色情報，発話情報を用いたビデオの自動要約手法

本多 光一郎† 藤村 憲市† 上原 邦昭†

ディジタルビデオの増加にともない，短時間で効率良く内容を把握したいという要求が高まっている．そのためには，内容を損なうことなく，ビデオを時間的に圧縮することが必要である．本稿では，ビデオに含まれる色情報および発話情報を利用して，自動的に要約を生成する手法を提案する．本手法の要約対象としては，ニュース番組などの比較的話題の区切れが明確なものではなく，ある一定のストーリーを持ったビデオを用いている．また，色情報としては画像の色に関する特徴量を，発話情報としては Closed Caption を利用している．生成された要約はオリジナルビデオの内容をほとんど損なうことなく，再生時間も 10〜30%に圧縮されている．

Video Summarization by Using Color and Speech Information

Koichiro Honda,† Kenichi Fujimura† and Kuniaki Uehara†

With increasing digital video, the user wants to grasp contents of video efficiently. We need to reduce the size of video without losing contents to satisfy with such request. In this paper, we propose the method that infers contents of video by using color and speech information of video and makes a summary automatically. Our targets for the summarization is video including consecutive story such as movie, not video such as TV news program which includes relatively clear topic boundaries. We use color frequency histograms of images as color information and Closed Caption as speech information. Generated summaries keep most of contents of original video, and compression rates of the summary are between 10% and 30%.

1. はじめに

近年，ディジタルビデオは数量，サイズともに増加の一途をたどっている．たとえば，ディジタルライブラリの分野では，大量のビデオデータを蓄え，インデックスなどを付加してユーザの効率的な検索を支援している．しかしながら，部分的なビデオデータの検索ではなく，ビデオ全体の概観を目的とした場合，インデックスのみでは，ビデオに登場する人物の関係などを把握することは困難である．このような場合には，長時間のビデオを内容を損なうことなく圧縮する必要がある．すなわち，ビデオの要約である．たとえば，ビデオライブラリにおいては，映像を概観するのみで短時間で内容を把握して，自分の気に入ったビデオを選択，鑑賞することができる．また，デジタル放送の普及にともなうビデオ配信では，ユーザにある程度の内容を理解してもらうために，ビデオの紹介となりうる要約映像を配信するという利用法が考えられる．さらに，今後ホームサーバが普及し，蓄積された大量のビデオから見たいビデオを選択する際，概観するための手段としても活用することができる．

一般的なビデオは，画像の色やカメラモーションなどの視覚的な要素と，登場人物のセリフや BGM などの聴覚的な要素からなる．視覚的な要素は動画像認識によって，聴覚的な要素は音声認識によって解析し，ビデオの内容を把握できれば，要約の作成が可能となる．しかしながら，これらの分野は未成熟であり，実用段階にはない．したがって，すべての視覚的な要素と聴覚的な要素を解析することは困難である．

本稿では，ビデオから色情報と発話情報を抽出し，画像処理と言語処理を行い，要約を生成する手法を提案する．色情報による要約では，ビデオの視覚的な要素のうち，画像の色を利用している．画像の色の変化を利用すれば，ビデオの場面変化を検出し，要約の候補となる区間を取り出すことができる．当然のことながら，色情報だけではビデオの内容を把握するには不十分である．たとえば，登場人物間の関係などは，色情報のみではとらえることができない．したがって，色情報に加えて，聴覚的な要素のうち，登場人物のセリフ，つまり発話情報を利用している．発話情報による要約では，ビデオの脈絡や登場人物間の関係の把握

† 神戸大学大学院自然科学研究科
Graduate School of Science and Technology, Kobe University

に重点を置いている．このようにして，色情報と発話情報が互いに補いながら，要約を生成している．

具体的には，色情報として画像の特徴量を算出し，確率モデルに基づいて画像の色の変化パターンを求めて，要約の候補を選び出している．また，発話情報として，ビデオに含まれる Closed Caption を抽出して会話分析を行っている．Closed Caption は発話の連続であり，通常のテキストのように情景や行動の記述を含まない．また，発話には，人名の繰返しや接続語を用いた冗長な表現が避けられることが多い．このため，発話間のつながりが薄くなり，単語の頻度や類似度を用いた従来のテキスト要約の手法では不十分である．したがって，Closed Caption の性質，すなわち会話の特徴に着目した言語モデルを導入し，会話の一貫性を発見している．言語モデルを導入すれば，単なる単語間のつながりだけでなく，発話間や登場人物間のつながりを要約に反映させることができる．本手法の特徴は，発話情報を積極的に利用し，色情報と統合させることでビデオの内容を重視した要約を作成している点にある．

2. 色情報を用いた要約の作成

2.1 基本概念

ビデオには，ショットと呼ばれる連続区間が存在しており，ショットを複数つなぎ合わせて場面を作成し，さらに場面をつなぎ合わせることにより，作品を組み立てている．場面の中では，たとえば登場人物や背景をほとんど変化させずに，登場人物の存在やストーリーの展開場所を強調している区間がある．このような区間を Unchanged と呼ぶ[1)]．また，個々の登場人物や背景がショットごとに交互に現れる区間では，登場人物同士のやりとり，ビデオ内で繰り広げられている出来事を強調している．このような区間を Multiplexing と呼ぶ[1)]．これらの区間を**図 1** に示す．Unchanged では，ショット $s_1, \ldots, s_j, \ldots, s_n$ がすべて色に関して類似している．また，Multiplexing では類似したショットが交互に出現する．たとえば，sequence 1 において，$s_1, s_3, \ldots, s_{2j-1}, \ldots, s_{n-1}$ はすべて色に関して類似している．逆に sequence 2 に関して，ショット $s_2, s_4, \ldots, s_n$ がすべて色に関して類似している．

ビデオを要約する際には，視聴者に対していかに視覚的に内容を理解させるかが重要である．逆にいうと，ビデオの視覚的な情報を用いれば，要約の候補となる場面を求めることができると考えられる．たとえば，以下のような部分は要約の候補として重要である．

- ストーリーが展開されるステージ

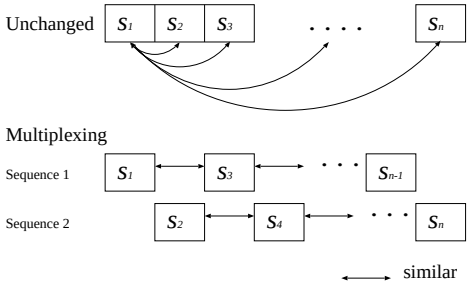


図 1 映像区間の形式
Fig. 1 Types of video intervals.

- 短時間で目まぐるしく画像が変化し，内容的に盛り上がりのあると思われる部分

ここで，ステージとは進行しているストーリーの舞台となっている場所のことである．ビデオでは，ある場所でのストーリーが始まるときには，最初にステージの風景が映し出されることが多い．この風景によって，視聴者はストーリーの展開される場所を知ることができる．ステージとなる場所を映し出す区間は，カメラワークが変化せず色変化が小さいため，Unchanged に相当すると考えられる．

逆に，ビデオ内で目まぐるしく画像が変化するような区間，たとえば戦闘の場面などでは，登場人物が演じる個々のアクションを短時間で交互に出現させて，ビデオ内で繰り広げられているやりとりが醸し出す緊迫感を表現している．このような区間は Multiplexing に相当すると考えられる．また会話の場面では，発話者がショットごとに交互に現れて，人物同士のやりとりを強調していることが多い．このような区間も Multiplexing と考えられる．このように，色情報の変化を考慮すれば，重要であると思われる部分を検出して要約の候補を求めることができる．

2.2 場面分割

本手法では，ビデオ内の場面の変わり目を検出するために，まずカット検出処理を行っている☆．さらに，各ショットのつながり関係，すなわちビデオの内容に則したショット間の関係を調べるために，確率モデルを導入している．確率モデルは，音声認識の研究で広く用いられており，単語列のあとにくる適切な単語の出現パターンを統計的に推測するために利用される．音声と同様に，画像にも色の変化パターンが存在していると考えられる．たとえば，色が類似したショット列からまったく異なる色で表現されるショットに推移

☆ 本研究では，ショットとショットの境界を示すものをカットとしている．

した場合は，場面が変わっている可能性が高い．このように，画像から得られる色変化パターンを学習データとすれば，場面の変わり目かどうかを統計的に推測することができる．

具体的には，まず各ショットから代表フレームの特徴量を算出している．特徴量は式 (1) で与えられる．

$$h_i = \sum_{j=1}^{64} C_j P_j \qquad (1)$$

i 番目のフレームの特徴量 h_i は，色レベル C_j とその画素数 P_j の積の総計である．本研究では，色空間として RGB 空間を用いて色を 64 段階の階級値に縮退している．これは，R，G，B 各々を 256 段階にすると色のとりうる値が非常に広範囲になり，類似した色でも別色として認識してしまうため，処理に大きな制限をかけてしまう恐れがあるからである．色レベル C_j は各階級値に対して設定される値であり，1 から 64 の値を割り当てている．また，代表フレームは，ショット内の連続する n フレームにおいて，特徴量の標準偏差が最小になる区間を選び，区間内のフレームのうち特徴量が中位数となるフレームとしている．

次に，隣接するショットとの特徴量の差分が場面境界を表すものとして妥当であるかどうかを判定するために，確率モデルを利用している．具体的には，大容量のビデオデータを扱うことを考慮して，学習データが少なくても信頼性のある統計的推測をするために，確率モデルのうち N-class モデルを導入している．N-class モデルは，データ集合をクラスと呼ばれるグループに分類し，データの推移を間接的に推測するものである．なお，本手法では，パラメータ推定と精度の観点から $N = 2$ の 2-class モデル[2)]を用いて，直前の特徴量の差分値から現在の特徴量の差分値への変化が場面境界を示す推移となりうるかを推測している．

場面境界でないと推測されるショット間は，内容的に連続していると考えられるため，これらのショットをつなぎ合わせて 1 つの場面を形成すれば，ビデオ全体を複数の場面に区切ることができる．しかしながら，色変化が小さく内容的に連続していると考えられる区間内にも，色変化が突発的に起こるようなショットが含まれている場合がある．たとえば，ある山の風景が映された区間があるとする．ここで，人物のショットが突発的に出現したとすると，この人のショットの前後では色変化が大きくなってしまう．このような場合，隣接ショットとの関係を調べる確率モデルを用いると，誤った場面境界を検出する可能性がある．このため，ショットをつなぎ合わせる際には，数ショット先に特徴量が非常に類似したショットがあるか調べて，もしあればそのショットまでを 1 つの場面としている．このようにして，確率モデルを用いた手法と特徴量の類似を用いた手法で得られる場面の境界が合致する箇所を，場面の境界と設定している．

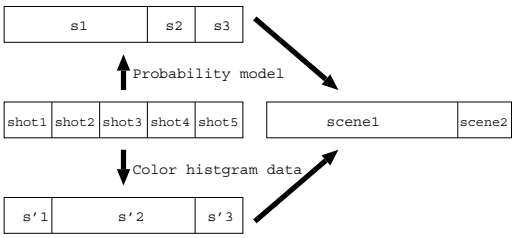


図 2 場面分割の手法
Fig. 2 Scene segmentation approach.

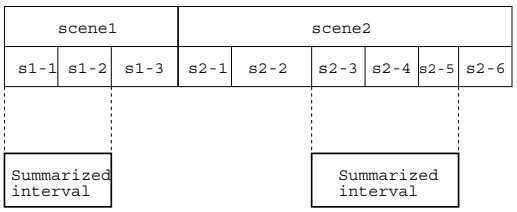


図 3 重要区間の抽出
Fig. 3 Extraction of significant sections.

ここまでの処理の流れを**図 2** に示す．ただし，図中の s1，s2，s3 は確率モデルを用いたときに得られる場面，s'1，s'2，s'3 は特徴量の類似を用いたときに得られる場面である．scene1，scene2 は 2 つの手法で検出される境界が合致する箇所で分割したものである．

2.3 重要な区間の抽出

場面分割が終わると，各場面でストーリー上重要と思われる区間を抽出する必要がある．これらの区間は，各場面内で重要と思われるショット列に相当している．たとえば，**図 3** の s1-1 は scene1 の第 1 番目のショット，s2-1 は scene2 の第 1 番目のショットを示し，Summarized interval は各場面から抽出された重要と思われる区間を示している．2.1 節で述べたように，ビデオにおける出来事や情景を強調している Unchanged，Multiplexing の区間では，特徴的な色変化が存在している．したがって，これらを検出できれば要約の候補となる区間を求めることができる．これらの区間を検出するために，ショットの代表フレームの特徴量の相関関係を用いている．具体的手法を以下に示す．

まず，ある場面において，先頭ショットから順に n 個のショットの代表フレームの特徴量を算出する．n は，各場面の内容を把握するために必要とされる時間に依存した値である．このため，本研究では 1 ショッ

トあたりの平均的な継続時間を約 6 秒とし，実験における被験者の意見から $n=4$ としている．各ショットの代表フレームの特徴量を $h_1, h_2, \ldots, h_n$ とすると，

$(h_1,\ h_2),\ (h_2,\ h_3),\ (h_3,\ h_4),\ \ldots,\ (h_{n-1},\ h_n)$

という $n-1$ 個の組を作成する．これらの組の第 2 項を $h'_i\ (i=1,2,\ldots,n-1)$ として，

$(h_1,\ h'_1),\ (h_2,\ h'_2),\ (h_3,\ h'_3),\ \ldots,\ (h_{n-1},\ h'_{n-1})$

とおく．これらの組を xy 座標平面にプロットすれば，$h_i\ (i=1,2,\ldots,n-1)$ と $h'_i\ (i=1,2,\ldots,n-1)$ の相関関係を求めることができる．各組の第 1 項の平均値を $\bar{h}$，第 2 項の平均値を $\bar{h'}$，さらに第 1 項の標準偏差を s_h，第 2 項の標準偏差を $s_{h'}$，相関係数を r とすると，各値は式 (2)，(3)，(4)，(5) で与えられる．

$$\bar{h}=\frac{1}{n-1}\sum_{i=1}^{n-1}h_i \quad \bar{h'}=\frac{1}{n-1}\sum_{i=1}^{n-1}h'_i \tag{2}$$

$$s_h=\sqrt{\frac{\sum_{i=1}^{n-1}(h_i-\bar{h})^2}{n-1}} \tag{3}$$

$$s_{h'}=\sqrt{\frac{\sum_{i=1}^{n-1}(h'_i-\bar{h'})^2}{n-1}} \tag{4}$$

$$r=\frac{1}{(n-1)s_h s_{h'}}\sum_{i=1}^{n-1}(h_i-\bar{h})(h'_i-\bar{h'}) \tag{5}$$

色変化が小さい Unchanged の区間では，各代表フレーム間の特徴量が類似しているため，上述した組 (h_1,h'_1)，(h_2,h'_2)，…，(h_{n-1},h'_{n-1}) の示す座標は局所的な領域に集まる．また，理想的な Multiplexing の区間では，色が交互に変化するという特徴があるため，各組の第 1 項 h と第 2 項の値 h' に大きな差がある組が多数存在する．したがって，これらの示す座標をプロットすると，全体的に傾きを持つ集合領域を形成する．いい換えると，Unchanged の区間では相関関係が無相関に近いため，相関係数 r は 0 に近い値となる．逆に，Multiplexing では負の傾きを持つ領域を形成するため，相関関係は強いと考えられる．したがって，相関係数 r が -1 に近い値を返す区間が Multiplexing の区間となる．

このようにして，相関係数 r が 0 に一番近い値となる区間と -1 に一番近い値となる区間を選び出す．これを**図 4** に示す．たとえば，Interval 1 の相関係数を $r1$ として，他の Interval も同様に $r2$，$r3$，$r4$ を求めておき，値が 0 あるいは -1 に近い Interval を選ぶ．このようにして，各場面で得られた区間が，最終的な要約の候補となる．

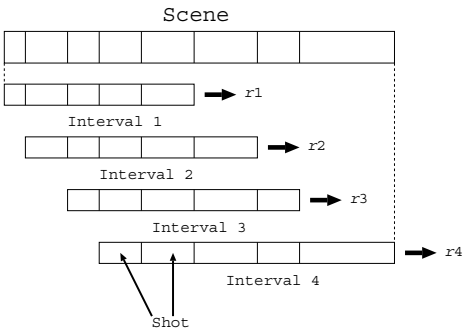


図 4 要約候補の選出
Fig. 4 Selection of candidates.

図 5 Closed Caption の抽出
Fig. 5 Extraction of Closed Caption.

3. 発話情報を用いた要約の作成

映画のようにストーリーを持つビデオでは，登場人物の存在が重要であり，色情報だけで登場人物の情報を推測することは困難である．また，要約には登場人物の会話のやりとりを含めるべきである．このため，本手法では登場人物の発話情報を取得し，会話分析を行って要約の候補となる区間を決定している．

3.1 Closed Caption の利用

本手法では，発話情報としてビデオの Closed Caption を利用している．Closed Caption は，キャプションデコーダを用いて，フレーム番号とともにテキストファイルとして抽出している．Closed Caption には登場人物のセリフや効果音などが記述されている．Closed Caption の抽出例を**図 5** に示す．

Closed Caption は，映画だけでなく，ニュース番組やスポーツ番組，さらにはホームビデオで撮影した作品にも付加することが可能である．特に，アメリカ合衆国などでは聴覚障害者を補助する手段として広く普及している．また，日本でも本年より普及実験が進められている．したがって，今後 Closed Caption が当然の技術として利用されることを想定すると，発話情報として Closed Caption を導入することは，妥当性があると思われる．

我々は，図5のようにして抽出されたClosed Captionを，小説などのようなストーリーを持った1つのテキスト文書であると仮定し，テキスト要約の手法をClosed Captionに適用してセリフの要約を行った．さらに，要約されたセリフの各部分に対応するビデオ区間をつなぎ合わせ，ビデオ全体の要約を試みたが，次節で述べる問題が生じた．

3.2 場面分割の困難性

一般的なテキスト要約の方法は，まずテキストを場面分割し，各場面から重要な文章のみを抽出し，最後にそれらの文章を連結する．したがって，テキストの要約において，場面分割は非常に重要な要素技術である．これまで，テキストの場面分割については多くの研究成果があげられている．Hearst[3]は，テキスト中の単語の出現頻度を用いて場面境界を決定するText-Tilingアルゴリズムを提案している．TextTilngはシンプルかつ領域に依存しないため，小説などの分割には有用である．しかしながら，あらかじめいくつかの話題ごとに区切られたテキストを分割の対象としており，Closed Captionのような連続したテキストには適用が困難である．

また小嶋ら[4),5)]は，辞書を用いてテキスト中の単語間の意味的な類似度を計算し，場面分割に利用している．我々も，電子シソーラスを用いてClosed Captionでの単語間の類似度を計算し，場面分割を試みた．しかしながら，セリフ（Closed Caption）には状況説明的な記述がほとんど含まれないため，意味的に類似した単語を発見することはごく稀であり，場面分割も失敗に終わった．

さらに，Beefermanら[6),7)]は，確率モデルを用いてテキスト中での単語の出現予測を行い，場面分割を行う手法を提案している．しかしながら，確率モデルを用いた手法は大量のコーパスを必要とするため，本手法で利用可能な映画のコーパスを用意することは難しい．このように，Closed Captionを用いてビデオの場面分割を行うことは困難である．

3.3 要約候補の選出

本稿でいうビデオの要約は，ストーリー全体を大雑把に把握できることが第1条件であり，映画の予告編のように，派手な部分のみを取り出して視聴者の関心を引くようなものである必要はない．このため，本手法では発話の連続性に注目し，Closed Captionから内容に一貫性がある会話を取り出し，それらを連結したものを要約の候補としている．これは，本研究におけるビデオの要約では，視聴するユーザの内容理解が重要であり，内容自体に一貫性がなければストーリーの流れがあいまいになる可能性が高いからである．また，色情報によって発見された会話の区間は必ずしも意味をともなったものではない．たとえば，一貫性のある内容を持った会話の区間があるとする．もし区間全体がUnchangedあるいはMultiplexingでないとすると，この会話の区間は未発見となってしまう可能性が高い．また，色変化パターンによる抽出のみでは，会話がまったくない区間が多く発見される可能性もあるため，登場人物間の関係などを知るためには不十分である．したがって，発話の連続性に着目することは非常に重要である．

本手法では，以下のような発話群を，内容に一貫性がある会話としている．また，それぞれの発話群と例をまとめたものを**表1**に示す．

Same speaker 同一の発話者による連続した発話．
理由：同一の発話者による発話は，時間，場所，内容について一定である．

Question-answer 疑問文を含む発話と直後の発話．ただし，疑問文は「?」を含む文とする．
理由：問いには必ず答えが存在し，問いかけられた話者は答えるのが自然である．

Second person 代名詞youの変化形を含んだ発話と近隣の発話．
理由：英語の会話において，you，yourなどの単語の出現は，1人以上の話し相手の存在を示している．つまり，ある話題についての会話が行われていると考えられる．

Conjunction and，thenなどの接続詞で始まる発話と1つ前の発話．
理由：接続詞は，直前の内容を受けて出現するのが自然であり，そこで内容が途切れることはありえない．

Pronoun he，she，his，herなどの代名詞を含む発話と1つ前の発話．
理由：代名詞の出現は，会話での第三者の存在，または直前の固有名詞の存在を意味し，3者に関係した内容の発話であると考えられる．

Co-occur 隣接する発話で単語の共起がある部分．ただし機能語は省く．
理由：単語の共起は，同じ内容についての発話をしている際に生じる．

3.4 要約の生成ルール

色情報，発話情報によって選び出された要約の候補から，生成ルールを用いて最終的な要約を生成する．たとえば，色情報，発話情報による要約候補が**図6**のように得られたとする．このとき，各候補の論理

表 1 一貫性のある会話の発見のための発話群の定義
Table 1 Definition of speech groups for finding coherent conversation.

発話群名	会話例（A，B は発話者）
Same speaker	**A** : I went to the shop.
	A : I like fishing very much.
Question-answer	A : Can I use the telephone**?**
	B : Yes, of course.
Second person	A : I'm sorry I'm late.
	B : What time did **you** get up?
	A : At nine.
Conjunction	A : I think the shop is open now.
	B : **But** I've not finished my homework yet.
Pronoun	A : Look. My brother.
	B : Please tell me **his** name.
Co-occur	A : I will go **tomorrow**.
	B : **Tomorrow**?

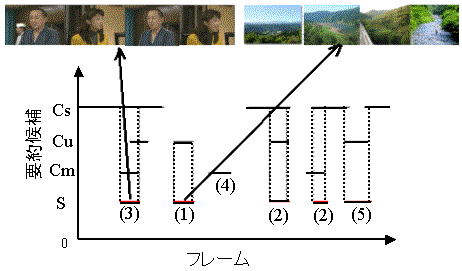


図 6 要約候補からの要約生成
Fig. 6 Making a summarization from candidates.

和（OR）または論理積（AND）を取って要約を生成する．

生成ルールは以下に示すとおりである．なお，発話情報，Unchanged，Multiplexing による要約の候補をそれぞれ，C_s，C_u，C_m と表し，最終的な要約は S と表すことにする．また，それぞれのルールを適用した例を，図 6 中に番号で示している．

(1) 色情報と発話情報による候補がお互いに重ならない場合，基本的には C_s と C_u の OR をとる．
理由：色や発話単独の情報では発見できなかった部分を互いに補い合うことを目的としている．たとえば，発話がなくても要約に含めるべき部分は多数存在し，そのような部分は色情報でしか発見できない．

(2) C_s と，C_u または C_m が重なった場合には，基本的にはそれらの AND をとる．
理由：C_s は時間的な制約を設けずに選んでいるため，再生時間が長くなることが多い．このため，時間的制約がある色情報による候補と AND をとり，要約をより簡潔にすることを目的としている．また，長い再生時間の中に複数の話題が含まれている可能性も考慮している．

(3) (2) において，C_u と C_m に重なりがある場合は，C_m を優先させる．
理由：C_m は頻繁な話者交代を表し，より重要な部分と見なせる．

(4) C_s がない部分に出現している C_m はすべて無視する．
理由：会話の一貫性がない部分では，C_m が話者交代を表している可能性は非常に低いと考えられる．

(5) C_u または C_m が複数の C_s にまたがっている場合は，C_s と重なりがない部分も採用する．
理由：隣り合う C_s 間の時間が非常に短いため，C_s と重なりがない部分を採用しても影響は少ない．

たとえば，図 6 中の (1) では，発話情報がないが，色情報の 1 つである Unchanged の区間が発見されており，これを要約に必要な意味的な区間（図中ではストーリーが展開している山の風景を提示している）として利用できる．

4. 実　　験

4.1 要約の評価

2，3 章で述べた方法を用いて，ビデオの要約の実験を行った．実験対象としては，SF 映画の Star Wars: Episode V およびコメディ映画の The Mask を用いている．3.4 節で述べたルールを用いて，2 つのビデオの開始から 10,000 フレームまでの部分（約 11 分間）の要約を作成して評価を行った．まず 25 人の被験者にオリジナルのビデオを視聴してもらい，続いて要約

表 2 被験者から得た回答結果
Table 2 Results of the answers from 25 subjects.

	時間圧縮の妥当性	重要部分の保持性	内容の一貫性
Star Wars: Episode V	3.0	2.7	2.6
The Mask	3.5	3.2	3.0

表 3 要約の時間圧縮率
Table 3 Compression rates of the summaries.

	オリジナルビデオ（秒）	要約（秒）	圧縮率（%）
Star Wars: Episode V	666	86	12.9
The Mask	666	178	26.7

表 4 登場人物のカバー率
Table 4 Cover rates of the characters at the summaries.

	オリジナルビデオ（人）	要約（人）	カバー率（%）
Star Wars: Episode V	11	11	100.0
The Mask	12	10	83.3

を視聴してもらった☆．これにより，すべての被験者はオリジナルビデオの内容は既知であるとし，条件の統一を行った．

要約は，オリジナルに比べて時間的に圧縮されていることが第 1 条件である．また，時間的に圧縮されても，内容の欠落は最小限に抑えるべきである．さらに，オリジナルのストーリーの流れを保持していることが望ましい．したがって，本実験では以下の 3 つの指標を考案した．

時間圧縮の妥当性 オリジナルビデオの再生時間に対して，要約の再生時間は妥当であるか．

重要部分の保持性 オリジナルビデオで重要と思われる部分は，要約中にも含まれているか．

内容の一貫性 要約の内容に一貫性があり，ストーリーの流れが理解できるか．

「時間圧縮の妥当性」では，要約の再生時間が短すぎたり長すぎたりせず，快適に視聴できるかどうかを判断基準としている．「重要部分の保持性」では，オリジナルビデオ内で被験者が重要と感じた部分が，要約にも現れているかどうかが判断基準である．この評価値が高いと，オリジナルビデオを見なくても，要約のみから短時間で要所が把握できることになる．また，「内容の一貫性」では，要約そのものが一貫したストーリーとして成り立っているかどうかが判断基準である．この評価値が高いと，オリジナルビデオを見なくても，大雑把な話の流れの把握が可能である．これらの 3 項目について，25 人の被験者から 5 段階評価で得た回答の結果を**表 2** に示す．ただし，評価値は 25 人の平均値である．

また，実験対象であるビデオについて，要約の実際の時間圧縮率および登場人物のカバー率を計算した．時間圧縮率は，オリジナルビデオの再生時間に対する要約の再生時間の比率である．登場人物のカバー率は，オリジナルビデオでの登場人物の数に対する要約での登場人物の数の比率であり，オリジナルビデオの内容を保持できているかを判断する際の参考としている．時間圧縮率および登場人物のカバー率の計算結果を，それぞれ**表 3** および**表 4** に示す．

4.2 結果および考察

3.4 節で述べたルールを適用した結果，特にルール (1) によって Unchanged の区間が効率良く発見された．これは，ストーリーが展開するステージの抽出に対して非常に有効であると考えられる．また，Star Wars: Episode V では，発見した区間に対して適用したルールは計 15 個であり，そのうち (1) のルール適用が 5 個，(2) が 6 個，(4) が 4 個であった．この結果から，色情報に関しては Unchanged，Multiplexing という特徴的な区間を各々独立して抽出できたことが分かる．また，発話情報と色情報が重複する区間はルール (2) によって約半数である 6 個が発見されている．これらは，視覚的な内容把握にさらに一貫性のある会話が付加された区間であるため，より内容理解に適していると考えられる．

また，発話情報による抽出では，映画によっては同一話者による回想シーンが挿入される場合があり，同一話者でも時間，場所が異なったシーンが抽出されてしまうという問題がある．このような場合，ストーリーが過去の出来事などに変わってしまい，内容の一貫性は保たれなくなるが，回想シーンは付随した形で

☆ Sundaram ら[8]は，彼らの手法の有意性を確かめるには，25 人以上の被験者が望ましいと記している．

抽出されるため，回想シーンとのつながり関係はある程度理解できると考えられる．次に，表2より，まず「内容の一貫性」についての評価が比較的低いことがあげられる．これは，本来，1つのストーリーとして凝縮されているビデオから要約を作成しようとしたため，部分的に内容の欠落が起こった結果と考えられる．一方，「内容の一貫性」を過度に重視すると，抽出する要約候補が時間的に長くなり，要約全体も冗長なものになってしまう恐れがある．実験では，発話の連続性を利用して内容に一貫性のある区間を抽出でき，内容理解の観点から見ると有効な手段と考えられる．しかしながら，ほとんどの区間が時間的に長くなってしまう傾向が見られた．本研究では，要約にストーリーとして一貫性を持たせることよりも，ビデオの内容を万遍なく抽出することに主眼を置いている．したがって，ビデオの要約に簡潔さと一貫性を共存させることは非常に困難であり，今後検討すべき課題である．

表3より，Star Wars: Episode Vの方がThe Maskに比べて時間圧縮率が低くなっている．また表2より，Star Wars: Episode VはThe Maskに比べて，すべての項目で評価値が低くなっている．これは，要約の再生時間が短かったため，重要と思われる部分をカバーしきれず，視聴者が時間的に物足りなさを感じたためと考えられる．しかしながら，評価値2.5は「適度である」という評価であるため，2つの映画ともに全体的には視聴者にとって理解しにくい内容ではなかったと考えられる．特に，表2のThe Maskにおいて，重要部分の保持性と内容の一貫性で各々3.2，3.0という比較的高い評価が得られた要因として，Star Wars: Episode Vよりも連続した発話の区間が多かったことがあげられる．具体的には，Star Wars: Episode Vでは56秒，The Maskでは135秒の発話の区間があり，それぞれ全体の65%，76%を占めていた．このことから，視聴者は視覚的な情報だけでなく，登場人物同士の対話という付加的な情報を得て，さらに理解を深めたことが分かる．すなわち，色情報と発話情報によって発見された区間に要約生成ルール(2)，(3)を適用した有効性が示されたと考えられる．また，「時間圧縮の妥当性」の評価値から，両ビデオともに不快と感じる要約ではなかったといえる．要約の再生時間については，2.3節で述べた重要区間の抽出の際に，抽出するショット列の長さを変化させて調節すればよい．したがって，今回は実装しなかったが，ユーザが要約の再生時間をある程度決定するようなシステムの実現も可能である．

次に，表4よりStar Wars: Episode Vでは要約中ですべての登場人物を再現していることが分かる．しかしながら，表2で「重要部分の保持性」の評価値はThe Maskの方が良く，「登場人物のカバー率」は要約の良し悪しに直接影響していないと考えられる．また，ビデオにおいて，どのような部分を重要と感じるかは個人の主観によるが，被験者から以下のような声が多く聞かれた．

- Star Wars: Episode Vにおいて，主人公が怪物に襲われる部分が欠落しており，その後，主人公がなぜ怪物に捕らわれているかが分からない．
- 「The Mask」というタイトルなのに，冒頭のマスクが登場する部分が要約に現れていない．

上記の部分は，いずれも色情報による重要な区間には該当せず，セリフもないため，発話情報による会話の一貫性も発見できなかった．これらの部分は重要であり，要約に含めたいという願望はある．しかしながら，これらの部分が要約に含まれていなくても，ストーリーの大筋を理解することは十分可能である．

このように，実際視聴者が受け取る重要な区間は個人で異なり，重要性の基準を設定することは非常に困難である．しかしながら，ビデオにはUnchanged, Multiplexingのような色変化パターンや，登場人物の発話の連続性という特徴を含んでいるため，これらを有効に活用することでユーザにある程度の内容理解を提供できると考えられる．

4.3 関連研究との比較

ビデオの要約についてはこれまで多くの研究がなされている．大きく分けると，ニュースのように構造化された映像と，映画のような非構造的な映像を対象にしたものがある．前者においては，あらかじめトピックごとにテロップやアナウンサの解説が付加されて構造化されているという特徴がある．このため，これらをインデックスとして活用すればビデオの検索が可能となる．たとえばInformedia Project[9]は，ニュース番組の内容を端的に示すテロップを活用し，テロップが表示される部分をインデックスの候補として検出している．また，画像処理と言語処理を行い，ビデオの場面やフレームを特徴付けしている．画像処理では，カメラモーションや登場人物の顔を利用している．言語処理では，字幕内の単語の頻度を用いてキーワードを抽出している．さらに，特徴付けされた個々のビデオデータを独自のルールで組み合わせて，インデックスを作成している．

一方，映画やドラマなどにはテロップなどの指標となるものが付加されておらず，構造化することが困難である．このため，ビデオに含まれる色情報や発話情

報を利用するアプローチがとられている．たとえば，DeMenthon ら[10]は，ビデオの色情報から特徴ベクトルを用いて色変化の推移をグラフとして表現している．グラフの形状はノイズによって複雑化していること，またビデオには多種多様な色変化が存在していることなどから，グラフの関数は高次元空間になっている．このため，低次元空間で表現できる関数に簡略化して，ビデオの大まかな色変化の境界を検出している．この手法に基づいて，低次元空間で表現される関数からセグメントを発見して，各セグメントに対応する区間を特定の場面としている．しかしながら，映画には，たとえば風景のシーンに人の顔の映像が突然挿入されるなど，突発的な映像が数多く存在するため，あやまった境界を検出する可能性が高い．また，複数の場面に分割して各場面の画像を提示するだけでは，登場人物間の関係やストーリーの流れを理解することは困難であると考えられ，言語の情報も考慮するべきである．

Sundaram ら[8]は，「visual complexity」および「film syntax」という 2 つの観点からビデオの要約を行っている．「visual complexity」では，Kolmogorov complexity を用いてショットの複雑さを計算している．ショットごとの Kolmogorov complexity からビデオの圧縮率を決定し，再生時間に上限と下限を設けて要約を生成している．「film syntax」では，「1 つの会話は，最低 3 つのショットから構成される」，「x 人による会話を表すためには，最低 $3x$ 個のショットが必要である」などのルールに基づいて，1 つの場面から不要なショットを削除して，要約を生成している．この要約手法は実験による評価も高く，有用であると考えられる．しかしながら，この手法はビデオ全体があらかじめ場面に分割されていることを想定しているが，このような前提条件は現実的ではないと考えられる．また，映画に適用する場合，対象となる映画の画像に対する処理はしないため，ショットの複雑さによる要約映像が視覚的に内容を把握できる映像とは限らない．

森山ら[11]は，テレビドラマの要約を生成する手法を提案している．この手法では，トラック構造と呼ばれる，カメラモーション，登場人物のセリフ，BGM，効果音などから，経験的知識に基づいて，心理的に重要と判断した部分のみを取り出して，要約を生成している．この研究では，これまであまり注目されなかった心理面に着眼点を置いており，実験によって手法の有効性も示されている．発話情報の利用については，登場人物のセリフの有無および密度を用いて重要と思われる部分を検出している．映画に適用した場合，効果音による意味的な区間の抽出は有効と考えられる．特に，アクションシーンでは爆発音などの効果音が多く用いられているので，このような区間の抽出はユーザにとって理解するうえである程度有益な情報になる．また，セリフの有無によって区間を限定すると，会話内容の把握に必要な発話者同士の会話のつながりまでは重視できず，内容の一貫性を考慮したことにはならないと考えられる．

一方，本研究では，ストーリー把握に必要とされる意味的な区間を色情報と発話情報の統合により抽出するという手法を用いている．たとえば，表 3 において，Star Wars: Episode V の要約時間 86 秒のうち計 16 秒の発話のないアクションシーンなどが色情報で発見でき，このような区間は特に内容理解に必要であると考えられる．また，色情報のみで計 35 秒の会話の区間が抽出されたが，いくつかの区間では会話の途中から抽出されているため，会話の前後関係があいまいになった．しかし，発話情報を補間することで計 70 秒の会話の区間が抽出された．この処理では特に会話の始まりから抽出される区間が増加するため，会話のつながり関係がより理解しやすくなる．また，発話情報を適用することで要約の時間が増加する要因としては，内容の一貫性を重視した会話の連続性によるものと考えられる．

5. おわりに

本稿では，ビデオに含まれる色情報および発話情報を用いて，自動的に要約ビデオを生成する手法を提案した．色情報としては，ショットの代表フレームの特徴量を算出し，確率モデルを導入して場面分割を行い，代表的な特徴を持つ 2 種類のビデオ区間（Unchanged および Multiplexing）を検出して，要約の候補を作成した．また，発話情報として Closed Caption を抽出し，会話の特徴に着目して会話分析を行い，一貫した発話群を取り出して要約候補を決定した．さらに，色情報と発話情報を用いて得られた要約の候補をルール化して組み合わせて，最終的な要約を生成した．

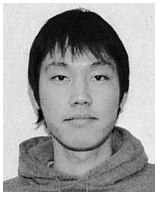
実験結果より，オリジナルビデオの内容を大幅に損なうことなく，時間的にも不快感を持たないような要約を生成することができた．しかしながら，要約の時間的な簡潔さと内容の一貫性を共存させることは実現されておらず，ユーザによる要約の再生時間の制御などとあわせて，今後の課題となっている．

謝辞 本稿で行った実験の評価の際，被験者として協力してくださった皆様に，心から感謝の意を表する．

参 考 文 献

1) 是津耕司，上原邦昭，田中克己：映像の意味的構造の発見，情報処理学会論文誌，Vol.41, No.1, pp.12–23 (2000).
2) 北　研二：確率的言語モデル，東京大学出版会 (1999).
3) Hearst, M. A.: Multi-Paragraph Segmentation of Expository Text, *Proc. 32nd Meeting of the Association for Computational Linguistics*, pp.9–16 (1994).
4) 小嶋秀樹，古郡延治：単語の意味的な類似度の計算，電子情報通信学会技術研究報告，AI92-100, pp.81–88 (1993).
5) 小嶋秀樹，古郡延治：単語の結束性にもとづいてテキストを場面に分割する試み，電子情報通信学会技術研究報告，NLC93-7, pp.49–56 (1993).
6) Beeferman, D., Berger, A. and Lafferty, J.: A Model of Lexical Attraction and Repulsion, *Proc. 35th Annual Conference of the Association for Computational Linguistics*, pp.182–189 (1997).
7) Beeferman, D., Berger, A. and Lafferty, J.: Text Segmentation Using Exponential Models, *Proc. 2nd Conference on Empirical Methods in NLP* (1997).
8) Sundaram, H. and Chang, S.: Condensing Computable Scenes Using Visual Complexity and Film Syntax Analysis, *Proc. ICME 2001* (2001).
9) Smith, M. A. and Kanade, T.: Video Skimming and Characterization through the Combination of Image and Language Understanding, *Proc. IEEE International Workshop on Content-Based Access of Image and Video Databases*, pp.61–70 (1998).
10) DeMenthon, D., Kobla, V. and Doermann, D.: Video Summarization by Curve Simplification, *Proc. ACM Multimedia 1998*, pp.211–218 (1998).
11) 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 (2001).

(平成 13 年 12 月 26 日受付)
(平成 14 年 9 月 5 日採録)

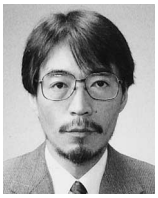


本多光一郎

昭和 52 年生．平成 12 年神戸大学工学部情報知能工学科卒業．平成 14 年同大学院自然科学研究科博士前期課程修了．現在，JR 西日本勤務．

藤村　憲市

昭和 53 年生．平成 13 年神戸大学工学部情報知能工学科卒業．現在，同大学院自然科学研究科博士前期課程在学中．



上原　邦昭（正会員）

昭和 29 年生．昭和 53 年大阪大学基礎工学部情報工学科卒業．昭和 58 年同大学院博士後期課程単位取得退学．大阪大学産業科学研究所助手，講師，神戸大学工学部情報知能工学科助教授，同都市安全研究センター教授を経て，現在，同大学院自然科学研究科教授．平成元年より 2 年まで Oregon State University，Visiting Assistant Professor．工学博士．人工知能，特に機械学習，データマイニング，マルチメディア処理の研究に従事．1990 年度人工知能学会研究奨励賞．2001 年度電子情報通信学会オフィスシステム研究賞．人工知能学会，電子情報通信学会，計量国語学会，日本ソフトウェア科学会，AAAI 各会員．