

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第4759745号
(P4759745)

(45) 発行日 平成23年8月31日(2011.8.31)

(24) 登録日 平成23年6月17日(2011.6.17)

(51) Int.Cl.		F I			
HO4N	5/76	(2006.01)	HO4N	5/76	B
HO4N	5/91	(2006.01)	HO4N	5/91	N
G06F	17/30	(2006.01)	G06F	17/30	170D
G11B	27/034	(2006.01)	G06F	17/30	210D
			G11B	27/034	

請求項の数 11 (全 23 頁)

(21) 出願番号	特願2006-171830 (P2006-171830)	(73) 特許権者	504173471
(22) 出願日	平成18年6月21日 (2006.6.21)		国立大学法人北海道大学
(65) 公開番号	特開2008-5167 (P2008-5167A)		北海道札幌市北区北8条西5丁目
(43) 公開日	平成20年1月10日 (2008.1.10)	(74) 代理人	110000338
審査請求日	平成21年4月16日 (2009.4.16)		特許業務法人原謙三国際特許事務所
特許法第30条第1項適用	2006年2月13日 社	(72) 発明者	長谷山 美紀
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」		北海道札幌市北区北14条西9丁目 北海
に発表		(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
		審査官	梅岡 信幸
			最終頁に続く

(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) 【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、

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上記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ものであり、

さらに、ショット内の各クリップの上記帰属確率の累積ヒストグラムのうち、最大の値を取るピンが属する上記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 1 または 2 に記載の映像分類装置。

【請求項 4】

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 ~ 3 の何れか 1 項に記載の映像分類装置。

【請求項 5】

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 3 または 4 に記載の映像分類装置。

10

【請求項 6】

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 3 ~ 5 の何れか 1 項に記載の映像分類装置。

【請求項 7】

上記表示手段は、分類されたシーンを類似したもの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 6 に記載の映像分類装置。

【請求項 8】

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 5 に記載の映像分類装置。

20

【請求項 9】

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、
を含み、

上記ショット間類似度判定ステップでは、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上記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し、

30

さらに、上記帰属確率の累積ヒストグラムのうち、最大の値を取るピンが属する上記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ステップを含むことを特徴とする映像分類方法。

【請求項 10】

請求項 1 ~ 8 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 11】

請求項 10 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。

40

【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み取り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

50

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、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。

10

【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。

20

【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.

30

【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.

40

【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, S

50

peech, 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、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割合を定量化し、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

10

20

30

40

50

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には、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）が可能であるといえることができる。

【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

10

20

30

40

50

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。

10

【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。

20

【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。

30

【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。

40

【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

50

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。

10

【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。

20

【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には、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)が可能であるといえることができる。

30

【発明を実施するための最良の形態】

40

【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(映像間類似度判定手段)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

50

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。

10

【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。

20

【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。

30

【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。

40

【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】

50

(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

10

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

20

【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。

30

【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は、 $W_f >$ 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(=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

40

【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】

50

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】

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

【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】

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

【0057】

ただし、 $\text{sign}\{\cdot\}$ は、以下で定義される関数である。

【0058】

【数3】

$$\text{sign}\{s_n(i)\} = \begin{cases} 1 & (s_n(i) \geq 0) \\ -1 & (s_n(i) < 0) \end{cases} \quad (3)$$

【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)を算出し、(1)の極小値のうち、最も1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】

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

【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 - 4500 Hz)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】

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

【0064】

ただし、 $S_n(\cdot)$ 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。

10

20

30

40

50

【 0 0 6 5 】

・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。

【 0 0 6 6 】

【数 6】

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

【 0 0 6 7 】

・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 - 630 Hz、630 - 1720 Hz、1720 - 4400 Hz、4400 - 11025 Hz）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。

10

【 0 0 6 8 】

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。

【 0 0 6 9 】

・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_v を設定し、次の式（7）を満たすフレームを無音と判断する。

20

【 0 0 7 0 】

【数 7】

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

【 0 0 7 1 】

・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）。

【 0 0 7 2 】

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。

30

【 0 0 7 3 】

（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現象（有限個の学習パターンか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。

40

【 0 0 7 4 】

（FCM の適用）

次に、上記 PCA の適用で得られた主成分に対し、FCM を施す。まず、処理対象信号（ $k = 1$ ）、参照信号 1（ $k = 2, \dots, N_c + 1$ ）、参照信号 2（ $k = N_c + 2, \dots, 2N_c + 1$ ）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。

【 0 0 7 5 】

【数 8】

$$\mathbf{f}_k \stackrel{\text{def}}{=} [p_1^k, p_2^k, \dots, p_M^k]^T \quad (8)$$

【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, $N_c + 2 \sim 2N_c + 1$: 参照信号 2) の第 i 主成分 ($i = 1, \dots, M$; 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} ($i = 1, 2$; $k = 1, \dots, 2N_c + 1$) 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 ($k = 1, \dots, 2N_c + 1$) を表す。

10

【0077】

この帰属度 μ_{ik} は、クリップ k がクラスタ i に属する割合を $[0, 1]$ 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_{i1} ($i = 1, 2$) 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、 μ^c_i を以下のように設定する。

【0078】

【数 9】

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

20

【0079】

【数 10A】

$$\begin{aligned} \mu_2^c &= \begin{cases} \mu_{21} & (\overline{\mu_{1\theta}} \geq \overline{\mu_{2\theta}}) \\ \mu_{11} & (\overline{\mu_{1\theta}} < \overline{\mu_{2\theta}}) \end{cases} \\ &= 1 - \mu_1^c \end{aligned} \quad (10)$$

【0080】

ただし、 c ($c = 1, \dots, 4$) は C L S # 1 から C L S # 4 の分類処理の番号を表す。

30

【0081】

【数 10B】

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

【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】

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

40

【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)

C L S # 1 から C L S # 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 ($i = 1, 2$; $c = 1, \dots, 4$)

50

を用いて、帰属確率を算出する。各クラス($S_i, S_p, Mu, SpMu, SpNo$)への帰属確率($P_{S_i}, P_{S_p}, P_{Mu}, P_{SpMu}, P_{SpNo}$)は、以下で定義される。

【0086】

【数12】

$$P_{S_i} = \mu_1^1 \quad (12)$$

【0087】

【数13】

$$P_{S_p} = \mu_2^1 \mu_2^2 \mu_1^4 \quad (13)$$

10

【0088】

【数14】

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

【0089】

【数15】

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

【0090】

【数16】

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

20

【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と同一のクラスに分類される確率とみなし、それらを積算することで、 $S_i, S_p, Mu, SpMu, SpNo$ 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 $P_{S_i}, P_{S_p}, P_{Mu}, P_{SpMu}, P_{SpNo}$ 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が可能となる。

30

【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】

40

本実施形態では、非特許文献5で提案されている分割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²検定法は、まずフレームを $4 \times 4 = 16$ 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 はヒストグラムのピンを表す。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 ($r = 1, \dots, 16$)を算出する。

50

【 0 0 9 5 】

【 数 1 7 A 】

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

【 0 0 9 6 】

さらに、算出された16個の評価値 C_r ($r = 1, \dots, 16$) において、 C_r 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 1 7 B] 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。

10

【 0 0 9 7 】

【 数 1 7 B 】

$$Th_{\chi^2}$$

【 0 0 9 8 】

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。

20

【 0 0 9 9 】

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 H_A (,) を作成する。ただし、 はショット番号、 は累積ヒストグラムのビン、すなわち S_i (= 0) , S_p (= 1) , M_u (= 2) , $S_p M_u$ (= 3) , $S_p N_o$ (= 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。

【 0 1 0 0 】

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 (,) を次式で定義する。

30

【 0 1 0 1 】

【 数 1 8 】

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

【 0 1 0 2 】

この距離 D (,) 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 (,) 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。

40

【 0 1 0 3 】

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。

【 0 1 0 4 】

50

(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、ヒストグラムの各ピ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ピ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

10

20

30

40

50

合は、上記特徴ベクトルは、ピンの値を次々と連結する方法で得られる。

【 0 1 1 1 】

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。

【 0 1 1 2 】

得られた特徴ベクトルに対して、k - m e a n s 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 - m e a n s 法を挙げたが、k - m e a n s 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 - m e a n s 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。

【 0 1 1 3 】

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、距離が長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。

【 0 1 1 4 】

（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、識別情報部呼部 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。

【 0 1 1 5 】

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。

【 0 1 1 6 】

（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。

【 0 1 1 7 】

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

10

20

30

40

50

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法を利用しては構わない。

【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 x 240 pixel、30 fps、44100 Hz、20 sec) 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】

10

20

30

40

【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]	

【 0 1 2 4 】

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。

10

【 0 1 2 5 】

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 $D(x_1, x_2)$ 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。

20

【 0 1 2 6 】

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、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。

30

【 0 1 2 7 】

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 $D(x_1, x_2)$ 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 $D(x_2, x_3)$ 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。

【 0 1 2 8 】

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。

40

【 0 1 2 9 】

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。

【 0 1 3 0 】

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

50

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、映像において複数の話者で、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】

10

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】

20

【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。

30

【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。

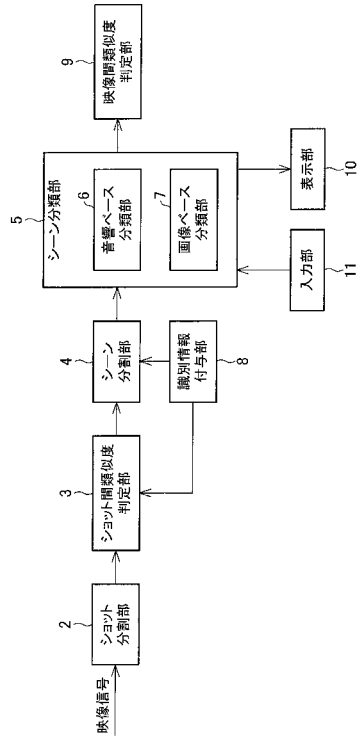
【符号の説明】

【0134】

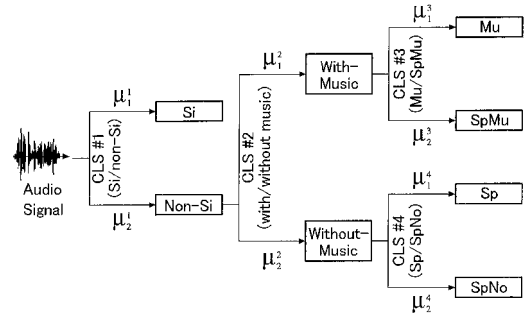
- 1 映像分類装置
- 2 ショット分割部(ショット分割手段)
- 3 ショット間類似度判定部(ショット間類似度判定手段)
- 4 シーン分割部(シーン分割手段)
- 5 シーン分類部
- 6 音響ベース分類部(音響ベース分類手段)
- 7 画像ベース分類部(画像ベース分類手段)
- 8 識別情報付与部(クラス識別情報付与手段、映像源識別情報付与手段)
- 9 映像間類似度判定部(映像間類似度判定手段)
- 10 表示部(表示手段)
- 11 入力部

40

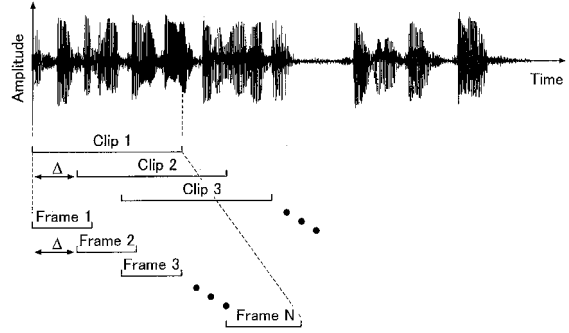
【図1】



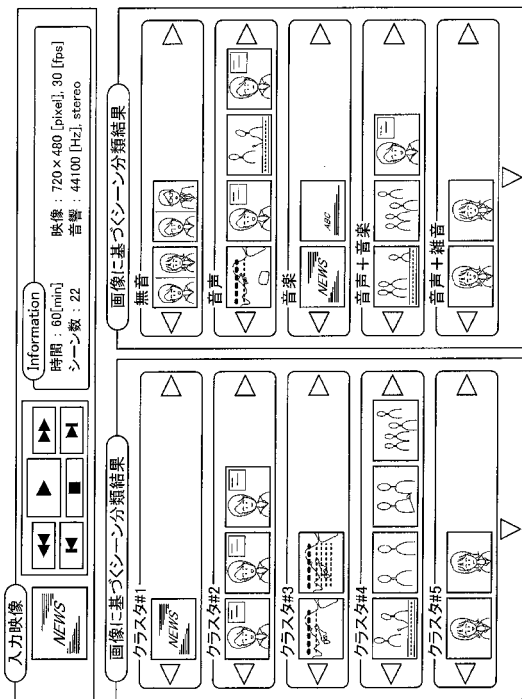
【図2】



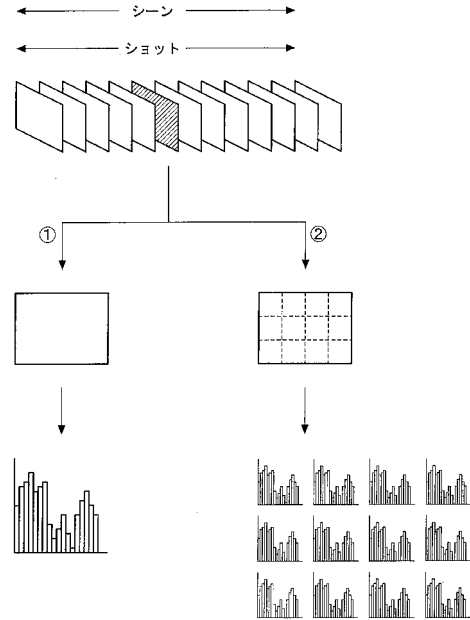
【図3】



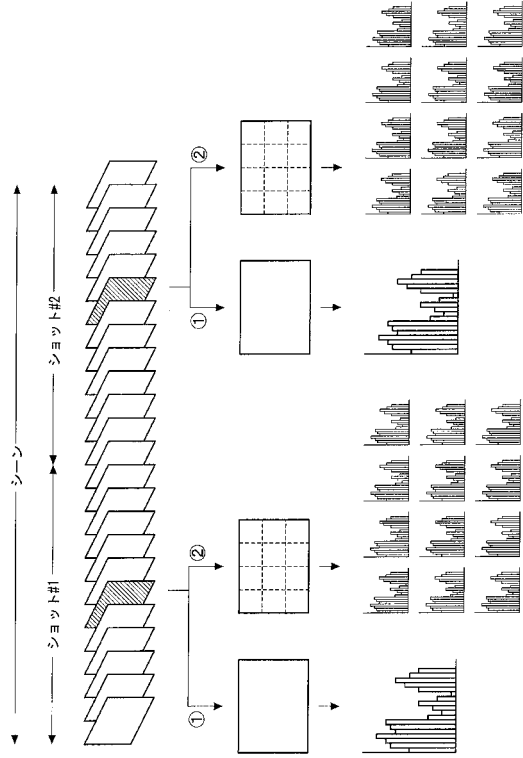
【図4】



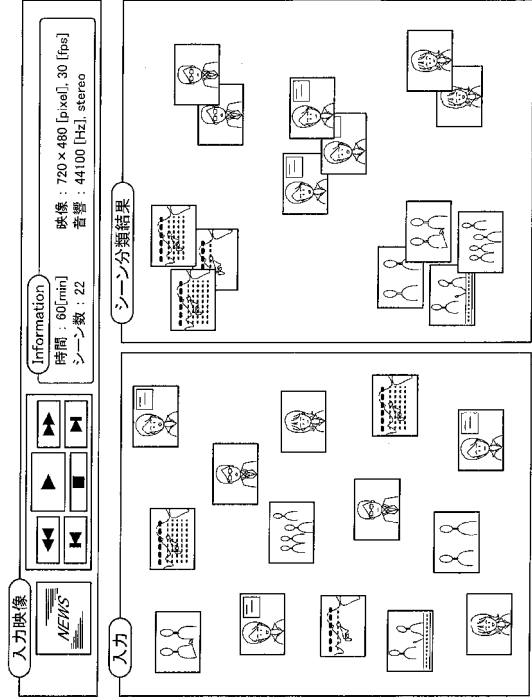
【図5】



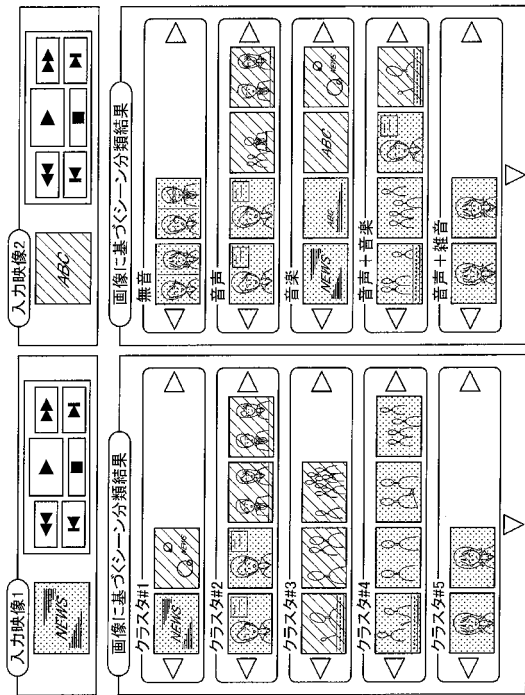
【図6】



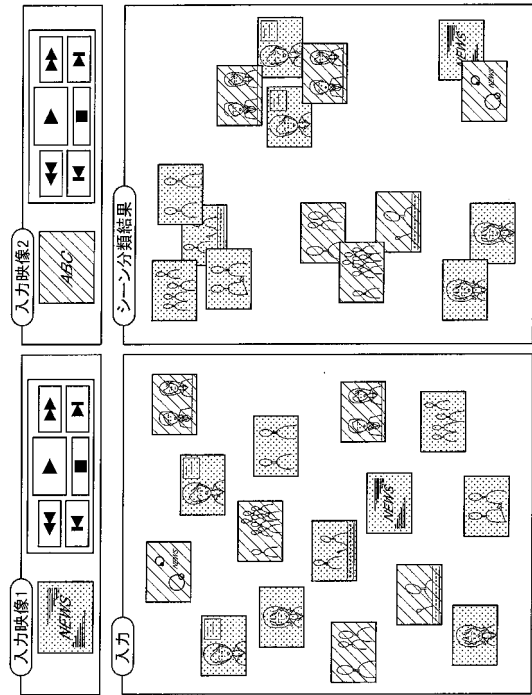
【図7】



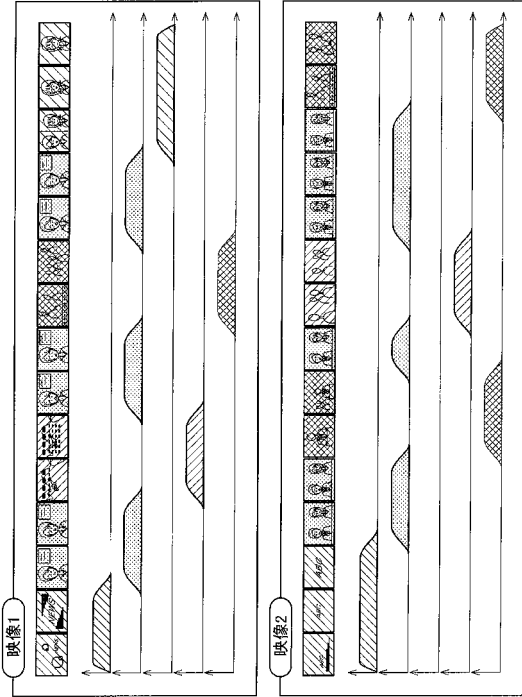
【図8】



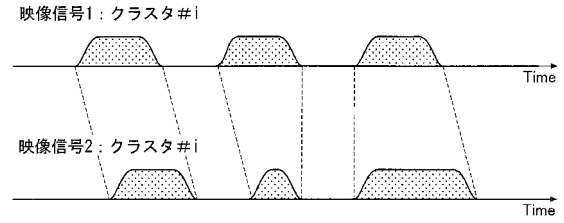
【図9】



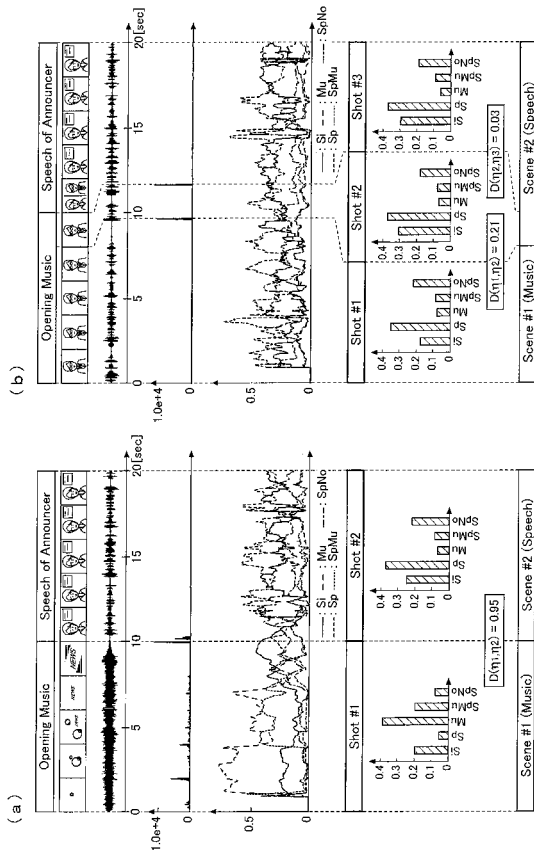
【図10】



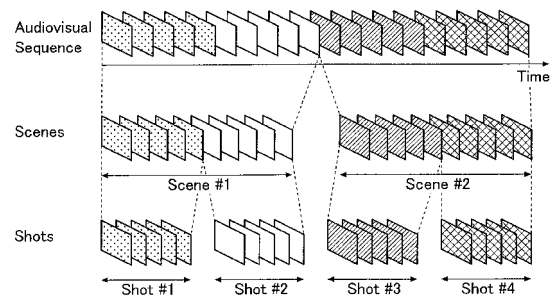
【図11】



【図12】



【図13】



フロントページの続き

- (56)参考文献 特開2000-285242(JP,A)
特表2004-517518(JP,A)
特表2006-510072(JP,A)
特開2003-330941(JP,A)
特開2002-124929(JP,A)
特開2005-236646(JP,A)

(58)調査した分野(Int.Cl., DB名)

H04N 5/76 - 5/956
G11B 20/10 - 20/16
G11B 27/00 - 27/34
G06F 17/30