ドラマ映像の心理的内容に基づいた要約映像の生成

森山　剛[†]　　坂内　正夫[†]

Video Summarization Based on the Psychological Unfolding of a Drama

Tsuyoshi MORIYAMA[†] and Masao SAKAUCHI[†]

あらまし　本論文では，テレビドラマ映像の心理的内容に基づいた要約映像の生成手法を提案する．従来の映像要約の研究では，カットなどの映像の構造を利用する方法やビデオ中のオブジェクトに基づく方法などがほとんどであった．一方テレビドラマに対しては，映像コンテンツの心理的な展開（盛り上がりなど）を視聴者は重要視していると考えられるが，映像の心理面に基づいた方法はまだ提案されていない．テレビドラマは，物理的構成要素としてトラック構造（カッティング，登場人物のセリフ，BGM，効果音）を有しているが，演出家などは特定の心理的な印象を表現するために経験的知識に基づいてその時間的構造を巧みに操作していると考えられる．そこで本研究では，この経験的知識に基づき，心理的に重要な箇所をそれに対応するトラック構造の時系列パターンとして検出し，要約映像を生成する手法を提案する．実験において本手法をテレビドラマ映像に適用した結果，約 14 分の 1 の時間長に圧縮でき，更に主観評価実験の結果，本編内容の保存性能及び切出しの自然性に関しても本手法の有効性が示唆された．

キーワード　マルチメディア，要約，トラック構造，テレビドラマ，心理的展開

1. まえがき

ディジタル放送やホームサーバの標準化並びに実用化などの動向を見ると，番組コンテンツとして大量の映像データが一般視聴者に提供されることは近い将来に現実となる．これは結果として，一般視聴者に膨大な量の映像データの中から得たい情報を選択することを強いることとなると考えられる．

その際，映像データを何らかの形で整理したり，記述を付加して従来のテキストベースのデータベース技術を適用できるようにするなどの技術が有効であると考えられ，従来より様々な研究が行われている [1]〜[5]．Rui ら [6] は，映像に関する研究を四つのカテゴリーに分類している．このうち，映像の要約に関する研究では，代表的な手法として，映像をショットに分割し，各ショットにおける代表フレームを決定するというような映像の物理的構造に基づく方法 [7], [8] や，オブジェクトの追跡を行って，その物理的特徴量や意味に対して付加する記述を利用するというようなオブジェクトに基づく方法 [9]〜[16]，また映像を意味的に構造化し，その構造に基づく方法 [17], [18] などがあった．これらの大半においては，ビジュアル情報のみが扱われており，オーディオ情報のうち，特に登場人物の音声に関しては，キャプションとして [4] やシナリオに時刻を付与するため [19]，またビジュアル情報に対して補助的に [20], [21] 用いられていた．これは映像の意味的な記述を付加することにより，映像の構造化，可視化，検索などを行うことを目的としていたためであると考えられる．

一方映像のうち，映画やテレビドラマに類するもの（以下，ドラマ映像）について考えると，製作者はシナリオをもとに，セリフ，BGM，効果音，カメラワークなどの編集要素を巧みに操ってシナリオの意味的内容の説明に加えて，心理的内容の展開をも演出していると考えられる．シナリオの意味的内容とは，人物名，出来事，人物間の関係などを指し，心理的内容とは，クライマックスや緊迫感などを指している．従来の研究は主に前者に関する記述を付加，利用することを目的としていたが，後者に関しては，映画の文法 [22] に基づいて映像の検索を行う研究はあったが [23], [24]，心理的な側面に注目して映像の要約を行った研究はいまだなかった．ここで視聴者の視点に立つと，ドラマ映像を視聴する際にはその後味や心理的印象に高いプラ

[†] 東京大学生産技術研究所，東京都
Institute of Industrial Science, University of Tokyo, 7–22–1 Roppongi, Minato-ku, Tokyo, 106–8558 Japan

イオリティをおいていることが多いと考えられる．従来の意味的アンカーに加え，心理的内容に関するアンカーを映像に付与することができれば，このような心理的側面からドラマ映像の構造化，可視化，検索などを行うことができ，視聴者にとってより自然でより本質的と感じられる映像技術に結び付くと考える．そこでドラマ映像における心理的内容の演出について考えると，前述の編集要素を，いわば製作者の経験的知識に基づいて構成することによって行っていると考えられる．このドラマ映像の印象がその物理的な構造と関係している点については Vasconcelos [16], [25] らも指摘している．この経験則を特定の心理的印象とトラック構造の時系列パターンとの対応関係に関するものであると仮定すると，第 1 段階として，ある短時間の映像に対する視聴者の印象とそのトラック構造との関係を学習し，次に第 2 段階では学習した特徴パターンを用いた応用演算として，パターンマッチングにより映像中に心理的記述を付加するなど，計算論的アプローチが可能となる．このように，従来扱われてきた情報に加えて，心理的な展開を表現するのに適したキューを映像の要約に応用できるような機能を提供することは有効であると考えられる．

本論文では，ドラマ映像の心理的に重要な箇所を検出して要約映像を生成する手法を提案する．次章ではまず，対象とするドラマ映像について明らかにした上でドラマ映像における心理的演出と物理的な構造の関係について述べ，**3.** で本手法について詳細に述べる．**4.** で実験における本手法の実際のテレビドラマ映像への適用とその評価について述べ，**5.** で本手法に関する検討を行っている．

2. 理　　論

2.1 対象とするドラマ映像

前章でドラマ映像に言及したが，これに類するものとして，文献 [22] ではニュース映画，ドキュメンタリー映画，劇映画（fiction film）を挙げている．本研究では当面，これらのうち劇映画に属するもの，特に日本のテレビで放送されているトレンディードラマを対象とする．これは映像リソースを容易に収集できるだけでなく，心理的印象を日本人を対象に観測するために邦画が適していると考えたためである．また，ここで提案する枠組みは，映像の構造という物理刺激に対する視聴者の心理的応答の関係に関するものであるから，他のドラマ映像にも応用が可能であると考えられる．

2.2 ドラマ映像のトラック構造

ドラマ映像のもととなるものとしてシナリオがある．シナリオは，登場人物のセリフ，ト書き，場面転換の指定，音響効果と音楽の指定，カメラの指定からなっている [26]．シナリオというテキスト形式から視聴覚メディアを駆使した形式のドラマ映像を生成する際の要素は大きく二つ，*mise-en-scene*（舞台装置，照明等）と *montage*（編集要素）に分けられる．ここで後者の *montage* としては，シナリオの指示には現れないカッティングにおける効果，BGM，効果音，セリフの間（ま）といった要素があり，これらを物理的搬送波として意味内容及び心理的内容が伝達されると考えられる．図 1 に，これらの要素（ここではトラックと呼ぶ）それぞれに関して時間を横軸にトラックの現れる範囲を方形で描画したドラマ映像のマルチトラック構造を示す．図 1 では，トラックの内容（例えば BGM で使用している楽曲の種類，音量など）は無視し，トラックが現れるか否か（on/off）の 2 値的な状態のみを記述している．

本研究では，ドラマ映像中のトラックが現れているという状態よりも，トラックが現れるか否かという状態の変化に注目している．これは，視聴者の直観にはトラックの内容（BGM における楽曲のジャンルやセリフの言葉の意味等）の分析を経るプロセスよりも，トラックの出現，消失という状態の変化を知覚するプロセスの方が 1 次的に作用すると考えたからである．

従来の研究では，映像のトラック構造に言及したものはあったが，主たる目的が映像の意味内容の記述，オブジェクトの動きや映像を構成する類似部分を一つの定常状態とみなした映像の構造モデル化であったために，状態の変化という物理的側面に注目したものはなかった [4], [8].

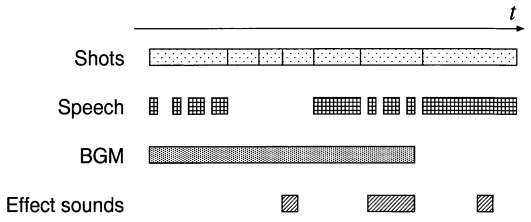

図 1　ドラマ映像のトラック構造
Fig. 1　Track structure of the drama video.

2.3 トラック構造による心理的展開の演出

ドラマ映像における様々な演出の手法，*montage* に関して，文献 [22] では主にカメラワークについて述べられているが，従来の知見のほとんどが視覚的なもののみであり，聴覚的なものと併せて体系的に一般化された明文規則はいまだない．

また，心理的な演出には 2 種類存在すると考えられる．一つは，視聴者間一般に共通する心理的応答に裏づけられたプリミティブな演出の体系であり，もう一つは，芸術表現や自己発現に用いられるような，個性化された，ある場合には記号化された演出体系である．前者の例としては，カッティングを頻発することである種の緊迫感が演出される場合や，大音量の効果音の出現によって驚きが演出される場合などがあり，後者の例としては，特定の演出家の作品の愛好家に共有された申し合わされた演出体系がある．映像のオーサリングなどの映像の個人編集に向けたアプリケーションなどでは後者が該当すると考えられるが，本研究では，ドラマ映像の要約という視聴者に共通的に用いられるアプリケーションの性質上，前者の演出体系を対象に考える．すなわち，ここでは図 1 のトラック構造の時系列パターンの主に変化部分と視聴者間に共通するようなプリミティブな心理的印象との対応関係に注目する．更にここでは簡単化し，心理的な印象の種類（「悲しい」や「盛り上がり」等）には言及せず，印象の強弱による重み付けのみを扱うこととする．

2.4 心理的な演出における経験則

特定のトラック構造が及ぼす心理的印象についての経験則を獲得する方法には次の二つが考えられる．

（1） 放送されているドラマの予告編に選択されている箇所を心理的重みの大きい場所とみなす方法

（2） 主観評価や発見的手段による方法

前者は i) ビデオ成分とオーディオ成分の時刻は必ずしも同期していない，ii) 予告編と本編それぞれを構成する各ショットの時刻順が必ずしも一致しない，更に iii) 予告編と本編は別々に編集されており，セリフなどが異なっている場合もあることがわかった．そこで基準とするのが困難と考え，ここでは後者を基準と考えることとする．

3. 提案するドラマ映像要約法

3.1 本手法の概要

図 2 に提案するトラック構造に基づくダイジェスト生成手法の概要を示す．ドラマ映像から各トラックの on/off 時刻を検出し，これをトラック構造とする．次にこれより近傍の統計量を計算し．各トラックごとに時系列データに変換する．この時系列データに対して，特徴的なトラック構造パターンとのマッチングを行い，経験則から決定される心理的に重要な箇所を代表区間として選択する．選択された代表区間から要約映像を生成する．

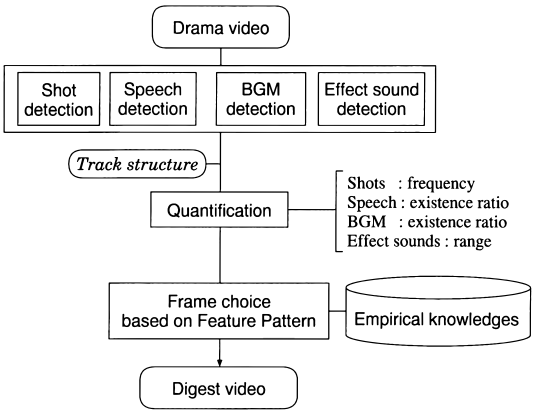

図 2 提案するドラマ映像要約法
Fig. 2 The proposed video summarization method.

3.2 トラック構造の抽出

図 2 では，最初にそれぞれのトラックを抽出する処理を示しているが，MPEG-4 ではこれらのトラックが AV オブジェクトとして別々に伝送され，STB（Set Top Box）側でそれを受信した後にシーン記述に基づいて再構成される枠組みについて標準化している．本研究ではこれを前提とし，ここではトラック分離に関してはすべてマニュアルで行うこととする．

カメラの切り換わる点をカットと呼び，カットで挟まれた，すなわち 1 台のカメラで撮影された部分をショットと呼ぶ．また，ショットを切り換える手法としては，通常のカットのほかに，ディゾルブ，オーバラップ，フェードイン，フェードアウト，ワイプなどがあるが，本研究ではショットの切換の頻度に注目するため，これらにおいて次のショットへ遷移する中間点をカット点とすることとする．このカット点の時刻のシーケンスをショットトラックとする．

登場人物がセリフを発している箇所の開始時刻と終了時刻の組のシーケンスをセリフ音声トラックとする．音声は厳密に始点と終点を決定することは困難であるが，ここではマニュアルで検出できる精度で検出することとする．なお，文中で間があく場合なども一続き

に話されているものとして扱うが，同一話者であっても間が数秒以上続く場合には，演出上の「無声」を意識していると考えられる．そこで本研究では，間の持続時間が一定のしきい値未満の場合は一続きとみなし，それ以上の場合をセリフ音声の切れ目とする．

BGM 及び効果音に関しては，マニュアルの精度で検出した開始時刻と終了時刻の組のシーケンスを BGM トラック及び効果音トラックとする．また *mise-en-scene* としての BGM，すなわち舞台装置として使用されているものは BGM トラックには含めないこととする．

以上のように抽出されたトラック構造をドラマ映像の物理的な状態記述として以下で用いる．トラック構造の抽出結果は，作業者の感覚の違いによって若干の差異が生じるので，本来は複数の作業者の抽出結果の多数決論理で決定すべきものである．しかし今回の実験の目的は，厳密な抽出結果を得ることではなく，後述する内容要約手法の有効性を検証することにあるので，以降は筆者 1 名の抽出結果を用いることとする．なお以下では，開始時刻から終了時刻までの区間をそのトラックがアクティブな区間と呼ぶこととする．

3.3 トラック構造の統計量への変換

本研究では，ある時刻の映像から受ける印象には，その時刻近傍のトラック構造の統計的な物理特徴が寄与していると仮定している．**3.2**で抽出したトラック構造は，各トラックのアクティブな区間を示しており，心理的印象に特徴的なパターンを見出すためには，ここから近傍の統計的特徴を反映した時系列データに変換する必要がある．そこで，表 1 に示すような統計量への変換を行う．

ショットトラックにおいては，注目時刻近傍 10 秒間のカット数をカウントし，5 秒間ずつシフトして全映像について求める．注目時刻の近傍区間の前後の秒数は等しいものとする．セリフ音声トラック及び BGM トラックにおいては，近傍 10 秒間でアクティブな区間が占める割合を計算し，5 秒間ずつシフトして全映像について求める．そして効果音トラックに関しては，近傍 5 秒間にアクティブな効果音が存在する区間を 1，それ以外を 0 とする 2 値の時系列パターンを求める．

表 1　トラック構造の統計量への変換
Table 1　Statistical quantity of the track structure.

トラック	説明
ショット	近傍 10 秒間のカットの回数
セリフ音声	近傍 10 秒間のアクティブな区間の割合
BGM	近傍 10 秒間のアクティブな区間の割合
効果音	近傍 5 秒間での有無（1/0）

3.4 経験則に基づいた代表区間の決定

3.3で得た各トラックの統計量の時間関数から，経験則に基づいた特徴パターンを抽出することによって代表区間を決定する．

2.4で述べたように，ここでは発見的手法によって経験則を構築するが，本来，各トラックの心理的内容を伝達する上での重要性は，ドラマの種類（ジャンル），またシーンの種類によっても変化すると考えられ，したがってそれぞれに対応した経験則セットが存在すると考えられる．本研究では，これらの体系化については今後の課題とし，まず手始めとして，**2.3**で述べた視聴者間一般に共有される心理的応答に裏づけられた演出の体系に基づいた，ドラマやシーンの種類に依存しないようなプリミティブな規則を導入する．具体的には，筆者による予備実験においてドラマを視聴して観察された結果から，以下のような規則を各トラックにおける心理的重要箇所に特徴的なパターンを与えるものとして用いる．ここで筆者 1 名の感覚に基づいた規則をプリミティブなものであると断定することはできないが，大きくかけ離れていないであろうと判断し，これを用いることとする．

3.4.1 ショットトラック

ショットトラックは，主に緊張感の調節に大きな影響をもっていると考えられ，カットが頻出する箇所はある種の緊迫感を強める効果がある．また予備実験として 10 秒間におけるカットの出現頻度のヒストグラムを 7 ドラマ映像について求めた結果，その形状が 2 項分布によって近似できる可能性が見出された（χ^2 検定で，有意水準 5%ですべて適合）（図 3）．正規分布の場合，非常にまれな発生率のしきい値として，両側危険率が 0.3%となる 3σ（σ：標準偏差）を選ぶことが多い．そこでここでは，分布は異なるが同じ 3σ をしきい値とし，これを超えたフレームをまれにみるカットの頻出箇所として検出する．すなわち，

（1） カットの近傍 10 秒間の累積頻度を 5 秒間ずつシフトしながら，全映像について計算する

（2） 累積度数分布を 2 項分布で近似し，3σ に相当する累積頻度値をしきい値として決定する

（3） 近傍のカット累積頻度がしきい値を超えたフレームを代表区間の先頭フレームとする

3.4.2 セリフ音声トラック

セリフ音声トラックでは，1) 事実を知らせる，2) 人

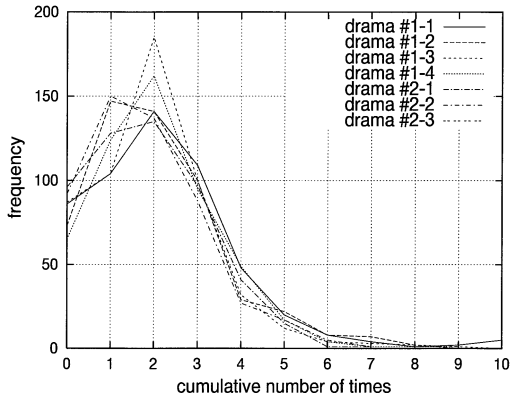

図 3　カット出現頻度のヒストグラム
Fig. 3　Histogram of cumulative number of cuts.

物の心理・感情を表す，3) ストーリーを進展させる，といった機能が知られており [26]，意味的内容の伝達に大きく寄与する．ここでは更にセリフ音声が存在することによって，視聴者の注意を喚起する機能を新たに考慮し，この点について経験則を導入する．セリフ音声では，注意を喚起するセリフの集中する箇所（アクティブな区間）の始端付近と終端付近により重要な箇所が存在する．これは，アクティブな区間の始端付近ではその後会話を継続するための“フリ”に相当する役割を担っており，一方終端付近ではストーリーの転換点（次の展開へのヒント）やそこを含むシーンのそれまでの内容の要約となっていて，視聴者がその後しばらくセリフがなくても展開を追えるようにする役割を担っていることが多いからである．

（1） 近傍 10 秒間の音声区間の占める割合を 5 秒間ずつシフトしながら，全映像について計算する

（2） 得られた曲線上で近傍 10 秒間における回帰直線を求め，傾きの経時特性を計算する

（3） その微分曲線の符号が正から負へ変化した直後のフレーム及び負から正へ変化した直後のフレームを代表区間の先頭フレームとする

3.4.3 BGM トラック

BGM トラックは，BGM が挿入される箇所の雰囲気の演出として（*mise-en-scene* として）用いられるほか，その箇所の心理的な色付けをすることも重要な機能であると考えられる．このように BGM が心理的な色付けをする場合には，アクティブな区間と非アクティブな区間の境界周辺に重要な箇所が存在することが多い．これは BGM はクライマックスなどの直後に，直前のシーンを色付けするために挿入されたり，同じシーン内で BGM が切れた後に重要な箇所が配置されたり，シーンと BGM が同時に終了した場合には直後に新たなシーンが開始したりというようなケースが多いからである．

（1） 全映像で，BGM の存在する箇所を 1，存在しない箇所を 0 とする

（2） 近傍 10 秒間に BGM の存在しない区間で，かつ BGM 部と非 BGM 部の境界となるフレームを代表区間の先頭フレームとする

3.4.4 効果音トラック

効果音トラックは，挿入される箇所に心理的に強烈なインパクトを付加するために用いられることが多く，ここでは表 1 に示した量をそのまま代表区間の選択に用いることとする．また，効果音は特定の登場人物や物などのオブジェクトが映像中に現れる場合に，その登場を知らせるというように記号的な意味で使用されることがあり，その場合には意味的にも重要な箇所であるといえると考えられる．

以上のように決定される先頭フレームをもとに，代表区間を決定する．区間長の設定によって，要約映像全体の長さを変化させることができ，要約の程度に関する要求に応じて変化させることができる．ここでは暫定的に，先頭フレームから 5 秒間を要約映像素片として切り出す．

4. 実　　験

本章では，提案のドラマ映像の心理的側面に注目した映像の要約手法に関して，実験によりその有効性を検証する．

4.1 実験の概要

提案手法を二つのトレンディードラマ（「危険な関係」（以下ドラマ#1）及び「隣人は秘かに笑う」（以下ドラマ#2））に適用し，ダイジェストを生成する実験を行った．これらのドラマは，generic なものを無作為に選択したものであり，以下のような内容である（2000 年 11 月 15 日現在の `http://www.asahi-net.or.jp/~RM1Y-FRSK/index.htm` より引用）．

> #1：「危険な関係」（CX, 1999/10/14〜12/23）：
> 他人に成り代わって別の人生を手に入れた男性と，彼を追いつめやがて愛してしまう女性刑事が繰り広げるサスペンス．タクシー運転手の新児（豊川悦司）は，大手スーパーの横領事件を追う刑事の有季子（藤原紀香）を降ろした後，高校時代の同級生，雄一郎（石黒

賢）を偶然，客として乗せたが…

#2：「隣人は秘かに笑う」（NTV, 1999/10/13〜12/15）：奈緒子（水野真紀）は，夫の邦彦（石橋凌）と娘の奈菜葉（坂野令奈），義母の治子（大谷直子）と平穏に暮らす主婦．最近，料理研究家として雑誌に取り上げられるようになった奈緒子は，近所の主婦のあこがれの的だった．しかし，奈緒子は変質者からの電話に悩まされていた．その矢先，深夜，奈緒子の家に何者かが侵入する．

ここでは，それぞれ全 11 回（11 週），10 回で放送されたものからそれぞれ第 4 回，第 3 回放送分の結果を例として示す．

ドラマ映像の要約法の評価を行う際に問題となる項目として次の 3 点が挙げられる．すなわち，

（1） 時間的長さの短縮率

（2） 内容の一覧性

（3） 内容の了解性，自然性

である．(1) に関しては，要約映像を生成することによって明らかとなるが，(2) 及び (3) に関しては，主観評価実験によって検証する必要がある．そこで，次に生成した要約映像の物理的評価を行った後に，主観評価実験において比較手法との比較を通した評価を行う．

4.2 要約映像の生成結果

抽出されたトラック構造の例を図 4 に示す．四つの方形からなる水平方向に長い短冊の一束は，ドラマ映像 9 分間の 4 トラックを表しており，黒色の方形領域が各トラックのアクティブな区間である．各束は，上からショット，セリフ音声，BGM，効果音トラックに該当する．なお，セリフ音声トラックにおいて間として決定する無声区間のしきい値は，2 秒間とした．

次にこのトラック構造を **3.3**で述べたように統計量へ変換したものを図 5 に示す．

これより **3.4**において述べたように，経験則に基づいて心理的に重要な箇所（代表区間）を検出した結果を図 6 の白色の方形で示す．中間処理の結果として，ショットトラックにおいて，フィッティングした 2 項分布の平均値及び標準偏差は，ドラマ#1 が平均 1.10 回，標準偏差 1.27 回，ドラマ#2 が平均 0.80 回，標準偏差 1.51 回であった．

生成したドラマ映像の時間的長さについて表 2 に示す．ショットトラックで選択された代表区間について見ると，ドラマ#1 では本編の長さが 44’28” であるので約 14 分の 1，ドラマ#2 では本編が 43’12” であるので約 24 分の 1 になっている．

4.3 主観評価実験

ここでは，要約映像の，本編内容に関する一覧性及び了解性，また自然性に関して評価するために，主観評価実験を行った．主観評価実験では主に次の点について検証した．

- grand truth に対する評価

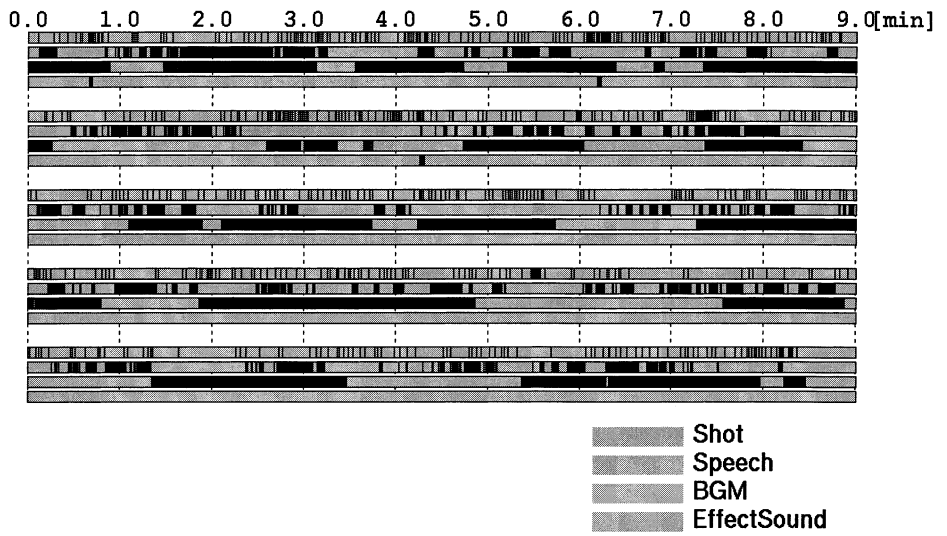

図 4 抽出されたトラック構造の例（ドラマ#1 に関して）
Fig. 4 Example of extracted track structure.

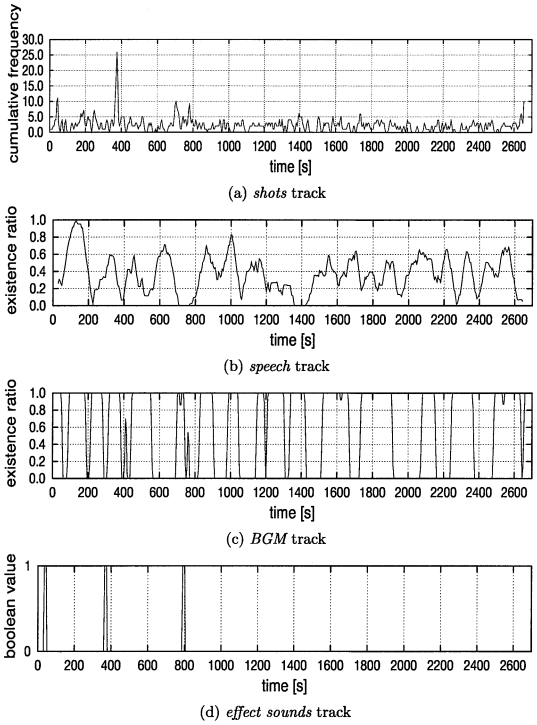

図 5　各トラックの統計量の時間関数（図 4 を変換したもの）
Fig. 5　Extracted time series patterns of each tracks.

- シーン（意味的なショットの集合）境界の保存性
- 切出しの自然性

ここでは比較対象として，一定間隔で代表区間（要約映像の総時間長が提案手法による要約映像と等しくなるようにそれぞれの要約映像素片の長さを決定）を切り出す方法（以下，STD: STandarD method）を用いることとする．提案手法（以下，TSB: Track Structure Based）による要約映像と同じ長さをなすように間隔を選んだ結果，本実験で用いたビデオは 10 fps であり，#1, #2 はそれぞれ 702 及び 1178 フレームごとに代表区間の先頭フレームを選ぶこととなった．

grand truth は，普段からトレンディードラマを視聴している大学生 3 名を被験者として，本編のみを視聴させ，重要とみなされる箇所を列挙させて作成した．本研究では，被験者ごとに grand truth が異なると仮定しており，ここで作成した三つの grand truth の集合のうち，各トラックから生成した要約映像が何%を網羅しているかをもって一覧性に関する評価とする．

また，シーン境界の保存性及び切出しの自然性に関しては，本手法によって生成した要約映像と比較手法

表 2　要約映像の時間的長さ
Table 2　Temporal length of summarization video.

ドラマ#	ショット	セリフ音声	BGM	効果音
#1	3’10”	2’55”	5’55”	0’15”
#2	1’50”	2’40”	4’40”	7’14”

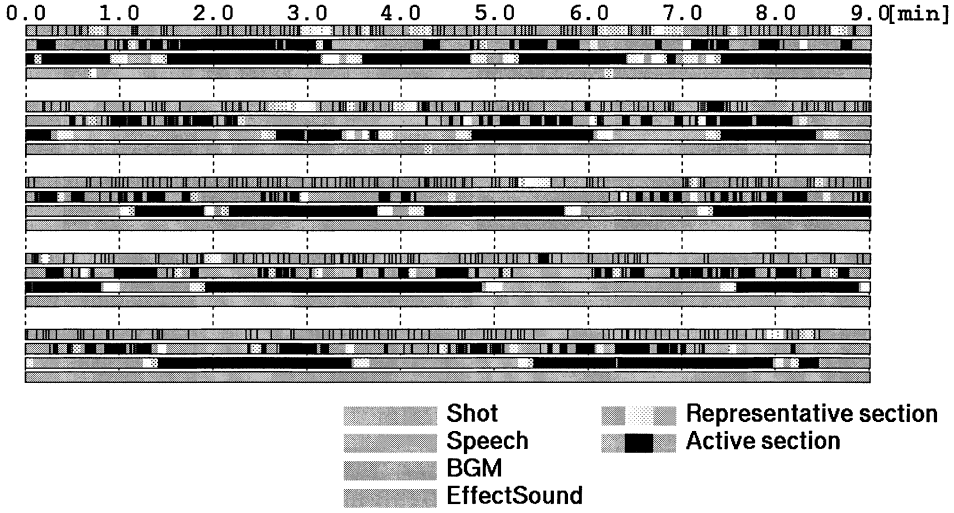

図 6　決定された代表区間の例（ドラマ#1 に関して）
Fig. 6　Example of determined representative frames.

によるものとを，生成した手法を知らせずに被験者 3 名に視聴させ，内観報告としてこれらの点に関してどちらの方法がどのように優れているかを答えさせた．

主観評価実験の結果，本編の内容の一覧性に関して，すべての被験者が TSB が STD より良いと判断した．grand truth に関しては，表 3（ドラマ#1 に関する結果）に示すように，被験者 1 ではショットトラックに現れる部分を比較的重要視しており，これに対して被験者 2 ではセリフ音声，被験者 3 ではセリフ音声及び BGM に現れる部分をより重視していると考えられる．これより，被験者間で重要な箇所として注目する側面が異なる可能性が示唆される．またショットトラックが心理的緊迫感の演出に関する寄与が大きく，セリフ音声が意味の伝達に関する寄与が大きいとすると，被験者によって心理的なものに注目する場合と意味的なものに注目する場合があると考えられ，すなわち，本研究の心理的な側面と従来行われてきた意味的な側面との両方を考慮することが有効であると考えられる．

シーン境界の保存性では，TSB はシーンの境界付近で区切る性質があることが明らかとなった．これは，TSB ではショットの出現頻度にしきい値を設けており，ショットの出現頻度とシーンを構成するショットの長さの間に相関があるためであると考えられる．すなわち，ショット長は長短分布しているが，シーンの中央付近ではショット長が短くなる傾向にあると同時にシーンの境界付近ではショットの長さが比較的長くなるために，しきい値を超えた箇所を抽出するとシーンの中央付近を切り出す結果となり，シーン境界をある程度保存するように切り出されると考えられる．

更に，切出しの自然性に関する内観報告では，STD では不自然な切出しが目立ったと評価されたが TSB はセリフ音声などの境界などを全く考慮していないにもかかわらず，自然な切出しを行えていると評価された．これは前述のように，シーン境界をある程度保存するため，セリフ音声や BGM などの切れ目もまた保存しているためと考えられる．

表 3　grand truth に対するカバー率（%）
Table 3　Cover ratio for the grand truth.

トラック	被験者 1	被験者 2	被験者 3
ショット	71	22	53
セリフ音声	57	89	76
BGM	57	56	71
効果音	14	22	0

5. 検　　討

本手法は，心理的な重要度に注目してドラマ映像を要約するものであるが，これには 2 通りのアプローチが存在する．すなわち一つは，心理量の個人によるばらつきを考慮し，個人に特化した心理的演出パターンに注目する，というものであり，もう一つは，統計による視聴者の平均的な印象に対応した心理的演出パターンに注目するというものである．これらは工学的にはアプリケーションに依存しており，前者は映像のオーサリングなどにおいて，複数の演出家の演出手法を選ばせる，というような場合に有効であり，後者は本研究で対象としているような一般視聴者向けの映像要約システムなどで有効である．本研究のアプリケーション例として，図 7 に示すような EPG（Electronic Program Guide）に実装した場合には，カーソルで指定されたドラマ番組の要約がサブウィンドウに表示され，ユーザは番組の本編の内容について，盛り上がり部分についての雰囲気をも知ることができる．この際，番組の宣伝を目的としている場合，物語の結末については隠す工夫が必要であるため，前述のように本手法を意味的内容に基づく方法と組み合わせる際にこの点を考慮する必要がある．

また，本研究では主観評価実験において grand truth と比較する方法をとったが，逆に grand truth を用いて重要な箇所のもつ物理的特徴を学習する，ということが考えられる．しかし実験の結果，被験者ごとに異なる grand truth が得られ，トラックや意味的内容あるいは心理的内容といった側面に関して注意を向ける点にあいまいさが観測された．したがって，grand

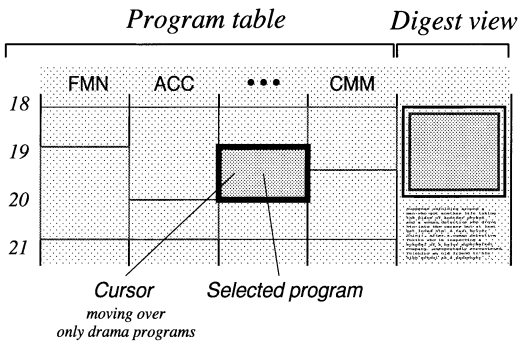

図 7　EPG に本手法を応用した場合の画面例
Fig. 7　Application example of the proposed method for EPG.

truth を学習の教師とするためには，被験者への教示において注目する点を限定したり，被験者の分類を行ったりする必要があると考えられる．

6. むすび

本論文では，ドラマ映像の要約を生成する新たな手法として，ドラマ映像の心理的内容に基づいた方法を新たに提案した．

本手法は，ドラマ映像の物理的様相としてのトラック構造に現れる心理的演出のパターンと映像が伝える心理的内容の対応関係を用いて，心理的に重要な箇所のみを切り出して要約映像を生成するものである．ショットトラックを基本に本手法を用いて要約映像を生成し，評価実験を行った結果，本編の長さを大幅に要約でき，また単純に一定間隔で代表区間を切り出す方法に比べ，自然性や本編の内容の一覧性をより保持し得ることが示された．

今後の課題として，心理的演出に関する経験則をより改善し，更に意味的内容との連携を行う方法について検討を行う必要がある．またここでは扱わなかったディゾルブなどの映像効果などについても検討していく．

謝辞 本研究（の一部）は，文部省科学研究費補助金（創成的基礎研究費）・課題番号 09NP1401：“人間主体のマルチメディア環境形成のための情報媒介機構の研究” による．

文　献

[1] F. Idris and S. Panchanathan, “Review of image and video indexing techniques,” J. Visual Communication and Image Representation, vol.8, no.2, pp.146–166, 1997.

[2] C.-W. Chang and S.-Y. Lee, “Video content representation, indexing, and matching in video information system,” J. Visual Communication and Image Representation, vol.8, no.2, pp.107–120, 1997.

[3] H. Mi, Y.I. Yoon, and C.K. Kio, “Unified video retrieval system supporting similarity retrieval,” Proc. 10th International Workshop on Database and Expert Systems Applications, pp.884–888, 1999.

[4] M.A. Smith and T. Kanade, “Video skimming and characterization through the combination of image and language understanding,” Proc. 1998 IEEE International Workshop on Content-Based Access of Image and Video Database, pp.61–70, 1998.

[5] Y. Li, 田中　譲, “メタデータの管理に基づくビデオデータベースの構成,” 情処学論, vol.39, no.4, pp.1137–1145, 1998.

[6] Y. Rui, S.X. Zhou, and T.S. Huang, “Efficient access to video content in a unified framework,” IEEE International Conference on Multimedia Computing and Systems, vol.2, pp.735–740, 1999.

[7] Z. Aghbari, K. Kaneko, and A. Makinouchi, “New indexing method for content-based video retrieval and clustering for MPEG video database,” Proc. International Symposium on Digital Media Information Base, pp.140–149, 1998.

[8] R. Ronfard, “Shot-level description and matching of video content,” Proc. SPIE, vol.3229, pp.70–78, 1997.

[9] B. Günsel, A.M. Tekalp, and P.J.L van Beek, “Content-based access to video objects: Temporal segmentation, visual summarization, and feature extraction,” Signal Processing, vol.66, pp.261–280, 1998.

[10] M.-K. Shan and S.-Y. Lee, “Content-based video retrieval via motion trajectories,” Proc. SPIE, vol.3561, pp.52–61, 1998.

[11] F. Coudert, J. Benois-Pineau, P.-Y. Le Lann, and D. Barba, “Binkey: A system for video content analysis “on the fly”,” Proc. IEEE International Conference on Multimedia Computing and Systems, vol.1, pp.679–684, 1999.

[12] A. Pope, R. Kumar, H. Sawhney, and C. Wan, “Video abstraction: Summarizing video content for retrieval and visualization,” Conference Record of Thirty-Second Asilomar Conference on Signals, Systems and Computers, vol.1, pp.915–919, 1998.

[13] M. Gelgon and P. Bouthemy, “Determining a structured spatio-temporal representation of video content for efficient visualization and indexing,” Proc. 5th European Conference on Computer Vision, vol.1, pp.595–609, 1998.

[14] D. Zhong and S.-F. Chang, “Spatio-temporal video search using the object based video representation,” Proc. Int. Conf. Image Process., vol.1, pp.21–24, 1997.

[15] H.J. Zhang, J.Y.A. Wang, and Y. Altunbasak, “Content-based video retrieval and compression: A unified solution,” Proc. Int. Conf. Image Process., vol.1, pp.13–16, 1997.

[16] N. Vasconcelos and A. Lippman, “Bayesian modeling of video editing and structure: Semantic features for video summarization and browsing,” Proc. 1998 International Conference on Image Processing, vol.3, pp.153–157, 1998.

[17] V. Roth, “Content-based retrieval from digital video,” Image and Vision Computing, vol.17, no.7, pp.531–540, 1999.

[18] 田中克己, “マルチメディア・コンテンツのアクセスアーキテクチャ,” 情処シンポ論, vol.97, no.11, pp.1–8, 1997.

[19] 谷村正剛, 中川裕志, “ドラマのビデオ音声トラックとシナリオのセリフの時刻同期法,” 情処学知能と複雑系研報, 99–ICS–118, pp.25–31, 1999.

[20] C. Saraceno and R. Leonard, “Audio as a support to

scene change detection and characterization of video sequences," Proc. IEEE International Conference on Acoustic, Speech and Signal Processing, MDSP3L.1, vol.4, pp.2597–2600, 1997.
[21] J. Nam, M. Alghoriemy, and A.H. Tewfik, "Audio-visual content-based violent scene characterization," Proc. International Conference on Image Processing, vol.1, pp.353–357, 1998.
[22] ダニエル・アリホン，岩本憲児，映画の文法, 紀伊国屋書店，1980.
[23] 石井孝和，吉高淳夫，平川正人，市川忠男，"映画の文法に基づくビデオ画像の内容検索," 情処学研報，97-DBS-111, pp.65–72, 1997.
[24] A.M. Ferman and A.M. Tekalp, "Editing cues for content-based analysis and summarization of motion pictures," Proc. SPIE, vol.3312, pp.71–80, 1997.
[25] N. Vasconcelos and A. Lippman, "A spatiotemporal motion model for video summarization," Proc. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.361–366, 1998.
[26] 大木英吉，鬼頭麟兵，鈴木通平，シナリオハンドブック, ダヴィッド社，1993.

（平成 12 年 8 月 25 日受付，12 月 22 日再受付）

森山　剛　（正員）

1994 慶大・理工・電気卒．1999 同大大学院博士課程了．その後，東大生産技術研究所にて日本学術振興会特別研究員 PD を経て，現在，米カーネギーメロン大学 post doctoral fellow．パターン認識，マルチメディア，感性情報処理に関する研究に従事．1998 本会学術奨励賞受賞．また，テノール歌手として種々の演奏会に出演するなど，音楽活動にも従事．工博．日本音響学会，IEEE 各会員．

坂内　正夫　（正員）

1975 東大大学院工学系研究科博士課程了．同年同大工学部電気工学科専任講師，その後，横国大学工学部情報工学科助教授，東大生産技術研究所助教授を経て，現在，同大学生産技術研究所教授．1998 より同大学生産技術研究所所長．マルチメディアデータベースなどの研究に従事．工博．