2007－AVM－56 (3)
2007／3／5

映像特徴と主観評価の関連付けによる映像要約手法

倪 嬋斌[†] 野村敏男[‡] 渡部秀一[‡] 岡田浩行[‡] 亀山渉[†]

[†] 早稲田大学 大学院国際情報通信研究科
〒367-0035 埼玉県本庄市西富田大久保山 1011-A213
[‡] シャープ株式会社 技術本部 先端映像技術研究所
〒261-8520 千葉市美浜区中瀬 1 丁目 9 番 2 号
Email: trist@akane.waseda.jp

あらまし 映像の撮影・編集の際，特定の意味や意図を強調する目的で使われる「映画の文法」は，視聴者の主観評価に影響を与えると考えられる．そこで，「映画の文法」を代表する映像パラメータを MPEG ビットストリームから抽出し，それを基に主観評価の結果を推定することで，制作者の意図と視聴者の嗜好の両方を反映した要約映像を自動生成する手法を提案する．具体的には，複数の視聴者から映像再生時の再生速度データを採集し，映像の特徴パラメータとの対応関係を機械学習により関数化することで，未知映像の特徴パラメータから自動的に可変速に再生速度を推定することを試みた．その結果と評価について報告する．

キーワード 映画の文法，映像特徴パラメータ，主観評価，MPEG-1， MPEG-7，機械学習

A Video Summarization Method Based on the Film Grammar and Subjective Rating

Chanbin NI[†] , Toshio NOMURA[‡] , Shuichi WATANABE[‡] , Hiroyuki OKADA[‡] , and Wataru KAMEYAMA[†]

[†] Graduate School of Global Information and Telecommunication Studies, Waseda University
A213 , 1011 Okuboyama, Nishi-Tomida, Honjo-shi, Saitama 367-0035, Japan
[‡] Advanced Image Research Laboratories, Corporate Research and Development Group, SHARP Corporation
1-9-2 Nakase, Mihama-ku, Chiba-shi, Chiba 261-8520, Japan
Email: trist@akane.waseda.jp

Abstract: Behind the fact that we understand not only the concept of objects appearing in a movie, but also the semantics that film makers intent to present, there is a certain relation between “the Film Grammar” and Audience’s media literacy. In this paper, we propose a video summarization method based on the film grammar and subjective rating, by mapping visual characteristic parameters extracted from compressed video to subjective rating. Additionally, in our prototype system, a proper play-rate is automatically generated, without any knowledge of the content.

Keywords: the Film Grammar, Visual Characteristic Parameters, Subjective Rating, MPEG-1, MPEG-7, Machine Learning

1 はじめに

放送と通信の融合が進み，数百タイトルからなる動画の自動録画，蓄積，視聴が，一般家庭でも可能となった．その結果，視聴者の視聴スタイルが，従来の「ながら視聴」から「目的視聴」へと転換しつつある．このような状況の中，録画したコンテンツの中から興味のある箇所だけを視聴する，いわゆる“ダイジェスト視聴”の需要が増加している．こうした用途には，映像を自動的に要約するシステムが有効であると考えられる．

映像メディアの現行の処理法においては解析の基となる事前知識が重要な要素であり，時間順の構成が類型化できるか否かは，映像のジャンルに依存する．そのため，映像メディアの自動解析についてはこれまで，知識のモデル化の容易性から，ニュース映像を対象にすることが圧倒的に多かった[1]．しかし，こうした知識ベースの処理法は，要約映像を作る対象が限定されている場合には妥当であるが，「汎用的な」映像要約問題には対応できない．ジャンルを問わない汎用的な映像要約技術は現在のところ Open Issue となっている．

また，視聴者が本当に必要とする要約映像を作るためには，主観や個性を反映させるメカニズムが必要となる．さらに，すでに流通している映像データに対応するためには，コンテンツの内容情報に関するメタデータではなく，MPEG 符号化データ情報に基づく手法が，計算量の観点から望ましい．

著者らはこれまでに，MPEG-1 形式のビデオビットストリームから符号化領域での時空間的な特徴パラメータを抽出し，視聴者の主観評価との関連づけを行うことにより映像を要約する手法について検討してきた[2]．この方式は，映像を制作する側に存在する「映画の文法」[3]と呼ばれるコモン・ルールと，それに基づく映像を視聴者が理解する背景であるコモン・ルール[4]を機械学習によりマッピングすることにより，上記の問題に対応する．即ち汎用的かつ視聴者の主観を直接反映可能な映像要約を実現する．

本稿では，この方式を発展させ，映像特徴パラメータ抽出の精度向上を図ると同時に，可変速早送りによる映像要約方式について検討したので報告する．具体的には，再生速度を被験者自身に実時間で調節させる実験を新たに行い，その結果と抽出した映像特徴パラメータを機械学習によって対応付けることで，ある映像区間の最適な再生速度を自動推定する手法を提案した．まず，2 章で提案方式の概要を述べる．3 章では新たに行った実験について述べる．4 章では，使用した映像特徴パラメータについて述べる．5 章では，最適再生速度の自動推定法とその結果について述べる．6 章で結論と今後の課題を述べる．

2 提案方式の概要

以下の手順により，要約映像の自動生成を実現した．図 1 にその概要を示す．

(1) 映像パラメータ抽出部において，映像特徴パラメータを抽出する．
(2) (1)で得られた映像特徴パラメータと，実験により得られた再生速度データの間を，機械学習によって関連づける．
(3) 最適再生速度推定部において，(2)で得られたマッピング結果に基づき，新たな映像データの映像特徴パラメータから，最適再生速度を自動推定し，再生制御データを生成する．
(4) 映像の再生速度を制御し，可変速早送りされた映像を出力する．

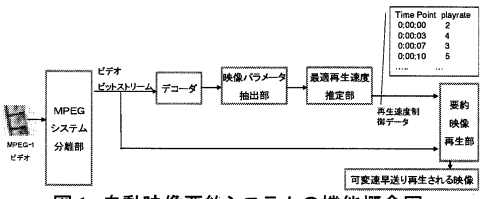


図 1 自動映像要約システムの機能概念図

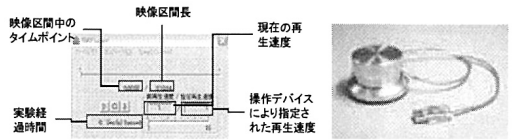
3 主観評価実験

著者らはこれまでの研究で，シーンを最小単位とする映像区間の重要度についての主観評価実験を行った[2]．その結果，異なる映像コンテンツ間で，被験者に共通する嗜好が存在することが確認された．そこで本稿では，映像の各箇所の適切な再生速度情報を得るため，可変速の再生速度を被験者自身に実時間で調節させる実験を新たに行った．詳細を以下に記す．

被験者：大学院生男女 14 名（24 歳～29 歳）．

実験の試料：4 つの代表的なジャンル (ドキュメンタリー，アニメ，バラエティ，映画) から選択した，5 件の映像刺激．

計測方法：初めに全体を通して等倍速固定で映像を呈示した．次に被験者自身が実時間で 等倍速～10 倍速の間で調節可能な状態で映像を呈示し，その際の再生速度を記録した (いずれも無音)．映像の再生に使用したアプリケーションと，再生速度の調節に用いたデバイスを図 2 に示す．



再生用アプリケーション　　GriffinTechnolog 社製 Powermate

図 2 映像の呈示と再生速度の調節に使用した装置

4 映像特徴パラメータ

先に提案した方式では，MPEG ビットストリームデータに含まれる動きベクトル情報，マクロブロックタイプ情報を用いて映像パラメータを抽出した．

本稿ではその拡張として，MPEG-7 Visual[5]で定義されている Motion Activity 記述子の一部に従った映像特徴パラメータを抽出した．これらの記述子は符号化データ領域の簡単な演算で抽出でき，使用された類似

映像の検索技術の精度には定評がある．

MPEG-7 で規定されている Motion Activity 記述子はいつくかの要素から構成され，P ピクチャ中のマクロブロックの動きベクトル値に基づきフレーム毎に計算される．本稿では動き強度・最頻方向・連続性の三つの記述子に従い，映像特徴パラメータを抽出した．図 3 に映像特徴パラメータ抽出の流れを示す．

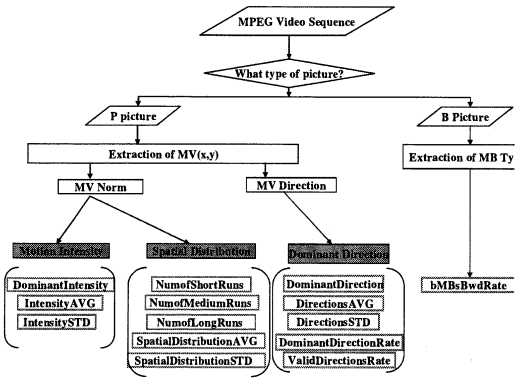


図 3 映像特徴パラメータ抽出の流れ

4.1 動き強度(Motion Intensity)

フレーム内の全動きベクトルの大きさに関する標準偏差を求め，図 4 の通り，5 段階に量子化し，当該フレームの動き強度とする．

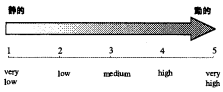


図 4 動き強度の量子化

映像区間内で，1 フレーム毎に，動き強度を算出し，カテゴリごとの数を集計し，区間内で Dominant なカテゴリ，全体の平均，標準偏差を算出する．

パラメータ 1：Dominant Intensity
パラメータ 2：Intensity AVG
パラメータ 3：Intensity STD

4.2 最頻方向(Dominant Direction)

フレーム内の全動きベクトルの角度成分を求め，図 5 の通り，8 方向（45 度刻み）に分類したものの中で最頻方向を示す値を，当該フレームにおける最頻方向とする．

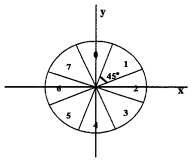


図 5 方向を 8 方向に量子化

実際に MPEG ビットストリームを解析したところ，Pan 区間では図 6 のように 1 つのカテゴリが突出した形で表され，Zoom 区間では図 7 のように 8 つのカテゴリが平均的な形で表わされることがわかった．

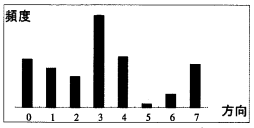


図 6 Pan 区間の分布

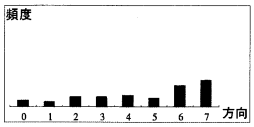


図 7 Zoom 区間の分布

映像区間内でカテゴリごとの出現数を集計し，区間の Dominant なカテゴリ，全体の平均，標準偏差，Active なブロック（動きベクトルが 0 でない）が全ブロック（1 フレーム計 300）中に占める割合，Dominant なカテゴリに属するブロックが Active なブロック中に占める割合を算出する．

パラメータ 4：Dominant Direction
パラメータ 5：Direction AVG
パラメータ 6：Direction STD
パラメータ 7：Dominant Direction Rate
パラメータ 8：Valid Directions Rate

4.3 連続性(Spatial Distribution)

1 フレーム毎に，動きベクトルのノルムの平均を求め，平均以下の動きベクトルが水平方向に連続する回数（Length）を数え，Short，Medium，Long の 3 カテゴリのヒストグラムを作成する．

また，画面の連続性を表す度合である Short，Medium，Long をそれぞれ，0，1，2 のラベルに変換し，数値化する．それを映像区間ごとに平均し，区間内での変化を表すために，標準偏差を算出する．

パラメータ 9：NumofShortRuns
パラメータ 10：NumofMediumRuns
パラメータ 11：NumofLongRuns
パラメータ 12：Spatial Distribution AVG
パラメータ 13：Spatial Distribution STD

4.4 後方参照マクロブロックの割合

映像区間内の B ピクチャにおいて，B タイプのマクロブロックが全マクロブロック中に占める割合を算出する．

パラメータ 14：BMB Rate

5 最適再生速度の自動推定

独立変数と従属変数が存在する場合，独立変数を入力，従属変数を出力として捉え，回帰分析により両者の関係を記述できる．MPEG ビットストリームの特徴パラメータを独立変数，被験者の再生速度を従属変数として捉えれば，MPEG 特徴パラメータからの最適再生速度の推定は，非線形回帰問題の一種と位置づけられる．

回帰分析により得られる数式 (モデル) の用途は主に 2 つある．一方は，独立変数の値が決まると従属変数を自動的に決定 (推定) する，装置としての用途．他方は，両者 (独立変数と従属変数) の間の関係 (分布の形状) を，人間が理解しやすい形で表現する用途である．バックプロパゲーション法に代表される，教師付き階層型ニューラルネットワーク (以下:NN) は，前者の意味で，非線形回帰分析の一種である．また，NN は非線形データに強く，通常の回帰分析では数式を当てはめにくい複雑なパターンに適用できるという強みがある．この点から，本稿では，MPEG 特徴パラ

メータからの最適再生速度の推定に，まずバックプロパゲーション法（以下 BP 法）を使用した．

5.1 NN による最適再生速度の推定

5.1.1 アルゴリズムの概要

初めに，使用した全 5 件のコンテンツに関して，映像区間ごとに MPEG 特徴パラメータを算出し，入力とした．また，再生速度から，全被験者分について，対応する区間長分を切り出し，区間内で平均したものを出力とした．

次に，階層型 NN を作成し，局所解に陥らないよう改善した BP 法[6] による学習を行った．なお，使用した MPEG ビットストリームは，全ての映像を前半 75%と後半 25%に分割し，前者を教師用に，後者を検証用に使用した．

5.1.2 結果と考察

概要で述べたアルゴリズムでは，以下の 4 項目が未決定である．そのため，これらを細かく変更しながら最適値を検討した．

- NN の係数 (学習項, 慣性項)
- NN の許容誤差
- NN の中間層のニューロン数
- MPEG ビットストリームの特徴パラメータ

様々な特徴パラメータの組み合わせを検討した結果，表 1 に示すものが最良の結果であったが，再生速度の予測値と実測値の相関係数は，0.29 に留まった．

表 1 使用した特徴パラメータ

区間長	10 秒 (20GOP)
使用したパラメータ	Dominant Intensity DominantDirection NumOfShortRuns NumOfMediumRuns NumOfLongRuns BMB Rate

推定された再生速度（予測値）と実際に測定された再生速度（実測値）を，図 8 に示す．

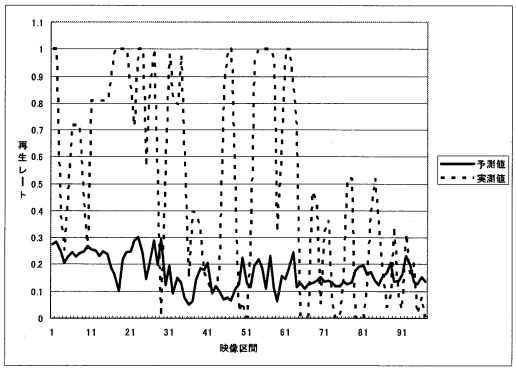


図 8 BP 法による予測値と実測値

5.2 SVR による最適再生速度の推定

NN には，局所最適解に収束する，標本数が増加した場合に学習が困難になる，中間層の素子数の選択の基準が曖昧である，といった問題点が存在する．Support Vector Machine (SVM) [7]は，1995 年に AT&T の Vapnik らにより統計的学習理論の枠組みで提案され，上記のような問題を解決した新しいパターン認識手法として知られている．元々は線形分離性をその前提としていたが，カーネル関数を用いて線形分離性を持つ特徴空間上にあらかじめ写像することで，線形分離不可能な問題にも適用可能となっている．

識別を行う SVM は Support Vector Classifier (SVC) と呼ばれることがあり，特に 2 クラスのパターン認識問題において，最も高い認識性能を示す学習モデルの 1 つとして，近年注目を集めている．文字認識や画像認識等の様々な応用分野において，従来法を上回る高い識別性能を示すことが，確認されている[8]．

本稿では，回帰推定を行う SVM である，Support Vector Regression (SVR) [9]を使用し，NN を使用した場合と同様に，MPEG 特徴パラメータからの最適な再生速度の推定を行った．

5.2.1 アルゴリズムの概要

初めに，5.1.1 で述べたのと同様の方法で，入出力データセットを作成した．

次に，SVR の学習を行った．なお，使用した全ての MPEG ビットストリームは，前半 75%と後半 25%に分割し，前者を教師用に，後者を検証用とした．

5.2.2 結果と考察

様々な特徴パラメータの組み合わせを検討した結果，表 2 に示すものが最良の結果であったが，再生速度の予測値と実測値の相関係数は，0.19 に留まった．

表 2 使用した特徴パラメータ

区間長	10 秒(20GOP)
使用したパラメータ	IntensityAVG IntensitySTD DirectionsAVG DirectionsSTD BMB Rate

推定された再生速度（予測値）と実際に測定された再生速度（実測値）を図 9 に示す．

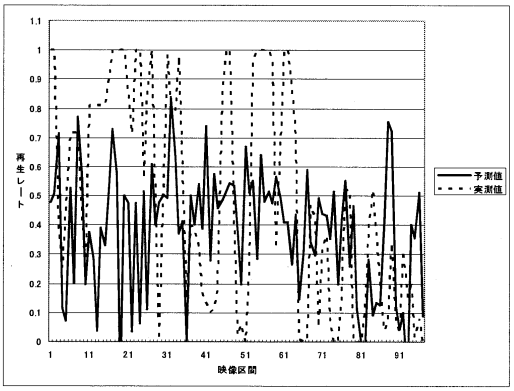


図 9 SVR による予測値と実測値

相関係数を見た場合，SVR による予測結果は NN による予測結果と比較して正答率が低いように見えるが，これは，SVR が算出した予測値の幅が NN によるものよりも大きいことが原因と考えられる．図 8 と図 9 に示したように，実測値と比較した場合，SVR による結果の方が相対的に良好であることがわかる．

5.3 クラスタ分析に基づく最適な再生速度の推定

NN には，特定の値を持つ入力ベクトルに対し，出力ベクトルの値が一意に決まるという性質がある．そのため，学習に用いる入出力データセット中に，入力ベクトルは似ているが出力ベクトルは全く異なる組み合わせが複数含まれる場合には，学習が収束せず，結果として，いずれの入力に対しても中途半端な値が出力される．

一方で，本稿で扱う入力データは，MPEG ビットストリーム中から抽出した特徴量であり，主に動きベクトルに関するものである．これは「映像の文法」が共通する区間では，ある程度の共通性が見られると予想されることから採用したものであり，したがって入力ベクトル群内に似通ったデータが見られることは，想像に難くない．さらに出力データに着目すると，被験者が実際に映像の再生速度を操作した結果であり，これには個人差や誤差が含まれることから，同じ映像区間に対してさえも，毎回厳密に一定となることはありえない．

NN や SVR による近似に失敗した背景として，こうした矛盾の存在が考えられることから，クラスタ分析によるデータの分類に基づく手法を考案した．

5.3.1 アルゴリズムの概要

出力となる映像再生速度は厳密には一定しないと前述したが，特定の映像区間に対しては，大まかな傾向は見られると考えられ，この傾向を記述することが目標となる．これを数理モデルのレベルに言い換えると，入力データセットの傾向により出力データセットの傾向に生じる偏りを，近似することが目標となる．本稿では，入出力ベクトルの分類に，クラスタ分析を使用した．図 10 にアルゴリズムの概要を示す．

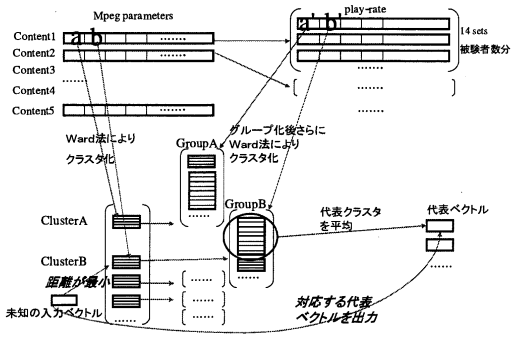


図 10 クラスタ分析による最適再生速度の自動推定

初めに，使用した全 5 件の MPEG ビットストリームに関して，映像区間ごとに特徴パラメータを算出した．再生速度に関しては，全被験者分について，対応する区間長分を切り出した．次に，クラスタ分析 (Ward 法) により，複数のクラスタに分類した．入力ベクトル (集計済み MPEG ビットストリームデータ) の分類に基づき，対応する出力ベクトル (再生速度) を分類した．例えば，入力ベクトル a がクラスタ A に分類された場合，a に対応する出力ベクトル a'はグループ A'に分類され，入力ベクトル b がクラスタ B に分類された場合，b に対応する出力ベクトル b'はグループ B'に分類される．分類された出力ベクトル群を，クラスタ分析により，グループ内で複数のクラスタに分類した．出力ベクトル群の最初のグループ化は，「パラメータ表現が似通った映像区間に対する再生速度」の分類を意味する．これをさらにクラスタリングすることで，同種のパラメータ表現に対し，どういった再生速度が主に選ばれているのか理解できると考えたためである．

次に，グループ化後さらにクラスタリングされた出力ベクトル群について，各グループ内で最大となるクラスタに属するものを平均し，そのグループの代表ベクトルとした．最後に，未知の (学習に使用していない) MPEG ビットストリームが与えられた場合には，4 章で述べた方法で映像区間ごとの特徴パラメータを算出し，Ward 法で定義された方法を基にした方法で各入力ベクトルクラスタとの距離を計算し，それが最小となるクラスタに対応する，出力ベクトル群のグループの代表ベクトルを，再生速度の予測値として出力する仕様とした．入力ベクトルのクラスタとの距離の計算方法は，以下の通りである．

- クラスタ内の全ベクトルと，新規のベクトルを併せ，全体集合とする．
- 全体集合内で通常通りクラスタ分析 (Ward 法) を行う (ただし，新規のベクトルは他のクラスタに結合せず，距離の計算のみ行う)．
- 新規のベクトル以外の全ベクトルが 1 つのクラスタにまとまった時点で，新規のベクトルと，他全てのベクトルが属するクラスタの間の距離が得られている．

概要で述べたアルゴリズムでは，以下の 4 項目が未決定である．そのため，これらを細かく変更しながら最適値を探索した．

- MPEG ビットストリームの特徴パラメータ
- 映像区間長
- 入力ベクトル群のクラスタ数
- 出力ベクトル群のクラスタ数 (グループ内)

5.3.2 実験結果

様々な特徴パラメータの組み合わせを検討した結果，表 3 に示すものの結果が最良であったが，再生速度の予測値と実測値の相関係数は，0.26 に留まった．

表 3 使用したパラメータ

区間長	10 秒(20GOP)
使用したパラメータ	IntensityAVG IntensitySTD DirectionsAVG DirectionsSTD BMB Rate
入力ベクトルのクラスタ数	50
出力ベクトルのクラスタ数	10

推定された再生速度（予測値）と実際に測定された再生速度（実測値）を図 11 に示す.

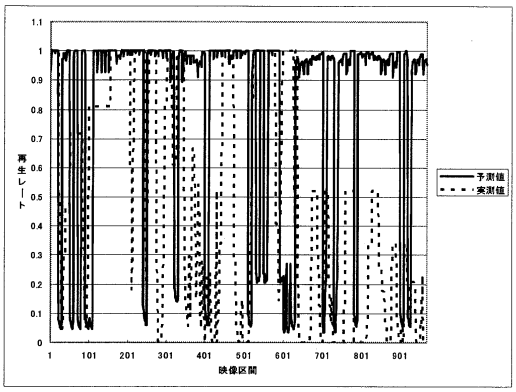


図 11 クラスタ分析による予測値と実測値

各コンテンツにおいて，異なるクラスタ（特徴パラメータのパターン）の割合を，入力ベクトルのクラスタ数を 8 に設定して調べたところ，各クラスタが複数のコンテンツ間にまたがり，程よく均等に分かれていることが分かった．これを図 12 に示す．多少の偏りは見られたものの，これは，今回抽出した特徴パラメータでは，異なる映像間に共通するパターンが存在していることを意味する.

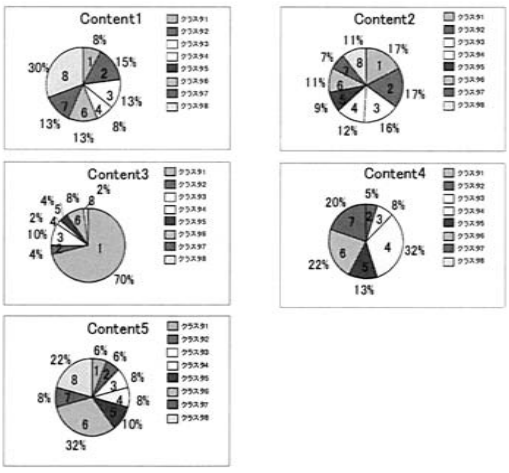


図 12 各コンテンツにおけるクラスタの割合

6 結論と今後の課題

本稿は，「映画の文法」と呼ばれる映像の操作・編集手法と，それらがもたらす心理的効果の間に明確な関連が存在するという仮説のもと，MPEG ビットストリームから，カメラワーク，ストーリ展開のテンポなど，代表的な映像操作・映像編集の手法を特徴パラメータとして抽出し，同時に，映像に対応する視聴者の再生速度を測定し，両者の間を機械学習によって対応付けることで，要約映像を自動作成するシステムの試作を行った.

具体的には，再生速度を被験者自身に実時間で調節させる実験によって得られた映像区間ごとの最適と思われる再生速度と，MPEG ビットストリームから抽出した「映像の文法」に対応する特徴パラメータとの対応づけを，ニューラルネットワーク，および SVR によって，学習させた．いずれも，既知のデータを推定する場合には高い正答率であったが，未知のデータを予測する場合には，正答率が低いままとなった.

そこで，Ward 法によるクラスタリングを行い，似通った特徴パラメータに対して大きく異なる再生速度を持つデータセットを排除し，残されたデータセットを基に実際に要約映像を自動生成するシステムを試作した．その結果，ジャンルが異なる映像間に共通するパターン（「映画の文法」）が抽出され，視聴者の主観をある程度反映した要約映像を生成することができた.

再生速度の高低には，MPEG ビットストリームの特徴パラメータに表れない要因や，視聴者の嗜好，操作ミス等によるノイズの影響が予想される．そのため，万人共通の要約映像の自動生成という課題の実現には，新たな特徴パラメータの模索，実験上の統制などの課題が残されている．視聴者の個性をより細かく反映させる仕組みの考慮や，主観評価以外の評価法の検討と合わせて，今後の課題としたい.

参考文献

[1]馬場口登，“ メディア理解による映像メディアの構造化，” 信学技報，IE-9918 PRMU99-42 MVE99-38(1999-07),pp39-46,Jul.,1999.
[2] 倪，渡部，野村，亀山，”映像の種類に依存しない映像要約手法に関する検討,” 情報処理学会第 68 会全国大会, 3C-5，Mar,2006
[3]Daniel Arijonf, “Grammar of the Film Language,”, London: Focal Press, 1976(邦訳：映画の文法)
[4]中島義明,“ 映像の心理学,” ,1996,サイエンス社
[5] ISO/IEC 15938-3/FDIS Information technology - Multimedia content description interface - Part 3 Visual
[6] Yutaka F., Hideo M., Haruyuki M., Akimasa I.: A modified back-propagation method to avoid false local minima, Neural Networks, 11, pp1059-1072, 1998.
[7] Corinna Cortes, Vladimir Vapnik, "Support-Vector Networks", Machine Learning, 20, 273-297, 1995.
[8]http://www.clopinet.com/isabelle/Projects/SVM/applist.html
[9]Vladimir Vapnik, Steven E. Golowich, Alex Smola, "Support Vector Method for Function Approximation Regression Estimation, and Signal Processing", Neural Information Processing Systems, Vol9 MIT Press, 1997