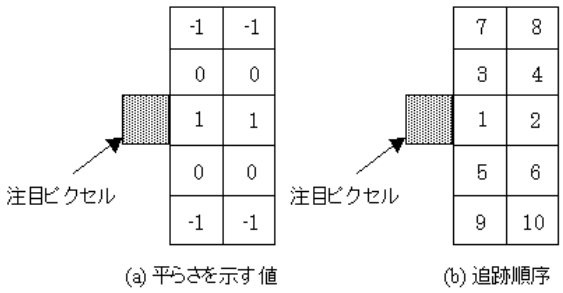

図 4　エッジの平らさを示す値とその追跡順序

3.3.3　音楽による落ち着き性

楽器音成分の継続時間により、落ち着き性を判定する。実験により、落ち着いた区間で流れている音楽は、楽器音成分の継続時間が長い傾向があることを確認している。

ショット s_kで楽器音成分の継続時間が長いことを判定する閾値を Th_{inst_C}[秒]とすると、音楽による落ち着き性を表す値 $MV_C(s_k)$を以下のように定義する。

$$MV_C(s_k)=\begin{cases}1, & if\ \dfrac{1}{10}\displaystyle\sum_{i=-5}^{4}IL(s_{k+i})>Th_{inst_C}\\ 0, & othewise\end{cases}$$

ただし、Th_{inst_C}は実験により求めた値で 1.40[秒]とした。

3.3.4　落ち着き性

以上で求めた各特徴による落ち着き性を表す値に基づき、ショット s_kでの落ち着き性を表す値 $Calm(s_k)$を以下の式のように定義する。以上で求めた 3 つの値に基づき、ショット s_kでの落ち着き性を表す値を求めるが、ある要素のみが必ず落ち着いた区間に表れるのではなく、各要素が満たされる可能性があるため、各要素の平均を求め落ち着き性を表す値としている。

$$Calm(s_k)=\frac{1}{3}\{SLV_C(s_k)+VTIV_C(s_k)+MV_C(s_k)\}$$

4.　要約映像の生成手順

4.1　強調された区間の抽出

3.1.4, 3.3.4 節で求めた、アクション性を表す値、落ち着き性を表す値に基づき、あるショットの値と、それ以前の 5 ショットの値の平均とを比較して高い値となっている区間を一連の話のつながったアクション区間、あるいは落ち着いた区間として抽出する。ただし、緊迫した区間に関しては、4.2 節で述べた条件に合致する区間のみとする。

区間を求める際、アクション性を表す値、落ち着き性を表す値において、注目する点とその前後の 2 点ずつを合わせた 5 点の平均をとり平滑化を行うことで、値の大まかな変動をみて、区間の抽出を行う。

4.2　主体の検出

画像内に輝度の変化が周囲と異なり、強調されたオブジェクトが存在する場合、そのショットは内容を伝える上で強調されているため重要である。各ショットの先頭フレームの輝度画像に対して、256 階調から 16 階調へ階調を落とすと、複雑なオブジェクトが存在する部分は画像上でエッジ密度が高くなる。この部分を主体が存在する可能性の高い部分とし、その部分に対し色相により領域分割を行う。その結果、類似した色相領域がその部分の 15%以上を占める場合、そのショットに主体が存在すると判定する。

4.3　区間の従属関係の検出

性質の異なる区間が連続している場合、それらは原因と結果を表す従属関係となる。よって、それらの関係を検出することにより、話の文脈を考慮することが可能となる。

原因と結果を表す映像区間には従属関係があるが、性質は異なっているため、それらの区間を同時に要約映像に採用することにより、印象を強めることができる。前後の区間の性質の差に着目し、アクション性を表す値、あるいは落ち着き性を表す値の平均値の差を求め、従属関係の度合いとする。従属関係の度合いを求めることで、編集上強調された区間と従属関係にある前後の区間のどちらから、要約映像に採用するかを決定する際の手がかりとする。これによって、より編集上強調された区間と従属関係が強い区間を要約映像として採用することが可能となる。

4.4　要約映像の生成

以下に要約映像の生成手順を述べる。要約映像の制約時間は、視聴者が 5, 10, 15, 20, 25, 30 分のいずれかを指定することにより決定される。

(1)　映画を n(= 20)等分する

(2)　指定された制約時間に対して、n 分割した各映像

区間で、アクション区間、緊迫した区間、落ち着いた区間の各々から要約映像として採用する映像の制約時間を決定する

(3) アクション区間を中心にして(2)で決定した制約時間を満たすまで(3.1), (3.2)の処理を繰り返す

(3.1) *n* 分割した映像区間の中で最大のアクション性をもつショットを含むアクション区間を求める

(3.2) (2)で決定した制約時間を満たすまで(3.2.1), (3.2.2)の処理を繰り返す

(3.2.1) (3.1)で求めたアクション区間からアクション性を表す値が高い順に、主体の存在するショットを要約映像として採用する

(3.2.2) そのアクション区間と従属する区間を従属度の高い順に参照し、その区間の中から区間の性質を表す値が高い順に、主体の存在するショットを要約映像として採用する

(4) 緊迫した区間を中心にして(2)で決定した時間を満たすまで(4.1), (4.2)の処理を繰り返す

(4.1) *n* 分割した映像区間の中で、最大の緊迫性をもつ緊迫した区間を求める

(4.2) (2)で決定した制約時間を満たすまで(4.2.1), (4.2.2)の処理を繰り返す

(4.2.1) (4.1)で求めた緊迫した区間を要約映像として採用する

(4.2.2) その緊迫した区間と従属する区間を従属度の高い順に参照し、その区間の中から区間の性質を表す値が高い順に、主体の存在するショットを要約映像として採用する

落ち着いた区間に関しては、アクション区間の要約映像の生成手法（手順(3)）と同様であるため、説明は省略する。ただし、手順(3), (4)において、制約時間を満たすまで処理を繰り返し行う際、一度要約映像として採用した区間は、次回からは候補から外す。

5. 実験と評価

以下に示す5本の映画に対して実験を行った。なお、使用したビデオデータの形式は、フレームサイズ 160×120[pixel]、フレームレート 30[frames/sec.]、24 ビットカラー、オーディオ形式はサンプリング周波数 22.050[kHz]、量子化 8 ビット、モノラルである。

- 「スピード 2」
  ヤン・デ・ボン監督, 1997 年, アクション
- 「ザ・ロック」
  マイケル・ベイ監督, 1996 年, アクション
- 「マイノリティ・リポート」
  スティーブン・スピルバーグ監督, 2002 年, SF／サスペンス
- 「A.I.」
  スティーブン・スピルバーグ監督, 2001 年, SF／ドラマ
- 「ダイヤル M を廻せ！」
  アルフレッド・ヒッチコック監督, 1954 年,サスペンス

5.1 編集上強調された区間の抽出

映画の始まりから 500 ショットを対象として、本手法によるアクション区間、緊迫した区間、落ち着いた区間の抽出結果と、主観によるそれらの抽出結果とをショット単位で照らし合わせて区間の抽出精度を算出した。表 1～3 にその結果を示す。ただし、*Nd* は提案手法によって抽出されたショット数、*Nc* は正しく抽出されたショット数、*Na* は主観評価により求めた区間のショット数とする。

表 1　アクション区間の抽出精度

	アクション区間	
	Precision *Nc*/*Nd*	Recall *Nc*/*Na*
Speed2	0.98(83/ 85)	0.70(83/119)
TheRock	0.83(100/120)	0.74(100/136)
MinorityReport	0.57(77/135)	0.84(77/ 92)
A.I.	0.75(43/ 57)	0.90(43/ 48)
DialMforMurder	0.00(0/ 31)	(*1)

(*1)：*Na* = 0 となる

表 2　緊迫した区間の抽出精度

	緊迫した区間	
	Precision *Nc*/*Nd*	Recall *Nc*/*Na*
Speed2	0.81(17/ 21)	0.71(17/ 24)
TheRock	0.58(28/ 48)	0.93(28/ 30)
MinorityReport	0.59(22/ 37)	0.32(22/ 69)
A.I.	0.70(57/ 81)	0.77(57/ 74)
DialMforMurder	0.48(26/ 54)	0.51(26/ 51)

表 3　落ち着いた区間の抽出精度

	落ち着いた区間	
	Precision *Nc/Nd*	Recall *Nc/Na*
Speed2	0.95(370/391)	0.99(370/372)
TheRock	0.93(295/318)	0.89(295/333)
MinorityReport	0.84(270/322)	0.84(270/321)
A.I.	0.93(320/343)	0.87(320/368)
DialMforMurder	0.93(374/401)	0.87(374/431)

アクション区間に関しては、全体的に高い精度で区間の抽出を行うことが出来ている。しかしながら、カメラワークの多用により画像の動きが激しくなってしまう場合に、誤検出される場合がある。

緊迫した区間の Recall の精度が下がっている原因は、音楽やその場面の雰囲気から緊迫していると感じられる場合があり、映画の文法に示されているようにショットの長さが徐々に短くなっていくといった特徴が見られないからである。本研究では、その場合は検出できないため、それに対する改善手法を考察する必要がある。また会話のシーンでは、セリフの長さによってショットが切り替わるため、ショットの長さが徐々に短くなり、緊迫した区間として誤検出される場合が多くみられた。この場合に関しては、音声とショット構成より会話シーンを検出し、この影響を小さくする必要がある。

落ち着いた区間に関しては、よい精度で抽出することが出来ている。

5.2　従属関係にある境界の検出

本手法によって検出された編集上強調された区間の境界と主観によって検出した区間の境界を照らし合わせ、それらの境界が一致している割合を算出した。その際、4 ショット以内の境界のズレは認める。これによって、どれだけ正確に従属関係にある境界が検出できているかを確認することができる。表 4 にその結果を示す。ただし、*Ld* は提案手法によって検出された従属関係にある境界の数、*Lc* は正しく検出された従属関係にある境界の数、*La* は主観評価により求めた従属関係にある境界の数とする。

表 4 において、比較的精度は高くなっているが、これは区間の抽出に大きく依存するため、精度が低いところに関しては区間の抽出精度を向上させる必要がある。

表 4　従属関係にある境界の検出精度

	Precision *Lc/Ld*	Recall *Lc/La*
Speed2	0.90(19/ 21)	0.95(19/ 20)
TheRock	0.57(17/ 30)	0.71(17/ 24)
MinorityReport	0.78(28/ 36)	0.90(28/ 31)
A.I.	0.63(22/ 35)	0.96(22/ 23)
DialMforMurder	0.36(11/ 30)	0.73(11/ 15)

5.3　要約映像の印象評価

大学生 6 名の被験者に、本手法により生成された要約映像と内容、文脈ともに考慮せずに生成された要約映像とを見比べてもらい、どちらの方が、映画の内容、話の流れが理解しやすい要約映像となっているかを評価した。

5.3.1　比較対象とする映像要約手法

比較対象として、以下のようなカットの頻度による要約映像を作成した。映画の先頭から 5 秒毎のフレームに対して、そこから 10 秒間に含まれるカットの数を求める。この 10 秒間に含まれるカット数が最も多いフレームから順にキーフレームとする。ここでキーフレームとは、要約映像を生成する際に着目するフレームのことである。キーフレームが含まれるショットを先頭ショットとして、先頭ショットから合計時間が 10 秒を越えるまでのショットを連結し、要約映像として採用する。要約映像の時間長が目的の時間に達するまでその処理を繰り返し、選択した区間を時間順に並べることで要約映像とした。

この手法により生成された要約映像は、ショットの長さが短く、映像として印象の強い区間のみをつなぎ合わせた映像となる。本研究では、ショットの長さが短い区間も考慮しているが、それ以外に編集上強調された区間も抽出し、さらにその区間と従属関係のある区間も検出することによって、映画の内容と話の流れを理解しやすい要約映像を生成している。ここで示した内容と文脈ともに考慮していない要約生成手法と本手法を比較することによって、本手法の方がどれほど映画の内容と文脈を理解しやすい要約映像が生成できているかを評価するために、この手法を比較手法として採用した。

5.3.2　評価方法

2 本の映画について本手法による 5 分と 10 分の要約映像と比較手法による同時間の要約映像を被験者に観

てもらい、話の内容の理解しやすさ、話の流れの理解しやすさの 2 つの観点について 5 段階評価をしてもらった。5 段階の内訳は、5 が本手法の方がよい、4 がどちらかといえば本手法の方がよい、3 がどちらともいえない、2 がどちらかといえば比較手法の方がよい、1 が比較手法の方がよいである。本研究の用途として、蓄積した映画のブラウジングや、映画評論家が話の内容を思い出す際での利用を想定しているため、どちらも映画を一度観たことがある人を対象としている。よって、本実験で用いた映画を観たことがない被験者に対しては、あらかじめ映画のあらすじを読んでもらうことによって、ある程度話の内容を理解してもらった上で実験を行った。

評価結果を図 4 に示す。図では、6 名の平均評価値をプロットしている。

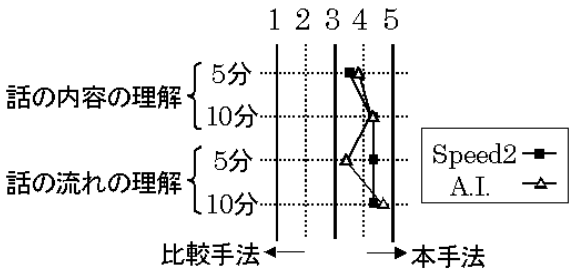

図 5　印象評価結果

全体的に本手法での要約映像の方が、話の内容、流れともに、理解のしやすい要約映像を生成することができている。本手法では、編集上強調された区間としてアクション区間、緊迫した区間、落ち着いた区間を抽出し、それに従属する区間も求めて要約映像を生成しているため、比較手法より話の内容、流れともに理解のしやすい要約映像が生成できたと考えられる。しかしながら「A.I.」の話流れの理解についての 5 分の要約映像では、少し評価値が下がっている。この映画は、ドラマ性の高い映画であるため、落ち着いた区間が長く抽出される場合が多い。そのため、その区間の中から要約映像としてショットを採用したとしても制約時間が短いため、その区間から 1 ショットのみ採用されてしまうといった場合がある。そのため、話の流れの理解に関しては少し評価値が下がっていると考えられる。

6.　まとめ

本研究では、映画の内容と文脈を考慮することにより、話の内容がより理解しやすい要約映像を生成する手法を提案した。映画の文法に基づき、アクション区間、緊迫した区間、落ち着いた区間を抽出することによって、内容が効果的に伝わるように編集上強調された区間を要約映像に含めることが可能となる。さらに、それらの区間との従属関係を求めることにより、前後の話のつながりもあまり失うことなく、要約映像を生成することが可能となる。

今後の課題として、区間の抽出精度を向上させるため、さらなる原因の追求と改善手法の考察が挙げられる。会話シーンの検出を行うことで緊迫した区間の誤検出を少なくすることである。また本研究では、効果音に対して考慮はしていないが、映画の要約映像を生成する上で、効果音も重要な要素と考えられるため、それも要約映像を生成する要素に含めることを考えている。

参考文献

[1] R. Lienhart, S. Pfeiffer, W. Effelsberg, “Video Abstracting”, Communications of the ACM, Vol. 40, No. 12, pp. 55-62, Dec. 1997.

[2] 森山剛, 坂内正夫, “ドラマ映像の心理的内容に基づいた要約映像の生成”, 電子情報通信学会論文誌, Vol. J84-D-II, No. 6, pp. 1122-1131, Jun. 2001.

[3] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, Mingjing Li, “A User Attention Model for Video Summarization”, Proc. of ACM Multimedia, pp. 533-542, Dec. 2002.

[4] Yihong Gong, Xin Liu, “Summarizing Video by Minimizing Visual Content Redundancies”, IEEE International Conference on Multimedia and Exposition, pp. 788-791, 2001.

[5] 加藤和也, 吉高淳夫, 平川正人, “文脈を考慮に入れた映画の要約作成”, 情報処理学会研究報告, Vol. 2002, No. 25, pp. 25-30, Mar. 2002.

[6] ダニエル・アリホン著, 岩本憲児, 出口丈人訳, “映画の文法”, 紀伊國屋書店, 1980.

[7] 阿久津明人, 外村佳伸, “投影法を用いた映像の解析手法と映像ハンドリングへの応用”, 電子情報通信学会論文誌, Vol. J79-D-II, No. 5, pp. 675-686, May 1996.

[8] 川崎智広, 吉高淳夫, 平川正人, 市川忠男, “映画における音楽、効果音の抽出及び印象評価手法の提案”, 信学技報, MVE97-96, pp. 23-29, 1998.