(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7299282号
(P7299282)

(45)発行日　令和5年6月27日(2023.6.27)

(24)登録日　令和5年6月19日(2023.6.19)

(51)国際特許分類

		FI		
G06T	*7/00 (2017.01)*	G06T	7/00	300F
G06F	*16/783 (2019.01)*	G06T	7/00	350C
G06F	*16/732 (2019.01)*	G06F	16/783	
		G06F	16/732	

請求項の数　13　（全20頁）

(21)出願番号　特願2021-166004(P2021-166004)
(22)出願日　令和3年10月8日(2021.10.8)
(65)公開番号　特開2022-20647(P2022-20647A)
(43)公開日　令和4年2月1日(2022.2.1)
審査請求日　令和3年10月8日(2021.10.8)
(31)優先権主張番号　202011358245.3
(32)優先日　令和2年11月27日(2020.11.27)
(33)優先権主張国・地域又は機関
中国(CN)

(73)特許権者　514322098
ベイジン バイドゥ ネットコム サイエンス テクノロジー カンパニー リミテッド
Ｂｅｉｊｉｎｇ Ｂａｉｄｕ Ｎｅｔｃｏｍ Ｓｃｉｅｎｃｅ Ｔｅｃｈｎｏｌｏｇｙ Ｃｏ．，Ｌｔｄ．
中華人民共和国 ペキン １０００８５，ハイディアン ディストリクト，シャンディ テンス ストリート，１０番，バイドゥ キャンパス ２階
２／Ｆ Ｂａｉｄｕ Ｃａｍｐｕｓ，Ｎｏ．１０，Ｓｈａｎｇｄｉ １０ｔｈ Ｓｔｒｅｅｔ，Ｈａｉｄｉａｎ Ｄｉｓｔｒｉｃｔ，Ｂｅｉｊｉｎｇ １０００

最終頁に続く

(54)【発明の名称】　ビデオ処理方法、装置、電子デバイス、記憶媒体、及びプログラム

(57)【特許請求の範囲】

【請求項１】

複数の第１ビデオフレームを取得し、前記複数の第１ビデオフレームに対してきめ細かい分割を行い、複数の第２ビデオフレームを取得することと、

前記複数の第２ビデオフレームに関するマルチモーダル情報に基づき、前記複数の第２ビデオフレームに対して特徴符号化を行い、前記マルチモーダル情報の融合を特徴付けるための特徴融合情報を取得することと、

前記特徴融合情報に基づき、前記複数の第２ビデオフレームに対して類似度のマッチングを行い、類似度のマッチング結果に基づいてターゲットビデオを取得することと、

予めトレーニングされた第１ニューラルネットワークモデルにより、前記複数の第２ビデオフレームから前記マルチモーダル情報を識別することと、

第２ニューラルネットワークモデルにより、前記マルチモーダル情報の中の各種類の情報に対して区別を行うことと、

第３ニューラルネットワークモデルにより、前記マルチモーダル情報に関する時系列情報に対して識別を行うことと、

前記第１ニューラルネットワークモデル、前記第２ニューラルネットワークモデル、前記第３ニューラルネットワークモデルの出力結果に対して融合を行い、前記特徴融合情報を取得することと、を含む

ことを特徴とするビデオ処理方法。

【請求項２】

複数の第１ビデオフレームを取得し、前記複数の第１ビデオフレームに対してきめ細かい分割を行い、複数の第２ビデオフレームを取得することは、
シーン及び色彩転換を特徴付けるためのパラメータに基づき、前記複数の第１ビデオフレームに対してきめ細かい分割を行い、複数の第２ビデオフレームを取得することを含む
ことを特徴とする請求項１に記載のビデオ処理方法。
【請求項３】
前記複数の第２ビデオフレームに関するマルチモーダル情報に基づき、前記複数の第２ビデオフレームに対して特徴符号化を行い、前記マルチモーダル情報の融合を特徴付けるための特徴融合情報を取得することは、
前記マルチモーダル情報に基づき、前記複数の第２ビデオフレームに対して特徴抽出及び特徴融合の処理を行い、前記特徴融合情報を取得することを含む
ことを特徴とする請求項１に記載のビデオ処理方法。
【請求項４】
前記特徴融合情報に基づき、前記複数の第２ビデオフレームに対して類似度のマッチングを行い、類似度のマッチング結果に基づいてターゲットビデオを取得することは、
前記特徴融合情報に基づき、前記複数の第２ビデオフレームの類似度に対して採点し、採点結果を前記類似度のマッチング結果とすることと、
前記類似度のマッチング結果として、同じイベントコンテンツに関する、隣接するビデオフレームが類似している場合、前記複数の第２ビデオフレームのそれぞれに対して、前記隣接するビデオフレームに対する結合がされるまで、前記隣接するビデオフレームに対してビデオ結合を行い、ビデオ結合結果に基づいて前記ターゲットビデオを取得することとを含む
ことを特徴とする請求項１に記載のビデオ処理方法。
【請求項５】
予めトレーニングされた第１ニューラルネットワークモデルにより、前記複数の第２ビデオフレームから前記マルチモーダル情報を識別することは、
前記第１ニューラルネットワークモデルの中のナレッジグラフ抽出器により、ナレッジグラフ情報を識別することと、
前記第１ニューラルネットワークモデルの中のテキスト抽出器により、テキスト情報を識別することと、
前記第１ニューラルネットワークモデルの中のオーディオ抽出器により、オーディオ情報を識別することと、
前記第１ニューラルネットワークモデルの中の色調抽出器により、色調情報を識別することと、
前記第１ニューラルネットワークモデルの中の物体抽出器により、物体情報を識別することと、
前記第１ニューラルネットワークモデルの中の動作抽出器により、動作情報を識別することとを含み、
前記マルチモーダル情報は、前記ナレッジグラフ情報、前記テキスト情報、前記オーディオ情報、前記色調情報、前記物体情報、前記動作情報の中の少なくとも１つを含む
ことを特徴とする請求項１に記載のビデオ処理方法。
【請求項６】
複数の第１ビデオフレームを取得し、前記複数の第１ビデオフレームに対してきめ細かい分割を行い、複数の第２ビデオフレームを取得するための分割モジュールと、
前記複数の第２ビデオフレームに関するマルチモーダル情報に基づき、前記複数の第２ビデオフレームに対して特徴符号化を行い、前記マルチモーダル情報の融合を特徴付けるための特徴融合情報を取得するための符号化モジュールと、
前記特徴融合情報に基づき、前記複数の第２ビデオフレームに対して類似度のマッチングを行い、類似度のマッチング結果に基づいてターゲットビデオを取得するためのビデオ処理モジュールと、

予めトレーニングされた第１ニューラルネットワークモデルにより、前記複数の第２ビデオフレームから前記マルチモーダル情報を識別するための識別モジュールと、
第２ニューラルネットワークモデルにより、前記マルチモーダル情報の中の各種類の情報に対して区別を行い、
第３ニューラルネットワークモデルにより、前記マルチモーダル情報に関する時系列情報に対して識別を行い、
前記第１ニューラルネットワークモデル、前記第２ニューラルネットワークモデル、前記第３ニューラルネットワークモデルの出力結果に対して融合を行い、前記特徴融合情報を取得するための融合モジュールと、を備える
ことを特徴とするビデオ処理装置。
【請求項７】
前記分割モジュールは、
シーン及び色彩転換を特徴付けるためのパラメータに基づき、前記複数の第１ビデオフレームに対してきめ細かい分割を行い、複数の第２ビデオフレームを取得することに用いられる
ことを特徴とする請求項６に記載のビデオ処理装置。
【請求項８】
前記符号化モジュールは、
前記マルチモーダル情報に基づき、前記複数の第２ビデオフレームに対して特徴抽出及び特徴融合の処理を行い、前記特徴融合情報を取得することに用いられる
ことを特徴とする請求項６に記載のビデオ処理装置。
【請求項９】
前記ビデオ処理モジュールは、
前記特徴融合情報に基づき、前記複数の第２ビデオフレームの類似度に対して採点し、採点結果を前記類似度のマッチング結果とし、
前記類似度のマッチング結果として、同じイベントコンテンツに関する、隣接するビデオフレームが類似している場合、前記複数の第２ビデオフレームのそれぞれに対して、前記隣接するビデオフレームに対する結合がされるまで、前記隣接するビデオフレームに対してビデオ結合を行い、ビデオ結合の結果に基づいて前記ターゲットビデオを取得することに用いられる
ことを特徴とする請求項６に記載のビデオ処理装置。
【請求項１０】
前記識別モジュールは、
前記第１ニューラルネットワークモデルの中のナレッジグラフ抽出器により、ナレッジグラフ情報を識別し、
前記第１ニューラルネットワークモデルの中のテキスト抽出器により、テキスト情報を識別し、
前記第１ニューラルネットワークモデルの中のオーディオ抽出器により、オーディオ情報を識別し、
前記第１ニューラルネットワークモデルの中の色調抽出器により、色調情報を識別し、
前記第１ニューラルネットワークモデルの中の物体抽出器により、物体情報を識別し、
前記第１ニューラルネットワークモデルの中の動作抽出器により、動作情報を識別するために用いられ、
前記マルチモーダル情報は、前記ナレッジグラフ情報、前記テキスト情報、前記オーディオ情報、前記色調情報、前記物体情報、前記動作情報の中の少なくとも１つを含む
ことを特徴とする請求項６に記載のビデオ処理装置。
【請求項１１】
少なくとも１つのプロセッサと、
前記少なくとも１つのプロセッサに通信接続されたメモリと、を備え、
前記メモリは、前記少なくとも１つのプロセッサによって実行可能な命令を記憶し、前

記命令は、前記少なくとも１つのプロセッサにより実行されると、前記少なくとも１つのプロセッサに請求項１～５のいずれか１項に記載のビデオ処理方法を実行させる
ことを特徴とする電子デバイス。
【請求項１２】
コンピュータに請求項１～５のいずれか１項に記載のビデオ処理方法を実行させるためのコンピュータ命令が記憶されている非一時的なコンピュータ可読記憶媒体。
【請求項１３】
コンピュータにおいて、プロセッサにより実行されると、請求項１～５のいずれか１項に記載のビデオ処理方法を実現することを特徴とするプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本開示は、人工知能分野に関し、特に、ディープラーニング、モデルトレーニング、ナレッジグラフ及びビデオ処理等の分野に関する。
【背景技術】
【０００２】
ポータブル機器、携帯電話端末等の電子デバイスは、以前よりもっとインテリジェントになり、チップの分析能力がより強く、特に、ビデオ情報の分析、画面のレンダリング等は、以前より高速且つ鮮明になり、ビデオ品質に対するユーザの要求が以前より高く、特に、高適時性シナリオ（例えば、軍事パレードシナリオ、スポーツイベント、リアルタイムビデオ生中継等）の場合、各ビデオの瞬間の素晴らしい画面に対し、ユーザは、キャプチャしたいと希望しているので、より正確で且つ鮮明なビデオ画面が必要になっている。
【０００３】
ビデオ処理においては、ビデオの分割を例とすると、手動の方式でビデオの分割を実現することができるが、多くの人件費を消費するだけではなく、上述した高適時性シナリオの要求を満たすことができない。一方、非手動での幾つかのビデオ分割方式では、ビデオフレームのコンテンツ情報（例えば、テキスト、ビデオ内の物体、動作等）を十分に理解することができず、ビデオイベントの連続性（例えば、シーン転換によるシナリオの切り替え等）も正しく制御することができないので、ビデオ画面に対する解釈の正確性が大幅に低下し、最終的なターゲットビデオにより提示されたビデオ品質の効果に影響を与えてしまう。
【発明の概要】
【０００４】
本開示は、ビデオ処理方法、装置、電子デバイス及び記憶媒体を提供する。
【０００５】
本開示の１つの側面では、
複数の第１ビデオフレームを取得し、前記複数の第１ビデオフレームに対してきめ細かい分割を行い、複数の第２ビデオフレームを取得することと、
前記複数の第２ビデオフレームに関するマルチモーダル情報に基づき、前記複数の第２ビデオフレームに対して特徴符号化を行い、前記マルチモーダル情報の融合を特徴付けるための特徴融合情報を取得することと、
前記特徴融合情報に基づき、前記複数の第２ビデオフレームに対して類似度のマッチングを行い、類似度のマッチング結果に基づいてターゲットビデオを取得することとを含むビデオ処理方法が提供される。
【０００６】
本開示のもう１つの側面では、
複数の第１ビデオフレームを取得し、前記複数の第１ビデオフレームに対してきめ細かい分割を行い、複数の第２ビデオフレームを取得するための分割モジュールと、
前記複数の第２ビデオフレームに関するマルチモーダル情報に基づき、前記複数の第２ビデオフレームに対して特徴符号化を行い、前記マルチモーダル情報の融合を特徴付ける

ための特徴融合情報を取得するための符号化モジュールと、
前記特徴融合情報に基づき、前記複数の第２ビデオフレームに対して類似度のマッチングを行い、類似度のマッチング結果に基づいてターゲットビデオを取得するためのビデオ処理モジュールとを備えるビデオ処理装置が提供される。
【０００７】
本開示のもう１つの側面では、
少なくとも１つのプロセッサと、
前記少なくとも１つのプロセッサに通信接続されたメモリとを備える電子デバイスが提供される。
【０００８】
ここで、前記メモリは、前記少なくとも１つのプロセッサによって実行可能な命令を記憶し、前記命令は、前記少なくとも１つのプロセッサにより実行されると、前記少なくとも１つのプロセッサに本開示の任意の１つの実施形態による方法を実行させる。
【０００９】
本開示のもう１つの側面では、コンピュータ命令が記憶されている非一時的なコンピュータ可読記憶媒体が提供される。ここで、前記コンピュータ命令は、前記コンピュータに本開示の任意の１つの実施形態による方法を実行させることに用いられる。
【００１０】
本開示によれば、複数の第１ビデオフレームを取得し、前記複数の第１ビデオフレームに対してきめ細かい分割を行い、複数の第２ビデオフレームを取得する。前記複数の第２ビデオフレームに関するマルチモーダル情報に基づき、前記複数の第２ビデオフレームに対して特徴符号化を行い、前記マルチモーダル情報の融合を特徴付けるための特徴融合情報を取得する。前記特徴融合情報に基づき、前記複数の第２ビデオフレームに対して類似度のマッチングを行い、類似度のマッチング結果に基づいてターゲットビデオを取得する。当該マルチモーダル情報に基づいて特徴符号化を行うことにより、より多くのビデオコンテンツの詳細を含む情報を取得し、類似度に基づいてマッチングされた後、得られたターゲットビデオの精度がより高いので、ビデオ分割の正確性を高めることができる。
【００１１】
当該部分に記載の内容は、本開示の実施形態の肝心又は重要な特徴を示すことを意図するものではなく、本開示の範囲を制限しないことが理解されたい。本開示の他の特徴は、以下の説明により、より理解しやすくなる。
【００１２】
添付の図面は、本実施形態をより良く理解するために用いられ、本開示に対する限定を構成するものではない。
【図面の簡単な説明】
【００１３】
【図１】本開示の実施形態によるビデオ処理の高適時性シナリオの複数のビデオフレームの模式図である。
【図２】本開示の実施形態によるビデオ処理方法のフローチャート模式図である。
【図３】本開示の実施形態によるビデオ処理方法を実現するシステムモジュールのアーキテクチャ図である。
【図４】本開示の実施形態による、マルチモーダル情報に基づいて特徴符号化を実現する模式図である。
【図５】本開示の実施形態による類似度のマッチングの模式図である。
【図６】本開示の実施形態によるビデオ結合の模式図である。
【図７】本開示の実施形態によるビデオ処理装置の構成構造模式図である。
【図８】本開示の実施形態によるビデオ処理方法を実現するための電子デバイスのブロック図である。
【発明を実施するための形態】
【００１４】

以下、図面を参照しながら、本開示の例示的な実施形態を説明し、理解を助けるために本開示の実施形態の様々な詳細を含んでいるが、これらは、単に例示的なものとみなされるべきである。よって、当業者は、本開示の範囲及び要旨から逸脱することなく、本明細書に記載の実施形態に様々な変更及び修正を加えることができることを識別すべきである。明瞭で且つ簡潔にするために、以下の説明では、周知の機能と構造の説明を省略している。

【００１５】

本文における用語「及び／又は」は、関連対象の関連関係を説明するものに過ぎず、３つの関係があっても良いことを表し、例えば、Ａ及び／又はＢは、Ａだけがあり、Ａ及びＢがあり、Ｂだけがあるという３つの場合を表すことができる。本文における用語「少なくとも１つ」は、複数の中の任意の１つ又は複数の中の少なくとも２つの任意の組み合わせを表し、例えば、Ａ、Ｂ、Ｃの中の少なくとも１つを含むことは、Ａ、Ｂ、Ｃからなる集合から選択された任意の１つ又は複数の元素を含むことを表すことができる。本文における用語「第１」、「第２」は、複数の類似する技術用語を指し、それらを区別するためのものであり、順序を制限する意味がなく、又は、２つだけに制限する意味がなく、例えば、第１特徴及び第２特徴は、２種類／２つの特徴を指し、第１特徴は１つ又は複数であっても良く、第２特徴も１つ又は複数であっても良い。

【００１６】

また、本開示をより良く説明するために、後述における具体的な実施形態においては、沢山の具体的な詳細が記載されている。当業者は、幾つかの具体的な詳細がなくても、本開示は、同様に実施することができることを理解すべきである。幾つかの実施形態においては、本開示の主旨を強調するために、当業者に周知されている方法、手段、素子及び回路に対しては詳しく説明していない。

【００１７】

ビデオ分割は、インターネットビデオ及びニューメディアの短いビデオコンテンツプラットフォームのニーズにより、従来のテレビメディア番組に対して行われた二次処理であり、即ち、元の完全な番組コンテンツを、ある論理的な思考又は特別のニーズに従って複数のビデオに分割するものである。インターネットビデオコンテンツの主なソースは、従来のテレビメディアの番組、様々な機関のビデオ製品、映画とテレビ会社の映画とテレビ作品を含み、これらのビデオを分割することにより、貴重な情報を深く掘り下げることができ、新たに編集された後、インターネットプロトコルテレビジョン（ＩＰＴＶ）、ＯＴＴ、モバイルＴＶ及びニューメディアの短いビデオプラットフォームに用いることができ、ニューメディアオーディオビジュアル番組の断片化の要求を満たすことができ、オーディオビジュアルの編集業界の１つの新しい試み及び探索となる。

【００１８】

従来の手動によるビデオ分割技術は、人手でビデオを編集及び分割するので、処理時間が長く、ビデオの量が多い場合、生産性の向上が遅いため、高適時性等のシナリオに応用することができず、大量の時間及び経験を消費するだけではなく、経済的コスト及び適時性の要求を満たすことができない。非機械学習の従来のビデオ分割アルゴリズムのような、非手動によるビデオ分割技術の場合、色調及びブロックマッチングに基づいてビデオ分割を行うことができるが、ピクチャー、シーンの間の視覚情報しか考慮せず、ビデオフレームのコンテンツ情報を十分に理解することができない。また、機械学習に基づくビデオ分割技術の場合、クラスタリング方法に基づいてビデオ分割（キーフレームの抽出、画像特徴の説明、クラスタリング等を含む）を行うことができるが、ビデオイベントの連続性を考慮せず、幾つかのシーン切り替えが比較的頻繁なシナリオ（図１に示す体育イベントの中の複数のビデオフレームから構成される素晴らしい瞬間等）、又は、複雑なイベントのシナリオにおいて複数のシーンの連続的な転換があるビデオの場合、ビデオ分割の効果が良くなく、ビデオ分割の正確率が低い。

【００１９】

本開示の実施形態では、ビデオ処理方法が提供される。図２は、本開示の実施形態によ

るビデオ処理方法のフローチャート模式図であり、当該方法は、ビデオ処理装置に用いることができ、例えば、当該装置は、端末、サーバ又は他の処理デバイスに展開でき、ビデオフレーム分割、ビデオフレーム特徴符号化、ビデオフレーム類似度のマッチングを実行することにより、最終的なターゲットビデオ等を取得することができる。ここで、端末は、ユーザ機器（ＵＥ、Ｕｓｅｒ　Ｅｑｕｉｐｍｅｎｔ）、モバイルデバイス、携帯電話、コードレス電話、パーソナルデジタルアシスタント（ＰＤＡ、Ｐｅｒｓｏｎａｌ　Ｄｉｇｉｔａｌ　Ａｓｓｉｓｔａｎｔ）、ハンドヘルドデバイス、コンピューティングデバイス、車載機器、ウェアラブル機器等であっても良い。幾つかの可能な実現方法においては、当該方法は、プロセッサがメモリに記憶されたコンピュータ可読命令を呼び出す方式により実現することができる。図２に示すように、ステップＳ１０１、ステップＳ１０２及びステップＳ１０３を含む。

【００２０】

ステップＳ１０１においては、複数の第１ビデオフレームを取得し、前記複数の第１ビデオフレームに対してきめ細かい分割を行い、複数の第２ビデオフレームを取得する。

【００２１】

ステップＳ１０２においては、前記複数の第２ビデオフレームに関するマルチモーダル情報に基づき、前記複数の第２ビデオフレームに対して特徴符号化を行い、前記マルチモーダル情報の融合を特徴付けるための特徴融合情報を取得する。

【００２２】

ステップＳ１０３においては、前記特徴融合情報に基づき、前記複数の第２ビデオフレームに対して類似度のマッチングを行い、類似度のマッチング結果に基づいてターゲットビデオを取得する。

【００２３】

ステップＳ１０１においては、シーン及び色彩転換を特徴付けるためのパラメータに基づき、前記複数の第１ビデオフレームに対してきめ細かい分割を行い、前記複数の第２ビデオフレームを取得する。シーン及び色彩転換を特徴付けるためのパラメータは、シーンの観点からは、ビデオエッジに対する分割、ビデオの中のブロックマッチングに基づくビデオ分割、統計的決定に基づくビデオ分割及び双閾値の比較（双閾値の比較を設定することにより、シーンの急変なのか、シーンの段階的な変化なのかを区別する）に基づくビデオ分割等を含むことができる。色彩転換の観点からは、色調に基づくビデオ分割を含むことができる。

【００２４】

ステップＳ１０２においては、前記マルチモーダル情報に基づき、前記複数の第２ビデオフレームに対して特徴抽出及び特徴融合処理を行い、前記特徴融合情報を取得する。ここで、特徴融合処理は、複数のニューラルネットワークモデル、又は、多機能抽出に統合された１つのニューラルネットワークモデルをエキスパートモデルとして用いることにより、第２ビデオフレームに関するマルチモーダル情報に対してそれぞれ特徴抽出を行う。ここで、マルチモーダル情報は、ナレッジグラフ情報、テキスト情報、オーディオ情報、色調情報、物体情報、動作情報の中の少なくとも１つを含む。

【００２５】

ステップＳ１０３においては、前記特徴融合情報に基づき、前記複数の第２ビデオフレームの類似度に対して採点し、採点結果を前記類似度のマッチング結果とし、前記類似度のマッチング結果として、同じイベントコンテンツに関する、隣接するビデオフレームが類似している場合、前記複数の第２ビデオフレームのそれぞれに対して、前記隣接するビデオフレームに対する結合がされるまで、前記隣接するビデオフレームに対してビデオ結合を行い、ビデオ結合の結果に基づいて前記ターゲットビデオを取得する。

【００２６】

本開示によれば、複数の第１ビデオフレームを取得し、前記複数の第１ビデオフレームに対してきめ細かい分割を行い、複数の第２ビデオフレームを取得する。前記複数の第２ビデオフレームに関するマルチモーダル情報に基づき、前記複数の第２ビデオフレームに

対して特徴符号化を行い、前記マルチモーダル情報の融合を特徴付けるための特徴融合情報を取得することができる。前記特徴融合情報に基づき、前記複数の第２ビデオフレームに対して類似度のマッチングを行い、類似度のマッチング結果に基づいてターゲットビデオを取得することができる。当該マルチモーダル情報に基づいて特徴符号化を行うことにより、より多くのビデオコンテンツの詳細を含む情報を取得し、類似度に基づいてマッチングされた後、得られたターゲットビデオの精度がより高いので、ビデオ分割の正確性を高めることができる。

【００２７】

１つの例においては、ビデオ分割モジュール、マルチモーダル特徴符号化モジュール、類似度マッチングモジュール（主に、隣接するビデオセグメントに対する類似度のマッチング）、ビデオフレームスプライシングモジュールにより、上述したステップＳ１０１～ステップＳ１０３を実現することができる。具体的には、当該ビデオ分割モジュールに入力された複数の第１ビデオフレームを取得した後、シーン及び色彩転換に従い、取得された複数の第１ビデオフレームに対してきめ細かい分割を行い、複数の第２ビデオフレームを取得することができる。当該マルチモーダル特徴符号化モジュールに入力された当該複数の第２ビデオフレームに対し、マルチモーダル情報に基づいて特徴符号化（例えば、マルチモーダル情報の特徴抽出及び特徴融合）を行い、マルチモーダル情報が融合された後の特徴情報を取得する。当該特徴情報を当該類似度マッチングモジュールに入力してビデオの類似度のマッチングを行い、類似度のマッチング結果（例えば、類似度の採点結果）を取得する。類似度のマッチング結果は、同じイベントコンテンツの２つの隣接するビデオフレームが類似している場合、同じイベント内のビデオセグメントに対して復元を行うという戦略に基づき、当該ビデオフレームスプライシングモジュールにより同じイベントコンテンツの２つの隣接するビデオフレームに対してそれぞれビデオ結合を行い、ビデオ処理後の最終的なビデオ分割結果を取得する。同じイベントコンテンツの類似度に従って結合し、ビデオコンテンツの詳細の類似度をより注目するので、ビデオ分割がより正確になり、最終的なビデオ分割結果の正確性を大幅に高めることができる。

【００２８】

１つの実施形態においては、予めトレーニングされた第１ニューラルネットワークモデルにより、前記複数の第２ビデオフレームから前記マルチモーダル情報を識別する。

【００２９】

１つの例においては、第１ニューラルネットワークモデルは、複数のエキスパートモデルから構成されても良く、複数のエキスパートモデルのそれぞれの機能を、１つのニューラルネットワークに集積されて構成されても良い。複数のエキスパートモデルのそれぞれの機能を、１つのニューラルネットワークに集積して構成される第１ニューラルネットワークモデルを例とする場合、当該第１ニューラルネットワークモデルは、ナレッジグラフ抽出器、テキスト抽出器、オーディオ抽出器、色調抽出器、物体抽出器及び動作抽出器を含んでも良い。ここで、第１ニューラルネットワークモデルの中のナレッジグラフ抽出器（又は、ナレッジグラフに基づく構造化ラベルベクトル抽出器と呼ばれる）により、ナレッジグラフ情報（例えば、ｋｎｏｗｌｅｄｇｅ特徴等）を識別することができ、第１ニューラルネットワークの中のテキスト抽出器（又は、テキストに基づくテキストベクトル抽出器と呼ばれる）により、テキスト情報（例えば、ｔｅｘｔ特徴）を識別することができ、第１ニューラルネットワークの中のオーディオ抽出器（又は、オーディオに基づくオーディオベクトル抽出器と呼ばれる）により、オーディオ情報（例えば、ａｕｄｉｏ情報）を識別することができ、前記第１ニューラルネットワークの中の色調抽出器（又は、画像に基づくＲＧＢ抽出器と呼ばれる）により、色調情報（例えば、ＲＧＢ特徴）を識別することができ、前記第１ニューラルネットワークの中の物体抽出器（又は、ターゲット検出に基づく物体特徴抽出器と呼ばれる）により、物体情報（例えば、ｏｂｊｅｃｔ特徴）を識別することができ、前記第１ニューラルネットワークの中の動作抽出器（動作識別に基づく動作ベクトル抽出器）により、動作情報（例えば、ａｃｔｉｏｎ特徴）を識別することができる。ここで、前記マルチモーダル情報は、前記ナレッジグラフ情報、前記テキス

ト情報、前記オーディオ情報、前記色調情報、前記物体情報及び前記動作情報の中の少なくとも１つを含む。

【００３０】

本実施形態によれば、本開示のインテリジェントビデオ分割技術と機械学習に基づく複数のエキスパートモデルを組み合わせることで、マルチモーダル情報の特徴識別、特徴抽出及び特徴融合を実現する。更に、融合された特徴情報（特徴融合情報と呼ばれる）に対して類似度の比較を実現する。よって、より詳細なビデオコンテンツ情報を取得し、ビデオコンテンツ及びイベント知識をより深く理解することができるので、最も正しいビデオ分割結果が得られ、最終的なビデオ分割結果の正確度を大幅に高めることができる。

【００３１】

１つの実施形態においては、予めトレーニングされた、ビデオ特徴抽出モデルF(ｕ)のような第１ニューラルネットワークモデルにより、前記複数の第２ビデオフレームから前記マルチモーダル情報を識別及び抽出することができる。ビデオ特徴識別モデルM(ｕ)のような第２ニューラルネットワークモデルにより、前記マルチモーダル情報のそれぞれの情報を区別することができる。ビデオに対応する時系列情報の抽出モデルT(ｕ)のような、第３ニューラルネットワークモデルにより、前記マルチモーダル情報に関する時系列情報を識別及び抽出し、前記ビデオ特徴抽出の時間オフセット表現を記録し、前記第１ニューラルネットワークモデル、第２ニューラルネットワークモデル及び第３ニューラルネットワークモデルの出力結果を融合し、前記特徴融合情報を取得する。特徴融合情報は、より多くのビデオコンテンツの詳細を記述することができるので、後続において類似度の比較を行う際にマッチング速度及び精度を高め、同じイベントコンテンツの２つの隣接するビデオフレームに対して類似度のマッチングが行われた後にビデオ処理の最終的なビデオ分割結果を取得し、結果がより正確であり、ビデオ分割がより正確であり、最終的なビデオ分割結果の正確度を大幅に高めることができる。

【００３２】

応用例

本開示の実施形態を応用する１つの処理流れは、以下の内容を含む。

【００３３】

図３は、本開示の実施形態によるビデオ処理方法を実現するシステムモジュールのアーキテクチャ図であり、ビデオ分割モジュール（主に、シーン及び色彩転換に従い、ビデオに対してきめ細かい分割を行う）、マルチモーダル特徴符号化モジュール（主に、マルチモーダル情報を用いてビデオに対して特徴符号化を行う）、類似度マッチングモジュール（主に、隣接するビデオセグメントに対して類似度のマッチングを行い、更に、同じイベントコンテンツに従い、ビデオセグメントに対してビデオ結合を行うことにより、最終的なビデオ分割結果を取得することができる）から構成されるシステムにより、本開示のインテリジェントビデオ分割の流れを実現することができる。マルチモーダル情報の融合により、ビデオコンテンツ及びイベント知識を深く理解し、ディープラーニングを組み合わせてビデオを分割する。図３に示すように、以下の内容を含む。

【００３４】

ビデオ分割モジュール

ビデオ分割モジュールにより、ビデオセグメントのきめ細かい分割を行うことができる。きめ細かい分割の原則は、主に、次の内容を含む。１）色調に基づく分割であり、２つのフレームのグレースケール差を直接に計算することができ、合計するフレーム差が設定されたある閾値より大きい場合、シーンの急変がある。２）エッジに基づく分割であり、エッジ特徴がシーンの分割に用いられることができ、まず、フレーム間の全体的な変位を計算し、これに基づいて位置合わせを行い、次に、エッジの数及び位置を計算する。３）ブロックマッチングに基づく分割であり、非圧縮ビデオに用いられるブロックマッチングシーンを例とする場合、動きの滑らかさの度量を用いてシーンの変化を検出することができる。４）統計的決定に基づく分割であり、動き補償特徴、適応閾値方式、ビデオシーケンスの時系列シーン急変モード及びシーン長さ分布情報を用い、統計的決定モデルを確立

し、当該統計的決定モデルが推定した基準により、シーン検出エラー率を最小限に抑えることができる。５）双閾値の比較に基づく分割であり、双閾値（例えば、Ｔｂ、Ｔｓ）を設定することができる。フレーム差がＴｂより大きい場合、シーンの急変があるが、フレーム差がＴｂより小さく且つＴｓより大きい場合、シーンの段階的な変化がある。接続するフレームのフレーム差がＴｓを超え始めると、このフレームは、シーンの段階的な変化の開始フレームと呼ばれ、これによって類推する。

【００３５】

二、マルチモーダル特徴符号化モジュール

図４は、本開示の実施形態による、マルチモーダル情報に基づいて特徴符号化を実現する模式図である。マルチモーダル特徴符号化は、主に、複数のエキスパートモデルにより若干のエキスパートベクトル（ｅｘｐｅｒｔｓ ｅｍｂｅｄｄｉｎｇ）を取得し、これらのｅｘｐｅｒｔｓ ｅｍｂｅｄｄｉｎｇによりビデオ全体のマルチモーダル情報の描画及び特徴抽出を完了する。図４に示すように、以下の内容を含む。

【００３６】

ビデオレベルベクトル（ｅｍｂｅｄｄｉｎｇ）表現は、マルチモーダルトランスフォーマー（ＭＭＴ、Ｍｕｌｔｉ－ｍｏｄｕａｌ Ｔｒａｎｓｆｏｒｍｅｒ）のようなマルチモーダルコーディングモジュールの１つの例により得られる。ＭＭＴは、Ｔｒａｎｓｆｏｒｍｅｒエンコーダーのアーキテクチャに従うことができ、Ｔｒａｎｓｆｏｒｍｅｒエンコーダーは、スタックされた自己注意機構（Ｓｅｌｆ－Ａｔｔｅｎｔｉｏｎ）層及び完全接続層で構成される。ＭＭＴの入力Ω（υ）は、ワンセットのｅｍｂｅｄｄｉｎｇ表現であり、全ての次元が同じであり、d_{model}次元として定義され、その中のそれぞれのｅｍｂｅｄｄｉｎｇは、何れもワンセットの特徴表現を表し、この入力は、式（１）に示されている。

【００３７】

【数１】

$$\Omega(\upsilon) = F(\upsilon) + M(\upsilon) + T(\upsilon) \quad (1)$$

【００３８】

ここで、式（１）の中のそれぞれのパラメータの意味は、次の通りである。

Ω（υ）は、ビデオフレーム特徴符号化の後のベクトル出力を表す。

【００３９】

Ｆ（υ）は、ビデオ特徴抽出モデルを表し、ビデオの中のマルチモーダル情報を抽出することに用いられる。ここで、F_{agg}^{kno}、F_1^{kno}、・・・F_k^{kno}は、ナレッジグラフ情報（例えば、ｋｎｏｗｌｅｄｇｅ）を抽出することを表し、「ｋｎｏ」は、ｋｎｏｗｌｅｄｇｅ特徴の略称であり、「ｋ」は、ベクトルの次元を表し、「ａｇｇ」は、平均ベクトルを表し、これによって類推し、テキスト情報（例えば、ｔｅｘｔ特徴）、オーディオ情報（例えば、ａｕｄｉｏ特徴）、色調情報（例えば、ＲＧＢ特徴）、物体情報（例えば、ｏｂｊｅｃｔ特徴）及び動作情報（例えば、ａｃｔｉｏｎ特徴）に対してそれぞれ抽出する。

【００４０】

Ｍ（υ）は、ビデオ特徴識別モデルを表し、マルチモーダル情報の中の異なるタイプの情報を区別することに用いられる。

【００４１】

Ｔ（υ）は、ビデオに対応する時系列情報の抽出モデルを表し、時系列情報（即ち、時系列ベクトル）を抽出及び記録することにより、特徴抽出の時間オフセット表現を記録す

るために用いられ、ここで、T_{agg}、T_1、・・・T_Dは、抽出する時系列情報を表し、「D」は、何秒を表す。

【００４２】

F（υ）により抽出された前記マルチモーダル情報の場合、ビデオデータの固有の異なる形式から効果的な表現を学習するために、様々なエキスパートモデルをビデオ特徴抽出器として用いることにより、前記マルチモーダル情報を抽出することができる。様々なエキスパートモデルは、ナレッジグラフに基づく構造化ラベルｅｍｂｅｄｄｉｎｇ抽出器、テキストに基づくｔｅｘｔ　ｅｍｂｅｄｄｉｎｇ抽出器、オーディオに基づくａｕｄｉｏ　ｅｍｂｅｄｄｉｎｇ抽出器、画像に基づくＲＧＢ抽出器、ターゲット検出に基づくｏｂｊｅｃｔ特徴抽出器及び動作識別に基づくａｃｔｉｏｎ　ｅｍｂｅｄｄｉｎｇ抽出器を主に含み、様々なエキスパートモデルがそれぞれ抽出された前記マルチモーダル情報に対して特徴融合を行うことにより、様々なコンテンツ次元でビデオ情報を完全に特徴付けることができる。特徴融合により、学習された、異なるエキスパートモデルにより抽出された前記マルチモーダル情報の間の関係に基づき、クロスモードとロングシーケンスの時間関係を利用して共同表現を行うことで、予めトレーニングされた異なるエキスパートモデル$\{F^n\}_{n=1}^{N}$を用いてより正確なビデオコンテンツの詳細を取得することができる。

【００４３】

それぞれのエキスパートモデルは、特別のタスクによるトレーニングによって取得されるものであり、その後、マルチモーダル情報の特徴抽出に用いられる。ビデオυに対し、それぞれのエキスパートモデルは、Ｋ個の特徴（ｆｅａｔｕｒｅｓ）を含むシーケンスを抽出することができ、$F^n(\upsilon)=[F_1^n,\cdots,F_k^n]$として表す。

【００４４】

様々なエキスパートモデルにより抽出されたビデオのｆｅａｔｕｒｅ特徴付けは、異なるエキスパートモデルを用いて特徴抽出を行うので、抽出された異なるエキスパートベクトルの特徴（又は、特徴ベクトルと呼ばれる）を共通のd_{model}次元にマッピングするために、Ｎ個のｌｉｎｅａｒ　ｌａｙｅｓ（それぞれのエキスパートごとに１つ）を用いて全ての特徴をR^{dmodel}に投影することができる。

【００４５】

Ｔｒａｎｓｆｏｒｍｅｒエンコーダーは、それぞれの特徴入力ごとに１つのｅｍｂｅｄｄｉｎｇを生成するので、複数の特徴に複数のｅｍｂｅｄｄｉｎｇ表現を提供する。それぞれの特徴の唯一のｅｍｂｅｄｄｉｎｇ表現を得るために、１つの纏めｅｍｂｅｄｄｉｎｇF_{agg}^nを定義することができ、当該ｅｍｂｅｄｄｉｎｇは、収集された特徴をコンテキスト化（現在の表現を集合して）し、最大プーリングにより当該埋め込みを初期化する：$F_{agg}^n=\mathrm{maxpool}(\{F_k^n\}_{k=1}^{N})$、よって、入力ｆｅａｔｕｒｅシーケンス全体の形式は、式（２）に示されている。

【００４６】

【数２】

$$\mathrm{F}(\upsilon)=\left[F_{agg}^1,F_1^1,\ldots,F_K^1\ldots,F_{agg}^N,F_1^N\ldots,F_K^N\right]\quad(2)$$

【００４７】

式（２）においては、Ｎは、エキスパートモデルの数（Ｎは、１より大きい正整数である）を表し、Ｋは、ベクトル次元（Ｋは、１より大きい正整数である）を表す。

【００４８】

Ｍ（υ）によりマルチモーダル情報の中の異なるタイプの情報を区別する場合、マルチモーダル情報をより良く処理及び区別するために、ＭＭＴは、現在処理しているｅｍｂｅｄｄｉｎｇ埋め込みがどのエキスパートモデルからの入力であるかを区別する必要があり、Ｎ個のd_{model}次元のｅｍｂｅｄｄｉｎｇ埋め込み$\{E_1,\cdots,E_n\}$を学習することにより、異

なるエキスパートのｅｍｂｅｄｄｉｎｇ表現を区別することができる。エキスパートモデルのｅｍｂｅｄｄｉｎｇからビデオエンコーダへのシーケンスは、次の式（３）に示す形式を採用することができる。

【００４９】

【数３】

$$\mathrm{E}(\upsilon) = [E^1, E^1, \ldots, E^1, \ldots, E^N, E^N, \ldots, E^N] \quad (3)$$

【００５０】

式（３）においては、Ｎは、エキスパートモデルの数（Ｎは、１より大きい正整数である）を表す。

【００５１】

Ｔ（υ）により時系列情報を記録する場合、ビデオのそれぞれの特徴が抽出したＭＭＴからの時間情報を提供する。１つのビデオの最長継続時間は、t_{max}秒であっても良く、「秒」を測定パラメータとし、$\{T_1,\cdots,T_D\}$のように、秒ごとに１つのd_{model}次元のＤ＝｜t_{max}｜を学習する。時間範囲［ｔ，ｔ＋１）から抽出されたそれぞれのエキスパートモードｆｅａｔｕｒｅをT_{t+1}として表す。例えば、ビデオにおいては、２．２秒で抽出された特徴は、時間埋め込みT_3により時間符号化される。２つの追加する時間埋め込みT_{agg}及びT_{unk}を学習し、集約特徴及び未知の時間情報特徴に対してそれぞれ符号化する。最後に、Ｔｅｍｐｏｒａｌ ｅｍｂｅｄｄｉｎｇｓＴは、次の式（４）に示す形式を採用する。

【００５２】

【数４】

$$\mathrm{T}(\upsilon) = [T_{agg}, T_1, \ldots, T_D \ldots., T_{agg}, T_1, \ldots, T_D] \quad (4)$$

【００５３】

式（４）においては、T_{agg}は、時間情報の平均ベクトルを表し、T_Dは、第Ｄ秒（Ｄは、１秒より大きい数値である）の時間情報を表す。

【００５４】

ＭＭＴの実現方式は、式（５）に示されている。

【００５５】

【数５】

$$\psi_{agg}(\upsilon) = MMT\big(\Omega(\upsilon)\big) = \left[\psi_{agg}^1, \ldots, \psi_{agg}^N\right] \quad (5)$$

【００５６】

式（５）においては、Ｎは、エキスパートモデルの数を表し、同時に、$\psi_{agg}(\upsilon)$は、ビデオ纏めの情報を表す。$\Omega(\upsilon)$は、ＭＭＴの入力を表す。

【００５７】

三、類似度マッチングモジュール

図５は、本開示の実施形態による類似度のマッチングの模式図である。類似度マッチングモジュールにより、２つの隣接するビデオセグメントの類似度の計算を主に完成し、類似度マッチングは、上下対称するモジュールの設計を採用する。複数のエキスパートモデルｅｍｂｅｄｉｎｇの類似性を計算し、重み計算（重みは、注意機構により自動的に学習することができる）を用いて類似度の採点を取得することにより、類似度のマッチング結果を取得する。また、損失関数は、双方向の最大結合ランキング損失関数（ｂｉ－ｄｉｒ

ｅｃｔｉｏｎａｌ ｍａｘ－ｍａｒｇｉｎ ｒａｎｋｉｎｇ ｌｏｓｓ）を採用でき、式（６）に示されている。

【００５８】

【数６】

$$\mathcal{L} = \frac{1}{B}\sum_{i=1}^{B}\sum_{j\neq i}[\max(0, s_{ij} - s_{ii} + m) + \max(0, s_{ji} - s_{ii} + m)] \quad (6)$$

【００５９】

式（６）においては、Ｌは、前記損失関数を表し、Ｂは、サンプルバッチ処理に用いられるハイパーパラメータ（ｂａｔｃｈ ｓｉｚｅ）を表し、s_{ij}=similarity(v_i,v_j)であり、s_{ij}は、２つのビデオセグメントの類似度を表し、ｍは、ｍａｒｇｉｎであり、値（０，１）を取ることができる。

【００６０】

四、類似度マッチングモジュール又は上述した図３を元に、類似度マッチングモジュールの後に、ビデオスプライシング処理に特別に用いられるビデオフレームスプライシングモジュールを追加する。

【００６１】

図６は、本開示の実施形態によるビデオ結合の模式図であり、当該ビデオスプライシング処理を統合する類似度マッチングモジュールを例とする。図６に示すように、類似度マッチングモジュールにより、隣接するビデオセグメントの結合とスプライシングを実現することができる。主に、同じイベント内の細かいビデオセグメントを復元する。２つの隣接するビデオセグメントが類似していると判断された場合、２つのビデオを結合し、順に比較し、最終的なビデオ分割結果を取得する。

【００６２】

本応用例によれば、マルチモーダル情報を抽出するための複数のエキスパートモデルの情報を融合し、マルチモーダル情報をキャプチャ及び融合することができるので、ビデオコンテンツ全体を完全に描画し、ビデオ画面の再現効果を高めることができる。ディープラーニングの方式により、大規模な且つ大量のビデオ分割を行うことができ、プロセス全体がより効率であり、コストがより低く、高適時性のビデオ要求を満たすことができる。ＫＧに基づく構造化ラベル技術（例えば、実体、主題等である）、テキストに基づく表現技術、ビジョンに基づく（ＲＧＢ、Ｏｂｊｅｃｔ Ａｃｔｉｏｎ）等を結合し、ビデオコンテンツの角度からビデオを分割し、複数のシーンの頻繁な切り替えによる分割効果の悪い問題を解決することができる。また、スケーラビリティが良く、使用シナリオは、ビデオ技術に限らず、ビデオ指紋識別、ビデオ短帯域長、同じビデオマッチング等のような、任意のビデオの類似度のマッチングシナリオに適用できる。

【００６３】

本開示の実施形態によれば、ビデオ処理装置が提供される。図７は、本開示の実施形態によるビデオ処理装置の構成構造模式図である。図７に示すように、当該ビデオ処理装置は、分割モジュール４１、符号化モジュール４２及びビデオ処理モジュール４３を備える。

【００６４】

分割モジュール４１は、複数の第１ビデオフレームを取得し、前記複数の第１ビデオフレームに対してきめ細かい分割を行い、複数の第２ビデオフレームを取得することに用いられる。

【００６５】

符号化モジュール４２は、前記複数の第２ビデオフレームに関するマルチモーダル情報に基づき、前記複数の第２ビデオフレームに対して特徴符号化を行い、前記マルチモーダル情報の融合を特徴付けるための特徴融合情報を取得することに用いられる。

【００６６】

ビデオ処理モジュール４３は、前記特徴融合情報に基づき、前記複数の第２ビデオフレ

ームに対して類似度のマッチングを行い、類似度のマッチング結果に基づいてターゲットビデオを取得することに用いられる。

【００６７】

１つの実施形態においては、前記分割モジュールは、シーン及び色彩転換を特徴付けるためのパラメータに基づき、前記複数の第１ビデオフレームに対してきめ細かい分割を行い、複数の第２ビデオフレームを取得することに用いられる。

【００６８】

１つの実施形態においては、前記符号化モジュールは、前記マルチモーダル情報に基づき、前記複数の第２ビデオフレームに対して特徴抽出及び特徴融合の処理を行い、前記特徴融合情報を取得することに用いられる。

【００６９】

１つの実施形態においては、前記特徴融合情報に基づき、前記複数の第２ビデオフレームの類似度に対して採点し、採点結果を前記類似度のマッチング結果とし、前記類似度のマッチング結果として、同じイベントコンテンツに関する、隣接するビデオフレームが類似している場合、前記複数の第２ビデオフレームのそれぞれに対して、隣接するビデオフレームに対する結合がされるまで、前記隣接するビデオフレームに対してビデオ結合を行い、ビデオ結合の結果に基づいて前記ターゲットビデオを取得することに用いられる。

【００７０】

１つの実施形態においては、予めトレーニングされた第１ニューラルネットワークモデルにより、前記複数の第２ビデオフレームから前記マルチモーダル情報を識別するための識別モジュールを更に備える。

【００７１】

１つの実施形態においては、前記識別モジュールは、前記第１ニューラルネットワークモデルの中のナレッジグラフ抽出器により、ナレッジグラフ情報を識別し、前記第１ニューラルネットワークモデルの中のテキスト抽出器により、テキスト情報を識別し、前記第１ニューラルネットワークモデルの中のオーディオ抽出器により、オーディオ情報を識別し、前記第１ニューラルネットワークモデルの中の色調抽出器により、色調情報を識別し、前記第１ニューラルネットワークモデルの中の物体抽出器により、物体情報を識別し、前記第１ニューラルネットワークモデルの中の動作抽出器により、動作情報を識別することに用いられる。前記マルチモーダル情報は、前記ナレッジグラフ情報、前記テキスト情報、前記オーディオ情報、前記色調情報、前記物体情報、前記動作情報の中の少なくとも１つを含む。

【００７２】

１つの実施形態においては、第２ニューラルネットワークモデルにより、前記マルチモーダル情報の中の各種類の情報に対して区別を行い、第３ニューラルネットワークモデルにより、前記マルチモーダル情報に関する時系列情報に対して識別を行い、前記第１ニューラルネットワークモデル、前記第２ニューラルネットワークモデル、前記第３ニューラルネットワークモデルの出力結果に対して融合を行い、前記特徴融合情報を取得するための融合モジュールを更に備える。

【００７３】

本開示の実施形態におけるそれぞれの装置の中の各モジュールの機能は、上述した方法の対応する記載を参照することができ、ここでは、繰り返して説明しない。

【００７４】

本開示の実施形態によれば、本開示は、電子デバイス及び可読記憶媒体が更に提供される。

図８は、本開示の実施形態による例示するビデオ処理方法を実現するための電子デイバスのブロック図である。当該デバイスは、上述した展開デバイス又はエージェントデバイスであっても良い。電子デバイスは、ラップトップコンピュータ、デスクトップコンピュータ、ワークステーション、パーソナルデジタルアシスタント、サーバ、ブレードサーバ、メインフレームコンピュータのような様々な形態のデジタルコンピュータ及び他の好適

なコンピュータを表すことを目的としている。また、電子デバイスは、また、様々な形態のモバイルデバイス、例えば、パーソナルデジタルアシスタント、携帯電話、スマートフォン、ウェアラブルデバイス、及び他の類似のコンピューティングデバイスを表すことができる。本明細書に記載のコンポーネント、それらの接続及び関係、ならびにそれらの機能は、例としてのみ意図されており、本明細書に記載及び／又は要求される本開示の実現を限定することを意図するものではない。

【００７５】

当該電子デバイスは、１つ以上のプロセッサ８０１、メモリ８０２、及び各コンポーネントを接続するための、高速インターフェース及び低速インターフェースを含むインターフェースを有する。様々なコンポーネントは、異なるバスを用いて相互に接続されており、共通のマザーボード上に実装されてもよいし、必要に応じて他の方式で実装されてもよい。プロセッサは、電子デバイス内で実行される命令を処理してもよく、当該命令は、メモリに又はメモリ上に記憶されることによって、外部入出力装置（例えば、インターフェースに結合されたディスプレイ装置）にＧＵＩのグラフィカル情報を表示させるための命令を含む。他の実施形態では、必要に応じて、複数のプロセッサ及び／又は複数のバスを複数のメモリと一緒に使用してもよい。同様に、複数の電子デバイスが接続されていてもよく、個々のデバイスが必要な操作の一部を提供する（例えば、サーバアレイ、ブレードサーバのグループ又はマルチプロセッサシステムとして）。図８は、一つのプロセッサ８０１を例としている。

【００７６】

メモリ８０２は、本開示によって提供される非一時的なコンピュータ可読記憶媒体である。ここで、前記メモリは、本開示により提供されるビデオ処理方法を前記少なくとも１つのプロセッサに実行させるために、前記少なくとも１つのプロセッサにより実行可能な命令を記憶している。本開示の非一時的なコンピュータ可読記憶媒体は、本開示によって提供されるタッチコマンドの処理方法をコンピュータに実行させるために使用されるコンピュータ命令を記憶している。

【００７７】

メモリ８０２は、非一時的なコンピュータ可読記憶媒体として、非一時的なソフトウェアプログラム、非一時的なコンピュータ実行可能プログラム及びモジュール、例えば、本開示の実施形態におけるビデオ処理方法に対応するプログラム命令／モジュール（例えば、図７に示す分割モジュール、符号化モジュール、ビデオ処理モジュール等のモジュール）を格納するために使用することができる。プロセッサ８０１は、メモリ８０２に記憶された非一時的なソフトウェアプログラム、命令及びモジュールを実行することにより、サーバの各種機能アプリケーション及びデータ処理を実行し、上述した方法の実施形態におけるビデオ処理方法を実現する。

【００７８】

メモリ８０２は、プログラム記憶領域とデータ記憶領域とを含んでもよく、プログラム記憶領域は、オペレーティングシステム、少なくとも１つの機能に必要なアプリケーションプログラムを格納してもよく、データ記憶領域は、タッチコマンドの処理方法の電子デバイスの使用により作成されたデータなどを格納してもよい。また、メモリ８０２は、高速ランダムアクセスメモリを含んでもよく、少なくとも１つのディスクメモリ装置、フラッシュメモリ装置又は他の非一時的なソリッドステートメモリ装置などの非一時的なメモリを更に含んでもよい。幾つかの実施形態では、メモリ８０２は、プロセッサ８０１に対して相対的に遠隔に配置されたメモリを含むことが好ましく、これらの遠隔メモリは、ネットワークを介して、タッチコマンドの処理方法の電子デバイスに接続されてもよい。前記ネットワークの例としては、インターネット、企業のイントラネット、ローカルエリアネットワーク、移動体通信ネットワーク及びそれらの組合せが挙げられるが、これらに限定されない。

【００７９】

ビデオ処理方法の電子デバイスは、入力装置８０３と出力装置８０４を更に含んでもよ

い。プロセッサ８０１、メモリ８０２、入力装置８０３及び出力装置８０４は、バスを介して接続されていてもよく、他の方式で接続されていてもよく、図８ではバスを介した接続を例に挙げている。

【００８０】

入力装置８０３は、入力された数値情報又は文字情報を受信するとともにタッチコマンドの処理方法の電子デバイスのユーザ設定及び機能制御に関連するキー信号入力を生成してもよく、例えば、タッチスクリーン、キーパッド、マウス、トラックパッド、タッチパッド、インジケータスティック、１つ以上のマウスボタン、トラックボール、ジョイスティック、その他の入力装置などが挙げられる。出力装置８０４は、表示装置、補助照明装置（例えば、ＬＥＤ）、触覚フィードバック装置（例えば、振動モータ）等を含んでもよい。当該表示装置としては、液晶ディスプレイ（Ｌｉｑｕｉｄ Ｃｒｙｓｔａｌ Ｄｉｓｐｌａｙ、ＬＣＤ）、発光ダイオード（Ｌｉｇｈｔ Ｅｍｉｔｔｉｎｇ Ｄｉｏｄｅ、ＬＥＤ）、プラズマディスプレイ等が挙げられるが、これらに限定されない。幾つかの実施形態では、表示装置は、タッチスクリーンであってもよい。

【００８１】

本明細書に記載のシステム及び技術の様々な実施形態は、デジタル電子回路システム、集積回路システム、専用集積回路（Ａｐｐｌｉｃａｔｉｏｎ Ｓｐｅｃｉｆｉｃ Ｉｎｔｅｇｒａｔｅｄ Ｃｉｒｃｕｉｔｓ、ＡＳＩＣ）、コンピュータハードウェア、ファームウェア、ソフトウェア、及び／又はそれらの組合せで実現することができる。これらの様々な実施形態は、以下を含み得る。１つ以上のコンピュータプログラムで実施し、当該１つ以上のコンピュータプログラムは、少なくとも１つのプログラマブルプロセッサを含むプログラマブルシステム上で実行及び／又は解釈され、当該プログラマブルプロセッサは、記憶システム、少なくとも１つの入力装置、及び少なくとも１つの出力装置からデータ及び命令を受信し、且つデータ及び命令を当該記憶システム、当該少なくとも１つの入力装置、及び当該少なくとも１つの出力装置へ転送することができる専用又は汎用のプログラマブルプロセッサであってもよい。

【００８２】

これらのコンピュータプログラムは、プログラム、ソフトウェア、ソフトウェアアプリケーション、又はコードとも呼ばれ、プログラマブルプロセッサのための機械命令を含み、高レベル手順及び／又はオブジェクト指向のプログラミング言語、及び／又はアセンブリ／機械語を使用してこれらのコンピュータプログラムを実装することができる。本明細書で使用されるように、「機械可読媒体」及び「コンピュータ可読媒体」という用語は、機械命令及び／又はデータをプログラマブルプロセッサに提供するために使用される任意のコンピュータプログラム製品、デバイス、及び／又は装置、例えば、磁気ディスク、光ディスク、メモリ、プログラマブルロジックデバイス（ｐｒｏｇｒａｍｍａｂｌｅ ｌｏｇｉｃ ｄｅｖｉｃｅ、ＰＬＤ）を指し、機械可読信号である機械命令を受け取る機械可読媒体を含む。「機械可読信号」という用語は、機械命令及び／又はデータをプログラマブルプロセッサに提供するために使用される任意の信号を指す。

【００８３】

ユーザとのインタラクティブを提供するために、本明細書に記載されているシステム及び技術は、ユーザに情報を表示するための表示装置（例えば、ＣＲＴ（Ｃａｔｈｏｄｅ Ｒａｙ Ｔｕｂｅ、陰極線管）又はＬＣＤ（液晶ディスプレイ）モニタ）と、ユーザがコンピュータに入力を提供するためのキーボード及びポインティング装置（例えば、マウス又はトラックボール）とを有するコンピュータ上に実装されてもよい。他の種類の装置もユーザとのインタラクティブを提供するためにも使用されてもよく、例えば、ユーザに提供されるフィードバックは、任意の形態の感覚フィードバック（例えば、視覚フィードバック、聴覚フィードバック、又は触覚フィードバック）であってもよく、ユーザからの入力は、任意の形態（音響入力、音声入力、又は触覚入力を含む）で受信されてもよい。

【００８４】

本明細書に記載されているシステム及び技術は、バックエンドコンポーネントを含むコ

ンピューティングシステム（例えば、データサーバー）、ミドルウェアコンポーネントを含むコンピューティングシステム（例えば、アプリケーションサーバー）、又はフロントエンドコンポーネントを含むコンピューティングシステム（例えば、グラフィカルユーザインターフェース又はウェブブラウザーを備えたユーザコンピューター。当該グラフィカルユーザインターフェース又は当該ウェブブラウザーを介して、ユーザはここで説明するシステム及び技術の実装とインタラクティブできる）、又はそのようなバックエンドコンポーネント、ミドルウェアコンポーネント、又はフロントエンドコンポーネントの任意の組合せを含むコンピューティングシステムで実装されてもよい。システムのコンポーネントは、任意の形態又は媒体のデジタルデータ通信（例えば、通信ネットワーク）を介して相互に接続されていてもよい。通信ネットワークの例としては、ローカルエリアネットワーク（Ｌｏｃａｌ　Ａｒｅａ　Ｎｅｔｗｏｒｋ、ＬＡＮ）、ワイドエリアネットワーク（Ｗｉｄｅ　Ａｒｅａ　Ｎｅｔｗｏｒｋ、ＷＡＮ）及びインターネット等がある。

【００８５】

コンピュータシステムは、クライアントとサーバを含むことができる。クライアントとサーバは一般的に互いに遠隔地にあり、通常は、通信ネットワークを介してインタラクティブする。クライアント－サーバ関係は、対応するコンピュータ上で実行され、互いにクライアント－サーバ関係を有するコンピュータプログラムによって生成される。サーバは、クラウドコンピューティングサーバ又はクラウドホストとも呼ばれるクラウドサーバであっても良く、クラウドコンピューティングサービスシステムのホスト製品であり、従来の物理ホスト及び仮想プライベートサーバー（ＶＰＳ）サービスにおける管理困難の問題及び業務拡大性が弱いという欠陥を解決する。サーバは、分散システムのサーバであっても良く、ブロックチェーンと組み合わせたサーバであっても良い。

【００８６】

本開示によれば、複数の第１ビデオフレームを取得し、前記複数の第１ビデオフレームに対してきめ細かい分割を行い、複数の第２ビデオフレームを取得し、前記複数の第２ビデオフレームに関するマルチモーダル情報に基づき、前記複数の第２ビデオフレームに対して特徴符号化を行い、前記マルチモーダル情報の融合を特徴付けるための特徴融合情報を取得し、前記特徴融合情報に基づき、前記複数の第２ビデオフレームに対して類似度のマッチングを行い、類似度のマッチング結果に基づいてターゲットビデオを取得する。本開示によれば、当該マルチモーダル情報に基づいて特徴符号化を行うことができるので、より多くのビデオコンテンツの詳細を含む情報を取得することができ、類似度のマッチングがされた後、取得されたターゲットビデオがより正確であり、ビデオ分割の正確性を高めることができる。

【００８７】

上述した処理の様々なプロセスを用い、順序を変えたり、ステップを追加又は削除したりすることができることが理解されるべきである。例えば、本開示に記載の各ステップは、並行して実行されてもよく、順次実行されてもよく、異なる順序で実行されてもよく、本開示に開示された技術案の所望の結果が達成される限り、限定されない。

【００８８】

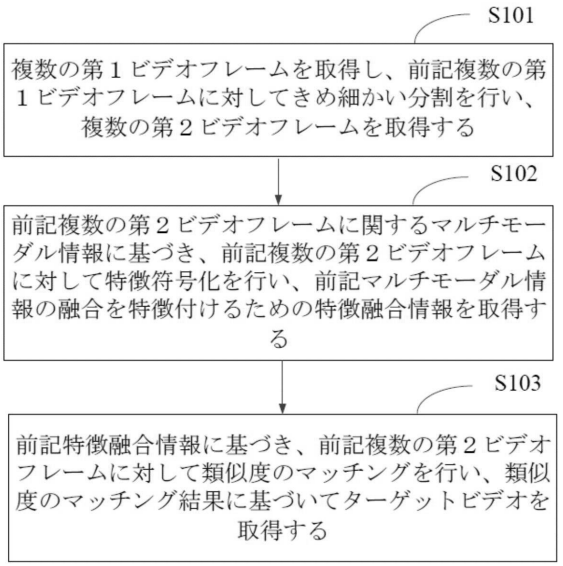
上記の具体的な実施形態は、本開示の保護範囲の制限を構成するものではない。設計要件及び他の要因に応じて、様々な変更、組み合わせ、サブ組み合わせ及び置換えが行われ得ることは、当業者によって理解されるべきである。本開示の要旨及び原則の範囲内で行われる如何なる修正、同等の代替、改良等は、すべて本開示の保護範囲に含まれる。

【図面】

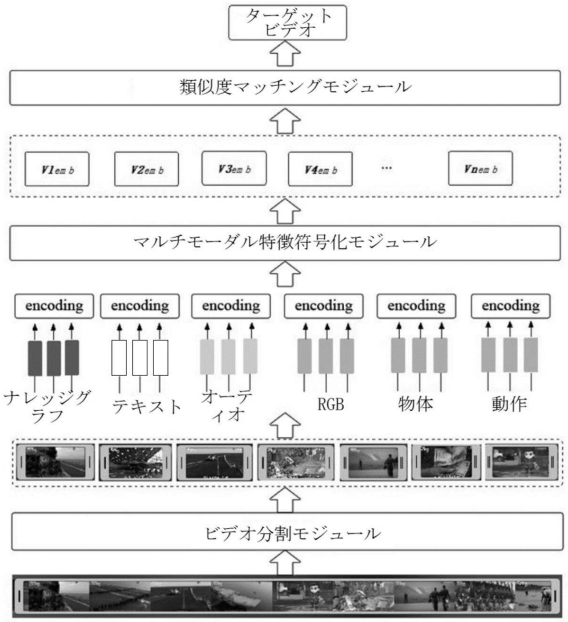
【図１】


継続する他のビデオフレーム

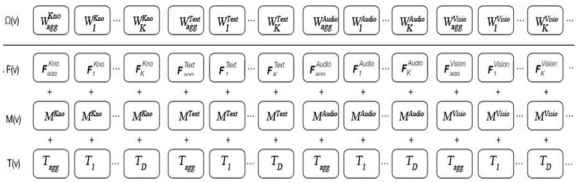
【図２】


S101
複数の第１ビデオフレームを取得し、前記複数の第１ビデオフレームに対してきめ細かい分割を行い、複数の第２ビデオフレームを取得する
S102
前記複数の第２ビデオフレームに関するマルチモーダル情報に基づき、前記複数の第２ビデオフレームに対して特徴符号化を行い、前記マルチモーダル情報の融合を特徴付けるための特徴融合情報を取得する
S103
前記特徴融合情報に基づき、前記複数の第２ビデオフレームに対して類似度のマッチングを行い、類似度のマッチング結果に基づいてターゲットビデオを取得する

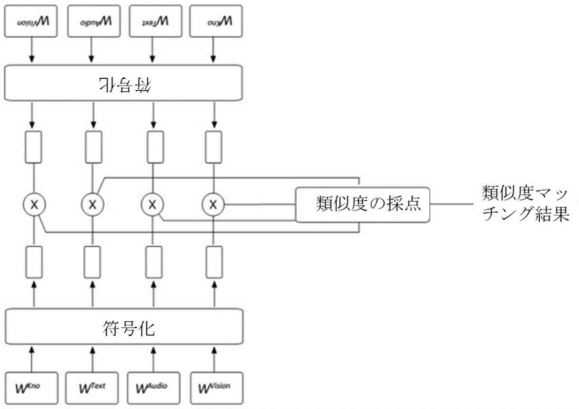
【図３】


ターゲットビデオ
類似度マッチングモジュール
マルチモーダル特徴符号化モジュール
encoding
encoding
encoding
encoding
encoding
encoding
ナレッジグラフ
テキスト
オーディオ
RGB
物体
動作
ビデオ分割モジュール

【図４】



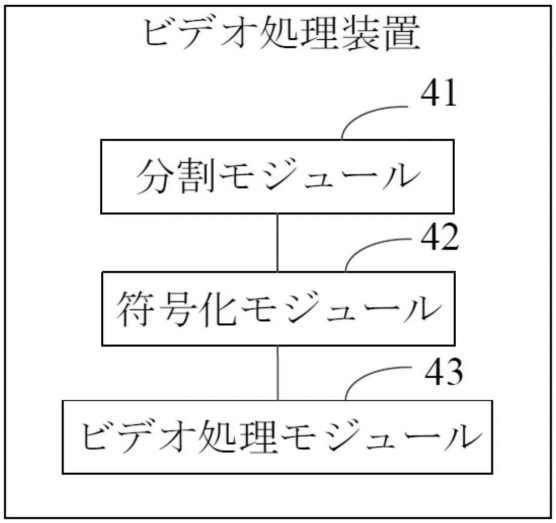
【図５】


符号化
類似度の採点
類似度マッチング結果
符号化

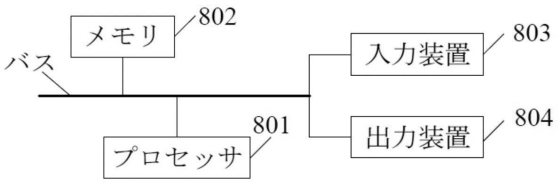
【図６】


最終結果
セグメント１２３
セグメント４
セグメント５６
セグメント７８
第４回の比較
セグメント１２３
セグメント４
セグメント５６
セグメント７
セグメント８
第３回の比較
セグメント１２３
セグメント４
セグメント５
セグメント６
セグメント７
セグメント８
第２回の比較
セグメント１２
セグメント３
セグメント４
セグメント５
セグメント６
セグメント７
セグメント８
第１回の比較
セグメント１
セグメント２
セグメント３
セグメント４
セグメント５
セグメント６
セグメント７
セグメント８

【図７】


ビデオ処理装置
41
分割モジュール
42
符号化モジュール
43
ビデオ処理モジュール

【図８】


メモリ
802
バス
入力装置
803
801
プロセッサ
出力装置
804

フロントページの続き

８５， Ｃｈｉｎａ
(74)代理人 110001519
弁理士法人太陽国際特許事務所
(72)発明者 チー ワン
中華人民共和国 ペキン １０００８５， ハイディアン ディストリクト， シャンディ テンス ストリート， １０番， バイドゥ キャンパス ２階
(72)発明者 チーファン フォン
中華人民共和国 ペキン １０００８５， ハイディアン ディストリクト， シャンディ テンス ストリート， １０番， バイドゥ キャンパス ２階
(72)発明者 フー ヤン
中華人民共和国 ペキン １０００８５， ハイディアン ディストリクト， シャンディ テンス ストリート， １０番， バイドゥ キャンパス ２階
(72)発明者 チュンコアン チャイ
中華人民共和国 ペキン １０００８５， ハイディアン ディストリクト， シャンディ テンス ストリート， １０番， バイドゥ キャンパス ２階
審査官 山田 辰美

(56)参考文献
特開平０９－０４４６３９（ＪＰ，Ａ）
特開２０１１－１２４６８１（ＪＰ，Ａ）
国際公開第２００７／０５２３９５（WO，Ａ１）
丹野 良介，マルチモーダル深層学習によるドライブレコーダーデータの分類，映像情報メディア学会誌 第７４巻 第１号，日本，一般社団法人映像情報メディア学会，2020年，第７４巻 第１号，p.44－p.48

(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ０６Ｔ ７／００
Ｇ０６Ｆ １６／７８３
Ｇ０６Ｆ １６／７３２