



HOKKAIDO UNIVERSITY

Title	階層ベイズモデルの導入による映像の構造に注目したシーン分割の高精度化に関する研究
Author(s)	宋, 妍
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第11764号
Issue Date	2015-03-25
DOI	https://doi.org/10.14943/doctoral.k11764
Doc URL	https://hdl.handle.net/2115/58919
Type	doctoral thesis
File Information	Yan_Song.pdf



博士論文

**階層ベイズモデルの導入による映像の構造に
注目したシーン分割の高精度化に関する研究**

A Study on Accurate Scene Segmentation Using Video
Structures Based on Bayesian Hierarchical Model

北海道大学 大学院情報科学研究科

メディアネットワーク専攻

メディアダイナミクス研究室

宋 妍

2015年3月

目次

1 章	序論	1
1.1	研究背景	1
1.2	本研究の目的	3
1.3	本論文の構成	3
2 章	映像のショット 分割に関する従来手法	5
2.1	はじめに	5
2.2	ショット 分割に関する用語の定義	5
2.3	映像のショット 分割に関する従来手法	7
2.4	まとめ	11
3 章	映像のシーン分割の従来手法	12
3.1	はじめに	12
3.2	映像の構造に注目したシーン分割に関する従来手法および階層ベイズモデルに基づくシーン分割に関する従来手法	13
3.2.1	映像の構造に注目するシーン分割手法	13
3.2.2	階層ベイズモデルに基づくシーン分割手法	16
3.3	従来手法における解決すべき課題	20
3.4	まとめ	21
4 章	階層ベイズモデルの導入による映像の構造に注目したシーン分割手法	22
4.1	はじめに	22
4.2	画像特徴量を用いたショット間の類似度算出	22
4.3	映像の構造を考慮したシーン境界候補の取得	23
4.4	階層ベイズモデルに基づいたシーン境界候補からシーン境界の推定	24
4.5	階層ベイズモデルの導入による映像の構造に注目したシーン分割手法の評価実験	26
4.6	まとめ	29
5 章	画像特徴量と音響特徴量を用いたシーン分割の高精度化手法	30
5.1	はじめに	30

5.2	画像・音響特徴量を用いたショット間の類似度算出	30
5.3	累積度数分布を用いた新たなショット間の類似度算出	32
5.4	画像特徴量と音響特徴量を用いた映像のシーン境界候補の取得	35
5.5	階層ベイズモデルに基づくシーン境界の推定	37
5.6	画像特徴量と音響特徴量を用いたシーン分割の高精度化手法の評価 実験	39
5.6.1	提案手法が比較手法と同様の特徴量を用いたシーン分割の高 精度化の有効性確認	39
5.6.2	画像・音響特徴量を用いた階層ベイズモデルの導入による映 像の構造に注目したシーン分割手法の有効性確認	42
5.7	まとめ	53
6 章	結論	54
	参考文献	57
	謝辞	63

第1章

序論

1.1 研究背景

近年，高速・大容量な通信回線の普及に伴い，インターネットによる動画配信サービスの利用者数が増加している．特に，YouTube¹や Dailymotion²などに代表される動画配信サービスの拡大は著しく，大量のユーザが膨大な量の映像を視聴している．具体的な数値として，YouTubeが2014年に公表した統計データ³では，映像を視聴する利用者数はひと月平均で10億人を超え，利用者のひと月の視聴時間は60億時間に上る．このような社会背景から，ユーザが望む映像を検索できるシステムの必要性が高まっている．一般に，映像は意味内容の異なる場面により構成されることから，意味内容の切れ目に注目して予め映像を分割することで，効率的な検索を行うことが可能となる．そのため，映像を同一意味内容のまとまりであるシーンに分割する技術が，映像解析の基礎技術として広く研究されている．従来より，シーンは同一もしくは類似する内容を表すショットが連続する区間として定義され，ショットは1台のカメラにより切れ間なく撮像される区間であり，映像のシーン分割は，隣接するショットの意味が同一もしくは類似であるか否かを判定し統合することで行われる．ところで，既存のシーン分割手法には，代表的なものとして2つの手法が挙げられる．1つは映像の構造に注目した手法であり，もう1つは階層ベイズモデルに基づく手法である．映像の構造に注目したシーン分割手法は，編集映像に現れる類似するショットの繰り返し構造を見つけることで，シー

¹<https://www.youtube.com>

²<http://www.dailymotion.com>

³<https://www.youtube.com/yt/press/ja/statistics.html>

ン分割を可能とする．Yeung らによって提案された手法 [1] は高精度なものとして知られ，拡張法も提案された [2]-[3]．文献 [1] では，映像のシーン分割問題をグラフの分割問題に置き換えることで，シーン分割を実現している．具体的に，ショットのクラスタリングを行い，得られる各クラスタをグラフのノードとし，隣接する 2 つのショットが同一クラスタ内に存在しないときに，クラスタ間に辺を与えることで有向グラフを取得する．次に，このグラフを複数のサブグラフに分割し，各サブグラフを 1 つのシーンとみなすことでシーン分割を実現している．文献 [1] と同様の考えで，類似するショット間を辺で結び，ショット間の類似関係に基づくシーン分割手法も提案されている [2]．この手法では，注目するショットあるいはその前後にあるショット間のリンク関係が存在しない場合，注目するショットを新たなシーンの先頭ショットとして用いることでシーン分割を可能とする．さらに，文献 [3] では，2 種類の特徴量を用いて，それぞれの特徴量を用いて映像の構造を取得し，分割結果を統合して用いることでシーン分割を実現している．これらの手法では，映像の構造を解析する際，注目する繰り返し構造を示す最長の時間幅を設定する必要がある．しかしながら，最長時間幅は映像の種類によっても編集者によっても異なるため，一意に決定することが困難であるという問題が存在する．一方，階層ベイズモデルに基づくシーン分割手法は，シーン数やシーン境界の位置に関する事後確率を算出するために事前確率，条件付き確率および尤度をそれぞれ定義し，事後確率が最大になるように，マルコフ連鎖モンテカルロ法 [4] を用いて最適なシーン境界を推定する [5]．しかしながら，この手法は，その事後確率をショット間の類似度のみに注目して定義しているため，同一シーン中の隣接するショット間で類似度が低い組み合わせが存在すると，誤ってシーン境界と推定する場合がある．つまり，映像の構造を考慮せずに，ショット間の類似度のみに注目しているため，シーン分割の精度が低下するという問題が存在する．

上に示す通り，既存の代表的な 2 種類のシーン分割手法には，各々に問題が存在する．映像の構造に注目した手法においては，映像の構造を取得する際にシーンの最長時間幅の設定が困難である．一方，階層ベイズモデルに基づく手法では，映像の構造を考慮していないため，シーン分割精度が低下する．このような従来手法の問題点を解決可能なシーン分割手法の提案が必要である．

1.2 本研究の目的

そこで、本論文では、映像の構造と階層ベイズモデルの両者を用いて、映像の構造に注目する手法が抱える繰り返し構造の最長時間幅を一意に決定することが困難な問題を解決可能なシーン分割手法の提案を行う。提案手法では、映像の構造に注目した手法において、シーンの最長時間幅を一つの値に設定するのではなく、複数の値に変化させてシーン境界の候補を求め、それらの中から階層ベイズモデルに基づく手法によって最適なシーン境界を選択する。具体的に、提案手法では、最長時間幅を異なる複数の値に設定して得られたシーン分割結果をシーン境界候補とし、その中からシーン境界の位置に関する事後確率を階層ベイズモデルに基づいて定義し、これをマルコフ連鎖モンテカルロ法により最適化することで、適切なシーン境界の選択を可能とする。さらに、本論文では、階層ベイズモデルにおける事後確率の算出において、映像を構成する画像信号と音響信号の両方の特徴量への変換を行う。本論文では、異なるショットの全ての組み合わせで類似度を算出して分布を求め、比較可能とすることで、事後確率を算出する。

1.3 本論文の構成

本論文は6章からなる。以下に、各章の概要を述べる。

本論文の第2章では、従来より提案されているショット分割手法について説明し、あわせて用語の定義を行う。

第3章では、映像のシーン分割を行う従来手法の概要を述べるとともに、本論文で解決すべき課題について明確にする。

第4章では、階層ベイズモデルの導入による映像の構造に注目したシーン分割手法を提案し、従来の映像の構造に注目したシーン分割手法の問題点の解決を可能とする。提案手法では、まず、予め取得される映像の各ショットに対し、そのショットの内容を代表するフレームの取得を行い、画像特徴量を得ることで、ショット間の画像特徴量を用いた際の類似度の算出を可能とする。次に、ショット間の類似度に注目しながら、映像の繰り返し構造の最長時間幅を一意に決定するのではなく、複数の値を与えることで、映像の構造を考慮したシーン境界候補の取得が可能となる。最後に、ショット間の類似度を用いて、シーン数やシーン境界の位置に関す

る事後確率を階層ベイズモデルで定義し，それをマルコフ連鎖モンテカルロ法を用いて最適化することで，映像の構造を考慮したシーン境界候補からシーン境界の推定を可能とする．第4章によって，映像の構造を考慮したシーン境界候補より，階層ベイズモデルを用いてシーン境界の推定を行うことから，従来の映像の構造に注目したシーン分割手法の問題点の解決が可能となる．また，実験により提案手法の有効性を示す．

第5章では，画像・音響特徴量を用いてショット間の類似度算出法を紹介すると共に，その分布を考慮しながら比較可能とする定式化を導入したシーン分割手法を提案し，高精度なシーン分割結果の取得を実現する．画像・音響のそれぞれの特徴量を用いてショット間の類似度を算出し，これを用いて映像の構造を考慮したシーン境界候補の取得を行うことで，網羅的なシーン境界候補の取得を可能とする．さらに，画像・音響特徴量を同時に利用することで事後確率の算出を可能とし，高精度なシーン分割が実現される．また，提案手法の有効性を示すため，比較実験を行う．

最後に，第6章において，本研究の成果と今後の課題についてまとめる．

第2章

映像のショット分割に関する従来手法

2.1 はじめに

本章では，映像のシーン分割の前処理として行われる映像のショット分割に関する従来手法について述べる．映像の分割単位には，シーン，ショットおよびフレームがある．シーンは幾つかのショットを含み，1つのショットの中に，複数枚のフレームがある．映像のシーンを用いて映像の検索を行う場合，まとまった内容の映像が検出され，大変効率が良いため，映像をあらかじめシーンに分割する必要がある．一般に，映像のシーン境界は，映像のショット境界から選出されるため，高精度なショット分割手法が必要とされる．映像のショット分割を行う際，フレーム間で各画素の輝度値を比較したり，テンプレートの比較をしたり，色ヒストグラムや χ^2 検定法を用いたショット分割手法が提案されている．そこで，本章では，映像のシーン，ショットおよびフレームの定義について述べ，これらの分割単位間の関係を明らかにする．さらに，映像のショット分割に関する従来手法について説明し，提案手法で用いられるショット分割に関する従来手法を明確にする．

2.2 ショット分割に関する用語の定義

本節では，映像の分割単位であるシーン，ショットおよびフレームについて定義する．シーンは，同一もしくは類似する内容を表すショットが連続する区間であり，ショットは，1台のカメラにより切れ間なく撮像される区間として定義されている [6].なお，フレームは静止画像であり，映像の最小分割単位である．映像のフレーム，ショットおよびシーンの定義では，複数枚のフレームがショットに含まれ，

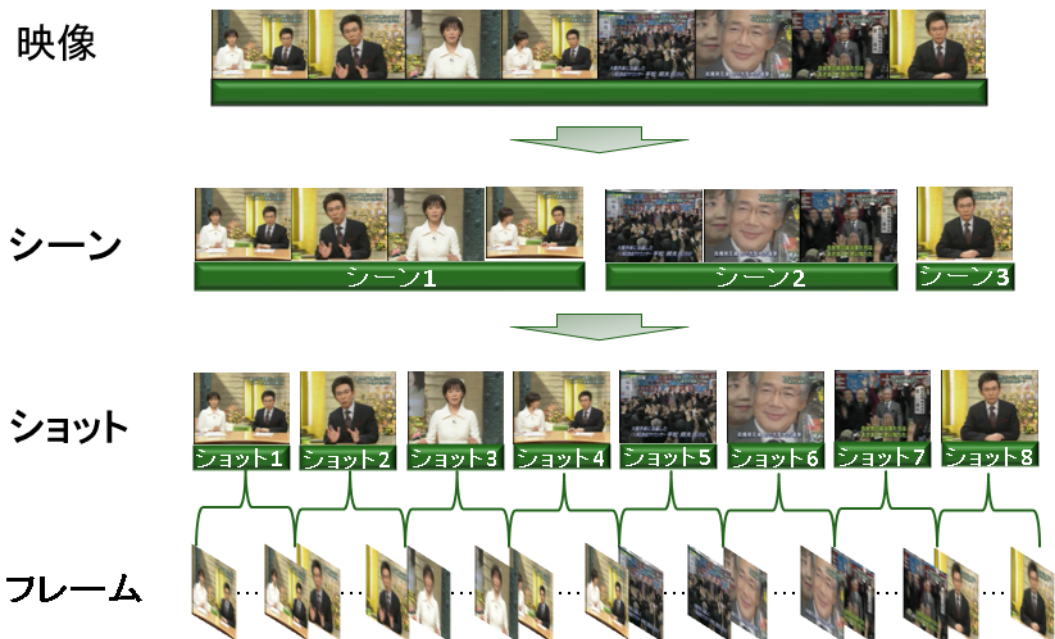


図 2.1: 映像，シーン，ショット及びフレームの例．

さらに，シーンの中に1つあるいは複数個のショットがある．シーンはショットより上位にある映像の分割単位であり，ショットはフレームの上位にある映像分割単位でもある．なお，シーン，ショットおよびフレーム間の関係をより直感的に理解するために，図 2.1 を用いて説明する．

図 2.1 より，映像はいくつかのシーンにより構成されており，1つのシーンには複数個のショットから構成され，1つのショットの中にも複数枚のフレームがあることがわかる．しかしながら，フレームが表す映像の内容に関する情報量は少ないため，一般的に，色特徴量を用いて類似する隣接のフレームを統合し，映像のショットを取得する．ここで，ショットは，一台のカメラにより切れ間なく撮像される区間として定義されているため，実際に映像のショットを検出するときに，隣接するフレーム間で特徴量の変化が大きい場合，ショット境界の検出が可能になる．文献 [7] で提案されている分割 χ^2 検定法を用いたショット境界の検出法は一般的に用いられているため，本研究でも同様に文献 [7] を用いて映像のショット分割を行う．なお，文献 [7] の詳細は本章の 2.3 節で述べる．

映像のショットは，一定程度に映像の内容を表現することができる．しかしながら，ショットは，図 2.1 のショット 2 とショット 8 のように同一人物が写されている

にもかかわらず，異なるカメラにより撮像されている映像であるため，異なる2つのショットとなる．このように，ショットを用いて，映像検索を行うことは効率的ではない．ここで，もう一つ映像の分割単位であるシーンについて考える．シーンは同一もしくは類似する内容を表すショットが連続する区間として定義されている．図2.1のショット1, ショット2, ショット3, およびショット4は，スタジオのシーン(シーン1)である．またショット5, ショット6 およびショット7は選挙に関する一連の報道シーン(シーン2)である．映像のシーン境界は一般に映像のショット境界の中から選択して用いるため，次節でショット分割の従来手法について説明する．また，映像のシーン分割に関する従来手法は，第3章で述べる．

2.3 映像のショット 分割に関する従来手法

映像のシーンは複数のショットより構成される．カメラのスイッチのオン/オフ，カメラ間での切り替え，或は映像の編集技術等の操作により，複数のショットは，異なるカメラの操作により撮像される．また，映画やテレビ番組において，ショットは異なるカメラから生成され，次々にショットを追加することで，ストーリーラインを構成する[5]．2.2節で，ショットは1台のカメラにより切れ間なく撮像される区間として定義されている．一般に，ショット境界では，前後の隣接するフレームの画像特徴が大きく変化する．したがって，ショット分割において，隣接するフレーム間の特徴が変化するか否かを捉えることが有効であり，隣接するフレーム間の特徴量間の変化に注目するショット分割手法が種々提案されている．具体的には，フレームの画素の輝度値に注目するものや，テンプレートを用いるもの，尤度やヒストグラムなどを比較することでショット分割を実現する手法が提案されている．以下に，各々について説明する．

(1) フレームの画素の輝度値の比較によるショット分割手法

フレームの画素の輝度値の比較によるショット分割に関する従来手法は，隣接する2フレームの同一位置にある画素間の輝度値の差を計算し，それに予め設定した閾値と比較することにより，ショット分割を可能とする．ここで，隣接する2フレーム間の全ての画素の輝度値の差 $D(f_i, f_{i+1})$ は以下の式により

与えられる．

$$D(f_i, f_{i+1}) = \frac{\sum_{x=1}^X \sum_{y=1}^Y |f_i(x, y) - f_{i+1}(x, y)|}{XY} \quad (2.1)$$

上式において， i 番目のフレーム f_i の画素 (x, y) の輝度値を $f_i(x, y)$ とし，各フレームのサイズは $X \times Y$ 画素である．また，カラー画像に適用する場合は，式 (2.1) を変形することで得られた次式を用いる．

$$D(f_i, f_{i+1}) = \frac{\sum_{x=1}^X \sum_{y=1}^Y \sum_c |f_i(x, y, c) - f_{i+1}(x, y, c)|}{XY} \quad (2.2)$$

この手法は隣接する2フレーム間の輝度値の差が簡便に計算できる一方，撮像される物体の位置変化などに大きく影響を受ける．特に，動きが激しい場合には，隣接する2フレーム間の輝度値の差は既定の閾値を超え，ショット境界に誤差が生じる．

(2) テンプレートの比較によるショット分割手法

(1) の手法では，隣接する2フレーム間において，対応する画素の輝度値の差を求めることによりショット境界の検出を可能とした．しかしながら，単純に同一位置に存在する画素間の輝度値の差を求めることは，撮像する物体の移動によりショット境界の誤検出が生じる．そのため，隣接する2フレームが存在する物体の対応を求めるため，物体を構成する画素を中心とする $n_x \times n_y$ の領域をマッチング対象とする．これにより，撮像される物体の移動量を推定し，位置を補正した後，2フレーム間の輝度値の差を算出する．具体的な隣接する2フレーム間の輝度値の差を，以下に示す．

$$D(f_i, f_{i+1}) = \sum_{y=-n_y/2}^{y=n_y/2} \sum_{x=-n_x/2}^{x=n_x/2} |f_i(x, y) - f_{i+1}(x, y)| \quad (2.3)$$

上式を用いて，まず，隣接する2フレームの各窓内において，輝度値の差を算出し，その値は既定の閾値を超えた数を算出する．次に，この数が一定の数を超えた際，映像の撮像内容が変化したと判断し，ショット分割を行う．

(3) 尤度比に基づくショット分割手法

隣接するフレーム間のサブ窓に対するグレースケールの二次統計により算出される尤度比に基づくショット分割手法も提案されている． $m[f_i(x, y)]$ と

$\sigma[f_i(x,y)]$ をそれぞれフレーム f_i のあるサブ画像の輝度値の平均と標準偏差とした際、2枚のサブ画像間の尤度比は以下の式により算出される。

$$D(f_i, f_{i+1}) = \left[\frac{\sigma[f_i(x,y)] + \sigma[f_{i+1}(x,y)]}{2} + \left(\frac{m[f_i(x,y)] - m[f_{i+1}(x,y)]}{2} \right)^2 \right]^2 \quad (2.4)$$

上式を用いることで、2枚のサブ画像間の輝度値の差が与えられ、これに閾値処理を施し、ショット分割が実現される。

(4) 色ヒストグラムに基づくショット分割に関する従来手法

撮像時にカメラや物体の移動により発生する影響は、隣接する2フレーム間の色ヒストグラムを比較することで、影響の抑制が可能である。なぜなら、隣接する2フレームにおいて、撮像される物体の場所が変化したとしても、色ヒストグラムを用いて比較すれば、その差が小さくなると考えられるからである。また、色ヒストグラムは、撮像物体の回転に影響を受け難く、撮像方向や撮像される物体のサイズの変化にもロバストである。さらに、隣接する2フレーム間の色ヒストグラムの比較は、簡便であり、実用的である。なお、ヒストグラムを用いた場合の隣接2フレーム間の色ヒストグラムの差は次式より算出される。

$$D(f_i, f_{i+1}) = \sum_j^n |h_i(j) - h_{i+1}(j)| \quad (2.5)$$

なお、 $h_i(j)$ は第 i 番目のフレームの第 j 番目のビンの値を表す。なお、カラー画像に対し、上式は以下の式に変形される。

$$D(f_i, f_{i+1}) = \sum_c \sum_j^n |h_{i,c}(j) - h_{i+1,c}(j)| \quad (2.6)$$

色ヒストグラムの利用により、撮像物体移動の影響によるショット分割の精度劣化を抑止することができるが、ズームインやズームアウトなどにおいて、依然としてショット分割に誤差が生じる。ショット分割手法において、あらかじめ設定される閾値は、一般に、映像の内容が大きく変化するショット境界のために設定された閾値である。しかしながら、カメラの操作により、徐々に変化するショットの境界も存在し、その場合は輝度値の差が小さくなり、あらかじめ設定する閾値と比較しても、正しいショット分割結果が得られない

場合がある．一方，この問題を回避するために，閾値を小さい値に設定すると，カメラや撮像物体の動きによる隣接するフレーム間の輝度値の差でも閾値を超えてしまい，誤ってショットの境界が検出され，過分割が発生する．このように，色ヒストグラムに基づくショット分割手法では，動きや雑音が少ない場合にロバストであるが，背景や撮像物体の移動やズームなどの映像の編集により，ショット境界が高精度に得られない場合がある．この問題を解決するために， χ^2 検定法を用いた色ヒストグラムに基づくショットの分割手法が提案されている．次に，それを説明する．

(5) χ^2 検定法に基づくショット分割手法

χ^2 検定法は，様々なデータに広く用いられている手法である． χ^2 検定法に基づくショット分割手法は，2つのヒストグラムの比較を，各度数の単純な差ではなく，カイ二乗検定を用いて行う．なお，各度数の差を2乗することで，ヒストグラム間の類似性を有効に表現することが可能になる．また，分母の $h_{i,c}(j)$ で割ることで正規化を行う．

$$D(f_i, f_{i+1}) = \sum_{j=0}^{63} \frac{\{h_{i,c}(j) - h_{i+1,c}(j)\}^2}{h_{i,c}(j)} \quad (2.7)$$

なお，各画素は64色種を取り， $h_{i,c}(j) - h_{i+1,c}(j)$ は2フレーム中の第 j 番目のヒストグラムの差を表す．式(2.7)に基づいて算出される値が大きければ，2フレーム間の差が大きくなる．この性質に基づいて，式(2.7)から算出される評価値を用いて，ショット分割を可能とする．しかしながら， χ^2 検定法に基づくショット分割手法を用いても，撮像時に，フラッシュやカメラ前に人が通過すると，ショット分割の精度が劣化する場合がある．

(6) 分割 χ^2 検定法に基づくショット分割手法

(5)の手法の問題を解決するために，文献[7]では，分割 χ^2 検定法に基づくショット境界の検出手法が提案されている．文献[7]では，まず，各フレームを 4×4 個のブロックに分割し，それぞれのブロックに位置を示す番号を付ける．次に，隣接する2フレームのそれぞれ同じ番号のブロックの画素の色ヒストグラムを比較し， χ^2 検定法に基づくショット分割手法と同様に計16個の評価値を得る．これらの16つの評価値のうち，値の小さいものから8つを選

択し，その値の総和を最終的な評価値とする．この評価値を用いてショット分割を行うことで，ノイズの影響が抑制され，高精度にショット分割を実現している．

2.4 まとめ

本章では，映像のシーン，ショットおよびフレームのそれぞれの用語を定義し，これらの用語の関係を，図を用いて説明した．また映像のシーン境界は，一般に映像のショット境界の中から選択することから，本章では，ショット分割に関する従来手法について説明した．2.3節で説明した従来に提案されているショット分割手法では，分割 χ^2 検定法に基づくショット分割手法は他の従来手法より高精度にショット分割を実現していると先行研究[7]で報告されている．このため，本研究でも，映像をあらかじめショットに分割する際，分割 χ^2 検定法に基づくショット分割手法を用いる．

第3章

映像のシーン分割の従来手法

3.1 はじめに

本章では，映像のシーン分割の従来手法の概要とそれらの問題点について述べる．

映像のシーン分割は従来より広く研究が行われ，映像のシーン分割手法は種々提案されている [1]-[3],[5],[8]-[50]．文献 [8]-[39] は，画像の特徴量のみを用いたシーン分割手法であり，[40]-[50] は画像と音響の特徴量を用いたシーン分割手法である．各々が高精度なシーン分割を目指しているが，特徴量の改良だけでは精度向上に限界がある．現在では，Yeung らの映像の構造に注目する手法 [1] やその改良手法 [2]-[3]，さらに Zhai らの階層ベイズモデルに基づく手法 [5] を用いるのが主流である．本章では，映像の構造に注目したシーン分割手法 [1]，および階層ベイズモデルに基づくシーン分割手法 [5] の従来研究の概要を説明し，問題点について述べる．さらに，これらの手法の問題点を踏まえて，映像のシーン分割手法の高精度化を実現するために，本論文で解決すべき課題について明確にする．

3.2 映像の構造に注目したシーン分割に関する従来手法および階層ベイズモデルに基づくシーン分割に関する従来手法

本節では、既存のシーン分割手法に代表的な映像の構造に注目するシーン分割手法および階層ベイズモデルに基づくシーン分割手法について説明する．文献[1]で提案されているシーン分割手法は映像の構造に注目するシーン分割手法のベンチマーキング手法である．また、階層ベイズモデルに基づくシーン分割手法は文献[5]で提案されている．本節では、3.2.1節で映像の構造に基づくシーン分割手法[1],およびその改良手法[51]について説明し、3.2.2節で階層ベイズモデルに基づくシーン分割手法[5]について述べる．

3.2.1 映像の構造に注目するシーン分割手法

本節では、図3.1(a)を用いて、映像の構造に基づくシーン分割手法(従来手法[1])および従来手法[1]の改良手法である従来手法[51]について説明する．従来手法[1]ではまず、対象の映像中に含まれるショットに対し完全連結法によるクラスタリングを行う．具体的に、 N_s 個のショット $s_i (i = 1, 2, \dots, N_s)$ をそれぞれ含むクラスタを $C_1 = \{s_1\}, C_2 = \{s_2\}, \dots, C_{N_s} = \{s_{N_s}\}$ とし、次式に示す異なるクラスタ C_u と C_v 間の類似度 $CSim(C_u, C_v)$ が最大となる2つのクラスタを順に統合していくことで実現する．

$$CSim(C_u, C_v) = \min_{s_i \in C_u, s_j \in C_v} Sim(s_i, s_j) \quad (3.1)$$

なお、 $Sim(s_i, s_j)$ はショット s_i および s_j 間の類似度を表す．具体的に、従来手法[1]では各ショットからキーフレーム¹を求め、2ショットそれぞれのキーフレームの色ヒストグラムからヒストグラムインターセクションを算出することで、ショット s_i およびショット s_j 間の類似度 $Sim(s_i, s_j)$ を求める．上式で算出されるクラスタ間の類似度 $CSim(\cdot, \cdot)$ があらかじめ設定された閾値 Th_C よりも小さくなる、またはクラスタ数が1となるまで統合を行い、得られるクラスタを $C_u (u = 1, 2, \dots, N_C; N_C$ は得

¹キーフレームとは、映像の内容を代表するフレームである．ショットの先頭フレーム、中間フレーム、最終フレームをキーフレームとして用いる場合がある．なお、ショットの先頭フレームをキーフレームとし、ショット中のフレームがキーフレームと比較して、輝度値の差が予め設定する閾値を超えた際のフレームを新たなキーフレームとして追加する手法も存在する．提案手法は、後者の手法を用いて、ショットのキーフレームの選出を行う．なお、映像のシーン分割において、計算量削減のために、一般にショット中のキーフレーム間の類似度を、ショット間の類似度として用いる．

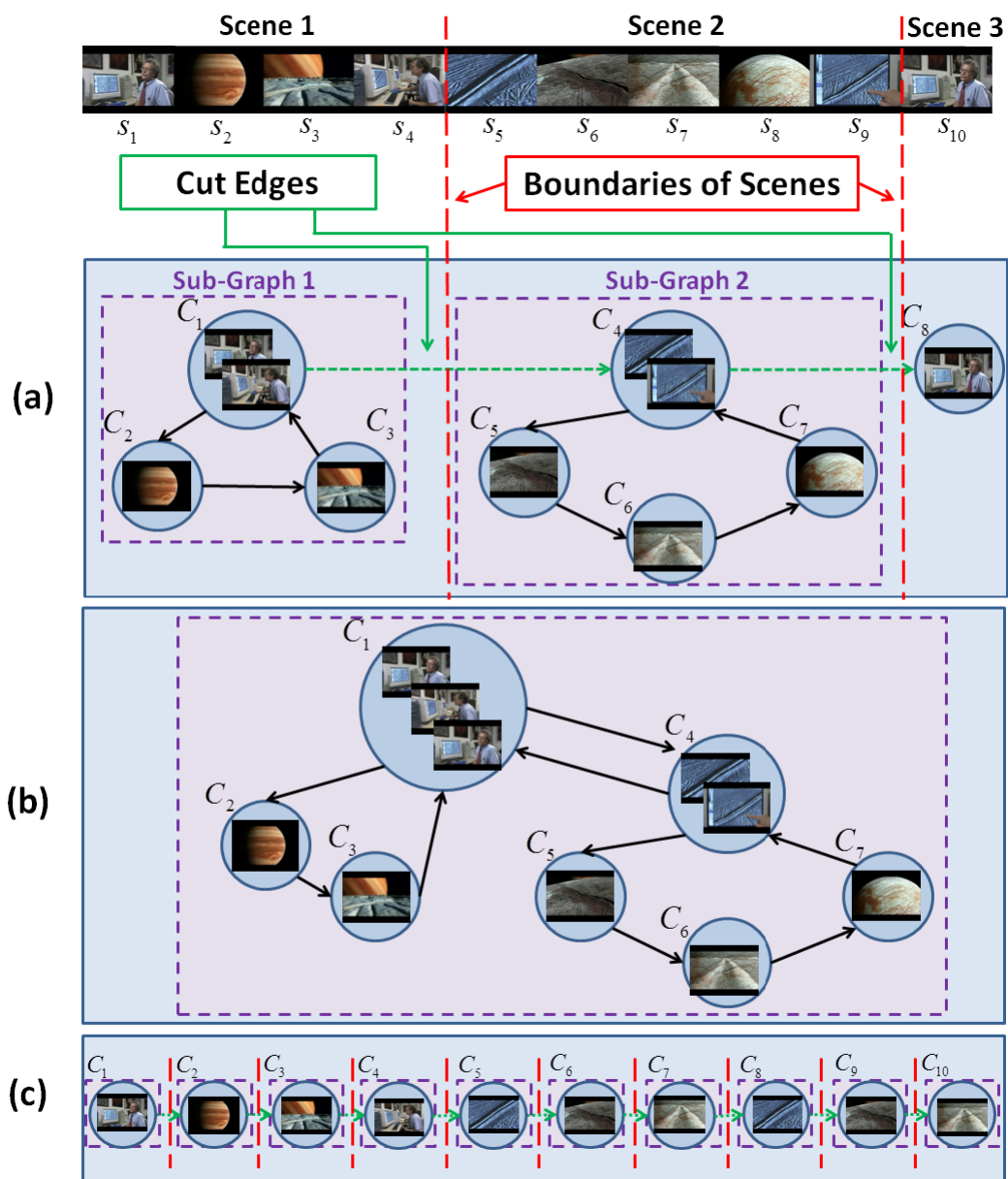


図 3.1: (a) シーンの最長時間幅を適切に設定した場合のシーン分割結果. (b) シーンの最長時間幅が適切に設定されていない場合 (シーンの最長時間幅が長すぎる場合) のシーン分割結果. (c) シーンの最長時間幅が適切に設定されない場合 (シーンの最長時間幅が短すぎる場合) のシーン分割結果. ただし, $s_1 \sim s_{10}$ はショット, $C_1 \sim C_{10}$ はクラスタであり, 緑点線はカット辺である.

られるクラスタの総数)とする．図 3.1(a) の例では，ショットのクラスタリングにより，得られるクラスタが 8 個存在し， $C_u(u = \{1, 2, \dots, 8\})$ で表されている．次に，得られたクラスタ C_u をノードとし，隣接するショット同士が異なるクラスタに属する場合，それらのクラスタ間に辺を与えることで，有向グラフを作成する．従来手法 [1] では，得られる有向グラフに対し，カット辺²(図 3.1 中における緑色の点線の辺)を除去することで，有向グラフを複数のサブグラフに分割する．このとき，各サブグラフを 1 つのシーンとみなすことでシーン分割を実現する．なお，図 3.1(a) では，緑の点線のカット辺を削除することで，Scene1~Scene3 が取得される．

以上に示す従来手法 [1] に対し，我々はその改良手法として従来手法 [51] を提案した．従来手法 [51] では，映像中の物体に注目し，算出される特徴点に関する 128 次元の SIFT 特徴量を画像特徴量として用いた．さらに，音響特徴量である音楽，音声，無音および雑音の 4 つのオーディオクラスへの帰属度を導入し，これら複数の特徴量を用いることで，従来手法の高精度化を行った．このとき，SIFT 特徴量を用いた際，2 ショットのキーフレーム間で一致したキーポイントの数に基づいてショット間の類似度を算出している．また，音響特徴量を用いた際には，ショットの音響信号が 4 つのオーディオクラスに帰属する確率(ヒストグラム)に関するヒストグラムインターセクションにより 2 ショット間の類似度を求めている．

従来手法 [1],[51] は，隣接するショット間で類似性が低い場合においても，類似するショットが出現する場合，その間のショットを含めて 1 つのシーンに統合することが可能な手法である．具体的に，図 3.1(a) の例に示すように，4 つのショットで構成される Scene1 において，ショット s_1 と s_4 の間のショット， s_2 と s_3 が互いに類似しない場合においても，ショット s_1 と s_4 が類似するため，図 3.1(a) に示すようにそれらが 1 つのクラスタにまとめられ，ショット $s_1 \sim s_4$ を 1 つのシーンとして取得することが可能となる．

しかしながら，十分に離れた時刻に存在するショットにおいて，類似したものが存在する場合，類似するショットおよびその間にある全てのショットが 1 つのシーンに統合されてしまう問題が存在する．したがって，従来手法 [1],[51] では，これを防ぐために，ショット間の類似度の算出において，シーンの最長時間幅を設定し，これ以上離れた時刻に類似したショットが存在した場合においては，類似度を 0 と

²有向グラフにおいて，1 本の辺が除かれた際に，有向グラフが 2 つのサブグラフに分かれる辺をカット辺という．

し、クラスタの統合を行わない処理を導入している．図3.1(a)は、シーンの最長時間幅が適切に設定された場合の例である．この場合、ショット s_1, s_4, s_{10} は互いに類似しているが、シーンの最長時間幅を用いているため、 s_1 および s_4 はクラスタ C_1 にまとめられるが、 s_{10} は別のクラスタに分かれ、正しいシーン分割が可能となる．

これに対して、シーンの最長時間幅が正しく設定されなかった例を2つ示す．まず、図3.1(b)は、シーンの最長時間幅が長すぎる例である．具体的に、シーンの最長時間幅を映像の長さと同じとした．この場合、類似ショット s_1, s_4 および s_{10} は1つのクラスタにまとめられ、従来手法 [1],[51]におけるシーン分割を行う際に注目するカット辺が存在しないため、映像全体が1つのシーンとなってしまう．一方で、図3.1(c)に示すようにシーンの最長時間幅が短すぎる場合、具体的にシーンの最長時間幅をショットの長さで設定した場合は、各々のショットが1つのクラスタになり、過剰な分割になることが確認できる．

このように、シーンの最長時間幅の設定によって、図3.1(a),(b),(c)に示すように、シーン分割の結果が大きく変わるため、その設定が重要である．しかしながら、これらの従来手法では、最長時間幅を適切に決定することが困難であるという問題が存在した．

3.2.2 階層ベイズモデルに基づくシーン分割手法

本節では、階層ベイズモデルに基づくシーン分割手法について説明する．階層ベイズモデルは、ベイズ統計を用いる手法の1つであり、一般に推定対象に関する事後確率を最大とすることで、問題の解を与える．このとき、事後確率は、尤度と事前確率から導かれるが、それらが複雑な定義によって与えられる場合、事後確率も同様に複雑な分布となる．このような場合、解を求めることは困難となる．この問題を解決するために、複雑な分布に対しても容易に解の取得を可能とするマルコフ連鎖モンテカルロ法 [4]が注目されている．マルコフ連鎖モンテカルロ法では、簡単な条件付き事後分布からのサンプリングを繰り返すことで、直接推定が困難である複雑な事後分布の再現を可能とする．マルコフ連鎖モンテカルロ法に基づいて事後分布を求める際には、Metropolis-Hastings アルゴリズム [52] や Metropolis-Hasting Green (MHG) アルゴリズム [53] 等が一般的に広く用いられている．そこで、従来手法 [5]では、映像のシーン分割問題を、ショット境界におけるシー

ン境界になりうる確率の算出問題に置き換える．具体的に，従来手法 [5] では，階層ベイズモデルを用いて，シーン数やシーン境界の位置に関する事後確率を定義し，マルコフ連鎖モンテカルロ法による最適化を行うことでシーン境界の推定を行う手法である．このとき，事後確率分布の関数が複雑であるため，複雑なままにサンプリングする技法の 1 つである MHG アルゴリズムを用いて，問題の解決を行う．

階層ベイズモデルに基づくシーン分割手法 [5] では，MHG アルゴリズム [53] を用いて繰り返しシーン数およびシーン境界位置に関する状態の更新を行うことで，最適なシーン境界の推定を実現する．なお，状態を $J = \{k, \theta_k\}$ と定義したとき， k はシーン数， θ_k はシーン境界位置をそれぞれ表す．具体的な処理を以下に示す．

(i) ショット境界より，ランダムに 1 つを選択し，映像を 2 つのシーンに分割することで，初期状態 J を得る．

(ii) 以下の処理 (a) ~ (c) をシーン境界位置が変化しなくなるまで繰り返す．なお，以降の説明のため， t 回目の更新におけるシーン数 k ，シーン境界位置 θ_k の状態を J_t ，アップデートにより得られるシーン数を k' ，シーン境界位置 $\theta_{k'}$ の新しい状態を J'_t とする．

(a) 閾値 $Th_\alpha (\in [0, 1])$ をランダムに生成する．

(b) J_{t-1} に対して，Diffusion および Jump によるアップデートを行い，新しい状態 J'_{t-1} を得る．このとき，アップデートの種類は文献 [53] に従って選択する．

(c) 状態 J_{t-1} と J'_{t-1} に関する確率 $\alpha(J_{t-1}, J'_{t-1})$ が閾値 Th_α よりも大きい場合， $J_t = J'_{t-1}$ とし，それ以外の場合， $J_t = J_{t-1}$ とする．また， $t \leftarrow t + 1$ とする．なお， $\alpha(J_{t-1}, J'_{t-1})$ は次式で算出される．

$$\alpha(J_{t-1}, J'_{t-1}) = \min \left\{ 1, \frac{p(J'_{t-1} | f) q(J'_{t-1}, J_{t-1})}{p(J_{t-1} | f) q(J_{t-1}, J'_{t-1})} \right\} \quad (3.2)$$

式 (3.2) には，事後確率 $p(J_{t-1} | f)$ および遷移確率 $q(J_{t-1}, J'_{t-1})$ が含まれており，以下ではそれぞれについて具体的に説明する．

まず，事後確率 $p(J_{t-1} | f)$ は $J_{t-1} = \{k, \theta_k\}$ であることより，

$$p(J_{t-1} | f) \propto \mathbb{L}(f | \theta_k) p(\theta_k | k) p(k) \quad (3.3)$$

が成り立つ．このとき，検出されたシーン数が k である事前確率 $p(k)$ ， k が与えられた際のシーン境界の位置 θ_k に関する条件付き確率 $p(\theta_k | k)$ および θ_k が与えられたときの尤度 $\mathbb{L}(f | \theta_k)$ は，事後確率と比例関係を持つ．ただし，式 (3.3) 中の f は映像全体の特徴量である．以下では， $p(k)$, $p(\theta_k | k)$, $\mathbb{L}(f | \theta_k)$ の具体的な算出法を説明する．

• **事前確率 $p(k)$**

シーン数 k が平均 λ のポアソン分布に従うと仮定し，事前確率 $p(k)$ は次式のように算出される．

$$p(k) = e^{-\lambda} \frac{\lambda^k}{k!} \cdot I_{[1, k_{\max}]}(k) \quad (3.4)$$

ただし， $I_{[1, k_{\max}]}(k)$ は， $1 \leq k \leq k_{\max}$ ならば 1，それ以外で 0 となる関数である．

• **条件付き確率 $p(\theta_k | k)$**

映像に N_s 個のショット境界および k 個のシーンが存在する場合， $p(\theta_k | k)$ は与えられた $N_s - 1$ 個のショット境界から $k - 1$ 個の部分集合を選択する確率となる．したがって，条件付き確率 $p(\theta_k | k)$ は以下のように定義される．

$$p(\theta_k | k) = \frac{1}{C_{k-1}^{N_s-1}} = \frac{(k-1)!(N_s-k)!}{(N_s-1)!} \quad (3.5)$$

• **尤度 $\mathbb{L}(f | \theta_k)$**

尤度 $\mathbb{L}(f | \theta_k)$ は次式のように定義される．

$$\mathbb{L}(f | \theta_k) = \left\{ \prod_{n=1}^{k_u} \mathbb{L}(f_n | S_n) \right\}^{\frac{1}{k_u}} \quad (3.6)$$

ここで， k_u はアップデートの際に取得されるシーンの数であり，シーン S_n に N_{S_n} 個のショット $\{s_n^1, \dots, s_n^{N_{S_n}}\}$ が存在する場合， $\mathbb{L}(f_n | S_n)$ は次式により算出される．

$$\mathbb{L}(f_n | S_n) = \frac{1}{N_{S_n}^2} \sum_{i=1}^{N_{S_n}} \sum_{j=1}^{N_{S_n}} \text{Sim}(s_n^i, s_n^j) \quad (3.7)$$

ただし， f_n はシーン S_n で算出される映像特徴量であり， $\text{Sim}(s_n^i, s_n^j)$ は，ショット s_n^i および s_n^j 間の類似度である．

次に，式 (3.2) 中の $q(J_{t-1}, J'_{t-1})$ は， J_{t-1} から J'_{t-1} への遷移確率を表す．文献 [5] では，Diffusion および Jump の 2 種類のアップデート処理が存在する．

Diffusion のアップデート処理は，隣接する 2 つのシーン間の境界を 2 つのシーン中に存在する別の境界に移動する処理である．また，Jump のアップデート処理は 2 種

類の処理が存在し、それぞれは1つのシーンを新たに2つのシーンに分割する split 処理、および2つのシーンを1つのシーンに統合する merge 処理である。Diffusion および Jump について詳しく説明する。

• Diffusion 処理

Diffusion 処理は隣接するシーン境界の位置を更新する処理である。この更新処理は以下のように行われる。離散一様分布 $[1, k-1]$ からランダムに映像中からシーン番号 m を決め、シーン $S_m = \{s_m^1, s_m^2, \dots, s_m^{n_m}\}$ と $S_{m+1} = \{s_{m+1}^1, s_{m+1}^2, \dots, s_{m+1}^{n_{m+1}}\}$ の間に境界候補を設定する。ただし、 s_m^l は m 番目のシーンに存在する l 番目のショットである。このとき、新しい境界候補 s^l は $[s_m^1, s_{m+1}^{n_{m+1}}]$ の範囲の中で更新前の境界 s_{m+1}^1 の映像全体におけるショット番号を平均とする以下の正規分布を最大とするもので与えられる。

$$p(t) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left\{-\frac{\Delta t^2}{2\sigma^2}\right\} I_{[s_m^1, s_{m+1}^{n_{m+1}}]}(t) \quad (3.8)$$

上式において、 Δt は新しい境界候補 s^l と更新前の境界 s_{m+1}^1 の映像全体におけるショット番号の差で表され、 σ は経験的に $\sigma = 2$ と設定する。また、 $I_{[s_m^1, s_{m+1}^{n_{m+1}}]}(t)$ は t が $[s_m^1, s_{m+1}^{n_{m+1}}]$ の範囲に存在する場合1であり、それ以外では $I_{[s_m^1, s_{m+1}^{n_{m+1}}]}(t) = 0$ となる。以上により、得られる境界候補 s^l により、更新されたシーン S'_m はショット $\{s_m^1, \dots, s^{l-1}\}$ を含み、 S'_{m+1} はショット $\{s^l, \dots, s_{m+1}^{n_{m+1}}\}$ を含む³。

次に、新しい境界 s^l が更新前の境界位置 $s_m^{n_m}$ に代わる確率は s^l が $s_m^{n_m}$ より前の時刻に存在する場合、 $q(J_{t-1}, J'_{t-1}) = (1/k - 1)p(t)$ であり、そうでない場合、 $q(J_{t-1}, J'_{t-1}) = (1/k - 1)(1 - p(t))$ となる。文献 [5] では、上記確率に基づき、MHG アルゴリズムにより s^l が $s_m^{n_m}$ に代わるか否かを決定している。

• Jump 処理

Jump 処理には、split および merge の2つの処理が存在する。split 処理は、シーン S_n を新たな2つのシーン S'_n および S'_{n+1} に分割する処理であり、式 (3.9) のように定義される。

$$q_{split}(J_{t-1}, J'_{t-1}) = \frac{1}{k} \mathbb{L}(S'_n | f'_n) \mathbb{L}(S'_{n+1} | f'_{n+1}) \quad (3.9)$$

³本文では、シーン境界の位置をカット点の後のシーンの先頭ショットで表現する。

また，merge 処理は，シーン S_n および S_{n+1} を統合して 1 つのシーン S'_n にする処理であり，次式のように定義する．

$$q_{merge}(J_{t-1}, J'_{t-1}) = \frac{1}{k-1} \mathbb{L}(S'_n | f'_n) \quad (3.10)$$

文献[5]では，split 処理を行う際には，処理(ii)-(c)の $q(J_{t-1}, J'_{t-1})$ に $q_{split}(J_{t-1}, J'_{t-1})$ を代入し，merge 処理を行う際には， $q(J_{t-1}, J'_{t-1})$ に $q_{merge}(J_{t-1}, J'_{t-1})$ を代入する．

階層ベイズモデルに基づくシーン分割手法[5]では，シーン数やシーン境界の位置に関する事後確率を与え，これを最大となるように，シーン境界の候補であるショット境界の中からシーン境界の位置を推定している．従来手法[5]では，映像の構造に基づくシーン分割手法における，シーンの最長時間幅等の設定が困難である問題に注目し，映像の構造を用いずにシーン分割を可能とする．しかしながら，事後確率の算出にあたり，ショット間の類似度のみに注目しているため，同一シーン中の隣接するショット間で類似度が低い組み合わせが存在すると，誤ってシーン境界と推定する場合がある．つまり，映像の構造を考慮せずに，ショット間の類似度のみに注目しているため，シーン分割の精度が低下するという問題が存在した．

3.3 従来手法における解決すべき課題

本節では，映像のシーン分割の高精度化を実現するために，本論文で解決すべき課題について明確にした．3.2 節で示したように，既存の代表的な 2 つのシーン分割手法として，1 つは映像の類似するショットが繰り返し現れる構造に注目したシーン分割手法であり，もう一つは階層ベイズモデルに基づくシーン分割手法である．映像の類似するショットが繰り返し現れる構造に注目したシーン分割手法では，映像のショットのクラスタをノードとし，隣接するショットが異なるクラスタにある場合，それらのクラスタ間に辺を与えることで，有向グラフを作成する．次に，カット辺を見つけて，除去することで，グラフをサブグラフに分割する．映像の構造に注目するシーン分割手法では，映像の分割問題をグラフの分割問題に置き換え，得られるサブグラフを一つのシーンとしてみなすことで，シーン分割を可能とする．このとき，映像の構造を取得する際，注目する繰り返し構造を示す最長の時間幅を設定する必要がある．しかしながら，最長時間幅は映像の種類

によっても編集者によっても異なるため、一意に決定することが困難であるという問題が存在した。一方、階層ベイズモデルに基づくシーン分割手法では、シーン数やシーン境界の位置に関する事後確率の最大化により、最適なシーン境界を検出することが可能となる。しかしながら、事後確率の算出は、ショット間の類似度のみに注目しているため、同一シーン中の隣接するショット間で類似度が低い組み合わせがある場合、誤ってシーン境界の推定を行う可能性がある。

上記の2種類の手法のそれぞれの問題に対し、映像の類似するショットが繰り返し現れる構造に注目したシーン分割手法のシーンの最長時間幅を一意に決定するのではなく、複数の値に変化させてシーン境界の候補を求め、それらの中から従来手法[5]により最適なシーン境界を推定することで、各々の手法の欠点を解決可能とし、互いの手法の利点を生かした高精度なシーン分割手法の実現が期待できる。

3.4 まとめ

本章では、映像のシーン分割に関する従来手法について説明した。また、従来手法における問題点について述べ、画像・音響特徴量を導入し、映像の構造と階層ベイズモデルの両者の利用によるシーン分割の高精度化を実現するために、本論文で解決すべき課題について明確にした。

第4章

階層ベイズモデルの導入による映像の構造に注目したシーン分割手法

4.1 はじめに

本章では，階層ベイズモデルの導入による映像の構造に注目したシーン分割手法を提案する．まず，提案手法では，あらかじめ取得された各ショットのキーフレームを抽出し，そのキーフレームの画像特徴量を算出し，ショット間の類似度を求める．次に，ショット間の類似度および複数の値に設定された映像のシーンの最長時間幅により映像の構造を取得し，シーン境界候補を得る．さらに，映像の構造を考慮したシーン境界候補に対し，階層ベイズモデルで定義された事後確率を，マルコフ連鎖モンテカルロ法を用いて最適化を行うことで，シーン境界の推定を可能とする．本章の最後では，実際にテレビで放送される映像に対して提案手法を適用し，その有効性を確認する．

4.2 画像特徴量を用いたショット間の類似度算出

本節では，色ヒストグラム特徴量を用いたショット間の類似度算出について説明する．提案手法では，あらかじめ映像にショット分割手法[7]を適用し，ショットを取得する．次に，従来手法[1]と同様にショット間の類似度を算出する．なお，従来手法[1]は，色ヒストグラム特徴量のみを用いているため，提案手法でも同様に色ヒストグラム特徴量を用いる．映像中の異なるショット s_i および $s_j(i, j = 1, 2, \dots, N_s; N_s$

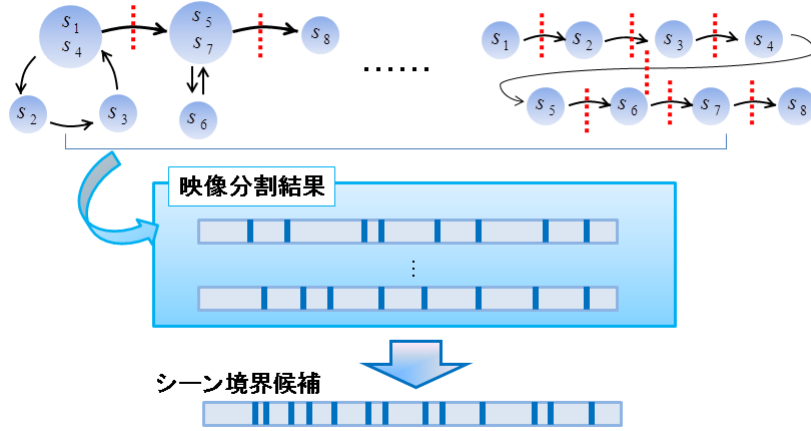


図 4.1: 映像の構造を考慮したシーン境界候補の取得

はショットの総数) 間の類似度 $Sim(s_i, s_j)$ を次式のように定義する.

$$Sim(s_i, s_j) = \begin{cases} \max_{f_l \in s_i, f_m \in s_j} Sim_f(f_l, f_m) & \text{if } D_T(s_i, s_j) \leq Th_T \\ 0 & \text{otherwise} \end{cases} \quad (4.1)$$

ただし, f_l, f_m はそれぞれショット s_i, s_j に含まれるキーフレームであり, $Sim_f(f_l, f_m)$ はフレーム f_l と f_m の色ヒストグラム間の類似度を表している. $D_T(s_i, s_j)$ はショット s_i, s_j およびその間に含まれる全てのショットに存在するフレームの総数であり, Th_T はあらかじめ設定された閾値である.

4.3 映像の構造を考慮したシーン境界候補の取得

本節では, 提案手法における映像の構造に注目したシーン境界候補の取得について説明する. 3.2.1 節で説明したように, 映像の構造に注目したシーン分割手法では, シーンの最長時間幅等の閾値を適切な値に設定することが困難である. このため, 提案手法では, ショットのクラスタリングを行う際の閾値 Th_C および最長時間幅 Th_T の閾値を様々な値に変化させ, 複数のシーン境界候補の取得を行う. なお, シーン境界候補の取得方法を図 4.1 に示す.

図 4.1 より, 提案手法ではショットのクラスタリングを行う際の閾値 Th_C および最長時間幅 Th_T の閾値を変化させ, n 個の映像のシーン分割結果を取得する. 次に, n 個のシーン分割結果に注目し, ショット境界において, 一回でもシーン境界とし

て選択された場合、そのショット境界をシーン境界候補として用いる。このようにシーン境界候補を選択する理由として、特徴量が大きく変化しない場合においても、正しくシーン境界をシーン境界候補として取得可能である。

以上により、映像の構造に注目したシーン境界候補が取得される。ただし、提案手法では、取得された全ての境界をシーン境界候補として用いているため、その結果は、正しいシーン境界の数より大きくなり、このままでは、映像の過分割となる。そこで、提案手法では、取得されたシーン境界候補の中から、階層ベイズモデルを用いて、最適なシーン境界を推定することで、シーンの最長時間幅等の閾値の設定が困難である問題の解決を試みる。

4.4 階層ベイズモデルに基づいたシーン境界候補からシーン境界の推定

本節では、提案手法における階層ベイズモデルに基づくシーン境界候補からシーン境界の推定について説明する。提案手法は、従来手法[5]と同様に、シーン境界を推定する際、Metropolis-Hasting Green アルゴリズムを用いている。このアルゴリズムを用いて、シーン境界を推定するとき、2つの値が用いられる。1つはシーン数やシーン境界の位置に関する事後確率であり、もう1つは、シーン境界に関する遷移確率である。

まず、提案手法と従来手法[5]で用いられる事後確率の算出について説明する。階層ベイズモデルにおける事後確率は、事前確率、条件付き確率および尤度に比例して算出される。このとき、提案手法は、従来手法と同様に事前確率および尤度を算出する。このため、条件付き確率の算出のみについて説明する。提案手法は、ショット境界の代わりに、シーン境界候補より最適と思われるシーン境界の推定を行う。このため、条件付き確率の算出について述べる必要がある。また、Metropolis-Hasting Green アルゴリズムを用いてシーン境界を推定する際、シーン境界に関する遷移確率の算出が必要である。ここで、従来手法[5]では、2つの遷移状態を用いている。1つはDiffusionであり、もう1つはJumpである。なお、DiffusionおよびJumpに関する詳細の説明は3.2.2節で行っている。従来手法において、Diffusionは隣接する2つのシーン境界をこの2つのシーン中に存在する他のショット境界に移動して、再度シーン境界に関する推定を行う処理である。一方、Jumpは、2つの

処理が存在する．split 処理は 1 つのシーンを 2 つのシーンに分割する処理であり，merge 処理は 2 つのシーンを 1 つのシーンに統合する処理である．提案手法において，Metropolis-Hasting Green アルゴリズムを用いて，繰り返しシーン境界の推定を行う際，Jump のみを用いる．そこで，本節において，新たに定義する必要がある条件付き確率について説明し，最後に，Diffusion の省略について述べる．

• 新たな条件付き確率の算出

提案手法で用いられる条件付き確率を， N_{cand} 個のシーン境界候補の中から $k-1$ 個のシーン境界を選択する組み合わせの問題として新たに定義する．したがって，提案手法で用いられる条件付き確率は，3.2.2 節の式 (3.5) に置き換えて，以下の式により算出される．

$$p(\theta_k | k) = \frac{1}{C_{k-1}^{N_{cand}}} = \frac{(k-1)!(N_{cand}-k+1)!}{N_{cand}!} \quad (4.2)$$

ただし， k はシーン数であり， θ_k はシーン境界位置を表し， N_{cand} はシーン境界候補の数である．

以上により，提案手法は，映像の構造を考慮したシーン境界候補を用いて，新たな条件付き確率の算出を可能とする．

• Diffusion 処理の省略

提案手法では，事前に映像の構造に注目してシーン境界候補を取得する際に，類似するショットを統合する．したがって，選択されるシーン境界候補の近辺において，シフト処理を行ってシーン境界の修正を行う処理が必要ないと考えられるため，本論文では，Diffusion 処理を用いずにアップデートを行う．

提案手法では，映像の構造に注目して取得されるシーン境界候補より，階層ベイズモデルに基づくシーン境界の推定を行うことで，映像の構造に注目した従来手法 [1],[51] において，シーンの最長時間幅を一意に決定することが困難である問題を解決し，高精度なシーン分割が期待できる．

4.5 階層ベイズモデルの導入による映像の構造に注目したシーン分割手法の評価実験

本章では，提案手法の性能評価を行うため，実際にテレビで放送された映像(320×240画素，24 bit，30 fps)を用いて実験を行う．なお，用いる映像は，ニュース映像5本(#1～#5)，ドキュメンタリー映像5本(#6～#10)，映画3本(#11～#13)およびアニメ映像2本(#14, #15)の4種類，合計15本の映像であり，各映像の時間長は約20分間となっている．事前のショット分割は，文献[7]を用いて行い，得られたショットに対して提案手法を適用する．なお，計算量削減のために，提案手法は従来手法[1]と同様の処理を行い，選出されたショットのキーフレームを用いる．また，提案手法の有効性を確認するため，従来手法[1]を比較手法1，従来手法[5]を比較手法2として採用し，提案手法と同様に上記の映像に適用した．なお，提案手法で用いられる特徴量は色ヒストグラム特徴量であり，2ショット間の類似度を式(4.1)により算出する．提案手法および比較手法を映像に適用して取得されるシーン分割結果から，以下で定義される Recall，Precision および F 値を求める．

$$\text{Recall} = \frac{\text{検出された正解シーン境界数}}{\text{正解シーン境界数}} \quad (4.3)$$

$$\text{Precision} = \frac{\text{検出された正解シーン境界数}}{\text{検出されたシーン境界数}} \quad (4.4)$$

$$\text{F 値} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (4.5)$$

Recallは，検出された正解シーン境界数が正解シーン境界数に占める割合として算出される．また，Precisionは，検出された正解シーン境界数が検出されたシーン境界数に占める割合である．これらの2つの値はトレードオフの関係を持ち，Recall および Precision を考慮して算出されるのがF値である．また，式(4.3)および(4.4)中の正解シーンについて，文献[54]では，同一環境で撮像される映像区間をシーンとして定義していることから，我々も同様の定義を用い，正解のシーン分割結果を手動で与えた．

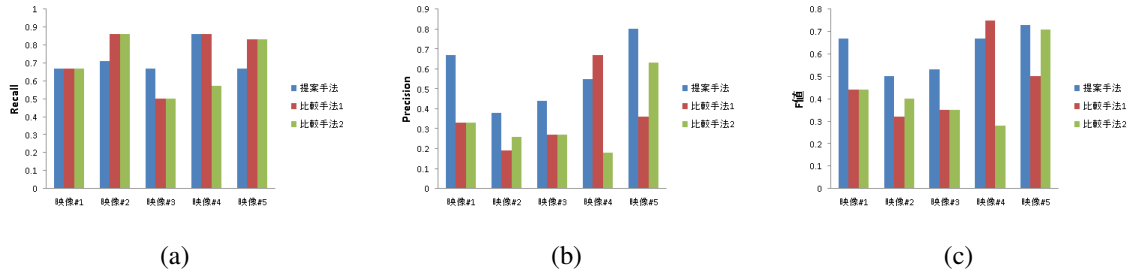


図 4.2: 各ニュース映像(映像#1～映像#5)に提案手法, 比較手法 1, 比較手法 2 をそれぞれ適用し, 各手法で F 値が最大となるようにパラメータの設定を行い, Recall, Precision および F 値を算出. (a)Recall (b)Precision (c)F 値

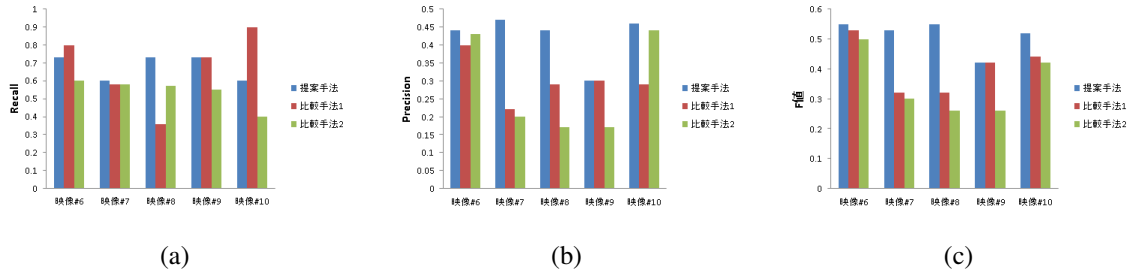


図 4.3: 各ニュース映像(映像#6～映像#10)に提案手法, 比較手法 1, 比較手法 2 をそれぞれ適用し, 各手法で F 値が最大となるようにパラメータの設定を行い, Recall, Precision および F 値を算出. (a)Recall (b)Precision (c)F 値

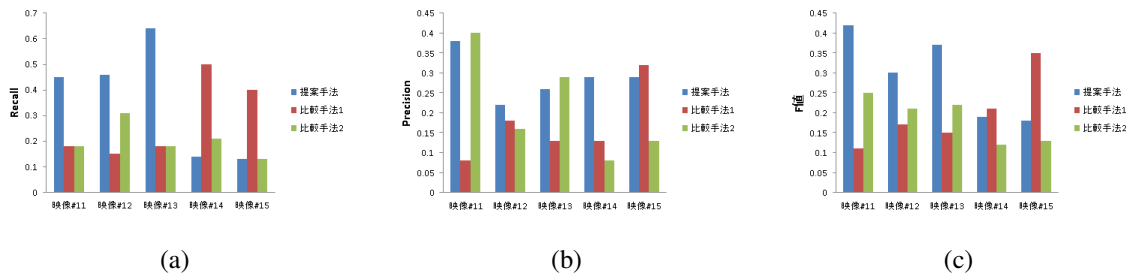


図 4.4: 各ニュース映像(映像#11～映像#15)に提案手法, 比較手法 1, 比較手法 2 をそれぞれ適用し, 各手法で F 値が最大となるようにパラメータの設定を行い, Recall, Precision および F 値を算出. (a)Recall (b)Precision (c)F 値

各実験映像に提案手法，比較手法1，および比較手法2を適用し，各手法では，F値が最大となるようにパラメータの設定を行った．なお，ニュース映像(#1～#5)の Recall，Precision および F 値をそれぞれ図 4.2(a),(b) および (c) に示す．これと同様にドキュメンタリー映像(#6～#10)の Recall，Precision および F 値をそれぞれ図 4.3(a),(b) および (c) に，映画(#11～#13)とアニメ映像(#14,#15)の Recall，Precision および F 値をそれぞれ図 4.4(a),(b),(c) に示す．実験で用いられる映像は，上記の4ジャンルであるため，各ジャンルの映像で得られる Recall，Precision および F 値について確認する．

まず，図 4.2 を用いてニュース映像(#1～#5)について確認する．各ニュース映像に対し Recall を求めた結果を図 4.2(a) に示す．映像 #3 において，提案手法は比較手法1および比較手法2より高精度に Recall の値を取得している．また，映像 #1 および映像 #4 において，提案手法と比較手法1，比較手法2と同様の値を取得していることが確認できる．また，図 4.2(b) において，提案手法および比較手法1，比較手法2で得られる Precision の比較を行う．映像 #1,#2,#3 および #5 において，提案手法はいずれの比較手法よりも高精度に Precision を取得している．これに対し，映像 #4 は，比較手法2より高精度にシーン分割結果が取得しているのに対し，比較手法1と同程度の Precision を取得していることが確認できる．ここで，Recall および Precision は互いにトレードオフの関係を持っているため，式 (4.5) で定義されている F 値を用いて，提案手法の確認を行う．提案手法は映像 #1～#3 および #5 において，F 値は比較手法1および比較手法2より高い値を持っていることが図 4.2(c) で確認できる．また，映像 #4 において，提案手法は従来手法2より高い値を持ち，比較手法1と同程度の値を持つことから，提案手法を5本のニュース映像に適用した際，比較手法1および比較手法2より高精度にシーン分割することを実現している．

次に，5本のドキュメンタリー映像(#6～#10)に対して，提案手法，比較手法1および比較手法2をそれぞれ適用し，各手法では，F値が最大となるようにパラメータの設定を行い，算出された Recall，Precision および F 値を図 4.3(a),(b),(c) にそれぞれ示す．提案手法を用いて5本のドキュメンタリー映像に対しシーン分割を行い，その分割結果に基づいて算出される Recall は比較手法1および比較手法2と同様の値であることが確認できる．一方，Precision を求める際，提案手法はいずれの比較手法よりも高精度にシーン分割を行っている．これにより，5本のドキュメンタリー映像に提案手法および比較手法を適用した際，提案手法は高精度にシーン分

割結果の取得を可能とする．

さらに，各映画（＃11～＃13）および各アニメ映像（＃14,＃15）に対して，提案手法の有効性を確認する．提案手法および比較手法1および比較手法2を映画に適用した際に，提案手法の Recall は比較手法より高く，Precision は比較手法と同程度である．このため，提案手法の F 値は比較手法より高い値を示し，高精度にシーン分割を実現していることが確認できる．一方，各アニメ映像に対し，提案手法を適用した際，比較手法と同様の Precision を持つ．しかしながら，Recall の値は比較手法より低い値である．これにより，提案手法は映像＃14 および映像＃15 において，比較手法2より高精度にシーン分割を可能としたが，比較手法1よりシーン分割精度が低下していることが確認できる．しかしながら，15本の映像に提案手法および比較手法1，比較手法2を適用した結果，提案手法が15本のうち11本の映像において高精度にシーン分割を実現している．ゆえに，階層ベイズモデルの導入による映像の構造に注目したシーン分割手法の有効性が確認できた．

一般に，映像を構成する画像信号と音響信号の両方の特徴量を用いることで高精度化が可能であるため，音響特徴量の導入による，映像のシーン分割の高精度化を行う必要があると考えられる．

4.6 まとめ

本章では，階層ベイズモデルの導入による映像の構造に注目したシーン分割手法を提案した．提案手法では映像の構造に注目したシーン境界候補を取得し，それらの中から階層ベイズモデルに基づいて最適なシーン境界を求めることで，従来手法の問題を改善した．また，実際にテレビで放送される映像に提案手法を適用し，その有効性を確認した．

第5章

画像特徴量と音響特徴量を用いた シーン分割の高精度化手法

5.1 はじめに

本章では，画像特徴量と音響特徴量を用いたシーン分割の高精度化手法を提案する．提案手法では，映像から得られる画像，音響に関する特徴量の類似度に関して，その分布を考慮しながら比較可能な数値への変換を行うことで，異なる信号の特徴量の統合を可能とする．また，得られる異なる特徴量により算出されるショット間の類似度を用いて，映像の構造に注目したシーン境界候補の取得を行うことで，画像特徴量のみ，または，音響特徴量のみ注目した際の結果と比較して高精度な結果の取得を可能とする．さらに，画像特徴量と音響特徴量を統合して，階層ベイズモデルにおける事後確率の算出を行うことで，高精度なシーン分割を可能とする．したがって，提案手法では，複数の特徴量を用いて，シーン境界候補の取得とシーン境界に関する事後確率の算出を行うことで，シーン分割精度の向上が実現される．

5.2 画像・音響特徴量を用いたショット間の類似度算出

本節では，ショット間の類似度算出について説明する．提案手法ではショット s_i および s_j 間において，従来手法 [1] と同様に色ヒストグラム特徴量によるショット間の類似度 $Sim^C(s_i, s_j)$ ，従来手法 [51] と同様に SIFT 特徴量および音響特徴量によるショット間の類似度 $Sim^S(s_i, s_j)$ ， $Sim^A(s_i, s_j)$ をそれぞれ算出する．なお， $Sim^C(s_i, s_j)$

は、4.2節で説明した $Sim(s_i, s_j)$ と同一のものであり、提案手法では3種類の特徴量を用いているため、色ヒストグラム特徴量に基づくショット間の類似度を $Sim^C(s_i, s_j)$ と改めて表記する．ここで、ショット間の類似度 $Sim^C(s_i, s_j)$ の再掲を行い、ショット間の類似度 $Sim^S(s_i, s_j)$ および $Sim^A(s_i, s_j)$ の算出について述べる．

・色ヒストグラム特徴量によるショット間の類似度算出

異なるショット s_i および s_j の類似度を算出する際に、各ショットのキーフレームを用いてショット間の類似度を算出する．なお、色ヒストグラム特徴量によるショット間の類似度は以下のように算出される．映像中の異なるショット s_i および $s_j (i, j = 1, 2, \dots, N_s; N_s \text{ はショットの総数})$ 間の類似度 $Sim^C(s_i, s_j)$ を次式のように定義する．

$$Sim^C(s_i, s_j) = \begin{cases} \max_{f_l \in s_i, f_m \in s_j} Sim_f(f_l, f_m) & \text{if } D_T(s_i, s_j) \leq Th_T \\ 0 & \text{otherwise} \end{cases} \quad (5.1)$$

ただし、 f_l, f_m はそれぞれショット s_i, s_j に含まれるキーフレームであり、 $Sim_f(f_l, f_m)$ はフレーム f_l と f_m の色ヒストグラム間の類似度を表している． $D_T(s_i, s_j)$ はショット s_i, s_j およびその間に含まれる全てのショットに存在するフレームの総数であり、 Th_T はあらかじめ設定されたシーンの最長時間幅に関する閾値である．

・SIFT 特徴量によるショット間の類似度算出

文献 [51] で定義されている SIFT 特徴量を用いたショット間の類似度について説明する．文献 [51] では、ショット s_i および $s_j (i, j = 1, 2, \dots, N_s)$ にそれぞれ含まれるキーフレーム f_l および f_m に対して、SIFT 特徴量を用いて対応付けを行い、ショット s_i および s_j 間の類似度 $Sim^S(s_i, s_j)$ を次式のように定義する．

$$Sim^S(s_i, s_j) = \begin{cases} \max_{f_l \in s_i, f_m \in s_j} \frac{N_{f_l, f_m}}{\min(N_{f_l}, N_{f_m})} & \text{if } D_T(s_i, s_j) \leq Th_T \\ 0 & \text{otherwise} \end{cases} \quad (5.2)$$

ここで、 N_{f_l} はフレーム f_l に SIFT 特徴量を適用することによって得られるキーポイント数であり、 N_{f_l, f_m} はフレーム f_l および f_m 間において対応付けされたキーポイント数である．また Th_T は予め設定されるシーンの最長時間幅に関する閾値である．

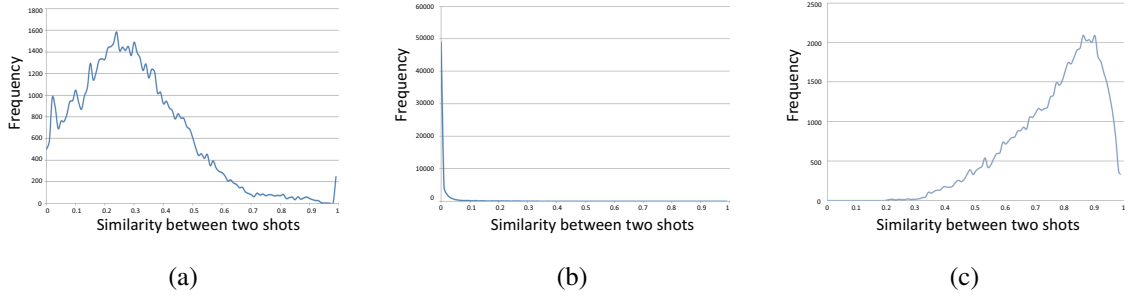


図 5.1: (a) 色ヒストグラム特徴量に基づいて算出されるショット間の類似度分布 (b) SIFT 特徴量に基づいて算出されるショット間の類似度分布 (c) 音響特徴量に基づいて算出されるショット間の類似度分布

・音響特徴量によるショット間の類似度算出

音響特徴量によるショット間の類似度算出について説明する．従来手法では，対象映像の音響信号に対して，文献 [46] で提案されているショット分類法を適用し，ショットに含まれる音響信号が無音 (silence:Si)，音声 (speech:Sp)，音楽 (music:Mu) および雑音 (noise:No) に属する確率 (以降，帰属確率) をそれぞれ算出する．このとき，ショット s_i における帰属確率を $p_i^e (e = 1, 2, 3, 4; \text{Si, Sp, Mu, No})$ とし，ショット s_i および s_j 間の類似度 $\text{Sim}^A(s_i, s_j)$ は文献 [51] で定義された式を用いて算出を行う．

$$\text{Sim}^A(s_i, s_j) = \begin{cases} \sum_{e=1}^4 \min(p_i^e, p_j^e) & \text{if } D_T(s_i, s_j) \leq Th_T \\ 0 & \text{otherwise} \end{cases} \quad (5.3)$$

ここで， $D_T(s_i, s_j)$ はショット s_i および s_j 間に存在するフレーム数で与えられ， Th_T は予め設定したシーンの最長時間幅に関する閾値である．

以上により，色ヒストグラム特徴量，SIFT 特徴量及び音響特徴量によるショット間の類似度算出を可能とする．

5.3 累積度数分布を用いた新たなショット間の類似度算出

本節では，累積度数分布を用いた新たなショット間の類似度算出について説明する．5.2 節で取得された 3 種類のショット間の類似度を比較可能な数値へ変換し，ショット間の類似度の統合処理を行う．図 5.1 に示すように，同一映像に対して，色ヒストグラム特徴量，SIFT 特徴量および音響特徴量のそれぞれの特徴量を用いた

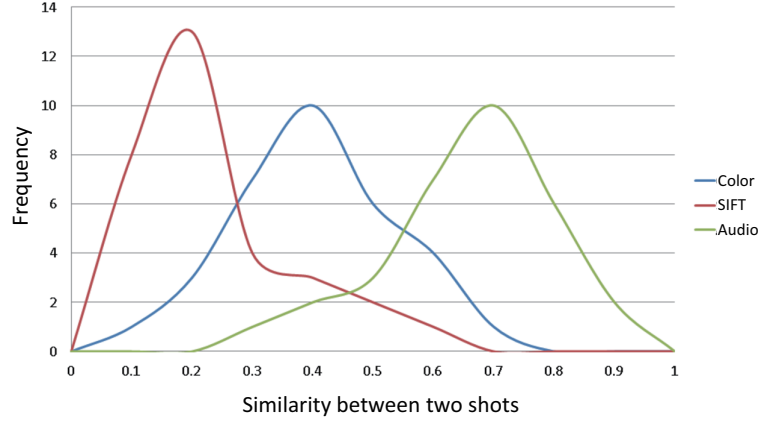


図 5.2: ショット間の類似度に関する分布

場合，算出されるショット間の類似度 $Sim^C(\cdot, \cdot)$, $Sim^S(\cdot, \cdot)$ および $Sim^A(\cdot, \cdot)$ に関するヒストグラムの分布は，互いに異なることが確認できる．このため，単純な重み付けによる統合方法等では，それぞれの特徴量に関するショット間の類似度の分布の形状を合わせることが困難であり，また映像毎に重みを決定することも困難である．そこで，提案手法では，対象の映像から各特徴量に基づき取得されるショット間の類似度に対してヒストグラムを生成し，その累積度数分布を算出することで，ショット間の新たな類似度の算出を行う．具体的に，各特徴量を用いて算出されるショット間の新たな類似度を $FSim^l(s_i, s_j) (l \in \{C, S, A\})$ と定義すると，その値は次式のように与えられる．

$$FSim^l(s_i, s_j) = \sum_{b=0}^{\lfloor B \cdot Sim^l(s_i, s_j) \rfloor} h^l(b) \quad (5.4)$$

このとき， B はショット s_i と s_j 間の類似度に関してヒストグラムを作成する際のビン数を表し， $h^l(b)$ は類似度 $Sim^l(s_i, s_j)$ が

$$\frac{b}{B} \leq Sim^l(s_i, s_j) < \frac{b+1}{B} \quad (5.5)$$

を満たす際の頻度を表す．ただし， $b = 0, 1, 2, \dots, B-1$ であり， $h^l(b)$ は $\sum_{b=0}^{B-1} h^l(b) = 1$ と正規化されているものとする．本研究では， B を単純に 100 と設定する．

提案手法では，式 (5.4) を用いて新たなショット間の類似度の算出を行っている．この処理は，特徴量毎に求められる全てのショット間の類似度について低い方からの順位を求めていることと等価である．また，式 (5.5) では，ショット間の類似度 $Sim^l(s_i, s_j) (l \in \{C, S, A\})$ が対応するビンを求めている．式 (5.4) および (5.5) の導入により，累積度数分布に基づくショット間の類似度の順位に注目した新たなショット

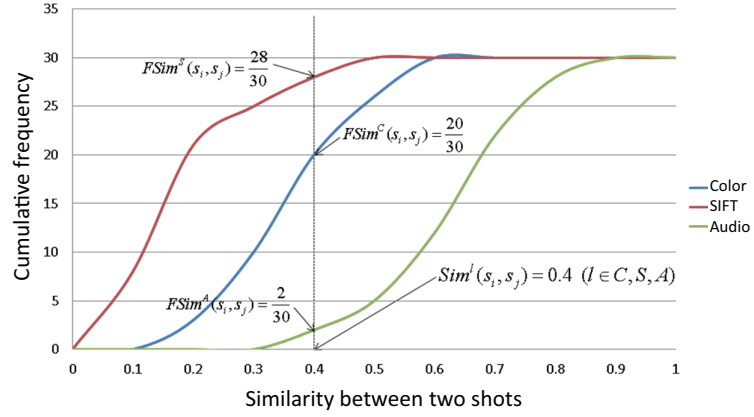


図 5.3: ショット間の類似度に関する累積度数分布

間の類似度 $FSim^l(s_i, s_j) (l \in \{C, S, A\})$ の算出が可能になる．ここで，ショット間の類似度に関する分布の例が図 5.2 に示されるように与えられ，各々の特徴量の類似度が全て $Sim^l(s_i, s_j) = 0.4 (l \in \{C, S, A\})$ の場合，図 5.3 に示すように，最終的な類似度 $FSim^l(\cdot, \cdot) (l \in \{C, S, A\})$ は，それぞれ $\frac{20}{30}$ ， $\frac{28}{30}$ および $\frac{2}{30}$ となる．したがって，上記に示す順位を用いる方法は，全体の分布 (ある特徴量を用いて得られるショット間の類似度に関する分布) 上でその類似度が相対的にどの程度高いかを判断することが可能である．つまり，異なる分布同士であっても，互いに比較可能な類似度を求めることが可能である．

式 (5.4) で得られる結果より，提案手法では次式により最終的なショット間の類似度 $FSim^O(s_i, s_j)$ を算出する．

$$FSim^O(s_i, s_j) = \sum_{l \in \{C, S, A\}} w_l \cdot FSim^l(s_i, s_j) \quad (5.6)$$

ただし， $w_l (l \in \{C, S, A\})$ は重み係数であり， $\sum_{l \in \{C, S, A\}} w_l = 1$ を満たす．先に述べた通り，画像特徴量と音響特徴量の利用を行うことで高精度にシーン分割が可能になることが報告されていることから [55]–[56]，提案手法も，画像と音響の特徴量を利用したシーン分割を行う．

5.4 画像特徴量と音響特徴量を用いた映像のシーン境界候補の取得

本節では、画像特徴量と音響特徴量を用いた映像のシーン境界候補の取得について述べる。提案手法では、累積度数分布に基づき、新たに算出されたショット間の類似度 $FSim^l(s_i, s_j) (l \in \{C, S, A\})$ を用いて、シーン境界候補を取得する。このとき、色ヒストグラム特徴量、SIFT 特徴量、および音響特徴量のそれぞれの特徴量を用いて、従来手法 [1] と同様の処理を行い、映像の分割結果を取得する。提案手法における映像の構造を考慮したシーン境界候補の取得は、図 5.4 を用いて説明する。3.2.1 節で説明したように、映像の構造に基づくシーン分割手法では、シーンの最長時間幅等の閾値を最適な値に設定することが困難である。このため、提案手法では、ショットのクラスタリングを行う際の閾値 Th_C およびシーンの最長時間幅 Th_T の閾値を様々な値に変化させる。ただし、提案手法で用いられる色ヒストグラム特徴量、SIFT 特徴量および音響特徴量に対して、同一の Th_C および Th_T を用いる。図 5.4 では、まず複数の Th_C および Th_T の組み合わせを設定し、 $FSim^C(s_i, s_j)$ を用いて従来手法 [1] と同様に、映像の分割を N^V 回を行う。得られた N^V 個の映像分割結果を統合し、これを色ヒストグラムに基づいて取得される境界候補とする。次に、色ヒストグラムと同様に、SIFT 特徴量および音響特徴量のそれぞれに基づき境界候補を得る。このように、色ヒストグラム特徴量、SIFT 特徴量および音響特徴量のそれぞれの特徴量に基づいて境界候補が得られる。取得される全ての境界候補をシーン境界候補として用いることにより、網羅的なシーン境界候補の選択を可能とする。この理由として、色ヒストグラム特徴量、SIFT 特徴量および音響特徴量はそれぞれシーン分割を行う際に、注目点が異なるためである。色ヒストグラム特徴量を用いてシーン境界を取得する際に、ショット境界において、前後のショットの色特徴量の変化が大きい場合、そのショット境界がシーン境界候補として選択されるが、色ヒストグラムが類似したショット間において、シーン境界が存在した場合、シーン境界候補の取得が困難となる。一方、SIFT 特徴量では、撮像される物体の変化に注目してシーン境界候補の取得を行っている。音響特徴量に基づきシーン境界候補を取得する際は、映像のオーディオ信号に注目する。このとき画像特徴量が類似しているにもかかわらず、オーディオ信号に変化が生じ

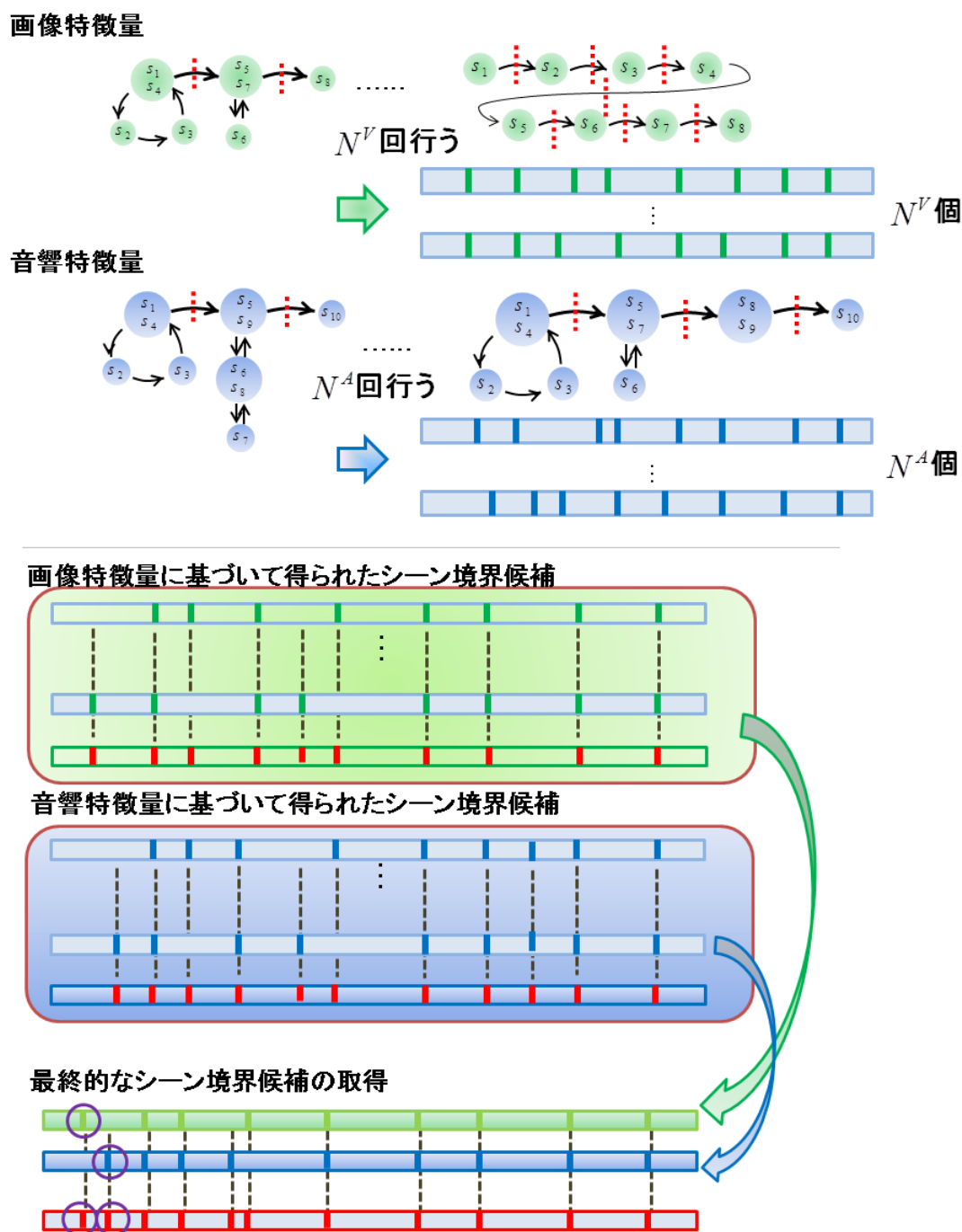


図 5.4: 画像特徴量および音響特徴量を用いて、映像のシーン境界候補の取得

た場合、シーン境界候補の取得を行う。これにより、画像の色ヒストグラム、撮像される物体、およびオーディオ信号の変化に注目することで、まんべんなくシーン境界候補の取得が可能になる。

以上により、映像の構造を取得する際に、ショットのクラスタリングに関する閾

値 Th_C および 最長時間幅である Th_T の組み合わせを用いて、映像の構造に注目したシーン境界候補が取得される。このとき、色ヒストグラム特徴量、SIFT 特徴量および音響特徴量のそれぞれの注目点が異なることから、取得されるシーン境界候補も変化する。このため、上記の異なる特徴量に基づき得られる全ての境界候補をシーン境界候補として用いることで、画像特徴量と音響特徴量に注目し、映像の構造を考慮した網羅的なシーン境界候補の取得が可能になる。

5.5 階層ベイズモデルに基づくシーン境界の推定

本節では、階層ベイズモデルに基づきシーン境界候補の中からシーン境界の推定を行う。従来手法 [5] では、色ヒストグラム特徴量が与えられた際のシーン境界に関する事後確率を算出し、シーン境界の推定を行った。一方、提案手法は、画像特徴量と音響特徴量を用いてシーン境界の推定を行うため、従来手法 [5] で用いられる事後確率に関する式の変形が必要である。提案手法は、4 章で提案した階層ベイズモデルの導入による映像の構造に注目したシーン分割手法と同様に、特徴量が与えられた際のシーン境界に関する事前確率は一様分布に仮定し、ポアソン分布を用いて算出する。一方、事後確率を算出するために必要である条件付き確率および尤度は、用いられるシーン境界候補および複数のショット間の類似度の利用に変化があるため、新たに算出する必要がある。そこで、以下では、提案手法で用いられる新たな条件付き確率および尤度の算出について説明する。

• 新たな条件付き確率の算出

提案手法では、条件付き確率の算出を、得られた N_{cand}^{all} 個のシーン境界候補の中から、 $k-1$ 個のシーン境界を選択する組み合わせの問題に変形する。条件付き確率は以下の式に基づき算出される。

$$p(\theta_k | k) = \frac{1}{C_{k-1}^{N_{cand}^{all}}} = \frac{(k-1)!(N_{cand}^{all} - k + 1)!}{N_{cand}^{all}!} \quad (5.7)$$

ただし、 k はシーン数であり、 θ_k はシーン境界位置を表し、 N_{cand}^{all} は画像特徴量と音響特徴量を用いた際に取得されるシーン境界候補の数である。

• 画像特徴量と音響特徴量を用いた尤度の算出

提案手法で用いられる尤度は、シーン S_n が与えられたときの特徴量 f_n に関

表 5.1: 提案手法および比較手法が用いられる特徴量の詳細.

	色ヒストグラム特徴量	SIFT 特徴量	音響特徴量
提案手法	○	○	○
比較手法 1	○		
比較手法 2		○	○
比較手法 3	○		

する条件付き確率である．従来手法 [5] では，同一シーン中に存在するショットの類似性に注目して尤度が定義されているため，提案手法では同様に特徴量 f_n に関する条件付き確率を求める．なお，提案手法では，5.3 節で得られるショット間の類似度を用いてシーン境界を推定するために，3.2.2 節の式 (3.6) と同様に尤度を算出する．ただし，シーン S_n に N_{S_n} 個のショット $\{s_n^1, \dots, s_n^{N_{S_n}}\}$ が存在する場合，式 (3.6) で用いられる $\mathbb{L}(f_n | S_n)$ は，式 (3.7) に換えて，次式により算出する．

$$\mathbb{L}(f_n | S_n) = \frac{1}{N_{S_n}^2} \sum_{i=1}^{N_{S_n}} \sum_{j=1}^{N_{S_n}} FSim^O(s_n^i, s_n^j) \quad (5.8)$$

ただし， f_n はシーン S_n で算出される映像の特徴量であり， $FSim^O(s_n^i, s_n^j)$ は，式 (5.6) で求められるショット間の類似度である．

• Diffusion 処理の簡略化

提案手法では，映像の構造を考慮したシーン境界候補を用いているため，シーン境界候補の近辺において，シフト処理を行って，シーン境界の修正を行う処理が必要ないと考えられる．このため，本提案手法においても，Diffusion 処理を用いずにアップデートを行う．

以上により，階層ベイズモデルの導入により，画像特徴量と音響特徴量に注目して得られたシーン境界候補から，シーン境界の推定を可能とする．従来手法 [1] に存在する最長時間幅などの閾値の決定が困難である問題に対し，複数の値を与えることで問題の解決を行う．さらに，階層ベイズモデルの導入により，シーン境界候補の中から，シーン境界の推定を可能とし，高精度なシーン分割が期待できる．

5.6 画像特徴量と音響特徴量を用いたシーン分割の高精度化手法の評価実験

本章では、提案手法の性能評価を行うため、実際にテレビで放送された映像(320×240画素, 24 bit, 30 fps)を用いて実験を行う。なお、用いる映像は、ニュース映像5本(#1～#5), ドキュメンタリー映像5本(#6～#10), 映画3本(#11～#13)およびアニメ映像2本(#14, #15)の4種類, 合計15本の映像であり, 各映像の時間長は約20分間となっている。事前のショット分割は, 文献[7]を用いて行い, 得られたショットに対して提案手法を適用する。提案手法の有効性を確認するため, 従来手法[1],[51]および[5]を比較手法として採用し, 提案手法と同様に上記の映像に適用した。なお, 従来手法[1]を従来手法1とする。次に, 従来手法[51]は従来手法1の高精度化手法であるため, これを比較手法2とする。さらに, 従来手法[5]を比較手法3とすることで, 比較実験を行う。なお, 提案手法で用いられる特徴量は色ヒストグラム特徴量, SIFT特徴量および音響特徴量であり, 各々の比較手法で用いられる特徴量は表5.1に示す。ただし, 特徴量の算出方法は, 手法によらず同一である。

そこで, 本節では, 評価実験を行い, 提案手法の有効性を確認する。なお, 提案手法を用いてシーン分割結果に関する実験を行う前に, まず, 従来手法と同様の特徴量を用いた際に, 階層ベイズモデルの導入による映像の構造を考慮したシーン分割手法の有効性を5.6.1節で確認する。次に, これらを導入した提案手法の有効性を5.6.2節で確認する。

5.6.1 提案手法が比較手法と同様の特徴量を用いたシーン分割の高精度化の有効性確認

本節では, 階層ベイズモデルの導入による映像の構造に注目したシーン分割手法の有効性を確認する。提案手法で, 従来手法1, 従来手法2, および従来手法3と同様の特徴量を用いて有効性を確認する。このとき, 実験映像#1～#15の映像を用いて, 式(4.6)-式(4.8)で定義されたRecall, PrecisionおよびF値を算出する。なお, Recall, PrecisionおよびF値の算出方法をここに再掲する。

$$\text{Recall} = \frac{\text{検出された正解シーン境界数}}{\text{正解シーン境界数}} \quad (5.9)$$

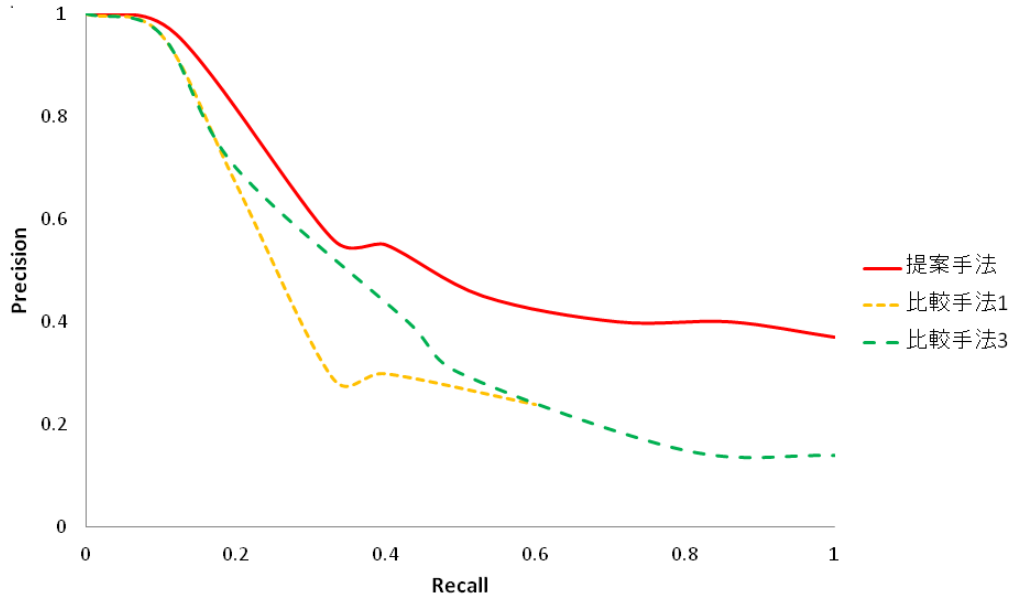


図 5.5: 提案手法および比較手法 1,3 に色ヒストグラム特徴量を適用した際に得られる Recall-Precision カーブ

$$\text{Precision} = \frac{\text{検出された正解シーン境界数}}{\text{検出されたシーン境界数}} \quad (5.10)$$

$$\text{F 値} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (5.11)$$

なお、式 (5.9) および (5.10) 中の正解シーンについて、文献 [54] では、同一環境で撮像される映像区間をシーンとして定義していることから、我々も同様の定義を用い、正解のシーン分割結果を手動で与えた。

実験対象映像 (#1 ~ #15) 全てを用いて、提案手法および比較手法 1,2,3 に、提案手法で用いられるショット間の類似度を用いた際の実験も行い、Recall-Precision カーブで示す。なお、本実験では、各手法で下記の値を変化させることで Recall-Precision カーブを求めた。

- ・提案手法：式 (3.4) に示すシーン数に関する事前確率中のパラメータ λ
- ・比較手法 1：クラスタリングの際に用いられる閾値 Th_c
- ・比較手法 2：シーン境界として採用する割合を表すパラメータ $r(\%)$ ¹
- ・従来手法 3：式 (3.4) に示すシーン数に関する事前確率中のパラメータ λ

提案手法は、比較手法 1 と比較手法 3 と同様に、色ヒストグラム特徴量を用いて

¹比較手法 2 では、各ショット境界においてシーン境界になる確率を算出し、確率が大きいショット境界からシーン境界として選択する。

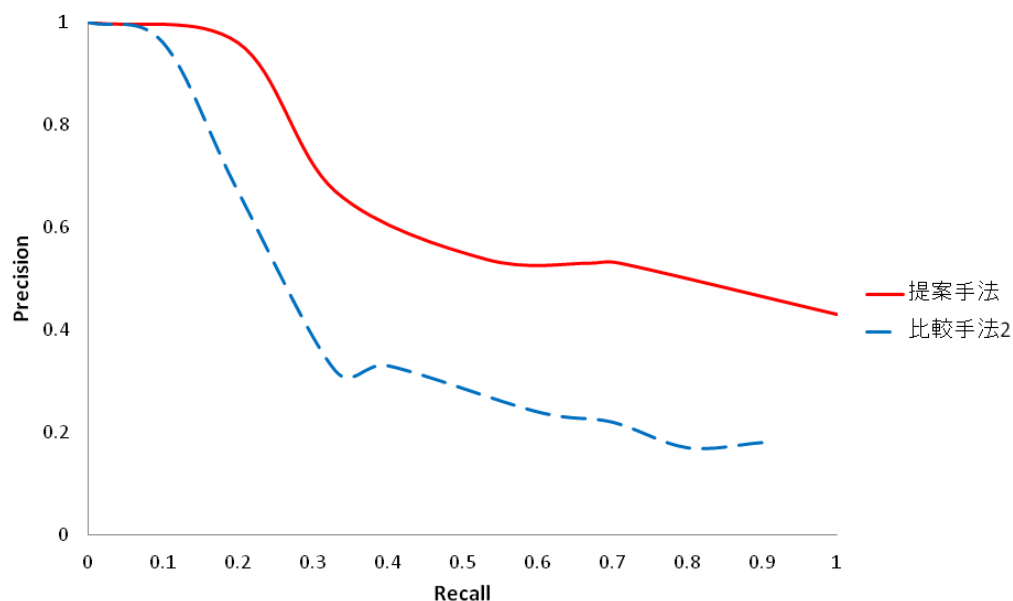


図 5.6: 提案手法および従来手法 2 に SIFT 特徴量および音響特徴量を適用した際に得られる Recall-Precision カーブ

シーン分割を行う。得られる Recall-Precision カーブを図 5.5 に示す。比較手法 2 では、SIFT 特徴量および音響特徴量を用いているため、提案手法は、従来手法と同様の特徴量を用いてシーン分割を行う。得られた Recall-Precision カーブを図 5.6 に示す。図 5.5、図 5.6 より、提案手法の結果は、比較手法 1~3 の結果を上回っており、提案手法が高精度なシーン分割を実現していることが確認できる。以下では、提案手法と比較手法との性能評価に関する考察を示す。

(i) 提案手法と比較手法 1 および比較手法 2 の比較

図 5.5、図 5.6 より、提案手法が高精度にシーン分割が可能であり、映像の構造に注目して取得されたシーン境界候補に対して、階層ベイズモデルに基づく最適なシーン境界の推定を行うことの有効性が確認できる。これにより、映像の構造に基づくシーン分割手法の共通の問題であるシーンの最長時間幅の設定が階層ベイズモデルの導入により不要となり、従来の映像の構造に基づくシーン分割手法より高精度なシーン分割が可能となった。

(ii) 提案手法と比較手法 3 の比較

図 5.5 より、提案手法が高精度にシーン分割が可能であり、提案手法においてあらかじめ映像の構造に注目したシーン境界候補を与えることが有効であることが確

表 5.2: Recall(R), Precision(P) および F 値 (F) の比較 (なお, 以下の Recall, Precision および F 値を取得するために, 各手法では, F 値が最大となるように, パラメータの設定を行った.)

評価	提案手法			比較手法 1(文献 [1])			比較手法 2(文献 [51])			比較手法 3(文献 [5])		
	R	P	F	R	P	F	R	P	F	R	P	F
ニュース映像 (#1)	0.67	0.67	0.67	0.67	0.33	0.44	1.00	0.23	0.38	0.67	0.33	0.44
ニュース映像 (#2)	0.57	0.57	0.57	0.86	0.19	0.32	0.71	0.31	0.44	0.86	0.26	0.40
ニュース映像 (#3)	0.50	1.0	0.67	0.5	0.27	0.35	0.5	0.50	0.50	0.5	0.27	0.35
ニュース映像 (#4)	0.71	1.0	0.83	0.86	0.67	0.75	0.86	0.75	0.80	0.57	0.18	0.28
ニュース映像 (#5)	0.83	1.0	0.91	0.83	0.36	0.50	0.83	0.71	0.77	0.83	0.63	0.71
ドキュメンタリー映像 (#6)	0.60	0.60	0.60	0.80	0.40	0.53	0.67	0.33	0.44	0.60	0.43	0.50
ドキュメンタリー映像 (#7)	0.67	0.67	0.67	0.58	0.22	0.32	0.33	0.67	0.44	0.58	0.20	0.30
ドキュメンタリー映像 (#8)	0.50	0.41	0.45	0.36	0.29	0.32	0.57	0.31	0.40	0.57	0.17	0.26
ドキュメンタリー映像 (#9)	0.27	0.60	0.37	0.73	0.30	0.42	0.55	0.23	0.32	0.55	0.17	0.26
ドキュメンタリー映像 (#10)	0.40	0.80	0.53	0.90	0.29	0.44	0.50	0.29	0.37	0.40	0.44	0.42
映画 (#11)	0.45	0.38	0.42	0.18	0.08	0.11	0.27	0.43	0.33	0.18	0.40	0.25
映画 (#12)	0.46	0.46	0.46	0.15	0.18	0.17	0.39	0.31	0.35	0.31	0.16	0.21
映画 (#13)	0.36	0.57	0.44	0.18	0.13	0.15	0.27	0.23	0.25	0.18	0.29	0.22
アニメ (#14)	0.14	0.67	0.24	0.50	0.13	0.21	0.43	0.13	0.20	0.21	0.08	0.12
アニメ (#15)	0.27	0.31	0.29	0.40	0.32	0.35	0.27	0.33	0.30	0.13	0.13	0.13

認できる.

以上により, 階層ベイズモデルの導入による映像の構造に注目したシーン分割は高精度に実現されることが確認できた.

5.6.2 画像・音響特徴量を用いた階層ベイズモデルの導入による映像の構造に注目したシーン分割手法の有効性確認

本節では, 提案手法の有効性を確認する. 提案手法では, まず画像特徴量と音響特徴量の各々を用いてショット間の類似度を算出し, それぞれのショット間の類似度に基づいて映像の構造に注目した映像分割を行い, 得られた全ての境界をシーン境界候補とする. 次に, 階層ベイズモデルで定義されたシーン数およびシーン境界に関する事後確率を, マルコフ連鎖モンテカルロ法を用いて最適化を行い, 最終的にシーン境界の推定を行う.

提案手法および比較手法を映像に適用して取得されるシーン分割結果から, 式 (5.9)-式 (5.11) で定義される Recall, Precision および F 値を求め, その値を表 5.2 に示す. また, 提案手法および比較手法 3 では事前確率を算出する際に用いられるシーン数に関するパラメータ λ を, 比較手法 1 はその文献中に記載された最適なシーン

表 5.3: 各実験映像に関する詳細.

	ショット 境界数	シーン境界 候補数	正解 シーン数	提案手法で取得 されるシーン数	比較手法1で取得 されるシーン数	比較手法2で取得 されるシーン数	比較手法3で取得 されるシーン数
映像 (#1)	728	30	3	3	6	13	6
映像 (#2)	249	30	7	7	31	16	23
映像 (#3)	259	25	6	3	11	6	11
映像 (#4)	134	12	7	5	9	8	22
映像 (#5)	117	15	6	5	14	7	8
映像 (#6)	103	26	15	15	30	30	21
映像 (#7)	70	24	12	12	32	6	35
映像 (#8)	126	31	14	17	17	26	47
映像 (#9)	113	28	11	5	27	26	35
映像 (#10)	101	30	10	5	31	17	9
映像 (#11)	259	24	11	13	25	7	5
映像 (#12)	252	40	13	13	11	16	25
映像 (#13)	322	51	11	7	15	13	7
映像 (#14)	730	20	14	3	53	46	37
映像 (#15)	803	26	15	13	19	12	15

表 5.4: パラメータ w_C, w_S, w_A を変化させた際の F 値 (ただし, $w_C + w_S + w_A = 1$ を満たす.)

$w_S \backslash w_C$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
0.1	0.50	0.52	0.56	0.50	0.54	0.50	0.56	0.50
0.2	0.48	0.58	0.58	0.61	0.51	0.52	0.52	-
0.3	0.52	0.49	0.58	0.56	0.52	0.54	-	-
0.4	0.52	0.56	0.61	0.50	0.59	-	-	-
0.5	0.54	0.56	0.63	0.56	-	-	-	-
0.6	0.54	0.54	0.50	-	-	-	-	-
0.7	0.49	0.58	-	-	-	-	-	-
0.8	0.50	-	-	-	-	-	-	-

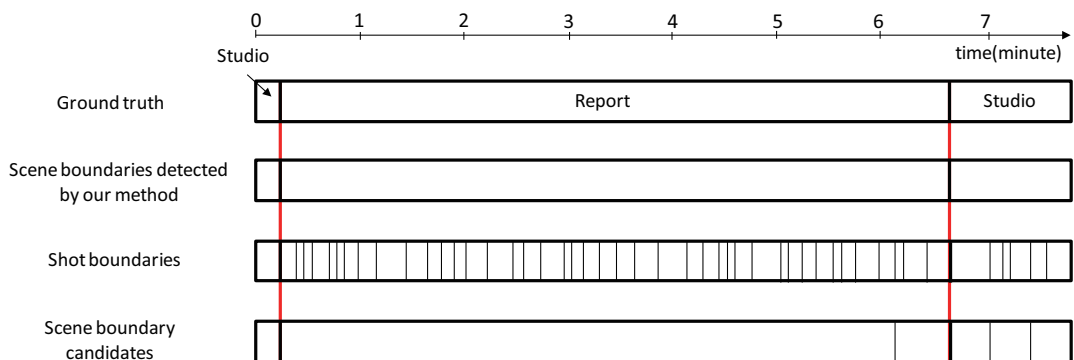


図 5.7: 映像 #5 のショット境界, シーン境界候補および正解シーン境界の関係図 (提案手法で得られたシーン分割結果も合わせて表示)

の最長時間幅 Th_T およびクラスタリングの際の閾値 Th_C の組み合わせを, また比較手法 2 は, シーン境界として採用する割合を表すパラメータを, それぞれ変化

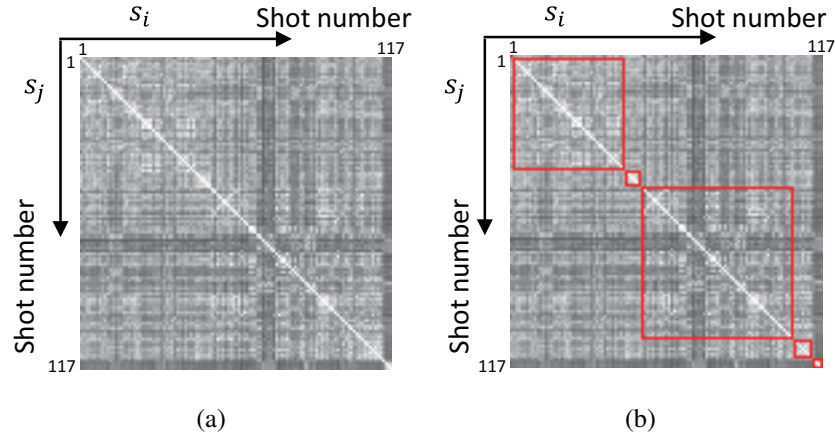


図 5.8: (a) ショット間の類似度行列に基づいたショット間の類似関係図 (b) 正解シーン境界付きショット間の類似度行列に基づいたショット間の類似関係図

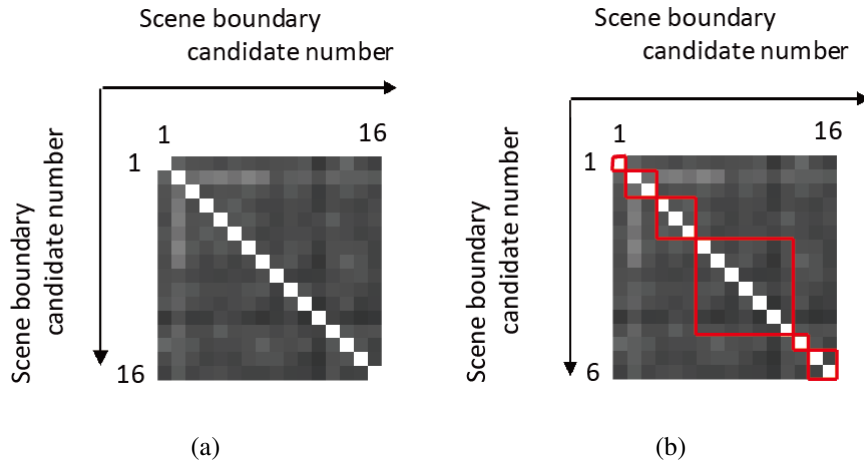


図 5.9: (a) シーン境界候補間の類似度行列の図 (b) 正解シーン境界付きシーン境界候補間の類似度行列の図

させて、F 値が最大となる時の値を表 5.2 に示している．各映像について，最大の Recall, Precision および F 値となる手法の値を太字で示した．表 5.2 より，15 本の映像の内 13 本の映像について，提案手法の F 値が比較手法よりも高い値を示しており，シーン分割が高精度に行われていることが確認できる．

提案手法において，様々な値を与える Th_C および Th_T は以下のように設定した．まず，クラスタリングの際の閾値 Th_C は従来手法 [1] と同様の値の範囲を用い，また，シーンの最長時間幅 Th_T は，500~3500 の範囲内において，500 刻みの値の全てに対して提案手法を適用し，事後的に性能が最も高くなるものに設定した．具体的に， $Th_C = \{0.3, 0.4, 0.5\}$ と $Th_T = \{2500, 3000\}$ の組み合わせを用いて，全ての実験対象

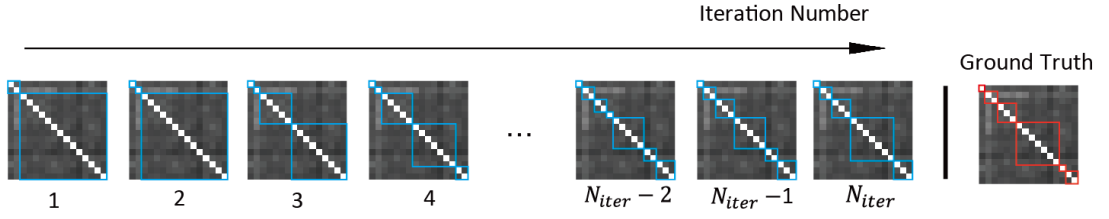


図 5.10: MCMC 法を用いた際のステップ毎の推定シーン境界の遷移図

映像に対して，シーン境界候補の取得を行った．

各実験映像に対して，ショット境界数，正解シーン数，および提案手法を適用して得られるシーン境界候補数，シーン数を表 5.3 に示す．また，各比較手法で得られたシーン数も合わせて表 5.3 に示す．表 5.3 より，シーン境界候補はショット境界に包含されており，35% 以下となっていることが確認できる．また，提案手法では，式 (5.6) におけるパラメータを実験的に $w_C = 0.3$ ， $w_S = 0.5$ ， $w_A = 0.2$ と設定した．ここで，実験対象映像 (#1 ～ #5) に対し提案手法を適用した際に，パラメータ w_C, w_S, w_A を変化させて得られた F 値を表 5.4 に示す．表 5.4 より，提案手法ではパラメータがロバストであり，また， $w_C = 0.3$ ， $w_S = 0.5$ ， $w_A = 0.2$ のとき，最も高精度となっていることが確認できるため，全実験対象映像に対して上記の値を用いて実験を行った．

次に，実験で用いられる 1 本の映像 (映像 #5)² に関するショット境界，シーン境界候補およびシーン境界の関係を例として，図 5.7 に示す³．さらに，従来手法 [5] と同様にショット間の類似度行列を図 5.8 に示す．また，提案手法ではショット境界ではなく，シーン境界候補を用いるため，シーン境界候補間の類似度行列を図 5.9 に示す．なお，同一シーン中に存在するショットを赤い線で囲んで示す．図 5.8 および図 5.9 の類似度行列の行および列はそれぞれ，ショット番号またはシーン境界候補番号に対応し，各ブロックはショット間の類似度またはシーン境界候補間の類似度を表す．さらにマルコフ連鎖モンテカルロ法を用いて最終的なシーン境界の推定を行うため，マルコフ連鎖モンテカルロ法を用いた際のステップ毎の推定シーン境界の遷移図を図 5.10 に示す．

²実験映像 #1 では，3 つのシーンのみ存在するため，例を示す際に，6 つのシーンを持つ映像 #5 を用いた．

³映像 #5 の全てのショット境界を図 5.7 に示すことが困難であるため，映像 #5 の正解のシーン 1 ～ シーン 3 を例とした．

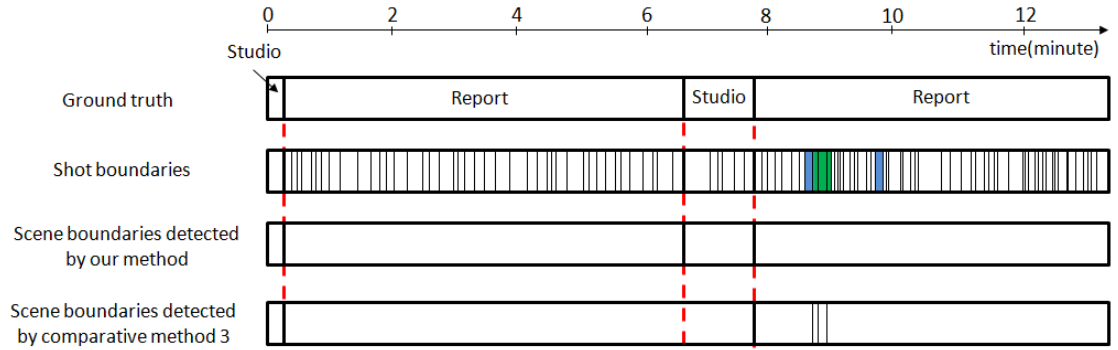


図 5.11: 提案手法と比較手法3におけるシーン分割結果の比較.

図 5.9 より, 同一シーン中に存在するショットが異なるシーン中に存在するショットより互いに類似していることが確認できる. また, 図 5.10 より, シーン境界候補より推定されるシーン境界位置が正解シーン境界に近づいていくことが確認できるため, マルコフ連鎖モンテカルロ法を用いてシーン境界候補よりシーン境界を推定することは有効であることが確認できる. 図 5.10 では, シーン数およびシーン境界の位置に関する状態は N_{iter} 回で更新を終了し, 提案手法で得られるシーン分割結果となっている. なお, 図 5.10 には正解シーン分割結果も合わせて表示した.

さらに, 映像 #5 の一部分に, 提案手法と従来手法 [5] を適用することで得られたシーン分割結果を図 5.11 に示し, その比較を行う. 図 5.11 より, 緑色で示される隣接ショットの類似性が低いのに対して, それらの両端に存在する青色で示されるショットの類似性が高い. 提案手法では, 青色で示される 2 ショットが類似していることに注目しているため, その間に含まれる隣接ショットの類似性が低い場合においても, この 2 ショットが同一クラスに統合され, 青色で示される 2 ショットおよびその間に存在するショットを同一シーンにまとめることが可能となっている. したがって, 提案手法では, 従来手法 [5] の問題点を解決し, 高精度なシーン分割が可能となっている.

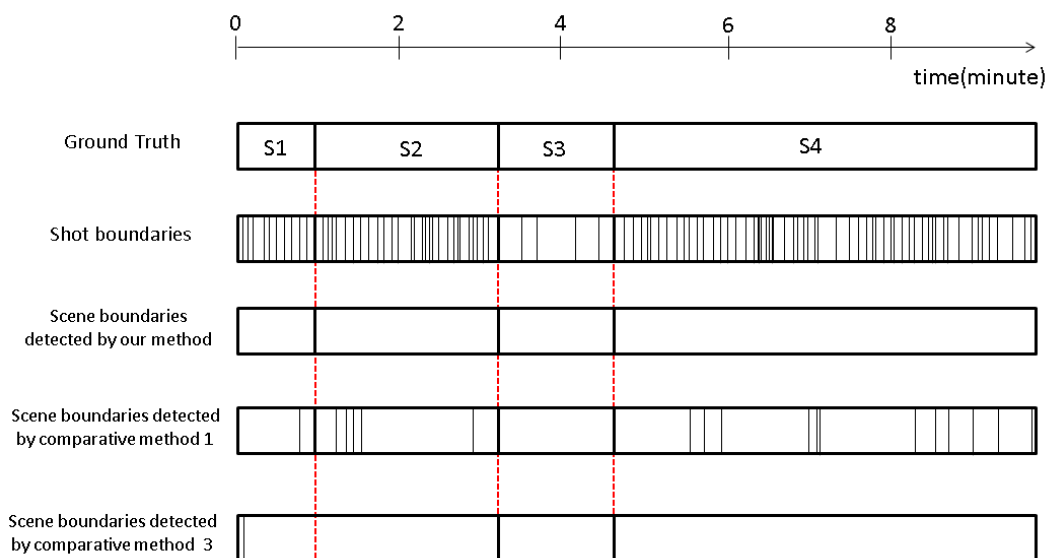


図 5.12: 映像 #2 に提案手法および従来手法 1,3 を適用して得られたシーン分割結果.

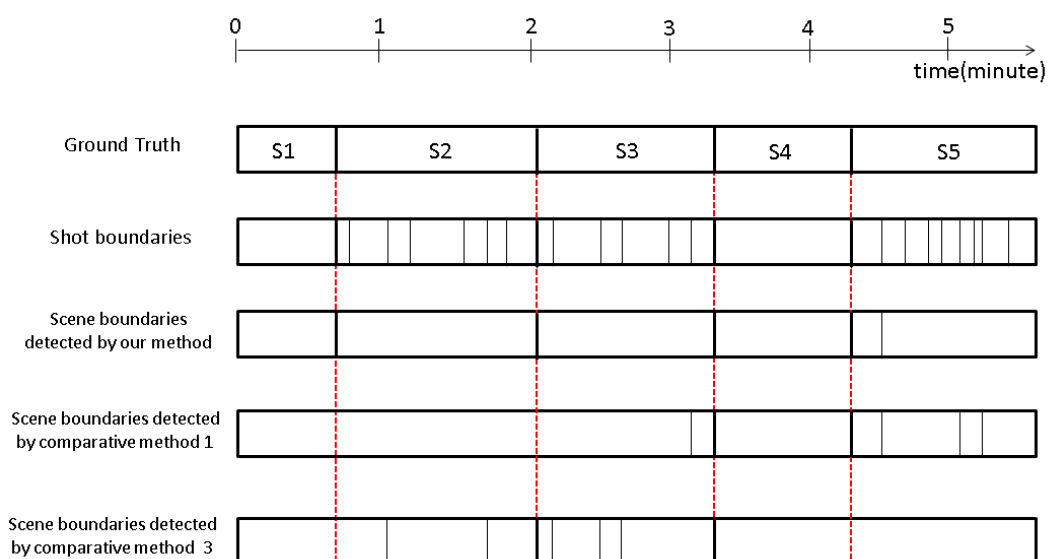


図 5.13: 映像 #6 に提案手法および従来手法 1,3 を適用して得られたシーン分割結果.

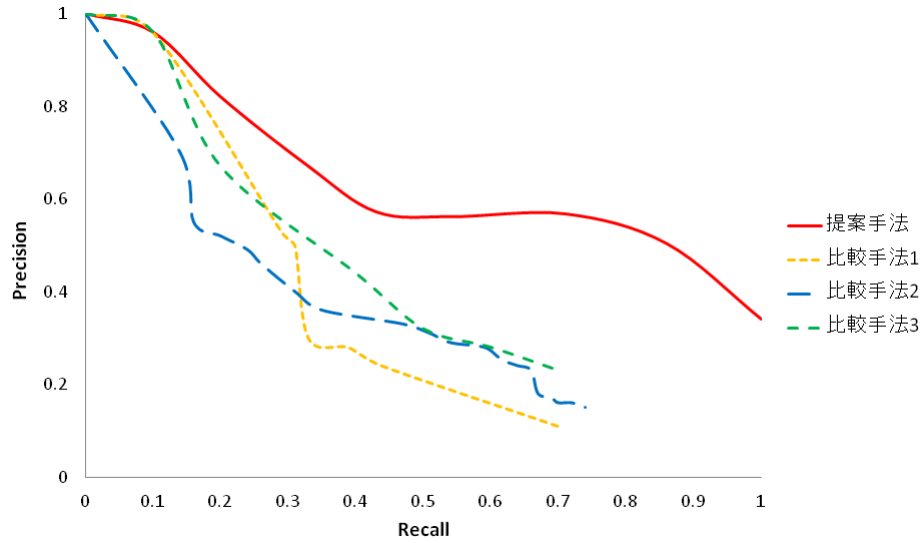


図 5.14: 提案手法および従来手法 1,2,3 に，提案手法で用いられるショット間の類似度 $FSim^O(s_i, s_j)$ を適用した場合の Recall-Precision カーブ.

さらに，映像 #2 および映像 #6 に対し，提案手法，比較手法 1 および比較手法 3 を適用し，得られたシーン分割結果の一部を図 5.12, 図 5.13 に示す．図 5.12 において，比較手法 1 では全ての境界を検出したが，過分割が発生していることが確認できる．一方，比較手法 3 では，比較手法 1 より過分割問題の抑制を可能としたが，未分割が発生した．これに対し，提案手法は，比較手法に存在する過分割問題や未分割問題を抑制し，高精度にシーン分割を可能とした．また，図 5.13 より，提案手法は過分割したものの，比較手法より高精度にシーン分割していることが確認できる．以上より，映像のシーン分割結果の例を用いて，提案手法の有効性が確認できる．

次に，実験対象映像 (#1 ~ #15) 全てを用いて，提案手法および比較手法 1,2,3 の Recall-Precision カーブを図 5.14 に示す．なお，Recall-Precision カーブを求める際，5.6.1 節と同様に取得する．図 5.14 より，比較手法 1,2,3 は提案手法と同様の特徴量を用いてシーン分割を行う際，提案手法が比較手法 1,2,3 より高精度にシーン分割が可能であることが確認でき，提案手法による高精度なシーン分割を実現している．以上により，図 5.5，図 5.6 および図 5.14 により，提案手法は比較手法と同様の特徴量を用いてシーン分割を行う際，いずれの比較手法よりも高精度にシーン分割を行うことから，提案手法において階層ベイズモデルの導入による映像の構造に注目

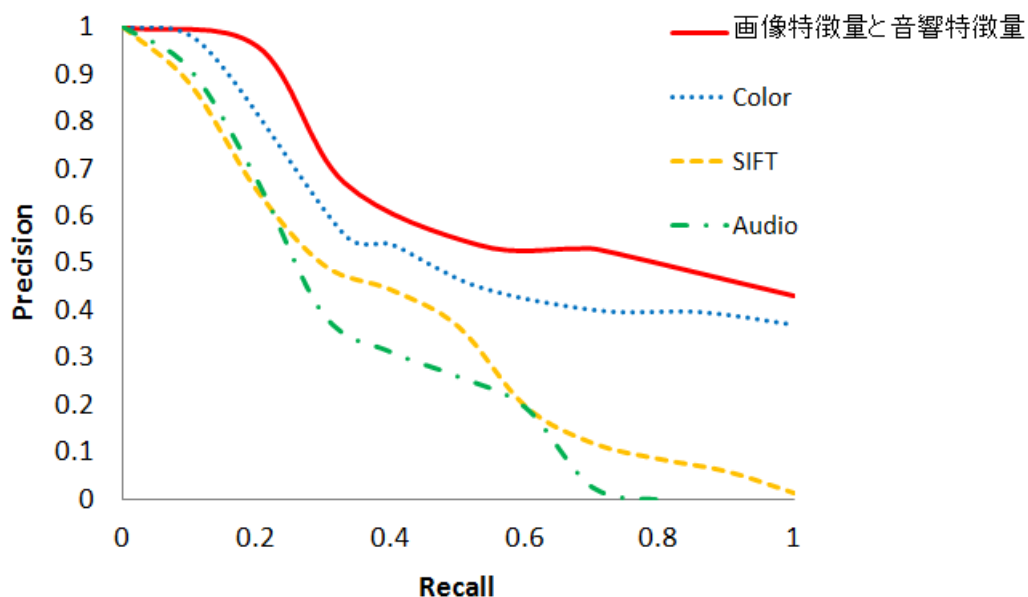
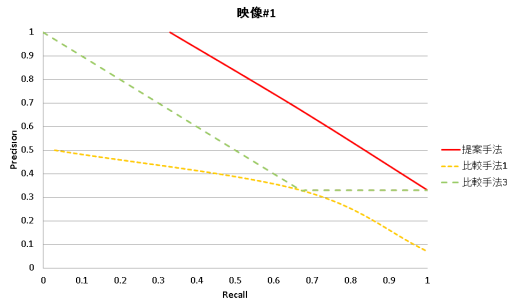


図 5.15: 提案手法に，画像特徴量と音響特徴量を用いた際に取得される Recall-Precision カーブと，単一特徴量のみを用いた際に得られる Recall-Precision カーブの比較．

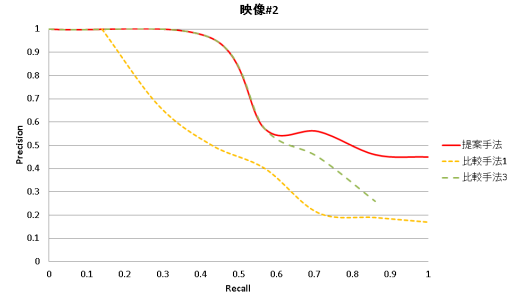
したシーン分割手法の有効性が確認できる．

さらに，提案手法において，単一特徴量のみを用いた際に得られる Recall-Precision カーブと，画像特徴量と音響特徴量を用いた際に得られる Recall-Precision カーブを比較し，画像特徴量と音響特徴量の利用の有効性を確認する．実験対象映像 (#1 ~ #15) に対し，提案手法に画像特徴量と音響特徴量を用いて得られる Recall-Precision カーブと提案手法に色ヒストグラム特徴量，SIFT 特徴量，および音響特徴量の各々を用いて得られる Recall-Precision カーブを図 5.15 に示す．図 5.15 より，画像特徴量と音響特徴量を用いた際に得られる Recall-Precision カーブは，単一特徴量のみを用いた際に得られる Recall-Precision カーブより高い値を持つことから，画像特徴量と音響特徴量の利用によるシーン分割の高精度化が確認できた．

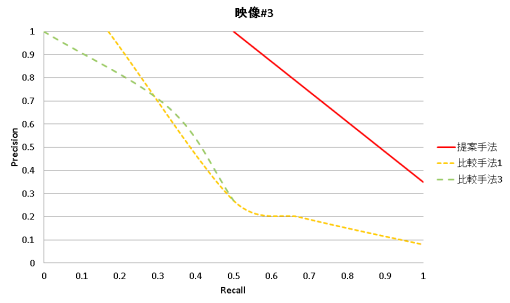
最後に，それぞれの実験映像に対して，Recall-Precision カーブを求め，提案手法とベンチマーキング手法である比較手法 1 および比較手法 3 に関する比較実験を行う．



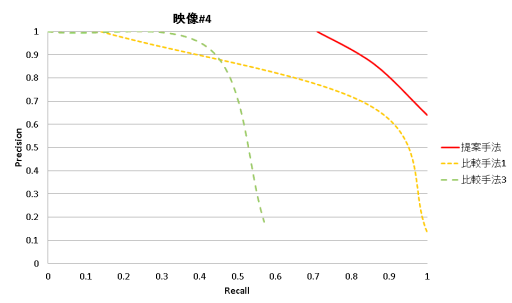
(a)



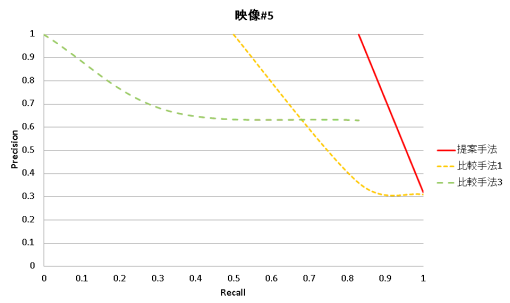
(b)



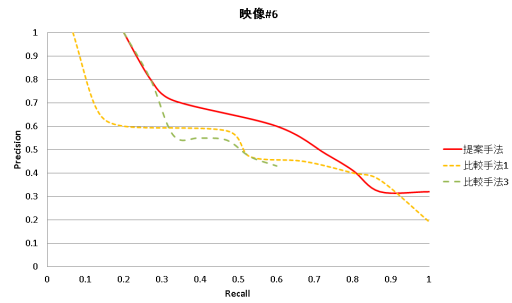
(c)



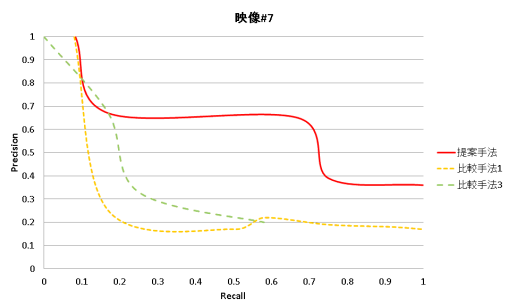
(d)



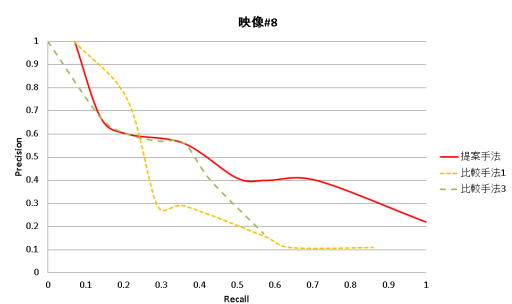
(e)



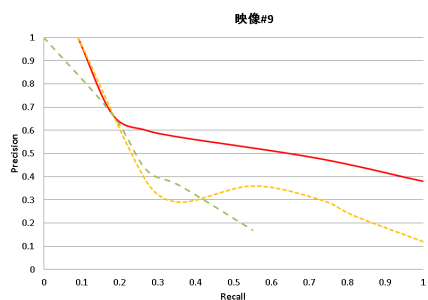
(f)



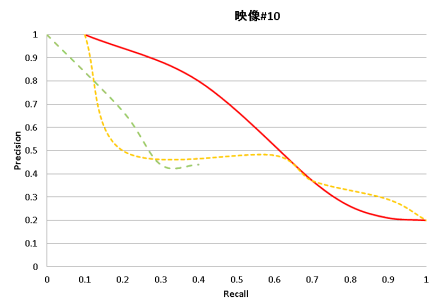
(g)



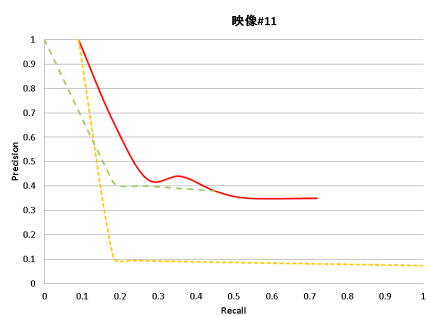
(h)



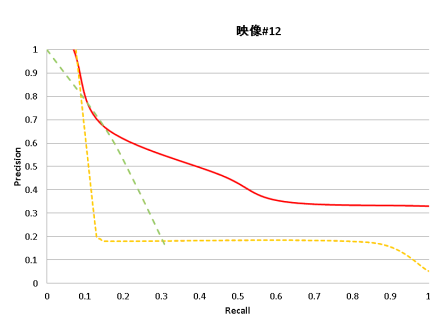
(i)



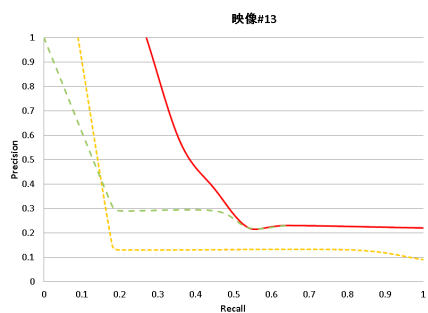
(j)



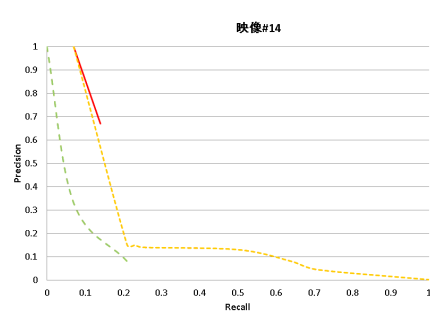
(k)



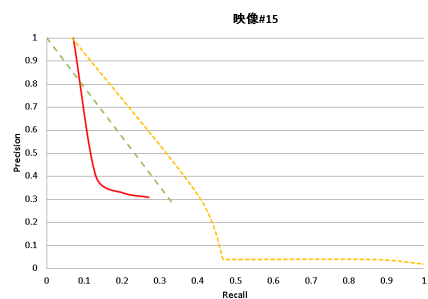
(l)



(m)



(n)



(o)

図 5.16: 各映像の Recall-Precision カurve.

図 5.16 により，提案手法は比較手法 1 および比較手法 3 に比べ，高精度にシーン分割を実現していることが確認できる．そこで，本節において，図 5.5，図 5.6 および図 5.14 で，提案手法が高精度にシーン分割を実現していることを確認した．このとき，提案手法と比較手法と同様の特徴量を用いているため，階層ベイズモデルを映像の構造に注目したシーン分割手法に導入する有効性が確認された．次に，図 5.15 では，画像特徴量と音響特徴量の両方を用いて得られる Recall-Precision カーブと単一特徴量のみを用いた Recall-Precision カーブを考察し，画像特徴量と音響特徴量の利用によるシーン分割のさらなる高精度化が確認された．さらに，画像特徴量と音響特徴量を用いた提案手法，色ヒストグラム特徴量を用いた比較手法 1 および比較手法 3 のそれぞれを，各実験対象映像に適用し，得られる Recall-Precision カーブを図 5.16 に示す．図 5.16 より，提案手法は比較手法 1 および比較手法 3 より高精度にシーン分割を実現していることを確認した．

最後に，異なるジャンルの映像に提案手法を適用して得られるシーン分割結果について考察を行う．まず，ニュース映像では，映像の構造が明確であるために，効果的にシーン境界の推定を行うことが可能であり，高精度なシーン分割が実現された．また，ドキュメンタリー映像についても，話題の変化が明確であるために，話題の切り替わりにおけるシーン境界の推定が容易であった．これに対し，映画では，異なるシーンに類似する景色や同一人物が撮像される場合が存在するため，シーン分割精度が劣化する場合が存在した．さらに，アニメ映像では，他の映像と特徴が大きく異なり，また映画と同様な問題を持つため，シーン分割精度が低下する場合が存在した．以上をまとめると，ニュース映像およびドキュメンタリー映像では，話題の切り替わりが明確であるため，高精度なシーン分割が可能である．一方，映画およびアニメ映像では，話題が変化しても，撮像される人物や物体が同一である場合が存在するために，シーン分割精度が劣化する可能性がある．

以上により，提案手法は映画およびアニメ映像においては，シーン分割精度に限界があったが，他手法に比べて，高精度にシーン分割を実現することが確認された．

5.7 まとめ

本章では，複数の特徴量を利用し，階層ベイズモデルの導入による映像の構造に注目したシーン分割の高精度化手法を提案した．提案手法では，異なるショット間の類似度に関する分布を考慮し，異なる信号の特徴量の統合を可能とした．また，複数の特徴量を用いて，映像の構造に注目するシーン分割手法により境界候補の取得を行うことにより，画像特徴量だけでなく，音響特徴量も同時に注目することで，高精度なシーン境界候補の取得を可能とした．さらに，定義される階層ベイズモデルの導入による映像の構造に注目したシーン分割精度の向上を可能とした．最後に，実験によりさらなるシーン分割の高精度化が実現されていることを確認した．

第6章

結論

本研究では，階層ベイズモデルの導入による映像の構造に注目したシーン分割手法について提案を行った．従来の映像の構造に注目したシーン分割手法は，編集映像に現れる，類似するショットの繰り返し構造を見つけることで，シーン分割を行う．この手法では，映像の構造を解析する際，注目する繰り返し構造を示す最長の時間幅を設定する必要がある．しかしながら，最長時間幅は映像の種類によっても編集者によっても異なるため，一意に決定することが困難であるという問題が存在した．一方，階層ベイズモデルに基づくシーン分割手法では，シーン数やシーン境界の位置に関する事後確率の最大化により，最適なシーン境界を検出することで，映像の構造に注目したシーン分割手法が抱える問題を回避している．しかしながら，この手法は，その事後確率をショット間の類似度のみに注目して定義しているため，同一シーン中の隣接するショット間で類似度が低い組み合わせが存在すると，誤ってシーン境界と推定する場合がある．さらに，従来のこの2種類の映像のシーン分割手法が，画像特徴量のみを用いることに対して，映像を構成する画像信号と音響信号の両方の特徴量を用いることで，シーン分割の高精度化が実現された．しかしながら，異なる特徴量を用いてショット間の類似度を算出する場合，その分布が異なるため，特徴量を比較可能な数値へ変換する必要があった．

そこで，本論文では階層ベイズモデルの導入による映像の構造を考慮したシーン分割手法を実現した．提案手法では，予め取得される各ショットのキーフレームを抽出し，異なるショットのキーフレーム間の類似度をショット間の類似度として算出する．次に，ショット間の類似関係に注目し，映像の構造を取得する際に必要

である映像のシーンの最長時間幅に複数の値を与え、映像の分割結果を取得する。このとき、得られた全ての分割結果を用いて映像のシーン境界の候補とする。提案手法では、映像の構造を考慮したシーン境界候補を階層ベイズモデルを用いてシーン境界の推定を行う。具体的に、シーン数やシーン境界の位置に関する事後確率を定義し、これをマルコフ連鎖モンテカルロ法を用いて最適化を行うことで、シーン境界候補の中からシーン境界を推定する。これにより、映像の構造と階層ベイズモデルの両者を利用することで、高精度なシーン分割の実現を可能とした。さらに、提案手法では、先行研究と同様に、複数の特徴量を用いてシーン分割のさらなる高精度化を試みる。このとき、画像特徴量と音響特徴量のそれぞれの特徴量に基づきシーン境界候補の取得を行うことで、網羅的にシーン境界候補の選択を可能とする。さらに、画像特徴量と音響特徴量の両者を用いて階層ベイズモデルにおける事後確率の算出を行うことで、高精度に事後確率の算出を可能とする。ただし、異なる特徴量を用いて算出されるショット間の類似度の分布が異なるため、提案手法では、異なるショットの全ての組み合わせで類似度を算出して分布を求め、比較可能とすることで、事後確率の算出を実現した。これにより、従来の映像の構造に注目したシーン分割手法の問題点を、階層ベイズモデルの導入により解決し、高精度なシーン分割が可能となる。さらに、画像と音響に関する特徴量の類似度に関して、その分布を考慮しながら比較可能とする定式化を導入することで、シーン分割精度のさらなる向上を可能とした。以下、各章で論じた内容について要約する。

第2章では、従来より提案されているショット分割法について説明し、あわせて用語の定義を行った。また、本論文で用いられるショット分割法を明記した。

第3章では、映像のシーン分割に関する従来手法の概要を述べるとともに、既存のシーン分割手法に代表的な映像の構造に注目したシーン分割手法および階層ベイズモデルに基づくシーン分割手法をそれぞれ説明し、映像のシーン分割の高精度化を実現するために、本論文で解決すべき問題について明確にした。

第4章では、階層ベイズモデルの導入による映像の構造に注目したシーン分割手法を提案し、従来の映像の構造に注目したシーン分割手法の問題点の解決を可能とした。提案手法は、まず、予め取得される映像の各ショットに対し、そのショットの内容を代表するキーフレームの取得を行い、ショット間の画像特徴量を用い

た際の類似度の算出を可能とした．次に，ショット間の類似度を用いて，映像の繰り返し構造の最長時間幅を一意に決定するのではなく，複数の値を与えることで，映像の構造を考慮したシーン境界候補の取得が可能となった．最後に，ショット間の類似度を用いて，シーン数やシーン境界の位置に関する事後確率を階層ベイズモデルで定義し，それをマルコフ連鎖モンテカルロ法を用いて最適化することで，映像の構造を考慮したシーン境界候補からシーン境界の推定を可能とした．第4章によって，映像の構造を考慮したシーン境界候補より，階層ベイズモデルを用いてシーン境界の推定を行うことから，従来の映像の構造に注目したシーン分割手法の問題が解決されることを可能とした．

第5章では，画像特徴量と音響特徴量を用いてショット間の類似度算出法を紹介すると共に，その分布を考慮しながら比較可能とする定式化を導入したシーン分割手法を提案し，高精度なシーン分割結果の取得を実現した．画像特徴量と音響特徴量のそれぞれ算出されたショット間の類似度を用いて，映像の構造を考慮したシーン境界候補の取得を行うことで，網羅的なシーン境界候補の取得を可能とした．また，画像特徴量と音響特徴量を用いて事後確率の算出を行うことで，画像特徴量および音響特徴量を同時に利用した高精度な事後確率の算出を可能とし，高精度なシーン分割が実現された．

本研究では，映像のシーン分割の高精度化を実現するため，従来に提案された映像の構造に注目したシーン分割手法および階層ベイズモデルに基づくシーン分割手法のそれぞれの手法が抱える問題点の解決によるシーン分割の精度向上と，画像特徴量と音響特徴量を用いるための比較可能な数値への変換を行うことによりシーン分割のさらなる向上に取り組んできた．そして，映像の構造に注目したシーン分割手法に，階層ベイズモデルの導入による高精度なシーン分割手法を実現した．また，複数の特徴量を利用してシーン分割を行い，さらなるシーン分割の精度向上を実現した．本研究の今後の課題として，ユーザの嗜好を考慮した映像の階層的なシーン分割手法の検討が挙げられる．

参考文献

- [1] M. Yeung, B. Yeo, B. Liu, “Segmentation of video by clustering and graph analysis,” *Computer Vision and Image Understanding*, vol. 75, no. 1, pp. 94-109, 1998.
- [2] A. Hanjalic, R. L. Lagendijk, “Automated High-level movie segmentation for advanced video-retrieval systems,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 9, no. 4, pp. 580-588, 1999.
- [3] H. Sundaram, S. F. Chang, “Video scene segmentation using video and audio features,” in *Int. Conf. Multimedia and Expo*, vol. 2, pp. 1145-1148, 2000.
- [4] C. J. Geyer, “Practical Markov Chain Monte Carlo,” *Statistical Science*, vol. 7, no. 4, pp. 473–483, 1992.
- [5] Y. Zhai, M. Shah, “Video Scene Segmentation Using Markov Chain Monte Carlo,” *IEEE Trans. on Multimedia*, vol. 8, no. 4, pp. 686–697, 2006.
- [6] P. Geetha, V. Narayanan, “A Survey of Content-Based Video Retrieval,” *Journal of Computer Science*, vol. 4, pp. 474–486, 2008.
- [7] 長坂 晃郎, 田中 譲, “カラービデオ映像における自動索引付け法と物体探索法,” *情報処学論*, vol. 33, no. 4, pp. 543–550, 1992.
- [8] J. Kender, B. L. Yeo, “ Video scene segmentation via continuous video coherence,” *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 367–373, 1998.
- [9] Z. Rasheed, M. Shah, “ Scene Detection in Hollywood Movies and TV Shows,” *IEEE Computer Society*, pp. 343, 2003.

- [10] Y. M. Kwon, C. J. Song, I. J. Kim, “ A new approach for high level video structuring,” IEEE International Conference on Multimedia and Expo, ICME (2000), vol. 2, pp. 773–776, 2000.
- [11] L. Zhao, S. Q. Yang, B. Feng, “ Video scene detection using slide windows method based on temporal constrain shot similarity,” IEEE International Conference on Multimedia and Expo, ICME (2001), pp. 1171–1174, 2001.
- [12] C. Wengang, X. De, “ A novel approach of generating video scene structure,” TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region, vol. 1, pp. 350–353, 2003.
- [13] X.Wang, S. Wang, H. Chen, M. Gabbouj, “ A shot clustering based algorithm for scene segmentation,” International Conference on Computational Intelligence and Security Workshops, CISW 2007, pp. 259–252, 2007.
- [14] I. Mikic, P. C. Cosman, G. T. Kogut, M. M. Trivedi, “ Moving Shadow and Object Detection in Traffic Scenes ”, Pattern Recognition, 2000. Proceedings. 15th International Conference on, vol. 1, pp. 321-324, 2000.
- [15] B. T. Truong, S. Venkatesh, C. Dorai, “ Scene extraction in motion pictures,” IEEE Trans. on Circuits Syst. Video Technol., vol. 13, no. 1, pp. 5–15, 2003.
- [16] Y. Rui, T. S. Huang, S. Mehrotra, “ Constructing table-of-content for videos,” Multimed. Syst., vol. 7, no. 5, pp. 359–368, 1999.
- [17] J. Zhou, W. Tavanapong, “ Shot Weave: A shot clustering technique for story browsing for large video databases,” XML-Based Data Management and Multimedia Engineering EDBT 2002 Workshops, Lecture Notes in Computer Science, vol. 2490, no. 17, pp. 529–533, 2002.
- [18] W. Tavanapong, J. Zhou, “ Shot Clustering Techniques for Story Browsing,” IEEE Trans. on Multimed., vol. 6, no. 4, pp. 517–527, 2004.
- [19] S. Zhu, Y. Liu, “ Video scene segmentation and semantic representation using a novel scheme,” Multimed. Tools Appl., vol. 42, no. 2, pp. 183–205, 2009.

- [20] J. M. Odobez, D. Gatica-Perez, M. Guillemot, “ Spectral structuring of home videos,” *Image and Video Retrieval, Lecture Notes in Computer Science*, vol. 2728, no. 31, pp. 85–90, 2003.
- [21] V. T. Chasanis, A. C. Likas, N. P. Galatsanos, “ Scene detection in videos using shot clustering and sequence alignment,” *IEEE Trans. on Multimed.*, vol. 11, no. 1, pp. 89–100, 2009.
- [22] O. Javed, Z. Rasheed, M. Shah, “ A framework for segmentation of talk and game shows,” *Eighth IEEE International Conference on Computer Vision, ICCV(2001)*, 2001.
- [23] C. W. Ngo, Y. F. Ma, H. J. Zhang, “ Video summarization and scene detection by graph modeling,” *IEEE Trans. on Circuits Syst. Video Technol.*, vol. 15, no. 2, pp. 296–305, 2005.
- [24] Y. Zhao, T. Wang, P. Wang, W. Hu, Y. Du, Y. Zhang, G. Xu “ Scene Segmentation and Categorization Using Ncuts,” *Computer Vision and Pattern Recognition, 2007. CVPR '07, IEEE Conference on*, pp. 1–7, 2007.
- [25] S. Benini, L. Q. Xu, R. Leonardi, “ Identifying video content consistency by vector quantization,” *Proceedings of the 2005 International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2005)*, 2005.
- [26] Z. Rasheed, M. Shah, “ Detection and representation of scenes in videos,” *IEEE Trans. on Multimed.*, vol. 7, no. 6, pp. 1097–1105, 2005.
- [27] H. Kang, “ A hierarchical approach to scene segmentation,” *Content-Based Access of Image and Video Libraries, 2001. (CBAIVL 2001). IEEE Workshop on*, pp. 65–71, 2001.
- [28] Z. Zhang, B. Li, H. Lu, X. Xue, “ Scene segmentation based on video structure and spectral methods,” *10th International Conference on Control, Automation, Robotics and Vision, 2008. ICARCV 2008*, pp. 1093–1096, 2008.
- [29] U. Sakarya, Z. Telatar, “ Graph-based multilevel temporal video segmentation,” *Multimed. Syst.*, vol. 14, no. 5, pp. 277–290, 2008.

- [30] C. Y. Weng, W. T. Chu, J. L. Wu, “RoleNet: Movie Analysis from the Perspective of Social Networks,” *IEEE Trans. on Multimed.*, vol. 11, no. 2, pp. 256–271, 2009.
- [31] U. Sakarya, Z. Telatar, “Video scene detection using graph-based representations,” *Signal Process. Image Commun.*, vol. 25, no. 10, pp. 774–783, 2010.
- [32] D. Gatica-Perez, A. Loui, M. T. Sun, “Finding structure in home videos by probabilistic hierarchical clustering,” *IEEE Trans. on Circuits Syst. Video Technol.*, vol. 13, no. 6, pp. 539–548, 2003.
- [33] L. Xie, “Structure analysis of soccer video with domain knowledge and hidden markov models,” *Pattern Recognit. Lett.*, vol. 25, no. 7, pp. 767–775, 2004.
- [34] T. Lin, H. J. Zhang, Q. Y. Shi, “Video scene extraction by force competition,” *IEEE International Conference on Multimedia and Expo*, pp. 192 , 2001.
- [35] Z. Gu, T. Mei, X. S. Hua, X. Wu, S. Li, “EMS: Energy Minimization Based Video Scene Segmentation,” *IEEE International Conference on Multimedia and Expo*, pp. 520–523, 2007.
- [36] J. R. Cao, “Algorithm of scene segmentation based on svm for scenery documentary,” *Third International Conference on Natural Computation, 2007 (ICNC 2007)*, vol. 3, pp. 95–98, 2007.
- [37] B. Han, W. Wu, “Video scene segmentation using a novel boundary evaluation criterion and dynamic programming,” *IEEE International Conference on Multimedia and Expo (ICME)*, 2011, pp. 1–6, 2011.
- [38] L. Lu, R. Cai, A. Hanjalic, “Audio elemes based auditory scene segmentation,” *IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings*, vol. 5, p. V, 2006.
- [39] F. Niu, N. Goela, A. Divakaran, M. Abdel-Mottaleb, “Audio scene segmentation for video with generic content,” *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, vol. 6820, 2008.

- [40] Y. Zhai, A. Yilmaz, M. Shah, “ Story segmentation in news videos using visual and text cues,” *Image and Video Retrieval, Lecture Notes in Computer Science*, vol. 3568, no. 13, pp. 92–102, 2005.
- [41] H. Sundaram, S. F. Chang, “ Video scene segmentation using video and audio features,” *IEEE International Conference on Multimedia and Expo, 2000. ICME (2000)*, vol. 2, pp. 1145–1148, 2000.
- [42] H. Sundaram, S. F. Chang, “ Computable scenes and structures in films,” *IEEE Trans. on Multimed.*, vol. 4, no. 4, pp. 482–491, 2002.
- [43] J. Huang, Z. Liu, W. Yao, “ Integration of audio and visual information for content-based video segmentation,” *International Conference on Image Processing, ICIP 98*, vol. 3, pp. 526–529, 1998.
- [44] R. Lienbart, S. Pfeiffer, W. Effelsberg, “ Scene determination based on video and audio features,” *IEEE International Conference on Multimedia Computing and Systems*, vol. 1, pp. 685–690, 1999.
- [45] S. C. Chen, M. L. Shyu, W. Liao, C. Zhang, “ Scene change detection by audio and video clues,” *Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference on*, vol. 2, pp. 365–368, 2002.
- [46] N. Nitanda, M. Haseyama, H. Kitajima, “ Audio signal segmentation and classification for scene-cut detection,” *IEEE International Symposium on Circuits and Systems, 2005. ISCAS 2005*, vol. 4, pp. 4030–4033, 2005.
- [47] A. Hauptmann, M. Witbrock, “ Story segmentation and detection of commercials in broadcast news video,” *Proceedings. IEEE International Forum on Research and Technology Advances in Digital Libraries, 1998. ADL 98*, pp. 168–179, 1998.
- [48] A. Velivelli, C. W. Ngo, T. S. Huang, “ Detection of documentary scene changes by Audio-Visual fusion image and video retrieval,” *Image and Video Retrieval, Lecture Notes in Computer Science*, vol. 2728, no. 23, pp. 227–238, 2003.

- [49] M. Kyperountas, C. Kotropoulos, I. Pitas, “Enhanced Eigen-Audioframes for audiovisual scene change detection,” *IEEE Trans. on Multimed.*, vol. 9, no. 4, pp. 785–797, 2007.
- [50] L. Chaisorn, T. S. Chua, C. H. Lee, “The segmentation of news video into story units,” *IEEE International Conference on Multimedia and Expo, 2002. ICME '02, 2002*, vol. 1, pp. 73–76, 2002.
- [51] M. Yamamoto, M. Haseyama, “An Accurate Scene Segmentation Method Based on Graph Analysis Using Object Matching and Audio Feature,” *IEICE Trans. on Fundamentals*, vol. E92-A, no. 8, pp. 1883–1891, 2009.
- [52] S. Chib, E. Greenberg, “Understanding the Metropolis-Hastings Algorithm,” *American Statistician*, vol. 59, no. 4, pp. 327–335, 1995.
- [53] P. Green, “Reversible jump Markov chain Monte Carlo computation and Bayesian model determination,” *Biometrika*, vol. 82, pp. 711–732, 1995.
- [54] L. Chen, Y. Lai, H. M. Liao, “Movie Scene Segmentation Using Background Information,” *Pattern Recognition*, vol. 41, no. 3, pp. 1056–1065, 2008.
- [55] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, I. Trancoso, “Temporal Video Segmentation to Scenes Using High-Level Audiovisual Features,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 21, no. 8, pp. 1163–1177, 2011.
- [56] Y. Zhu, Z. Ming, “SVM-Based Video Scene Classification and Segmentation,” *International Conference on Multimedia and Ubiquitous Engineering*, pp. 407–412, 2008.

謝辞

本研究は、著者が北海道大学大学院情報科学研究科メディアネットワーク専攻に在学中の、2010年4月より現在にいたる約5年に渡って行ったものである。

本研究の遂行のみならず、終始御指導、御鞭撻を頂いた、北海道大学大学院情報科学研究科、長谷山美紀教授に深く感謝致します。

本論文をまとめるにあたり、御助言を頂き、副査を御引き受け頂いた、北海道大学大学院情報科学研究科、荒木健治教授ならび山本強教授に深く感謝致します。

さらに、著者の研究期間中、御指導を頂き、公私にわたり多大なる御助力を頂いた、北海道大学大学院情報科学研究科、小川貴弘助教、高橋翔特任助教に深く感謝致します。

最後に、御協力を頂いた、北海道大学大学院情報科学研究科メディアネットワーク専攻メディアダイナミクス研究室諸先輩並びに学生の皆様に感謝致します。

著者の研究業績

- [1] 宋 妍, 小川 貴弘, 長谷山 美紀, “映像の構造に基づいたシーン分割の高精度化に関する検討,” 電気・情報関係学会北海道支部連合大会講演論文集, p. 155 (2009-10)
- [2] 宋 妍, 小川 貴弘, 長谷山 美紀, “映像の構造に注目した MCMC 法によるシーン分割の高精度化に関する検討,” 電子情報通信学会技術報告, vol. 110, no. 86, pp. 115–120 (2010-06)
- [3] 宋 妍, 小川 貴弘, 長谷山 美紀, “音響特徴量を導入した映像の構造に基づく MCMC 法によるシーン分割の高精度化に関する検討,” 電気・情報関係学会北海道支部連合大会講演論文集, p. 141 (2010-10)
- [4] Yan Song, Takahiro Ogawa, Miki Haseyama, “MCMC-Based Scene Segmentation Method Using Structure of Video,” 10-th International Symposium on Communications and Information Technologies 2010(ISCIT), vol. FM-1-2-3, pp. 862–866 (2010-10)
- [5] 宋 妍, 小川 貴弘, 長谷山 美紀, “撮像物体の対応を考慮した MCMC 法に基づくシーン分割の高精度化に関する検討,” 電子情報通信学会技術報告, vol. 110, no. 420, pp. 131–135 (2011-02)
- [6] 宋 妍, 小川 貴弘, 長谷山 美紀, “複数特徴量の協調利用による MCMC 法に基づくシーン分割に関する検討–映像の構造に注目したシーン境界候補の導入–,” 第 13 回 DSPS 教育者会議予稿集, pp. 10–20 (2011-09)
- [7] Yan Song, Takahiro Ogawa, Miki Haseyama, “Audio-Visual Feature-based Scene Segmentation and Its Application to Multi-genre Videos,” The 27th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC 2012) (2012-07)
- [8] 宋 妍, 小川 貴弘, 長谷山 美紀, “映像のシーン分割における複数特徴量の協調利用の有効性に関する評価,” 電気・情報関係学会北海道支部連合大会講演論文集, p. 123 (2012-10)

- [9] 宋 妍, 小川 貴弘, 長谷山 美紀, “タグ付き画像を用いた同一意味を表現するタグの選択に関する検討,” 映像情報メディア学会技術報告, vol. 38, no. 7, pp. 211-216 (2014-02)
- [10] 宋 妍, 小川 貴弘, 長谷山 美紀, “映像の構造に注目した MCMC 法に基づくシーン分割法,” 電子情報通信学会論文誌 D, vol. J97-D, no. 3, pp. 560 - 573 (2014-03)