(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2008-5167
(P2008-5167A)
(43)公開日　平成20年1月10日(2008.1.10)

(51)Int.Cl.		F I			テーマコード（参考）
H04N 5/76	(2006.01)	H04N	5/76	B	5B075
H04N 5/91	(2006.01)	H04N	5/91	N	5C052
G06F 17/30	(2006.01)	G06F	17/30	170D	5C053
G11B 27/034	(2006.01)	G06F	17/30	210D	5D110
		G11B	27/034		

審査請求 未請求　請求項の数 13　ＯＬ　（全 23 頁）

(21)出願番号　特願2006-171830 (P2006-171830)
(22)出願日　平成18年6月21日 (2006.6.21)

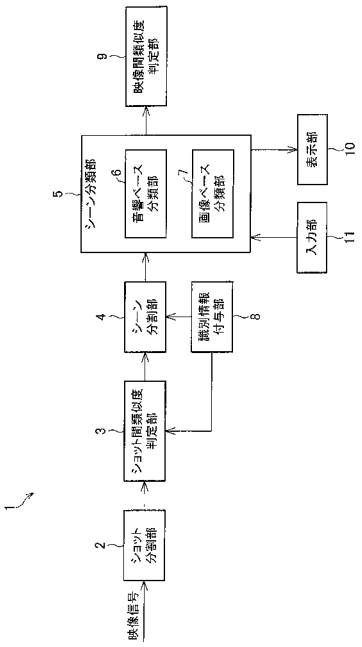
特許法第３０条第１項適用申請有り　２００６年２月１３日　社団法人　電子情報通信学会発行の「電子情報通信学会技術研究報告　信学技報Ｖｏｌ．１０５　Ｎｏ．６０８」に発表

(71)出願人　504173471
国立大学法人　北海道大学
北海道札幌市北区北８条西５丁目８番地
(74)代理人　110000338
特許業務法人原謙三国際特許事務所
(72)発明者　長谷山　美紀
北海道札幌市北区北１４条西９丁目　北海道大学大学院情報科学研究科内
(72)発明者　二反田　直己
北海道札幌市北区北１４条西９丁目　北海道大学大学院情報科学研究科内

Ｆターム(参考)　5B075 ND12 NR02 PQ48 UU35
5C052 AB03 AB04 AC08 DD10
5C053 FA30 GB09 GB11 HA29 LA14
5D110 AA13 AA14 AA27 AA29 CB06
CD03 CD04 CD22 DA15

(54)【発明の名称】映像分類装置、映像分類方法、映像分類プログラムおよびコンピュータ読取可能な記録媒体

(57)【要約】
【課題】効果的なシーン間の境界を検出することのできる映像分類装置を実現する。
【解決手段】映像分類装置１は、映像信号に含まれるビデオ信号に基づきショット間の境界を検出して映像を各ショットに分割するショット分割部２と、分割されたショット毎に、ショット内のオーディオ信号について、音の種類で分類された各クラスにどの程度属しているかを示す帰属確率を算出し、この帰属確率を用いて隣接するショット間の類似度を判定するショット間類似度判定部３と、判定されたショット間類似度が所定値より高いショット同士は統合させ上記映像を各シーンに分割するシーン分割部４と、を備えている。よって、オーディオ信号の帰属確率からショット間の類似の度合を算出するので、類似したショットをシーンとしてまとめることができ、その結果、効果的なシーン間の境界を検出することができる。
【選択図】図１



【特許請求の範囲】
【請求項１】
映像信号に含まれるビデオ信号に基づきショット間の境界を検出して映像を各ショットに分割するショット分割手段と、
分割されたショット毎に、ショット内のオーディオ信号について、音の種類で分類された各クラスにどの程度属しているかを示す帰属確率を算出し、この帰属確率を用いて隣接するショット間の類似度を判定するショット間類似度判定手段と、
判定されたショット間類似度が所定値より高いショット同士は統合させ、上記映像を各シーンに分割するシーン分割手段と、
を備えたことを特徴とする映像分類装置。
【請求項２】
上記音の種類とは、無音、音声、音楽、音楽付き音声、雑音付き音声の５つの種類であることを特徴とする請求項１に記載の映像分類装置。
【請求項３】
上記ショット間類似度判定手段は、
ショット内のオーディオ信号を分割した各クリップについて上記帰属確率を算出し、ショット内の各クリップの帰属確率の累積ヒストグラムを基に、隣接するショット間の類似度を判定することを特徴とする請求項１または２に記載の映像分類装置。
【請求項４】
上記帰属確率の累積ヒストグラムのうち、最大の値のクラスを示す識別情報を処理対象のショットに付与するクラス識別情報付与手段を備えることを特徴とする、請求項３に記載の映像分類装置。
【請求項５】
上記分割された各シーンに含まれるショットに付与された上記識別情報に基づき、各シーンを分類する音響ベース分類手段を備えたことを特徴とする請求項４に記載の映像分類装置。
【請求項６】
上記分割された各シーンを当該シーンに含まれる画像の特徴に基づき、各シーンを分類する画像ベース分類手段を備えたことを特徴とする請求項１～５の何れか１項に記載の映像分類装置。
【請求項７】
同一の映像信号源から得られる各シーンに、同一の識別情報を付与する映像源識別情報付与手段を備えたことを特徴とする請求項５または６に記載の映像分類装置。
【請求項８】
分類されたシーン毎にまとめて表示を行う表示手段を備えたことを特徴とする請求項５～７の何れか１項に記載の映像分類装置。
【請求項９】
上記表示手段は、分類されたシーンを類似したもの同士をかためて近距離に配置する表示と、類似したもの同士を列毎に配置する表示とで、切り替え可能に表示することを特徴とする請求項８に記載の映像分類装置。
【請求項１０】
映像間の類似度を判定する映像間類似度判定手段を備えたことを特徴とする請求項７に記載の映像分類装置。
【請求項１１】
映像信号に含まれるビデオ信号に基づきショット間の境界を検出して映像を各ショットに分割するショット分割ステップと、
分割されたショット毎に、ショット内のオーディオ信号について、音の種類で分類された各クラスにどの程度属しているかを示す帰属確率を算出し、この帰属確率を用いて隣接するショット間の類似度を判定するショット間類似度判定ステップと、
判定されたショット間類似度が所定値より高いショット同士は統合させ、上記映像を各

シーンに分割するシーン分割ステップと、
を含むことを特徴とする映像分類方法。
【請求項１２】
請求項１～９の何れか１項に記載の映像分類装置を動作させるための制御プログラムであって、コンピュータを上記映像分類装置における各手段として機能させるための映像分類プログラム。
【請求項１３】
請求項１２に記載の映像分類プログラムが記録されているコンピュータ読み取り可能な記録媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、映像信号をシーン毎に分類する映像分類装置、映像分類方法、映像分類プログラムおよびコンピュータ読取可能な記録媒体に関するものである。
【背景技術】
【０００２】
近年、地上波デジタル放送や光ファイバーによる高速通信網を介した映像配信が開始され、また、Blu-ray DiskやHD DVD（High Definition Digital Versatile Disk）等の大容量の記録媒体が出現している。これらのことからも、ユーザが保持する映像コンテンツは急速に増加することが容易に予測される。このような状況において、蓄積された映像コンテンツの中から所望の映像を得るためのツールとして、映像信号の検索システムが必要となる。このような映像信号の検索システムを構築する場合、前処理として映像信号を分割し、内容を表すインデックスを付加する必要がある。
【０００３】
ここで、図１３に示すように映像信号は一般に１台のカメラで連続的に撮影された区間であるショット、及び内容に関連のあるショットを統合したシーンにより構成される（例えば、非特許文献１参照）。そのため映像信号は、ショットあるいはシーンが切り換わる時刻を境界として分割することが望ましい。
【０００４】
以上のような背景のもと、映像信号より得られるビデオ信号を用いて隣接するショット間の境界（以降、ショットカットと呼ぶ）を検出する手法が提案されている（例えば、非特許文献２～５参照）。これらの手法では、ショットカットの前後で画像の輝度値や動きベクトルが急激に変化するという特徴に基づき、ショットカットを検出する。
【０００５】
他方、隣接するシーン間の境界（以降、シーンカットと呼ぶ）は、ショットカットの一部として検出される。しかしながらショットカットとシーンカットとの両者において、輝度値や動きベクトルの変化の様子に明確な差異は存在せず、ビデオ信号を処理しただけではショットカットの中からシーンカットを検出することは困難となる。そこでシーンカットでは、ビデオ信号だけでなくオーディオ信号も同時に切り換わることに着眼し、ビデオ信号とオーディオ信号とを併せて使用することで、シーンカットを検出する手法が提案されている（例えば、非特許文献６～１０参照）。これらの手法では，オーディオ信号を音声や音楽等のクラスに分類し、分類されたクラスが切り換わる時刻とショットカットが一致した場合、その時刻をシーンカットであると判断する。
【非特許文献１】長谷山美紀，「ユーザが望む映像を提供するために－画像認識とクラスタリングそして意味理解への発展－」，映像情報メディア学会技術報告，vol.29，no.47，pp.49-52，2005.
【非特許文献２】S.J.F. Guimaraes et al., Video segmentation based on 2D image analysis, Pattern Recognition Letters, vol.24, no.7, pp.947-957, 2003.
【非特許文献３】鈴木賢一郎，中嶋正臣，坂野鋭，三部靖夫，大塚作一，「動き方向ヒストグラム特徴を用いた映像データからのカット点検出法」，情報通信学会論文誌（D-II）

, vol.J-86-D-II, no.4, pp.468-478, 2003.
【非特許文献４】中島康之, 氏原清乃, 米山暁夫, 「部分復号を用いた MPEG データからのカット点検出」, 情報通信学会論文誌（D-II）, vol.J81-D-II, no.7, pp.1564-1575, 1998.
【非特許文献５】長坂晃朗, 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」, 情報処理学会論文誌, vol.33, no.4, pp.543-550, 1992.
【非特許文献６】T. Zhang and C.-C. J. Kuo, Audio content analysis for online audiovisual data segmentation and classification, IEEE Transactions on Speech and Audio Processing, vol.9, no.4, pp.441-457, 2001.
【非特許文献７】R. Wang, Z. Liu, and J. Huang, Multimedia content analysis using both audio and visual clues, IEEE Signal Process Mag., vol.17, no.6, pp.12-36, 2000.
【非特許文献８】Z. Liu and Y. Wang, Audio feature extraction and analysis for scene segmentation and classification, J. VLSI Signal Process., vol.20, pp.61-79, 1998.
【非特許文献９】C. Saraceno and R. Leonardi, Audio as a support to scene change detection and characterization of video sequences, Proc. Int. Conf. Acoustics, Speech, and Signal Processing, vol.4, pp. 2597--2600, 1997.
【非特許文献１０】中島康之, 陸洋, 菅野勝, 柳原広昌, 米山暁夫, 「MPEG 符号化データからのオーディオインデキシング」, 情報通信学会論文誌（D-II）,vol.J83-D-II, no.5, pp.1361-1371, 2000.
【非特許文献１１】G.F. Hughes, On the mean accuracy of statistical pattern recognizers, IEEE Trans. Information Theory, vol.IT-14, no.1, pp.55--63, 1968.

【発明の開示】

【発明が解決しようとする課題】

【０００６】

しかしながら、従来のシーンカット検出手法は, 隣接するシーンのオーディオ信号が, 同一のクラスである場合、シーンカットの未検出が発生する危険性がある。このような状況は、例えばニュース番組において、男性があるニュースを読み上げ、その後別のニュースを読み上げる際に生じる。この場合、話題が変化しており、シーンカットが存在するが、どちらのオーディオ信号も音声のクラスに分類されるため、両者の境界はショットカットと判別され、その結果、シーンカットが得られないことになる。あるいは、例えば、男性が会話をしている場面から女性が話す場面に切り換わるとする。この場合、話者が男性から女性に変わる時刻にシーンカットが存在するが、どちらのオーディオ信号も音声のクラスに分類されるため、両者の境界はショットカットと判別され、その結果、シーンカットが得られないことになる。

【０００７】

これらのように従来の技術では、異なるシーンであるにも関わらず、映像処理を用いても、音響信号処理を用いても、どちらのオーディオ信号も音声のクラスに分類されるため、両者の境界はショットカットと判別され、その結果、シーンカットが得られない。しかしながら、ユーザにとっては、多数のショットカットよりもシーンカットが重要であり、従来の技術では、大容量メディアの到来を前に、魅力的な映像シーンの提供は不可能である。

【０００８】

そこで、本発明は、上記の問題点に鑑みてなされたものであり、その目的は、効果的なシーン間の境界を検出することのできる、映像分類装置、映像分類方法、映像分類プログラムおよびコンピュータ読取可能な記録媒体を実現することにある。また、本発明は、上記問題を解決する技術と、その技術により得られるシーンの効果的なユーザへの提示システムの実現を目的とする。

【課題を解決するための手段】

【０００９】
本願発明者等は、上記課題を解決するために、鋭意検討し、オーディオ信号から算出されたボリュームや零交差率等の特徴量に主成分分析（Principal Component Analysis：ＰＣＡ）を適用することで、分類に有効である主成分を得、その後に、得られた主成分にファジィｃ－ｍｅａｎｓ法（Fuzzy c-Means：ＦＣＭ）を適用し、その結果算出される帰属度を用いることで、処理対象であるオーディオ信号が音声や音楽等の各クラスに属する度合を定量化し、その値を用いてインデックスを付加した。さらに、オーディオインデキシング結果と、ビデオ信号より得られるショットカットを組み合わせ、隣接するショット間の類似度を定義した。これらのことにより、従来手法の問題を解決し、高精度なインデキシングを実現できることを見いだし、本発明を完成させるに至った。

【００１０】
本発明に係る映像分類装置は、上記課題を解決するために、映像信号に含まれるビデオ信号に基づきショット間の境界を検出して映像を各ショットに分割するショット分割手段と、分割されたショット毎に、ショット内のオーディオ信号について、音の種類で分類された各クラスにどの程度属しているかを示す帰属確率を算出し、この帰属確率を用いて隣接するショット間の類似度を判定するショット間類似度判定手段と、判定されたショット間類似度が所定値より高いショット同士は統合させ、上記映像を各シーンに分割するシーン分割手段と、を備えたことを特徴としている。

【００１１】
また、本発明に係る映像分類方法は、上記課題を解決するために、映像信号に含まれるビデオ信号に基づきショット間の境界を検出して映像を各ショットに分割するショット分割ステップと、分割されたショット毎に、ショット内のオーディオ信号について、音の種類で分類された各クラスにどの程度属しているかを示す帰属確率を算出し、この帰属確率を用いて隣接するショット間の類似度を判定するショット間類似度判定ステップと、判定されたショット間類似度が所定値より高いショット同士は統合させ上記映像を各シーンに分割するシーン分割ステップと、を含むことを特徴としている。

【００１２】
上記構成および方法によると、まず、映像信号に含まれるビデオ信号に基づきショット間の境界を検出して映像を各ショットに分割するショット分割する。そして、この分割されたショット毎に、ショット内のオーディオ信号について、音の種類で分類された各クラスにどの程度属しているかを示す帰属確率を算出する。そして、この帰属確率を用いて隣接するショット間の類似度を判定するショット間類似度判定し、ショット間類似度が所定値より高いショット同士は統合させて上記映像を各シーンに分割する。よって、オーディオ信号の帰属確率からショット間の類似の度合を算出するので、類似したショットをシーンとしてまとめることができ、その結果、効果的なシーン間の境界を検出することができる。なお、映像信号は、音響の信号であるオーディオ信号と、画像の信号であるビデオ信号とを含むものとする。

【００１３】
ここで、映像（映像信号）において同一の話者で、短時間の無音が存在する場合には、上記構成および方法、従来技術、共に、無音を検出することで、シーンカット（シーン間の境界）を得ることが可能である。また、映像において同一の話者で、短時間の無音が存在しない場合には、上記構成および方法、従来技術、共に、シーンカットの検出は困難となる。また、映像において複数の話者で、短時間の無音が存在する場合では、上記構成および方法、従来技術、共に、無音を検出することで、シーンカットを得ることが可能である。また、映像において複数の話者で、短時間の無音が存在しない場合では、上記構成および方法はシーンカットの検出が可能であるが、従来技術では検出が困難となる。ただし、実際に話題が変化しているにも関わらず、同一の話者で、短時間の無音が存在しない場合が発生することは稀であると考えられるので、本発明に係る上記構成および方法は、高精度なシーンの分割（シーンカットの検出）が可能であるということができる。

【００１４】

また、上記構成および方法によると、従来技術において雑多に用いられてきたパラメータ（特徴量）の中から、分類対象の映像がどのジャンル（ドラマ、音楽番組、ニュースなど）に含まれるかを与えれば、自動的に有効な特徴量を選択し、そのジャンルに適したインデキシングを実現することができる。

【００１５】

また、本発明に係る映像分類装置では、上記構成に加え、上記音の種類とは、無音、音声、音楽、音楽付き音声、雑音付き音声の５つの種類であってもよい。これら５種類は、日常によくある音の種類であり、これらのクラスにどの程度属しているかを示す帰属確率を求めるので、的確に映像（映像信号）についてシーン分割を行うことができる。もちろん、これ以上の種類、これら以外の種類に分けてもかまわない。

【００１６】

なお、本発明に係る映像分類装置では、上記ショット間類似度判定手段は、ショット内のオーディオ信号を分割した各クリップについて上記帰属確率を算出し、ショット内の各クリップの帰属確率の累積ヒストグラムを基に、隣接するショット間の類似度を判定する。

【００１７】

また、本発明に係る映像分類装置は、上記構成に加え、上記帰属確率の累積ヒストグラムのうち、最大の値のクラスを示す識別情報を処理対象のショットに付与するクラス識別情報付与手段を備えていてもよい。

【００１８】

上記構成によると、帰属確率の累積ヒストグラムのうち、最大の値のクラスを示す識別情報が処理対象のショットに付与される。そして、本発明に係る映像分類装置は、上記構成に加え、上記分割された各シーンに含まれるショットに付与された上記識別情報に基づき、各シーンを分類する音響ベース分類手段を備えていてもよい。

【００１９】

上記構成によると、分割された各シーンに含まれるショットに付与された上記識別情報に基づいて、各シーンを分類することができる。よって、各シーンを、音響に基づいて的確に分類することができる。

【００２０】

また、本発明に係る映像分類装置は、上記構成に加え、上記分割された各シーンを当該シーンに含まれる画像の特徴に基づき、各シーンを分類する画像ベース分類手段を備えていてもよい。

【００２１】

上記構成によると、分割された各シーンを当該シーンに含まれる画像の特徴に基づいて、各シーンが分類される。よって、ユーザが視認したときに確認が行いやすくなり、ユーザによって利便性の高い表示を行うことができる。

【００２２】

また、本発明に係る映像分類装置は、上記構成に加え、同一の映像信号源から得られる各シーンに、同一の識別情報を付与する映像源識別情報付与手段を備えていてもよい。

【００２３】

上記構成によると、同一の映像信号源から得られる各シーンには、同一の識別情報が付与される。よって、付与された識別情報毎にシーンが分類されるように、例えば同一の識別情報のシーンには同一の色の網がけを行って表示した場合に、ユーザは、どのシーンが同じ映像源からのものであるかを容易に確認することができる。

【００２４】

また、本発明に係る映像分類装置は、上記構成に加え、上記分類されたシーン毎にまとめて表示を行う表示手段を備えていてもよい。

【００２５】

上記構成によると、表示手段により、上記分類されたシーン毎にまとめて表示が行われる。よって、ユーザは、どのシーンがどのように分類されているのかを、容易に把握する

ことができる。

【００２６】

また、本発明に係る映像分類装置では、上記構成に加え、上記表示手段は、分類されたシーンを類似したもの同士をかためて近距離に配置する表示と、類似したもの同士を列毎に配置する表示とで、切り替え可能に表示するようになっていてもよい。

【００２７】

上記構成によると、分類されたシーンを類似したもの同士をかためて近距離に配置する表示と、類似したもの同士を列毎に配置する表示とで、切り替え可能に表示することができるので、ユーザの好みに応じて切り替えることができる。ユーザは、分類されたシーンを類似したもの同士をかためて近距離に配置された表示では、類似性が高いことを直感的に把握することができる。ユーザは、類似したもの同士を列毎に配置する表示では、系統立てて把握することができる。

【００２８】

また、本発明に係る映像分類装置は、上記構成に加え、映像間の類似度を判定する映像間類似度判定手段を備えていてもよい。

【００２９】

上記構成によると、映像間（映像信号間）の類似度を判定することができ、映像（映像信号）の分類を的確に行うことができる。

【００３０】

ところで、上記映像分類装置は、ハードウェアで実現してもよいし、プログラムをコンピュータに実行させることによって実現してもよい。具体的には、本発明に係るプログラムは、上記いずれかの構成の映像分類装置の各手段としてコンピュータを動作させるプログラムであり、本発明に係るコンピュータ読み取り可能な記録媒体には、当該プログラムが記録されている。

【００３１】

このプログラムがコンピュータによって実行されると、当該コンピュータは、上記映像分類装置として動作する。したがって、上記映像分類装置と同様に、効果的なシーンカット検出しシーンを分類することができる。

【００３２】

なお、本発明は、従来技術の問題を解決する技術と、その技術により得られるシーンの効果的なユーザへの提示システムを含むものである。

【発明の効果】

【００３３】

本発明に係る映像分類装置は、以上のように、映像信号に含まれるビデオ信号に基づきショット間の境界を検出して映像を各ショットに分割するショット分割手段と、分割されたショット毎に、ショット内のオーディオ信号について、音の種類で分類された各クラスにどの程度属しているかを示す帰属確率を算出し、この帰属確率を用いて隣接するショット間の類似度を判定するショット間類似度判定手段と、判定されたショット間類似度が所定値より高いショット同士は統合させ、上記映像を各シーンに分割するシーン分割手段と、を備えている。

【００３４】

上記構成によると、オーディオ信号の帰属確率からショット間の類似の度合を算出するので、類似したショットをシーンとしてまとめることができ、その結果、効果的なシーン間の境界を検出することができる。

【００３５】

ここで、映像（映像信号）において同一の話者で、短時間の無音が存在する場合には、上記構成および方法、従来技術、共に、無音を検出することで、シーンカット（シーン間の境界）を得ることが可能である。また、映像において同一の話者で、短時間の無音が存在しない場合には、上記構成および方法、従来技術、共に、シーンカットの検出は困難となる。また、映像において複数の話者で、短時間の無音が存在する場合では、上記構成お

よび方法、従来技術、共に、無音を検出することで、シーンカットを得ることが可能である。また、映像において複数の話者で、短時間の無音が存在しない場合では、上記構成および方法はシーンカットの検出が可能であるが、従来技術では検出が困難となる。ただし、実際に話題が変化しているにも関わらず、同一の話者で、短時間の無音が存在しない場合が発生することは稀であると考えられるので、本発明に係る上記構成および方法は、高精度なシーンの分割（シーンカットの検出）が可能であるということができる。

【発明を実施するための最良の形態】

【００３６】

本発明の一実施形態について図１～図１１に基づいて説明すると以下の通りである。図１に示すように、本実施の形態の映像分類装置１は、ショット分割部（ショット分割手段）２、ショット間類似度判定部（ショット間類似度判定手段）３、シーン分割部（シーン分割手段）４、音響ベース分類部（音響ベース分類手段）６と画像ベース分類部（画像ベース分類手段）７とを備えたシーン分類部５、識別情報付与部（クラス識別情報付与手段、映像源識別情報付与手段）８、映像間類似度判定部（映像間類似度判手段）９、表示部（表示手段）１０、入力部１１を備えている。

【００３７】

ショット分割部２は、映像信号に含まれるビデオ信号に基づきショット間の境界を検出して映像を各ショットに分割する。ここで映像信号は、音響の信号であるオーディオ信号と、画像の信号であるビデオ（ビジュアル）信号とを含むものである。

【００３８】

ショット間類似度判定部３は、分割されたショット毎に、ショット内のオーディオ信号について、音の種類で分類された各クラスにどの程度属しているかを示す帰属確率を算出し、この帰属確率を用いて隣接するショット間の類似度を判定する。具体的には以下で説明するが、ショット内のオーディオ信号を分割した各クリップについて、音の種類で分類された各クラスにどの程度属しているかを示す帰属確率（後段で説明する）を算出し、ショット内の各クリップの帰属確率の累積ヒストグラムを基に、隣接するショット間の類似度を判定する。なお、上記音の種類とは、本実施形態では、無音、音声、音楽、音楽付き音声、雑音付き音声の５つの種類とするが、これ以外であってもよい。

【００３９】

シーン分割部４は、判定されたショット間類似度が所定値より高いショット同士は統合させ、上記映像を各シーンに分割する。

【００４０】

シーン分類部５は、分割されたシーンを分類するものであり、音響ベース分類部６と画像ベース分類部７とを備えている。音響ベース分類部６は、分割された各シーンに含まれるショットに付与された下記識別情報に基づき、各シーンを分類する。画像ベース分類部７は、分割された各シーンを当該シーンに含まれる画像の特徴に基づき、各シーンを分類する。

【００４１】

識別情報付与部８は、上記帰属確率の累積ヒストグラムのうち、最大の値のクラスを示す識別情報を処理対象のショットに付与する。また、同一の映像信号源から得られる各シーンに、同一の識別情報を付与する。本実施形態では、クラス識別情報の付与と映像源識別情報の付与とを識別情報付与部８が両方行うものとするが、別々に行うものが設けられていてもよい。

【００４２】

映像間類似度判定部９は、後段で詳しく説明するが映像間の類似度を判定する。

【００４３】

表示部１０は、ユーザにユーザインターフェイスを提供するものであり、各種画像や各種操作ボタン等の表示を行う。表示部１０は、例えば、液晶表示素子等のフラットパネルディスプレイやＣＲＴなどのから構成されている。表示部１０は、分類されたシーンを表示する際、シーン毎にまとめて表示を行う。また、分類されたシーンを類似したもの同士

をかためて近距離に配置する表示と、類似したもの同士を列毎に配置する表示とで、切り替え可能に表示する。

【００４４】

入力部１１は、映像分類装置１に対する操作をユーザが行うための指示信号を入力する入力デバイスである。例えば、テンキーや十字キーなどが設けられたリモコンや、キーボードなどの入力デバイスとして構成してもよいし、表示部１０と一体としたタッチパネルとして実現してもよい。後者の場合、表示部１０に、操作ボタンなどのＧＵＩ画面を表示し、ユーザの指（または、タッチペンなどのポインティングデバイス）により押下されることにより、その位置に対応するボタンが示す指示信号が、映像分類装置１内部に入力される。

【００４５】

以下に本実施形態の映像分類装置１における処理について詳細に説明する。以下では、帰属確率を求める処理、映像を各ショットに分割する処理、分割された各ショットに含まれるクリップの帰属確率を基にショット間類似度を判定し映像をシーンに分割する処理（オーディオビジュアルインデキシング）、分割されたシーンを分類する処理、映像間の類似度を判定する処理、の順に説明する。

【００４６】

（ＰＣＡとＦＣＭとを用いたオーディオインデキシング）

ここでは、ＰＣＡとＦＣＭとを用いて、オーディオ信号が以下の（１）～（５）に定義する５種類のクラスに属する程度（以降、帰属確率と呼ぶ）を算出する。

（１）無音（Silence：Ｓｉ）：準静的な背景音のみを含むオーディオ信号

（２）音声（Speech：Ｓｐ）：会話等の音声を含むオーディオ信号

（３）音楽（Music：Ｍｕ）：楽器の演奏等の音を含むオーディオ信号

（４）音楽付き音声（Speech with Music：ＳｐＭｕ）：背景に音楽が存在する環境下での音声を含むオーディオ信号

（５）雑音付き音声（Speech with Noise：ＳｐＮｏ）：背景に雑音が存在する環境下での音声を含むオーディオ信号

各クラスへの帰属確率は、図２に示すＣＬＳ＃１からＣＬＳ＃４の４つの分類処理を施し、それらの分類結果を用いて算出される。ここで、ＣＬＳ＃１からＣＬＳ＃４までの各分類処理は、全て同一の手順であり、処理対象信号及び２種類の参照信号に対し、「特徴量の算出」、「ＰＣＡの適用」、及び「ＦＣＭの適用」の３つの処理を行う。ただし、表１に示すように、参照信号は分類処理の目的に応じてＳｉ，Ｓｐ，Ｍｕ，ＳｐＭｕ，ＳｐＮｏのいずれか（あるいは複数）のオーディオ信号を含む。

【００４７】

【表１】

	参照信号1	参照信号2
CLS#1	Si	Sp,Mu,SpMu,SpNo
CLS#2	Mu, SpMu	Sp,SpNo
CLS#3	Mu	SpMu
CLS#4	Sp	SpNo

【００４８】

以下では、各特徴量の算出、ＰＣＡの適用、ＦＣＭの適用について説明し、その後ＣＬＳ＃１～＃４の分類結果を用いた帰属確率の算出法について説明する。

【００４９】

（特徴量の算出）

まず、処理対象であるオーディオ信号、及び表１に示した２種類の参照信号から、特徴

量を算出する。ここで、特徴量は、フレーム（フレーム長：W_f）とクリップ（クリップ長：W_c）と呼ばれる、大きさの異なる２種類の分析窓を用いて算出される。ただし、図３に示すように、フレームの大きさは、クリップに比べて、十分に小さいものとする。また、フレーム及びクリップの移動幅Δは、$W_f > \Delta$を満たすものとする。図３の点線で示しているように、クリップ１には、フレーム１，２，３，…，Ｎが含まれる。また、図には記載されていませんが、フレーム及びクリップは移動幅Δ（＝フレーム長の半分）で移動することから、クリップ２にはフレーム２，３，４，…，Ｎ＋１が、クリップ3にはフレーム３，４，５，…，Ｎ＋２が含まれることとなる。通常、フレームやクリップは、隣接するフレーム／クリップが重なるように移動させる。これは、隣接するフレームが重なりを許すことで、ハニング窓やハミング窓等の窓関数を用いて切り出された信号を、元の信号に復元できることに起因している。多くの場合、この窓の移動幅は窓長の１／２が使用されるため、本実施形態でもフレーム長の半分に設定している。しかし、これに限定されることはない。

【００５０】

以下では、フレーム単位で算出する特徴量、及びクリップ単位で算出する特徴量について説明する。

【００５１】

初めに、フレーム単位で算出される特徴量について説明する。フレーム単位で算出される特徴量は、以下に示す９種類である。

【００５２】

・ボリューム：ｎ番目のフレームにおけるボリュームＶＯ（ｎ）を次式で定義する。

【００５３】

【数１】

$$VO(n) = \sqrt{\frac{1}{W_f}\sum_{i=0}^{W_f-1}\{s_n(i)\}^2} \qquad (1)$$

【００５４】

ただし、s_n（ｉ）はｎ番目のフレームにおけるｉ番目のサンプルを表す。

【００５５】

・零交差率：ｎ番目のフレームにおける零交差率ＺＣ（ｎ）を次式で定義する。

【００５６】

【数２】

$$ZC(n) = \frac{1}{2}\sum_{i=0}^{W_f-1}\left|sign\{s_n(i)\} - sign\{s_n(i-1)\}\right| \quad (2)$$

【００５７】

ただし、ｓｉｎｇ{・}は、以下で定義される関数である。

【００５８】

【数３】

$$sign\{s_n(i)\} = \begin{cases} 1 & (s_n(i) \geqq 0) \\ -1 & (s_n(i) < 0) \end{cases} \qquad (3)$$

【００５９】

・ピッチ：ｎ番目のフレームにおけるピッチをＰＴ（ｎ）で表す。ピッチの推定方法について、従来より様々な手法が提案されているので何れかを採用すればよいが、本実施形態では、非特許文献８で提案されている推定手法を採用する。この手法は、以下の式（４）で定義される Average Magnitude Difference Function (AMDF) γを算出し、γ（ｌ）の極小値のうち、最もｌの小さな値を検出することで、ピッチの推定を実現する。

【００６０】
【数４】

$$\gamma(l) = \frac{\sum_{i=0}^{W_f - l - 1} |s_n(i+l) - s_n(i)|}{W_f - l} \qquad (4)$$

【００６１】
ただし、非特許文献８では、音声のピッチのみを得るため、音声のピッチが存在する周波数帯（４０－４５０Ｈｚ）のみを処理対象とし、上記周波数帯にピッチが存在しない場合は、ＰＴ（ｎ）＝０とする。
【００６２】
・周波数中心位置：ｎ番目のフレームにおける周波数中心位置ＦＣ（ｎ）を次式で定義する。
【００６３】
【数５】

$$FC(n) = \sqrt{\frac{\int_0^{\pi} \omega |S_n(\omega)|^2 d\omega}{\int_0^{\pi} |S_n(\omega)|^2 d\omega}} \qquad (5)$$

【００６４】
ただし、S_n（ω）は、ｎ番目のフレームにおける短時間フーリエ変換を表す。
【００６５】
・周波数帯域幅：ｎ番目のフレームにおける周波数帯域幅ＦＢ（ｎ）を次式で定義する。
【００６６】
【数６】

$$FB(n) = \sqrt{\frac{\int_0^{\pi} \{\omega - FC(n)\}^2 |S_n(\omega)|^2 d\omega}{\int_0^{\pi} |S_n(\omega)|^2 d\omega}} \qquad (6)$$

【００６７】
・サブバンドエネルギー比率：非特許文献８に記載されている４種類の周波数帯（０－６３０Ｈｚ、６３０－１７２０Ｈｚ、１７２０－４４００Ｈｚ、４４００－１１０２５Ｈｚ）における、全周波数帯に対するエネルギーの割合をサブバンドエネルギー比率と定義し、それぞれSER_1（ｎ）、SER_2（ｎ）、SER_3（ｎ）、SER_4（ｎ）で表す。
【００６８】
次に、クリップ単位で算出される特徴量について説明する。クリップ単位の特徴量としては、以下に示す非無音率、及び零比率を使用する。
【００６９】
・非無音率：クリップ内において、無音であるフレームを１、無音以外であるフレームを０としたときの、０の割合を非無音率と定義する。ただし、閾値Th_{vo}を設定し、次の式（７）を満たすフレームを無音と判断する。
【００７０】
【数７】

$$VO(n) < Th_{vo} \qquad (7)$$

【００７１】
・零比率：同一の周波数帯に一定時間連続してパワースペクトルの極大値が存在する場合を１、それ以外を０とし、クリップ内の０の割合を零比率と定義する（非特許文献６参照）。
【００７２】
さらに、上記で得たフレーム単位の特徴量の、クリップ内での平均値、及び標準偏差を

算出し、それらをクリップ単位の特徴量とする。

【００７３】

（ＰＣＡの適用）

次に、処理対象信号のクリップから算出された特徴量、及び２種類の参照信号から算出されたクリップ単位の特徴量（参照信号のクリップ数は共にN_cとする）を正規化し、ＰＣＡを施す。ＰＣＡを施すことで、相関の高い特徴量間の影響を軽減することが可能となる。また、ＰＣＡより得られた主成分のうち、その固有値が１以上であるものを下記で説明するＦＣＭに使用することで、計算量の増加やヒューズの現象（有限個の学習パターンあら識別器を設計する際、特徴空間の次元を高くすると識別性能が低下する現象）（非特許文献１１参照）を回避することが可能となる。

【００７４】

（ＦＣＭの適用）

次に、上記ＰＣＡの適用で得られたで主成分に対し、ＦＣＭを施す。

まず、処理対象信号（ｋ＝１）、参照信号１（ｋ＝２，…，N_c＋１）、参照信号２（ｋ＝N_c＋２，…，２N_c＋１）の各クリップから得られた特徴量を用いて、特徴ベクトルf_kを次式で定義する。

【００７５】

【数８】

$$\boldsymbol{f}_k \stackrel{\mathrm{def}}{=} [p_1^k, p_2^k, \cdots, p_M^k]^{\mathrm{T}} \tag{8}$$

【００７６】

ただし、$p^k{}_i$は、クリップｋ（１：処理対象信号，２～N_c＋１：参照信号１，N_c＋２～２N_c＋１:参照信号２）の第ｉ主成分（ｉ＝１，…，Ｍ；Ｍは固有値が１以上の主成分の総数）を表す。また、Ｔは転置を表す。これら２N_c＋１個の特徴ベクトルを２つのクラスタに分類するＦＣＭを適用し、得られる帰属度μ_{ik}（ｉ＝１，２；ｋ＝１，…，２N_c＋１）を観察することで、処理対象信号が参照信号１、参照信号２のどちらに類似した信号であるかを判別することが可能となる。ただし、ｉはクラスタ番号（ｉ＝１，２）、ｋはクリップの番号（ｋ＝１，…，２N_c＋１）を表す。

【００７７】

この帰属度μ_{ik}は、クリップｋがクラスタｉに属する度合を［０，１］の実数で表す。しかしながら、参照信号１（あるいは参照信号２）のクリップがどちらのクラスタに属するかは分からず、処理対象信号の帰属度μ_{il}（ｉ＝１，２）を観察しただけでは、処理対象信号がどちらの参照信号と同一のクラスタに属しているかを知ることはできない。そこで、帰属度μ_{ik}を用いて、$\mu^c{}_i$を以下のように設定する。

【００７８】

【数９】

$$\mu_1^c = \begin{cases} \mu_{11} & (\overline{\mu_{1\theta}} \geqq \overline{\mu_{2\theta}}) \\ \mu_{21} & (\overline{\mu_{1\theta}} < \overline{\mu_{2\theta}}) \end{cases} \tag{9}$$

【００７９】

【数１０Ａ】

$$\mu_2^c = \begin{cases} \mu_{21} & (\overline{\mu_{1\theta}} \geqq \overline{\mu_{2\theta}}) \\ \mu_{11} & (\overline{\mu_{1\theta}} < \overline{\mu_{2\theta}}) \end{cases} \tag{10}$$

$$= 1 - \mu_1^c$$

【００８０】

ただし、ｃ（ｃ＝１，…，４）はＣＬＳ＃１からＣＬＳ＃４の分類処理の番号を表す。

【００８１】

【数１０Ｂ】

$$\overline{\mu_{i\theta}}\ (i=1,2)$$

【００８２】

また、上記数式（１０）における、上記［数１０Ｂ］は、参照信号１のクリップにおけるクラスタｉ(ｉ＝１,２)への帰属度の平均値であり、次式より算出される。

【００８３】

【数１１】

$$\overline{\mu_{i\theta}}=\frac{1}{N_c}\sum_{j=2}^{N_c+1}\mu_{ij} \qquad (11)$$

【００８４】

このようにμ$^{c}{}_{i}$を定義することで、処理対象信号が参照信号１及び参照信号２と同一のクラスタに属する帰属度が、それぞれμ$^{c}{}_{1}$及びμ$^{c}{}_{2}$で表されることとなる。

【００８５】

（帰属確率の算出）

ＣＬＳ＃１からＣＬＳ＃４の各分類処理において、上記で説明した特徴量の算出、ＰＣＡの適用、ＦＣＭの適用の処理を行い、得られたμ$^{c}{}_{i}$（ｉ＝１，２；ｃ＝１，…，４）を用いて、帰属確率を算出する。各クラス(Ｓｉ，Ｓｐ，Ｍｕ，ＳｐＭｕ，ＳｐＮｏ)への帰属確率（P_{Si}，P_{Sp}，P_{Mu}，P_{SpMu}，P_{SpNo}）は、以下で定義される。

【００８６】

【数１２】

$$P_{Si}=\mu_1^1 \qquad (12)$$

【００８７】

【数１３】

$$P_{Sp}=\mu_2^1\mu_2^2\mu_1^4 \qquad (13)$$

【００８８】

【数１４】

$$P_{Mu}=\mu_2^1\mu_1^2\mu_1^3 \qquad (14)$$

【００８９】

【数１５】

$$P_{SpMu}=\mu_2^1\mu_1^2\mu_2^3 \qquad (15)$$

【００９０】

【数１６】

$$P_{SpNo}=\mu_2^1\mu_2^2\mu_2^4 \qquad (16)$$

【００９１】

上式は、ＣＬＳ＃１からＣＬＳ＃４の各分類結果において、μ$^{c}{}_{i}$（ｉ＝１，２）を、参照信号１、２と同一のクラスタに分類される確率とみなし、それらを積算することで、Ｓｉ，Ｓｐ，Ｍｕ，ＳｐＭｕ，ＳｐＮｏの各クラスに属する確率を算出することを表す。従って、クリップごとに算出される帰属確率P_{Si}，P_{Sp}，P_{Mu}，P_{SpMu}，P_{SpNo}から、そのクリップがどのクラスにどの程度属しているか知ることが可能となる。また、帰属確率の変動を観察することにより、処理対象であるオーディオ信号がどのように変化するかを知ることも可能となる。

【００９２】

（オーディオビジュアルインデキシング）

上記で説明したオーディオ信号に基づく分類と、ビデオ信号から得られるショットカットを用いた、オーディオビジュアルインデキシング（分類）について説明する。本実施形態では、代表的なショットカット検出法である分割χ^2検定法（非特許文献５参照）を用いてショットカットを検出し、得られたショットカットと、上記で得られたオーディオ信号のインデキシング結果を組み合わせることで、オーディオビジュアルインデキシングを実現する。そこで、以下で分割χ^2検定法によるショットカット検出、及びオーディオビジュアルインデキシングについて説明する。

【００９３】

本実施形態では、非文献特許文献５で提案されている分割χ^2検定法を用いて、ショットカットを得る。この処理はショット分割部２が行う。しかし、ショットカットを得る手法としては、これに限定はされない。なお、ショットカットの精度を向上させるために、フェードやディゾルブ等の特殊効果も検出可能な手法を導入するのがよい。

【００９４】

分割χ^2検定法は、まずフレームを４×４＝１６個の同じ大きさの矩形領域に分割し、各領域毎に６４色種の色ヒストグラムＨｖ（ｆ，ｒ，ｉ）を作成する。ただし、ｆはビデオ信号のフレーム番号、ｒは領域番号、ｉはヒストグラムのビンを表す。隣接する２枚のフレームの色ヒストグラムから、次式で定義される評価値Ｃｒ（ｒ＝１，…，１６）を算出する。

【００９５】

【数１７Ａ】

$$C_r = \sum_{i=0}^{63} \frac{\{H_V(f,r,i) - H_V(f-1,r,i)\}^2}{H_V(f,r,i)} \qquad (17)$$

【００９６】

さらに、算出された１６個の評価値Ｃｒ（ｒ＝１，…，１６）において、Ｃｒの中で値の小さい８つの総和C_{sum}を算出し、C_{sum}が予め設定した以下の［数１７Ｂ］に示す閾値よりも大きな値を示す時刻に、ショットカットが存在すると判断する。以上の処理はショット分割部２が行う。

【００９７】

【数１７Ｂ】

$$Th_{\chi^2}$$

【００９８】

次に、ショット間の類似度を用いたオーディオビジュアルインデキシングについて説明する。この処理はショット間類似度判定部３が行う。上記したオーディオインデキシングは、クリップごとに５種類のクラスへの帰属確率を算出する。そこで、ショット分割部２により得られたショットカットを併せて用いることで、ショット単位でのインデキシングを行う。なお、１つのショットが長時間のものであれば、このショットに含まれるクリップ数も多数になる。

【００９９】

まず、単一のショット内における帰属確率の累積ヒストグラムＨＡ（η，φ）を作成する。ただし、ηはショット番号、φは累積ヒストグラムのビン、すなわちＳｉ（φ＝０），Ｓｐ（φ＝１），Ｍｕ（φ＝２），ＳｐＭｕ（φ＝３），ＳｐＮｏ（φ＝４）を表す。また、累積ヒストグラムの各ビンは、そのショット内におけるクリップの総数で除することにより、正規化されている。この累積ヒストグラムにおいて、最大値を持つビンのクラスを、そのショットのインデックスとする。

【０１００】

各ショットで累積ヒストグラムを定義することで、ショット間の距離を定義することが可能となる。すなわち、ショット間の距離Ｄ（η_1，η_2）を次式で定義する。

【０１０１】

【数１８】

$$D(\eta_1,\eta_2)=\sum_{\phi=0}^{4}|H_A(\eta_1,\phi)-H_A(\eta_2,\phi)| \qquad (18)$$

【０１０２】

この距離Ｄ（η_1，η_2）が予め設定した閾値Th_Dよりも高い値を示す場合、ショット間の類似度は低く、両者は異なるシーンに属すると判断する。逆に、距離Ｄ（η_1，η_2）が閾値Th_Dよりも低い値を示す場合、ショット間の類似度は高く、両者は同一のシーンに属すると判断する。同一のシーンに属すると判断した両者を統合するよう、統合処理を行うことで、シーンを得ることが可能となる。言い換えれば、映像信号をシーン毎に分割する。この処理はシーン分割部４が行う。これにより、従来にはない、ショット間の類似度を考慮したシーンカット検出が可能となり、従来技術の問題点を解決することが可能となる。

【０１０３】

このように、本実施の形態の映像分類装置１では、隣接するショット間の類似度を定義するため、従来技術の問題を解決し、高精度なオーディオビジュアルインデキシングが可能となる。

【０１０４】

（音響信号に基づくシーン分類）

以上のような処理を行うことで、映像信号はシーン単位に分割される。各シーンに無音、音声、音楽、音楽付き音声、雑音付き音声のインデックスが付加されると、付加されたインデックスに基づき、図４の右側に示すような音響に基づくシーン分類が可能となる。ここでの処理は、識別情報付与部８、音響ベースシーン分類部６、表示部１０が主に行う。また、入力部１１からの指示により行われてもよい。

【０１０５】

（画像信号に基づくシーン分類）

次に、上記音響（オーディオ）に基づくシーン分類で得られたシーンを、画像（ビジュアル）の特徴に基づき分類する。ここでの処理は画像ベースシーン分類部７および表示部１０が行う。また、入力部１１からの指示により行われてもよい。分類には、画像から算出されるヒストグラム（色ヒストグラム）を利用する。ただし、ヒストグラムの各ビンは画素数によって正規化されており、画像サイズによる影響はないものとする。

【０１０６】

ここでは、図５に示す２種類のヒストグラムを使用する。一方は、フレーム全体を用いて算出される画像ヒストグラムである。これは、画像全体の特徴を捉えたヒストグラムとなっており、画像の回転に対して頑健である。他方は、フレームを複数のブロックに分割し、各ブロックで算出された複数のヒストグラムである。ブロックに分割することで、フレーム中に存在するオブジェクトの位置等、画像の構造を考慮することが可能となる。

【０１０７】

ここで、画像全体から１つのヒストグラムを作成した場合は、使用されている色の割合が等しいため、例えば、青白赤の帯が縦に並んだ（フランス国旗）画面と、赤白青の帯が横に並んだ（オランダ国旗）画面とを区別することはできない。他方、画像を複数の領域に分割し、同じ位置の領域から算出されたヒストグラムを比較すると、２つは異なる画像であると判断することが可能となる。画像を回転させた場合（番組制作側の映像効果の１つとして想定される）も、上記に例示した２種の国旗の画像のような状況が発生する。これらを区別したくない場合、上記の、フレーム全体を用いて算出される画像ヒストグラムを区別したい場合には、フレームを複数のブロックに分割し、各ブロックで算出された複数のヒストグラムを選択すればよいことになる。本実施の形態の映像分類装置１では、この選択をユーザが設定できるものとする。

【０１０８】

また、オブジェクトの位置を考慮する理由として、次のようなケースが考えられる。白

い背景に１台の青い自動車が表示されている２つの画像があり、この２つの画像は自動車（オブジェクト）の位置だけが異なるものとする。これらを異なる画像として区別したい場合は、オブジェクトの位置を考慮する必要がある。このようなケースでは、上記の後者（他方）の場合のように、画像（フレーム）を複数の領域分割し、同じ位置にある領域から得られたヒストグラムを考慮しなければならない。逆に、どちらも同じ自動車であることから、両者を区別したくない場合は、オブジェクトの位置を考慮しないようにしなければならない。このようなケースでは、上記の前者（一方）の場合のように、画像（フレーム）全体から得た色ヒストグラムを使用する必要がある。

【０１０９】

なお、映像信号は連続しているため、１つのシーンから複数の色ヒストグラムが得られることになる。そのため、映像信号の分割／インデキシング（可視化）に用いる特徴量として、各フレームの色ヒストグラムを全て使用する、シーン内における平均ヒストグラムを算出し使用する等、複数の方法が考えられる。また、色ヒストグラム以外のものを用いて、シーンを分類してもかまわない。

【０１１０】

図５に示す２つのうち、どちらのヒストグラムも、シーン内での平均を算出し、その結果得られる各ビンの値を要素とするベクトルを、そのシーンにおける特徴ベクトルとする。なお、画像（フレーム）を分割した各領域から得られる色ヒストグラム群を使用する場合は、上記特徴ベクトルは、ビンの値を次々と連結する方法で得られる。

【０１１１】

ただし、シーンは、単一のショットで構成される場合と、複数のショットで構成される場合がある。後者の場合は、図６に示すように、各ショット内で特徴ベクトルを生成し、それらを個別に使用する。

【０１１２】

得られた特徴ベクトルに対して、ｋ－ｍｅａｎｓ法を適用することで、画像の特徴が類似したシーン群を得ることができる。これにより、図４左側に示す画像に基づくシーン分類が可能となる。図４左側に示すように、類似した特長ベクトルを有する映像群は近い距離（あるいは、同じクラスタ）に、大きく異なる特徴ベクトルを有する映像群は遠い距離（あるいは、異なるクラスタ）に配置して表示することで、ユーザが映像を検索・選択する労力を軽減することが可能となる。なお、特徴ベクトルの分類の代表的なものとして上記のようにｋ－ｍｅａｎｓ法を挙げたが、ｋ－ｍｅａｎｓ法以外の方法で画像の特徴が類似したシーン群を得てもよい。ｋ－ｍｅａｎｓ法は、特徴ベクトル間のユークリッド距離を算出し、この距離が近いものを１つのクラスタとして分類するものである。「クラスタ」とは、互いが類似した映像であると判断され、１つに分類された塊を指すものとする。

【０１１３】

また、各クラスタの距離を可視化することにより、図７右側に示すように、類似した映像は１つの塊のように近い位置に配置され、異なる映像は遠い位置に配置されるように、表示させることも可能となる。各クラスタの距離を可視化するとは、クラスタの中心間の距離を算出し、距離が短いクラスタを近い位置に、距離が遠いクラスタを遠い位置に配置することで、どの映像が類似しているかを直感的に理解できるように、２次元平面上に可視化することを意味している。なお、図７左側は、入力された画像を示している。

【０１１４】

（複数の映像信号に対する分類）

次に、複数の映像信号に対し、それぞれ上記で説明したシーン分類を行う。ここでの処理は、識別情報府呼部８、画像ベースシーン分類部７、表示部１０が行う。また、入力部１１からの指示により行われてもよい。

【０１１５】

この場合、各シーンに映像信号のソースを示すＩＤを付与することで、異なる映像信号間においても、類似したシーンが同一のクラスタに属する様子を可視化することが可能となる。分類の様子は、上記したものと同様に、例えば、図８に示すようにユーザが所望す

るシーンを目的別に選択する方法と、例えば図９に示すようにシーン間の距離を可視化する方法とがある。これは、図８および９に示す例では、点でハッチングされた入力映像１における各シーンには同じＩＤが付され、斜線でハッチングされた入力映像２における各シーンには同じＩＤ（ただし点でハッチングされた入力画像１のＩＤとは異なる）が付されていることを示している。なお、これらは単なる例示であり、ユーザ入力により、異なる映像信号間の分類の可視化を行えるようになっていてもよい。例えば、ユーザが入力を行うためのボタン等を用意しておき、ボタンを押して、チェックを入れると、同一のソースの映像に、同色の網掛けが施されたり、チェックを外すと網掛けも外れるように表示されるようになっていてもかまわない。

【０１１６】

（複数の映像信号間の類似度の測定）

また、上記のように得られた分類結果に基づき、映像信号間の類似度を測定してもよい。ここでの処理は、映像間類似度判定部９、表示部１０が主に行う。また、入力部１１からの指示により行われてもよい。

【０１１７】

この場合、まず、図１０に示すように、各シーンが分類されるクラスタの帰属度を時系列に並べる。このとき、例えばニュース番組であれば、オープニングミュージック→アンカーショット→レポート→アンカーショット→…のように、番組の構成がある程度定められている。この番組の構成は、図１０に示すように、各クラスタへの帰属度として表現される。従って、この番組の構成を比較することで、異なる映像信号間の類似度を定義することが可能となる。具体的な処理としては、各クラスタにＩＤを割り当て、異なる映像信号間のＩＤに対し、ＤＰマッチングを適用することで、図１１のように類似度を得ることができる。図１１は、単一ショットで構成されたシーンの画像ヒストグラムである。ここでは、ＤＰマッチングを適用しているが、自己組織化マップや、上述のｋ－ｍｅａｎｓ法を利用しても構わない。

【０１１８】

最後に、映像分類装置１の各ブロックは、ハードウェアロジックによって構成してもよいし、次のようにＣＰＵを用いてソフトウェアによって実現してもよい。

【０１１９】

すなわち、映像分類装置１は、各機能を実現する制御プログラムの命令を実行するＣＰＵ（central processing unit）、上記プログラムを格納したＲＯＭ（read only memory）、上記プログラムを展開するＲＡＭ（random access memory）、上記プログラムおよび各種データを格納するメモリ等の記憶装置（記録媒体）などを備えている。そして、本発明の目的は、上述した機能を実現するソフトウェアである映像分類装置１の制御プログラムのプログラムコード（実行形式プログラム、中間コードプログラム、ソースプログラム）をコンピュータで読み取り可能に記録した記録媒体を、上記映像分類装置１に供給し、そのコンピュータ（またはＣＰＵやＭＰＵ）が記録媒体に記録されているプログラムコードを読み出し実行することによっても、達成可能である。

【０１２０】

上記記録媒体としては、例えば、磁気テープやカセットテープ等のテープ系、フロッピー（登録商標）ディスク／ハードディスク等の磁気ディスクやＣＤ－ＲＯＭ／ＭＯ／ＭＤ／ＤＶＤ／ＣＤ－Ｒ等の光ディスクを含むディスク系、ＩＣカード（メモリカードを含む）／光カード等のカード系、あるいはマスクＲＯＭ／ＥＰＲＯＭ／ＥＥＰＲＯＭ／フラッシュＲＯＭ等の半導体メモリ系などを用いることができる。

【０１２１】

また、映像分類装置１を通信ネットワークと接続可能に構成し、上記プログラムコードを通信ネットワークを介して供給してもよい。この通信ネットワークとしては、特に限定されず、例えば、インターネット、イントラネット、エキストラネット、ＬＡＮ、ＩＳＤＮ、ＶＡＮ、ＣＡＴＶ通信網、仮想専用網（virtual private network）、電話回線網、移動体通信網、衛星通信網等が利用可能である。また、通信ネットワークを構成する伝送

媒体としては、特に限定されず、例えば、ＩＥＥＥ１３９４、ＵＳＢ、電力線搬送、ケーブルＴＶ回線、電話線、ＡＤＳＬ回線等の有線でも、ＩｒＤＡやリモコンのような赤外線、Ｂｌｕｅｔｏｏｔｈ（登録商標）、８０２．１１無線、ＨＤＲ、携帯電話網、衛星回線、地上波デジタル網等の無線でも利用可能である。なお、本発明は、上記プログラムコードが電子的な伝送で具現化された、搬送波に埋め込まれたコンピュータデータ信号の形態でも実現され得る。

【０１２２】

〔実施例〕

上記実施の形態で説明した映像分類装置を用いて、映像信号の分類を行った。本実施例では、テレビのニュース番組から得た２種類の映像信号（３２０×２４０pixel、３０fps、４４１００Hz、２０sec）を使用した。また、本実施例で用いたパラメータは、表２に示す値を用いた。

【０１２３】

【表２】

W_f : 2048[sample]	Th_{vol} : 500
W_c : 40960[sample]	Th_{χ^2} : 10000
Δ : 1024[sample]	Th_D : 0.2
N_c : 175[clip]	

【０１２４】

上記実施の形態で説明した映像分類装置では、ピッチや周波数中心位置等の特徴量を使用するため、これらの特徴量の算出には、スペクトル解析が必要であり、通常５０ｍｓ程度の分析窓が使用される。また、計算機上でスペクトル解析を行う場合、通常高速フーリエ変換（ＦＦＴ）を使用するが、このとき分析窓の窓長を２のべき乗に設定する必要がある。そこで、本実施では、フレーム長を２０４８サンプル（サンプリング周波数が４４．１ｋＨｚなので、およそ４６ｍｓとなる）とした。また、クリップ長は、フレームを整数個含み、かつ約１秒となるように設定した。

【０１２５】

各映像信号に対する分類の結果を図１２（ａ），（ｂ）に示す。ただし、図は上から映像コンテンツ、ビデオ信号、オーディオ波形、分割χ^2検定法によるショットカット検出結果、オーディオインデキシング結果、各ショットにおける累積ヒストグラム、ショット間の距離、最終的なインデキシング結果を表している。実施例１（図１２（ａ））で用いた映像信号は、前半に番組のオープニングミュージックが、後半にアナウンサーの音声が存在する構成となっており、両者の境界である９．９秒にシーンカットが存在する。図１２（ａ）からわかるように、ショットカットは正しく検出され、いずれのショットも正しいクラスに分類されていることが確認できる。また、ショット間の距離Ｄ（η_1，η_2）は０．９５と高い値を示しており、両者は異なるシーンに属していることが確認できる。

【０１２６】

また、実施例２（図１２（ｂ））で用いた映像信号は、前半に男性アナウンサーの音声が、後半に女性アナウンサーの音声が存在する構成となっており、両者の境界である９．６秒にシーンカットが存在する。図１２（ｂ）からわかるように、この映像信号は３つのショットに分割され、いずれのショットも音声のクラスに分類されていることが確認できる。

【０１２７】

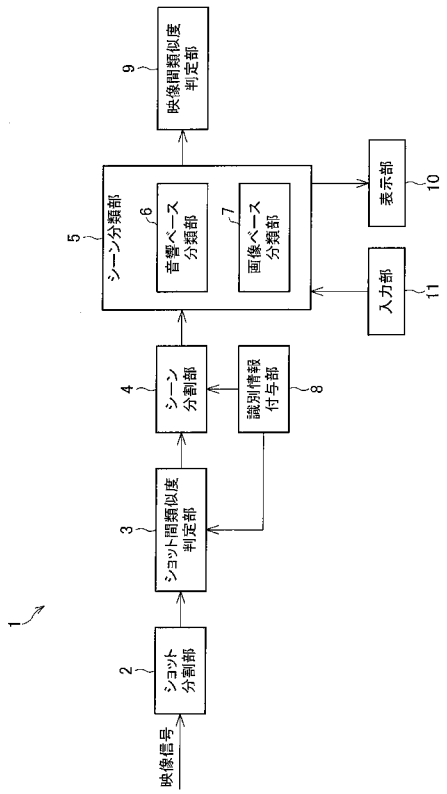
一方、累積ヒストグラムより、ショット＃１とショット＃２間の距離Ｄ（η_1，η_2）は０．２１であるのに対し、ショット＃２とショット＃３間の距離Ｄ（η_2，η_3）は０．０３であり、シーンカットはショット＃１とショット＃２との境界に存在することが分

かる。
【０１２８】
ここで、ショット#３では女性アナウンサーの音声が存在している。ショット#２からショット#３にかけて、女性アナウンサーの音声は連続して存在している（＝話題の変化がない）。図１２（ｂ）に示すように、ショット#２とショット#３は同一のシーンと判断していることから、上記実施形態の映像分類装置はシーンカットを正しく検出していることが分かる。
【０１２９】
以上のことから、上記実施形態の映像分類装置では、効果的にシーンを分類することができることがわかる。
【０１３０】
なお、映像（映像信号）において同一の話者で、短時間の無音が存在する場合には、上記実施形態の映像分類装置、従来技術、共に、無音を検出することで、シーンカットを得ることが可能である。また、映像において同一の話者で、短時間の無音が存在しない場合には、上記実施形態の映像分類装置、従来技術、共に、シーンカットの検出は困難となる。また、映像において複数の話者で、短時間の無音が存在する場合では、上記実施形態の映像分類装置、従来技術、共に、無音を検出することで、シーンカットを得ることが可能である。また、映像において複数の話者で、短時間の無音が存在しない場合では、上記実施形態の映像分類装置はシーンカットの検出が可能であるが、従来技術では検出が困難となる。ただし、実際に話題が変化しているにも関わらず、同一の話者で、短時間の無音が存在しない場合が発生することは稀であると考えられるので、上記実施形態の映像分類装置は、高精度なシーンカット検出が可能であると言える。なお、商品として魅力的なユーザインターフェイスがあると好ましい。
【０１３１】
本発明は上述した実施形態および実施例に限定されるものではなく、請求項に示した範囲で種々の変更が可能である。すなわち、請求項に示した範囲で適宜変更した技術的手段を組み合わせて得られる実施形態についても本発明の技術的範囲に含まれる。
【産業上の利用可能性】
【０１３２】
本発明によると、映像を画像の類似度に基づいてクラスタリングすることができるので、デジタル画像の中から所望のシーンを選択するユーティリティソフトおよび、各シーンの提示装置の実現に利用することができる。
【図面の簡単な説明】
【０１３３】
【図１】本発明の実施形態を示すものであり、映像分類装置の要部構成を示すブロック図である。
【図２】オーディオインデキシングの処理の概要を示す図である。
【図３】オーディオ信号をフレーム及びクリップへ分解することを示す図である。
【図４】シーンを分類した図である。
【図５】単一のショットで構成されたシーンの画像ヒストグラムを表す図である。
【図６】複数のショットで構成されたシーンの画像ヒストグラムを表す図である。
【図７】クラスタの距離を可視化してシーンを分類した図である。
【図８】シーンを目的別に分類した図である。
【図９】シーン間の距離を可視化した図である。
【図１０】各シーンが分類されるクラスタの帰属度を時系列に並べらべた図である。
【図１１】単一ショットで構成されたシーン画像のヒストグラムを表す図である。
【図１２】（ａ）は一実施例の結果を示す図であり、（ｂ）他の実施例の結果を示す図である。
【図１３】映像信号の階層構造を示す図である。
【符号の説明】

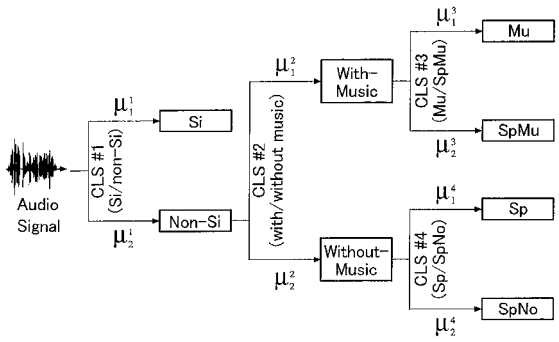
【０１３４】

１　映像分類装置
２　ショット分割部（ショット分割手段）
３　ショット間類似度判定部（ショット間類似度判定手段）
４　シーン分割部（シーン分割手段）
５　シーン分類部
６　音響ベース分類部（音響ベース分類手段）
７　画像ベース分類部（画像ベース分類手段）
８　識別情報付与部（クラス識別情報付与手段、映像源識別情報付与手段）
９　映像間類似度判定部（映像間類似度判手段）
１０　表示部（表示手段）
１１　入力部

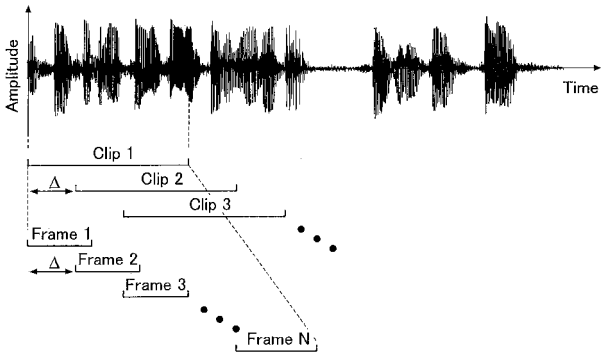
【図１】



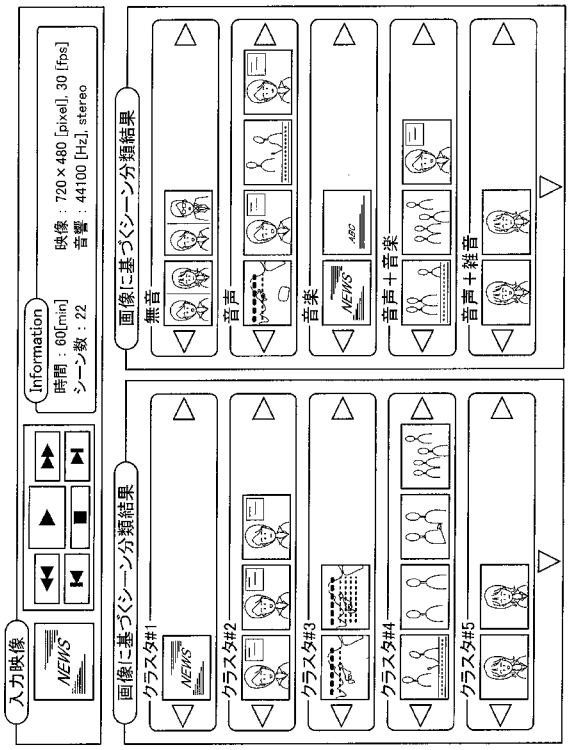
【図２】



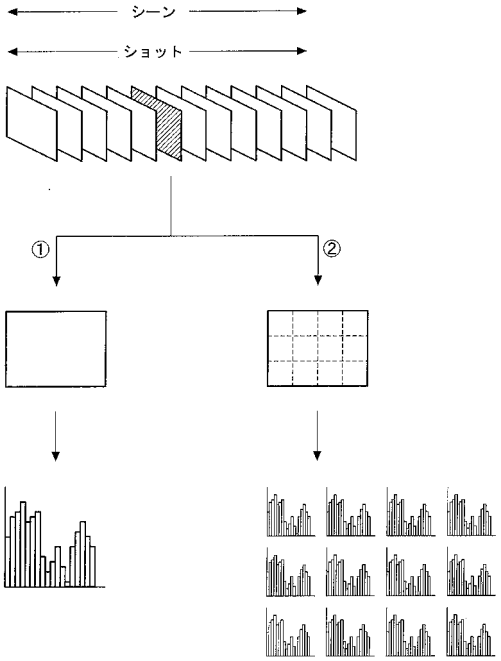
【図３】



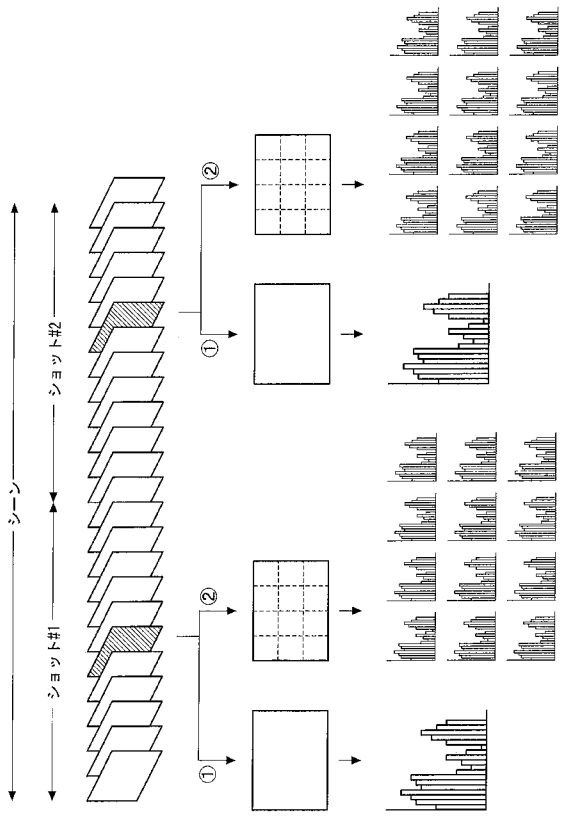
【図４】


入力映像
Information
時間：60[min]
シーン数：22
映像：720×480 [pixel], 30 [fps]
音響：44100 [Hz], stereo
画像に基づくシーン分類結果
無音
音声
音楽
音声＋音楽
音声＋雑音
画像に基づくシーン分類結果
クラスタ#1
クラスタ#2
クラスタ#3
クラスタ#4
クラスタ#5

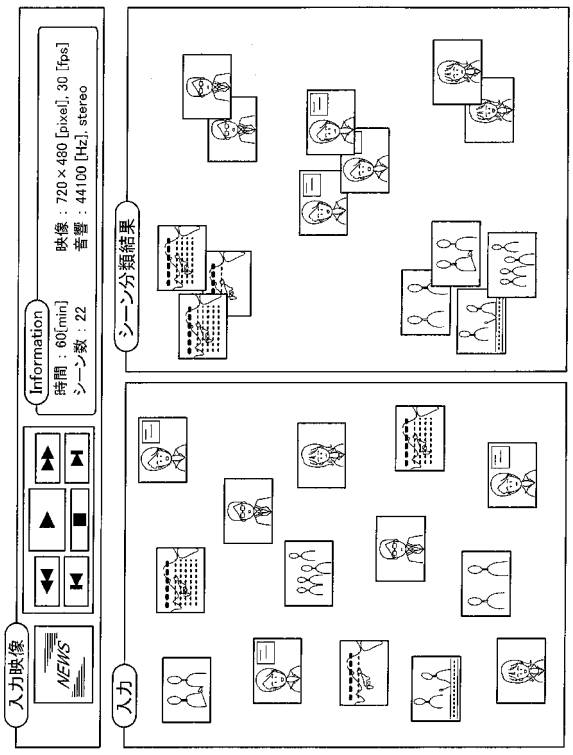
【図５】


シーン
ショット
①
②

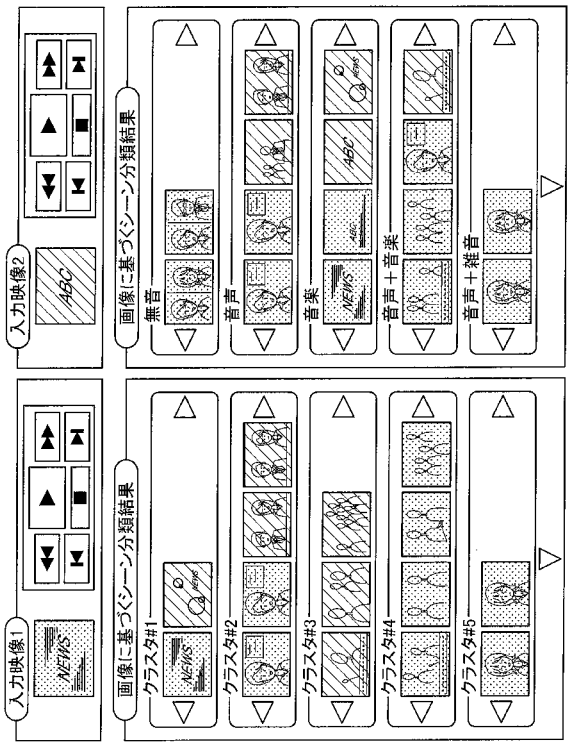
【図６】


シーン
ショット#1
ショット#2
①
②
①
②

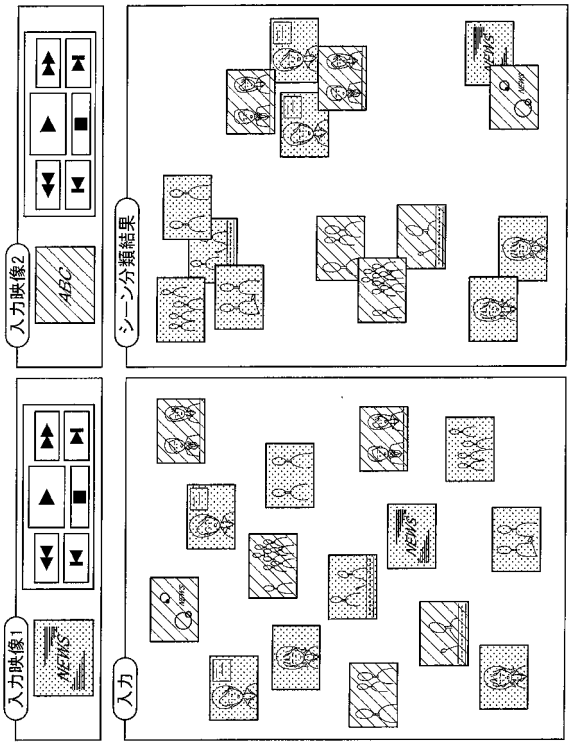
【図７】


入力映像
Information
時間：60[min]
シーン数：22
映像：720×480 [pixel], 30 [fps]
音響：44100 [Hz], stereo
シーン分類結果
入力

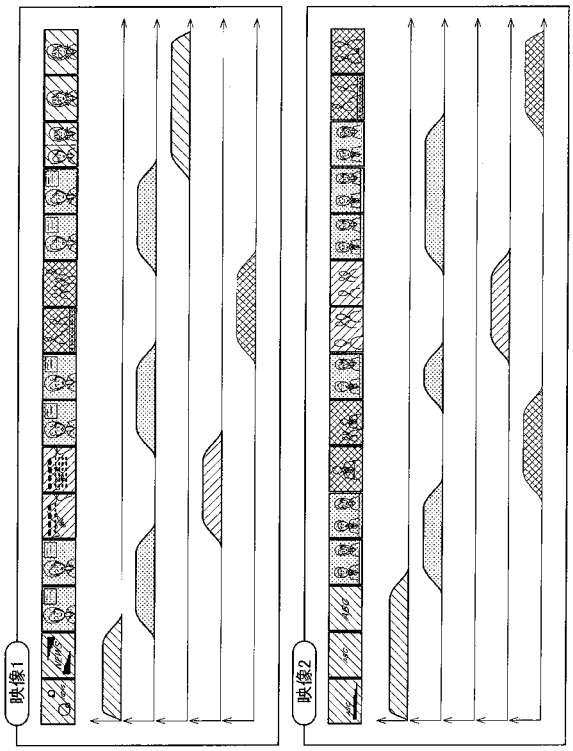
【図８】


入力映像2
画像に基づくシーン分類結果
無音
音声
音楽
音声＋音楽
音声＋雑音
入力映像1
画像に基づくシーン分類結果
クラスタ#1
クラスタ#2
クラスタ#3
クラスタ#4
クラスタ#5

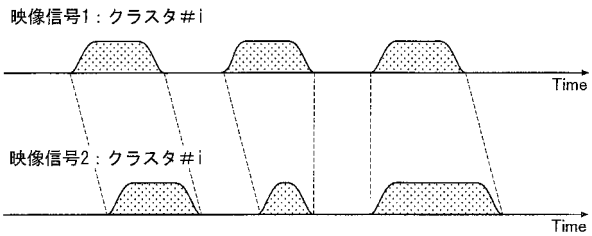
【図９】


入力映像2
シーン分類結果
入力映像1
入力

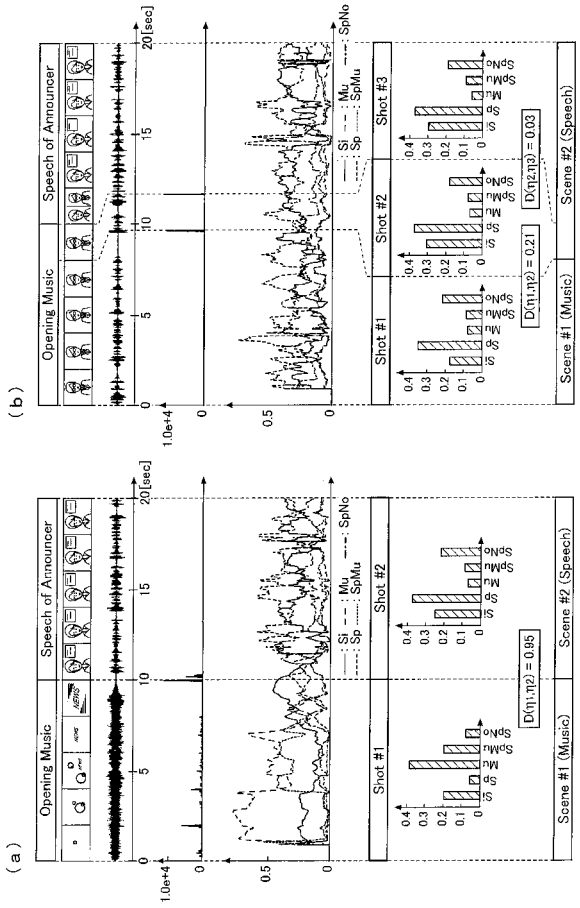
【図１０】


映像1
映像2

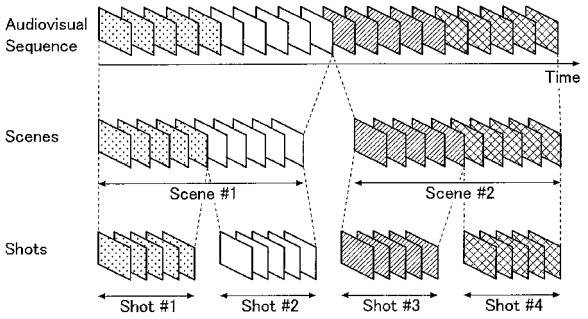
【図１１】


映像信号1：クラスタ＃i
Time
映像信号2：クラスタ＃i
Time

【図１２】


(a)
(b)
Opening Music
Speech of Announcer
20[sec]
Si
Sp
Mu
SpMu
SpNo
Shot #1
Shot #2
Shot #3
Scene #1 (Music)
Scene #2 (Speech)
D(η1,η2) = 0.95
D(η1,η2) = 0.21
D(η2,η3) = 0.03

【図１３】


Audiovisual Sequence
Time
Scenes
Scene #1
Scene #2
Shots
Shot #1
Shot #2
Shot #3
Shot #4