(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3848319号
(P3848319)
(45)発行日　平成18年11月22日(2006.11.22)　(24)登録日　平成18年9月1日(2006.9.1)

(51)Int.Cl.　　F I
H04N 5/91 (2006.01)　　H04N 5/91 R
G10L 15/00 (2006.01)　　G10L 15/00 200G

請求項の数 25　(全 42 頁)

(21)出願番号　特願2003-381637(P2003-381637)
(22)出願日　平成15年11月11日(2003.11.11)
(65)公開番号　特開2005-150841(P2005-150841A)
(43)公開日　平成17年6月9日(2005.6.9)
審査請求日　平成15年11月11日(2003.11.11)

(73)特許権者　000001007
キヤノン株式会社
東京都大田区下丸子３丁目３０番２号
(74)代理人　100076428
弁理士　大塚　康徳
(74)代理人　100112508
弁理士　高柳　司郎
(74)代理人　100115071
弁理士　大塚　康弘
(74)代理人　100116894
弁理士　木村　秀二
(72)発明者　深田　俊明
東京都大田区下丸子３丁目３０番２号　キヤノン株式会社内

最終頁に続く

(54)【発明の名称】情報処理方法及び情報処理装置

(57)【特許請求の範囲】
【請求項１】
画像データと音声データを対応付ける情報処理方法であって、
文字を含む前記画像データから文字領域を検出する検出工程と、
検出された前記文字領域に含まれる文字情報を認識して複数の文字情報を取得する第１の認識工程と、
前記音声データ中の複数の音声区間のそれぞれに対応する音声認識結果を取得する第２の認識工程と、
前記第１の認識工程により取得された前記複数の文字情報と前記第２の認識工程により取得された前記複数の音声認識結果とをそれぞれ照合することによって、該文字情報と該音声認識結果を対応付ける対応付け工程と
を有することを特徴とする情報処理方法。
【請求項２】
前記対応付け工程は、前記複数の文字情報と前記複数の音声認識結果の対応付けに基づいて、該文字情報に対応する前記文字領域と該音声認識結果、該文字情報と該音声認識結果に対応する音声区間、又は該文字情報に対応する前記文字領域と該音声認識結果に対応する音声区間のいずれかを対応付けることを特徴とする請求項１記載の情報処理方法。
【請求項３】
前記文字情報を発音列に変換する発音列変換工程をさらに有し、
前記対応付け工程は、前記文字情報に基づく発音列と前記音声認識結果の発音列との照

合結果に基づいて前記文字情報と前記音声認識結果を対応付ける
ことを特徴とする請求項１に記載の情報処理方法。
【請求項４】
前記音声認識結果を文字列に変換する文字列変換工程をさらに有し、
前記対応付け工程は、前記文字情報の文字列と前記音声認識結果に基づく文字列との照合結果に基づいて前記文字情報と前記音声認識結果を対応付ける
ことを特徴とする請求項１に記載の情報処理方法。
【請求項５】
前記文字情報を発音列に変換する発音列変換工程と、
前記音声認識結果を文字列に変換する文字列変換工程をさらに有し、
前記対応付け工程は、前記文字情報に基づく発音列と前記音声認識結果の発音列との照合結果と、前記文字情報の文字列と前記音声認識結果に基づく文字列との照合結果とに基づいて、前記文字情報と前記音声認識結果を対応付ける
ことを特徴とする請求項１に記載の情報処理方法。
【請求項６】
前記第１の認識工程が、前記複数の文字情報のそれぞれについて、候補と該候補の度合いを取得し、
前記第２の認識工程が、前記複数の音声認識結果のそれぞれについて、候補と該候補の度合いを取得し、
前記文字情報の候補の度合いと前記音声認識結果の候補の度合いとに基づいて、それぞれの候補間の関連の度合いを算出する算出工程とをさらに有し、
前記対応付け工程が、前記関連の度合いの高さに応じて、前記文字情報の候補と前記音声認識結果の候補とを対応付ける
ことを特徴とする請求項１に記載の情報処理方法。
【請求項７】
前記候補の度合いは、前記候補の認識確率又は認識尤度であることを特徴とする請求項６に記載の情報処理方法。
【請求項８】
前記算出工程が、前記文字情報の候補又は前記音声認識結果の候補に重み付けを付与して前記候補間の関連の度合いを算出することを特徴とする請求項６に記載の情報処理方法。
【請求項９】
前記第２の認識工程は、前記音声データ中の複数の音声区間それぞれに対応する音声認識結果を文字列に変換し、該音声認識結果の文字列中の前記第１の認識工程で取得した文字情報に含まれない文字列を除外したものを、前記音声認識結果として取得することを特徴とする請求項１記載の情報処理方法。
【請求項１０】
前記第２の認識工程は、前記第１の認識工程で取得した文字情報を音声認識対象として音声認識を行い、前記音声認識結果を取得することを特徴とする請求項１記載の情報処理方法。
【請求項１１】
前記第２の認識工程は、前記音声データ中の複数の音声区間それぞれに対応する音声認識結果から、前記第１の認識工程で取得した文字情報を発音列に変換したものに含まれない音声認識結果を除外したものを、前記音声認識結果として取得することを特徴とする請求項１記載の情報処理方法。
【請求項１２】
前記第２の認識工程は、前記第１の認識工程で取得した文字情報を発音列に変換したものを音声認識対象として音声認識を行い、前記音声認識結果を取得することを特徴とする請求項１記載の情報処理方法。
【請求項１３】

少なくとも自立語を含む重要語を抽出するためのデータに基づいて、前記文字情報に含まれる重要語を抽出する重要語抽出工程をさらに有し、
前記第２の認識工程は、前記重要語をキーワードスポッティングの対象とするか、前記重要語の音声認識用言語モデルの確率値を増加させるかの少なくともいずれかを行って音声認識を行うことで、前記音声認識結果を取得する
ことを特徴とする請求項１に記載の情報処理方法。
【請求項１４】
前記第１の認識工程により認識された前記文字情報について、該文字情報のフォントサイズ、色、アンダーライン、太字、斜体、又はフォント種の少なくとも何れか１つを含むフォント情報を抽出するフォント情報抽出工程をさらに有し、
前記第２の認識工程が、前記フォント情報を利用して特定された文字列をキーワードスポッティングの対象するか、特定された文字列の統計的言語モデルの確率値を増加させるかの少なくともいずれかを行って音声認識を行い、前記音声認識結果を取得する
ことを特徴とする請求項１に記載の情報処理方法。
【請求項１５】
前記第１の認識工程は、前記第２の認識工程により取得された音声認識結果を文字列に変換したものに含まれない文字列を、前記検出された文字領域に含まれる文字情報を認識して取得した複数の文字情報から除外したものを、前記複数の文字情報として取得することを特徴とする請求項１記載の情報処理方法。
【請求項１６】
前記第１の認識工程は、前記第２の認識工程により取得された音声認識結果を文字列に変換したものを文字認識対象として文字認識を行い、前記文字情報を取得することを特徴とする請求項１記載の情報処理方法。
【請求項１７】
前記画像データを複数の領域に分割して分割画像を取得する画像分割工程をさらに有し、
それぞれの分割画像に関して文字情報を認識する
ことを特徴とする請求項１に記載の情報処理方法。
【請求項１８】
前記第１の認識工程により認識された前記文字情報を文字概念表現に変換する文字概念変換工程と、
前記第２の認識工程により認識された前記音声認識結果を音声概念表現に変換する音声概念変換工程と、
前記文字概念表現と前記音声概念表現とを照合する概念対応工程とをさらに有し、
前記対応付け工程が、前記概念対応工程によって得られる概念間の照合結果に基づいて、前記文字情報と前記音声認識結果とを対応付ける
ことを特徴とする請求項１に記載の情報処理方法。
【請求項１９】
画像データと音声データを対応付ける情報処理方法であって、
前記画像データに含まれるオブジェクト領域を検出する第１の検出工程と、
検出された前記オブジェクト領域からオブジェクト情報を認識する第１の認識工程と、
前記音声データ中の複数の音声区間のそれぞれに対応する音声認識結果を取得する第２の認識工程と、
前記第１の認識工程により認識された前記オブジェクト情報の特徴情報に対応する文字情報と前記第２の認識工程により認識された前記音声認識結果とを照合することによって該オブジェクト情報と該音声認識結果とを対応付ける対応付け工程と
を有することを特徴とする情報処理方法。
【請求項２０】
前記オブジェクト情報は図形情報であり、
前記オブジェクト情報の特徴情報は、前記図形情報の形状、色の少なくともいずれかで

あることを特徴とする請求項１９記載の情報処理方法。
【請求項２１】
画像データと音声データを対応付ける情報処理方法であって、
前記画像データに含まれる人物領域を検出する第１の検出工程と、
検出された前記人物領域から人物又は少なくとも人物の性別、年代のいずれかを含む人物のクラスを認識する第１の認識工程と、
前記音声データ中の複数の音声区間それぞれに対応する話者又は少なくとも話者の性別、年代のいずれかを含む話者クラスを認識する第２の認識工程と、
前記第１の認識工程により認識された人物又は人物のクラスと、前記第２の認識工程により認識された話者又は話者クラスとを対応付ける対応付け工程と
を有することを特徴とする情報処理方法。
【請求項２２】
請求項１乃至２１に記載の情報処理方法をコンピュータに実行させるための制御プログラム。
【請求項２３】
画像データと音声データを対応付ける情報処理装置であって、
文字を含む前記画像データから文字領域を検出する検出手段と、
検出された前記文字領域に含まれる文字情報を認識して複数の文字情報を取得する第１の認識手段と、
前記音声データ中の複数の音声区間それぞれに対応する音声認識結果を取得する第２の認識手段と、
前記第１の認識手段により取得された前記複数の文字情報と前記第２の認識手段により取得された前記複数の音声認識結果とをそれぞれ照合することによって、該文字情報と該音声認識結果を対応付ける対応付け手段と
を有することを特徴とする情報処理装置。
【請求項２４】
画像データと音声データを対応付ける情報処理方法であって、
前記画像データに含まれるオブジェクト領域を検出する第１の検出手段と、
検出された前記オブジェクト領域からオブジェクト情報を認識する第１の認識手段と、
前記音声データ中の複数の音声区間それぞれに対応する音声認識結果を取得する第２の認識手段と、
前記第１の認識手段により認識された前記オブジェクト情報の特徴情報に対応する文字情報と前記第２の認識手段により認識された前記音声認識結果とを照合することによって該オブジェクト情報と該音声認識結果とを対応付ける対応付け手段と
を有することを特徴とする情報処理装置。
【請求項２５】
画像データと音声データを対応付ける情報処理方法であって、
前記画像データに含まれる人物領域を検出する第１の検出手段と、
検出された前記人物領域から人物又は少なくとも人物の性別、年代のいずれかを含む人物のクラスを認識する第１の認識手段と、
前記音声データ中の複数の音声区間それぞれに対応する話者又は少なくとも話者の性別、年代のいずれかを含む話者クラスを認識する第２の認識手段と、
前記第１の認識手段により認識された人物又は人物のクラスと、前記第２の認識手段により認識された話者又は話者クラスとを対応付ける対応付け工程と
を有することを特徴とする情報処理装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、画像データと音声データを対応付ける情報処理方法及び情報処理装置に関する。

【背景技術】
【０００２】
近年、デジタルカメラで静止画像を撮影するとともに、撮影された当該静止画像に対して音声メモ機能を用いて当該静止画像についてのコメント等を録音するといった、画像データと音声データを関連付ける技術が開発されている。例えば、Ｅｘｉｆ（EXchangeable Image File Format）と呼ばれるデジタルカメラ用画像ファイルの標準規格では、１枚の静止画像ファイルの中に付属情報として音声データを関連付けることができる。このようにして静止画像に関連付けられた音声データは、単に静止画像に音声データが付与されたものであるだけではなく、当該音声データを音声認識することによって付与された音声データを認識して文字情報に変換し、文字又は音声をキーとして、複数の静止画像の中から所望の静止画像を検索するといった目的に利用することができる。
【０００３】
また、ボイスレコーダー機能を搭載したデジタルカメラや、デジタルカメラ機能を搭載したボイスレコーダーでは、最大で数時間程度の音声データを録音することが可能である。
【発明の開示】
【発明が解決しようとする課題】
【０００４】
しかしながら、上述したような従来の技術は、１枚の静止画像全体に対して１つ又は複数の音声データを関連付けるに留まっており、１枚の静止画像中の特定の部分領域と、それに対する音声データ中の特定の音声区間とを関連付ける技術ではない。すなわち、デジタルカメラで撮影された静止画像中の部分領域とボイスレコーダーで録音された音声データ中の部分音声データとを関連付けるような技術については、現時点において出願人は発見していない。
【０００５】
ここで、展示会等において１枚のパネルを用いて、発表者が口頭で製品のプレゼンテーションをしている場面を想定する。このとき、当該プレゼンテーションの聴講者は、ボイスレコーダーで発表者のプレゼンテーションに関する音声を録音する一方で、展示されているポスター（例えば、ポスター全体）をデジタルカメラで静止画像として撮影することがある。そして、その後、当該聴講者が、プレゼンテーション終了後、自宅等において、プレゼンテーション時に撮影した静止画像と録音した音声を再生して、撮影した静止画像中のある部分領域（例えば、展示されていたポスター中の一部に記載されている「製品の特徴」に関する部分）に関するプレゼンテーションを聞く場合を考える。
【０００６】
この場合、当該聴講者は、録音した音声データから該当する部分領域についての録音音声を人手によって探す必要があるため、非常に時間のかかる作業となるという問題がある。特に、当日プレゼンテーション会場に居合わせておらず、自宅等で初めて当該プレゼンテーションを視聴する人にとっては、撮影されたポスターの上記部分領域に対するプレゼンテーションが、録音された音声データ全体のどのあたりに記録されているのかが全く分からないため、所望の部分音声を探すためには、録音音声を最初から聞いていく必要があり非常に手間がかかるという問題がある。
【０００７】
本発明は、このような事情を考慮してなされたものであり、画像データ中の部分画像領域と音声データ中の部分音声データとを好適に対応付けることができる情報処理方法及び情報処理装置を提供することを目的とする。
【課題を解決するための手段】
【０００９】
上記課題を解決するために、本発明は、画像データと音声データを対応付ける情報処理方法であって、
文字を含む前記画像データから文字領域を検出する検出工程と、

検出された前記文字領域に含まれる文字情報を認識して複数の文字情報を取得する第１の認識工程と、
前記音声データ中の複数の音声区間のそれぞれに対応する音声認識結果を取得する第２の認識工程と、
前記第１の認識工程により取得された前記複数の文字情報と前記第２の認識工程により取得された前記複数の音声認識結果とをそれぞれ照合することによって、該文字情報と該音声認識結果を対応付ける対応付け工程と
を有することを特徴とする。

【００１１】

また、上記課題を解決するために、本発明は、画像データと音声データを対応付ける情報処理装置であって、
文字を含む前記画像データから文字領域を検出する検出手段と、
検出された前記文字領域に含まれる文字情報を認識して複数の文字情報を取得する第１の認識手段と、
前記音声データ中の複数の音声区間それぞれに対応する音声認識結果を取得する第２の認識手段と、
前記第１の認識手段により取得された前記複数の文字情報と前記第２の認識手段により取得された前記複数の音声認識結果とをそれぞれ照合することによって、該文字情報と該音声認識結果を対応付ける対応付け手段と
を有することを特徴とする。

【発明の効果】

【００１２】

本発明によれば、画像データ中の部分画像領域と音声データ中の部分音声データとを好適に対応付けることができる。これにより、例えば、画像データと音声データとから画像データ中の部分画像領域に関連する音声データ中の部分音声区間を人手によって探す手間が省け、大幅に時間を節約することができる。

【発明を実施するための最良の形態】

【００１３】

以下、図面を参照して、本発明の好適な実施例について詳細に説明する。

【実施例１】

【００１４】

図１は、本発明の第１の実施例に係る画像データと音声データの部分データ同士を対応付ける静止画像・音声処理装置の構成を示すブロック図である。図１において、ＣＰＵ１０１は、ＲＯＭ１０２に記憶された制御プログラム或いは外部記憶装置１０４からＲＡＭ１０３にロードされた制御プログラムに従って、本実施例の静止画像・音声処理装置の各種制御・処理を行う。ＲＯＭ１０２は、各種パラメータやＣＰＵ１０１が実行する制御プログラム等を格納している。ＲＡＭ１０３は、ＣＰＵ１０１による各種制御の実行時に作業領域を提供するとともに、ＣＰＵ１０１により実行される制御プログラムを記憶する。

【００１５】

外部記憶装置１０４は、ハードディスク、フレキシブルディスク、ＣＤ－ＲＯＭ、ＤＶＤ－ＲＯＭ、メモリカード等で実現される固定式記憶装置或いは着脱可能な可搬記憶装置であり、例えば、外部記憶装置１０４がハードディスクの場合には、ＣＤ－ＲＯＭやフレキシブルディスク等からインストールされた各種プログラムが記憶される。１０５は、マイクロフォン等による音声入力装置であり、音声入力装置１０５から取り込まれた音声は、ＣＰＵ１０１による音声認識又は音響信号分析によって静止画像に関連した音声が認識又は分析される。１０６は、デジタルカメラ等による画像入力装置であり、取り込まれた画像は静止画像に変換され、文字認識やオブジェクト認識が行われる。

【００１６】

１０７は、ＣＲＴ、液晶ディスプレイ等の表示装置であり、処理内容の設定・入力に関する表示・出力を行う。１０８は、ボタン、テンキー、キーボード、マウス、ペン等の補

助入出力装置である。１０９は、上記各部を互いに接続するバスである。尚、静止画像及び当該静止画像に対応付けられるための音声データは、それぞれ画像入力装置１０６及び音声入力装置１０５によって入力してもよいし、別の装置等によって獲得したものをＲＯＭ１０２、ＲＡＭ１０３、外部記憶装置１０４若しくはネットワークを介して接続された外部装置に記憶しておいてもよい。

【００１７】

図２は、第１の実施例で互いに部分データ同士の対応付け処理の対象となる静止画像（ａ）と当該静止画像に関連する音声（ｂ）の一例について示す図である。図２に示すように、この静止画像には、白地に「春」、「夏」、「秋」、「冬」という４つの文字が撮像されている（以降、静止画像の左下を原点として、水平方向をｘ軸、垂直方向をｙ軸とする座標軸を用いる。尚、座標単位にはピクセルを用いることができるが、特にこれに限定されることはない。）。また、この静止画像に関連した音声は、「フユ」、「ハル」、「アキ」、「ナツ」という４つの発声がこの順で録音されている（以降、音声の開始時間を０とした時間軸を用いる。尚、時間単位としてはサンプル数や秒を用いることができるが、特にこれらに限定されることはない。）。また、この音声は、各発声間に十分な無音区間を含んでいるものとする。

【００１８】

尚、この音声は、発声場所、発声時間、発声者に制限はない。すなわち、静止画像を撮影した場所、時間、撮影者は、当該音声の発声場所、発声時間、発声者と同じでもあっても異なっていてもよい。また、音声データは、Ｅｘｉｆ等のように静止画像ファイルの一部として含まれていてもよいし、静止画像とは別のファイルであってもよい。さらに、静止画像データと音声データは、同じ装置又は同じ記憶媒体に記憶されていてもよいし、ネットワーク等を介して別の場所に格納されているものであってもよい。

【００１９】

図３は、本発明の第１の実施例において静止画像と音声を入力して静止画像と音声との対応関係（画像音声対応情報）を求める際のモジュール構成を示すブロック図である。図２において、２０１は文字検出部であり、静止画像から文字部分を含む所定領域（文字領域）を検出する。図２の例では、「春」、「夏」、「秋」、「冬」の４つの文字領域が矩形の部分画像として、座標情報（図２のｘ、ｙの値）と共に検出される。尚、文字検出部２０１で検出される当該部分画像は、あくまで画像データであって、文字データではない。ここで、図７は、図２に示す静止画像と音声の例に対する文字認識結果情報と音声認識結果情報を対応させた結果を示す図である。図７（ａ）に示すように、各部分画像データの座標情報は、各文字領域（部分画像）の中心座標を表している。

【００２０】

また、図３において、２０２は文字認識部であり、文字検出部２０１で検出された各文字領域に対して文字認識を行う。尚、文字認識処理自体については、既存の技術を用いることが可能である。図２の例では、４つの文字領域の部分画像データから、文字認識部２０２によって、「春」、「夏」、「秋」、「冬」の４文字の文字データが認識される。ここで、図６は、図７に示す文字認識結果情報と音声認識結果情報の例を示す図である。図６（ａ）に示すように、文字認識部２０２によって、各文字データと中心座標とが認識結果から対応付けられている。

【００２１】

図３において、２０３は音声検出部であり、音声データから例えば人が発声した部分（音声区間）を検出する。図２の例では、「フユ」、「ハル」、「アキ」、「ナツ」の４つの音声区間が部分音声データとして、時間情報（図２のｔの値）と共に検出される。図７（ｂ）に示すように、各音声区間の時間情報は、各音声区間の開始及び終了時間を表している。

【００２２】

図３において、２０４は音声認識部であり、音声検出部２０３で検出された各音声区間に対して音声認識を行う。尚、音声認識処理自体については、既存の技術を用いることが

できる。ここでは、簡単のため、「春（ハル）」、「夏（ナツ）」、「秋（アキ）」、「冬（フユ）」の４単語のみを認識対象語彙とする単語音声認識を行った場合について考える。この場合、図２の例では、４つの音声区間の音声データが、音声認識部２０４によって、「冬」、「春」、「秋」、「夏」の４単語の文字データに変換される。図６（ｂ）に示すように、音声認識部２０４によって、各音声区間の音声データと時間情報とが認識結果から対応付けられている。

【００２３】

図３において、２０５は静止画音声対応部であり、文字検出部２０１と文字認識部２０２の処理結果として得られる静止画像内の文字認識結果及びその座標情報（文字認識結果情報）と、音声検出部２０３と音声認識部２０４の処理結果として得られる音声内の音声認識結果及びその時間情報（音声認識結果情報）を用いて、静止画像と音声データの対応付けを行う。例えば、図２に示す静止画像と音声の例では、図６（ａ）に示される文字認識結果情報による文字列と、図６（ｂ）に示される音声認識結果情報に基づく文字列とを比較・照合する。図８は、第１の実施例における静止画像と音声との対応付けの一例を示す図である。

【００２４】

図９は、静止画像と音声との対応結果を用いたアプリケーションの例である。図９に示す例では、静止画像中の文字が位置する部分（例えば、図９では座標（ｘ１，ｙ１）付近）にマウスカーソル（図９の矢印マーク）を持っていくと、この文字に対応した音声データ（図９では、図７（ｂ）に示す時刻ｓ２からｅ２までの音声データ）が再生され、スピーカー等の音声出力装置から出力される。

【００２５】

尚、図９に示す例とは逆に、音声を先頭から、或いはマウス、キーボード等で任意の時間を指定することによってその間の音声を再生し、再生されている音声区間に対応する静止画像の対応部分に枠を付与して表示することも可能である。図６０は、図２に示す静止画像と音声との対応結果を用いた別のアプリケーションに基づく表示例を示す図である。図６０に示す例では、利用者が「フユ」と音声認識された音声区間（図７のｓ１からｅ１）にマウスカーソル（図６０の矢印マーク）を持っていくと、当該音声区間に対応した文字領域（すなわち、「冬」）に文字領域分の外枠が生成・表示される。この結果、本装置の操作者は、出力されている音声が静止画像のどの部分に対応しているかを容易に理解することができる。

【００２６】

以下、図３に示す文字検出部２０１から静止画音声対応部２０５の各モジュールの動作についてさらに詳細に説明する。

【００２７】

文字検出部２０１は、静止画像中から写真、絵、文字、図形、図表等の所定領域を切り出す技術（セグメンテーション）を用いる。セグメンテーションの方法としては、文書中に存在する文字の部分を他の図表や画像等の部分と区別するための技術である文書認識技術といった既存の技術を用いることができる。尚、上述した文字領域の検出に関する説明では、簡単のため、文字領域の座標情報として、図７（ａ）に示したように文字領域の中心座標としているが、矩形領域を表すことが可能な座標（２点の座標）とするものが一般的であり、融通性があるため好適である。

【００２８】

文字認識部２０２は、文字検出部２０１で検出された文字領域からなる部分画像データを入力として、これに含まれる文字を認識する。文字認識の方法としては、既存の文字認識技術を用いればよいが、本実施例では静止画像を入力としているため、オンライン文字認識技術は適用することはできず、オフライン文字認識又はＯＣＲ（Optical Character Recognition）技術を用いる必要がある。また、文字の種類が文字認識を行う前に分かっている場合、或いは文字認識時に利用者等によって与えることが可能な場合には、その文字の種類に応じた文字認識方法を適用することができる。

【００２９】
ここでいう文字の種類とは、例えば、手書き文字と印刷活字文字である。手書き文字は、さらに、制限付き手書き文字（点線上にそって文字が書かれる文字等）、常用手書き文字、自由手書き文字に分類することもできる。また、印刷活字文字は、さらに、フォント種が１つのシングルフォント、複数のフォント種が混在するマルチフォントに分類することもできる。また、文字の種類が予め分からない場合には、これらの全ての手法を適用して最も信頼度やスコアの高い結果を利用する方法や、各文字の種類を文字認識前に判定して、判定結果に基づいた文字認識方法を適用する方法等を用いればよい。
【００３０】
図４は、第１の実施例における文字認識部２０２の細部モジュール構成を示すブロック図である。図４において、３０１は前処理部であり、文字認識処理を行い易くするための各種処理を施し、正規化データとして出力する。具体的には、雑音成分の除去、文字の大きさの正規化等を行う。３０２は特徴抽出部であり、その文字が表わす特徴を抽出する。これは正規化データを、よりその文字の特徴を捉えた次元数の低いデータへ変換・圧縮する。例えば、２値画像の輪郭におけるchain-code等を特徴として抽出する。
【００３１】
また、３０３は識別部であり、特徴抽出部３０２で得られた入力の特徴量を文字認識用テンプレート３０５と比較・照合（マッチング）することによって、入力特徴量の文字の識別を行う。マッチング方法としては、ＤＰマッチング法や２次元ＨＭＭ（Hidden Markov Model）法等を用いればよい。ここで、文字間の言語的な関係を言語知識として確率的に利用することにより、文字認識性能が向上する場合がある。３０６は、この場合に用いる文字認識用の言語モデルであり、具体的には、２つ組み文字の出現確率（文字バイグラム）等である。しかし、文字認識用言語モデル３０６は必ずしも必要なものではない。３０４は文字認識結果情報出力部であり、識別部３０３で得られる文字認識の結果と、対応する文字領域の静止画像における座標情報を文字認識結果情報として出力する。
【００３２】
２０４は、音声検出部２０３で検出された音声区間からなる音声データを入力としてこれを音声認識する音声認識部である。音声認識部２０４における音声認識の方法としては、ＨＭＭに基づく方法等の既存の音声認識技術を用いればよい。音声認識の方法としては、単語音声認識、文法ベースの連続音声認識、Ｎ－ｇｒａｍベースの大語彙連続音声認識、単語単位を用いない音素認識もしくは音節認識を用いることがある。上述した音声認識の説明では、簡単のため、単語音声認識を用いたが、実際には、単語単位で発声される保障はなく、発声内容も事前に分からないため、大語彙連続音声認識又は音素認識（音節認識）による方法を利用することが望ましい。
【００３３】
図５は、第１の実施例における音声認識部２０４の細部モジュール構成を示すブロック図である。図５において、４０１は、音声分析部で、音声をスペクトル分析し、特徴量を求める。音声分析の方法としては、ＭＦＣＣ分析（Mel-Frequency Cepstrum Coefficient）や線形予測分析等を用いればよい。４０５は、音声認識を行う際の辞書（表記と読み）及び言語制約（単語N-gramや音素N-gram等の確率値）が格納されている。４０２は、探索部で、４０１で得られた入力音声の特徴量を４０４の音声認識用音響モデルと４０５の音声認識用言語モデルを用いることによって音声認識結果を得る。４０３は、音声認識結果情報出力部で、４０３で得られる音声認識の結果と、対応する音声区間の音声における時間情報を音声認識結果情報として出力する。
【００３４】
２０５は、文字認識部２０２から得られる文字認識結果情報と、音声認識部２０４から得られる音声認識結果情報を入力として静止画像と音声を対応付け、静止画音声対応情報を出力する。対応付けは、文字認識の結果得られる文字もしくは文字列と、音声認識の結果得られる表記（単語）から得られる文字もしくは文字列のマッチングを取ることによって行う。或いは、文字認識の結果得られた文字列の発音列と音声認識の結果得られた発音

列との照合によって行う。尚、これらの詳細については、以降の実施例において詳細に説明する。図２の例では、説明を簡単にするため、静止画像中の文字と音声の発声が１対１に対応している例を示した。

【００３５】

よって、文字列のマッチングは、完全に一致するものを探すことにより対応付けが行える。しかしながら、実際にプレゼンテーション等で録音される音声は、静止画像の文字をそのまま発声することはほとんどないと考えられる。このような場合には、文字認識の結果得られる文字列を音声認識の結果得られる文字列に対して部分マッチングさせて対応付けを行う。

【００３６】

例えば、「ここにある春は、…」や、「つまり、これは夏になると…」という発声がなされていた場合、文字認識結果の「春」は前者、「夏」は後者の音声認識結果の部分文字列と一致するため、これらを対応付ける。さらに一般的には、文字領域に対する音声区間がない、文字領域とは関係のない音声区間がある、文字認識結果に誤りがある、音声認識結果に誤りがあることが考えられるため、一致するか否かといった決定的なマッチングではなく、どの程度マッチングするかといった確率的な柔軟なマッチングを行う必要がある。

【００３７】

以上の説明から明らかなように、本実施形態によれば、静止画像データから静止画像の部分画像領域と抽出し、音声データから音声の部分音声区間を抽出し、お互いに関連のあるものを好適に対応付けることができるようになり、その結果として、画像データ中の部分画像領域に関連した音声データ中の音声区間（部分音声データ）を従来のように人手によって探す手間が省け、大幅に時間を節約することが可能となる。

【実施例２】

【００３８】

上述した第１の実施例における静止画音声対応部２０５では、文字認識の結果として得られる文字列と、音声認識の結果として得られる文字列とを直接比較して対応付けていた。しかし、音声認識方法が音素（音節）認識であったり、同音異表記が出力された場合には、文字列の直接比較を行うことができない。例えば、文字認識結果が「春」であり、音声認識結果が「ｈａｒｕ」、「ハル」、「張る」等の場合である。そこで、一般に、音声認識では入力音声の読み情報（発音列）が分かっていることから、文字認識結果を読み情報（発音列）に変換した後に、発音列同士でマッチングを取ることによって、文字列同士の比較ができないような場合においても文字認識結果情報と音声認識結果情報の対応を取ることが可能となる。

【００３９】

図１０は、本発明の第２の実施例における発音列マッチングによる静止画音声対応部の細部モジュール構成を示すブロック図である。図１０において、５０１は、文字認識部２０２から得られる文字認識結果情報の文字認識結果を発音列に変換する文字認識結果発音列変換部である。５０２は、文字列を発音に変換するために文字認識結果発音列変換部５０１で用いられる発音変換辞書である。ここで、文字と発音の対応は、一般に、１対１ではなく１対多となるため、１つの文字列に対する発音列は多くの場合１種類ではなく、発音列候補として１つ又は複数出力される。

【００４０】

具体的には、図６（ａ）に示される文字認識結果情報の、「春」、「夏」、「秋」、「冬」という文字列から、それぞれ「ハル／シュン」、「ナツ／カ」、「アキ／シュウ」、「フユ／トウ」というような発音列候補を得る。図１１は、第２の実施例における文字認識結果と音声認識結果に対する発音列の例を示す図である。すなわち、図６（ａ）に示される文字認識結果情報から図１１（ａ）に示されるような発音列候補を得る。

【００４１】

図１０において、５０３は、音声認識部２０４から得られる音声認識結果情報から発音

列を抽出する音声認識結果発音列抽出部である。具体的には、図６（ｂ）に示される音声認識結果情報から、図１１（ｂ）に示すように、「フユ」、「ハル」、「アキ」、「ナツ」という発音列を抽出する。

【００４２】

また、図１０において、５０４は発音列マッチング部であり、文字認識結果の文字列を発音列に変換したものと音声認識結果の発音列とのマッチングを取る。このマッチング処理によって、図１１に示す例では、文字認識結果の複数の発音列候補から「ハル」、「ナツ」、「アキ」、「フユ」が選択され、音声認識結果の発音列と対応付けられる。

【００４３】

さらに、図１０において、５０５は静止画音声対応情報出力部であり、マッチング結果を図８に示すような静止画音声対応情報として出力する。尚、この例では、発音列としてカタカナ表記を用いているが、これに限らず音素表現等別の表記を用いてもよいことは言うまでもない。また、文字認識結果の発音列候補は「シュウ」や「トウ」と書き言葉の発音列を生成していたが、「シュー」や「トー」といった話し言葉の発音列に変換した結果や、これを書き言葉の発音列に加えた結果を用いてもよい。

【００４４】

以上の説明から明らかなように、本実施例によれば、文字認識結果の文字列と音声認識結果の文字列が直接比較できない場合においても、静止画像と音声の対応付けを行うことが可能となる。

【実施例３】

【００４５】

上述した第２の実施例では、文字認識の結果として得られる文字列を発音列に変換し、音声認識の結果として得られる発音列とマッチングしていたが、これとは逆に、音声認識の発音列を文字列に変換し、文字認識結果の文字列とマッチングすることも可能である。

【００４６】

図１２は、本発明の第３の実施例における文字列マッチングを行う静止画音声対応部２０５の細部モジュール構成を示すブロック図である。図１２において、６０１は、文字認識部２０２から得られる文字認識結果情報の文字認識結果から文字列を抽出する文字認識結果文字列抽出部である。具体的には、図６（ａ）に示される文字認識結果情報から、図１３（ａ）に示されるように「春」、「夏」、「秋」、「冬」という文字列を抽出する。すなわち、図１３は、第３の実施例における文字認識結果と音声認識結果に対する文字列の例である。

【００４７】

図１２において、６０２は、音声認識部２０４から得られる音声認識結果情報の音声認識結果（発音列）を文字列に変換する音声認識結果文字列変換部である。また、６０３は、音声認識結果文字列変換部６０２で発音列を文字列に変換する際に用いられる文字変換辞書である。ここで、発音と文字の対応は、一般に、１対１ではなく１対多となるため、１つの発音列に対する文字列は１種類ではなく、文字列候補として複数出力する。

【００４８】

具体的には、図６（ｂ）に示される音声認識結果情報の、「フユ」、「ハル」、「アキ」、「ナツ」という発音列から、図１３（ｂ）に示すように、それぞれ「冬／不輸」、「春／張る／貼る」、「空／飽き／秋」、「夏／奈津／捺」という文字列候補を得る。

【００４９】

６０４は、文字列マッチング部であり、文字認識結果の文字列と音声認識結果の発音列を文字列に変換したものとのマッチングを行う。このマッチング処理によって、図１３に示す例では、音声認識結果の複数の文字列候補から「冬」、「春」、「秋」、「夏」が選択され、文字認識結果の文字列と対応付けられる。また、６０５は、静止画音声対応情報出力部であり、文字列マッチング部６０４によるマッチング結果を図８に示すように静止画音声対応情報として出力する。

【００５０】

以上の説明から明らかなように、本実施例によれば、文字認識結果の文字列と音声認識結果の文字列が直接比較できないような場合においても、発音列でのマッチングを行うことによって、静止画像と音声の対応付けを行うことが可能となる。

【実施例４】

【００５１】

上述した実施例では、文字認識結果及び音声認識結果はいずれも１つの認識結果のみであり、また、静止画像と音声との対応付け処理では、認識結果の文字列又は発音列のみを用いて対応付けを行っていたが、認識結果に尤度や確率等のスコア情報を保持した複数候補を出力し、このスコア付きの複数候補を用いて文字認識結果と音声認識結果を対応付けることも可能である。

【００５２】

ここで、Ｎ個の文字領域Ｉ１，…，ＩＮに対して、Ｍ個の音声区間Ｓ１，…，ＳＭの１つと対応付けを行った結果をＣ１，…，ＣＮ（但し、Ｃｎ＝（Ｉｎ，Ｓｍ），１≦ｎ≦Ｎ，１≦ｍ≦Ｍ）とするとき、Ｃｎは、

Ｃｎ＝ａｒｇｍａｘ（ＰＩｎｉ，ＰＳｍｊ，δＲＩｎｉ，ＲＳｍｊ）

によって求めることができる。

【００５３】

ここで、ＰＩｎｉは文字領域Ｉｎのｉ番目の文字認識結果候補のスコア（１≦ｉ≦Ｋ，但し、Ｋは文字認識結果の候補数。）、ＰＳｍｊは音声区間Ｓｍのｊ番目の音声認識結果候補のスコア（１≦ｊ≦Ｌ，但し、Ｌは音声認識結果の候補数）である。また、Ｉｎの第ｉ位の文字認識結果の文字列（又は、発音列）をＲＩｎｉ、Ｓｍの第ｊ位の音声認識結果の文字列（又は、発音列）をＲＳｍｊとするとき、δＲＩｎｉ，ＲＳｍｊは、ＲＩｎｉ＝ＲＳｍｊの場合はδＲＩｎｉ，ＲＳｍｊ＝１、それ以外の場合はδＲＩｎｉ，ＲＳｍｊ＝０という関数で与えられる。さらに、ａｒｇｍａｘは、ＰＩｎｉ，ＰＳｍｊ，δＲＩｎｉ，ＲＳｍｊを最大にするｉ，ｍ，ｊの組を求める演算を表し、これを求めることによって、Ｉｎに対するＳｍ、すなわちＣｎを決めることができる。

【００５４】

以下、図１４、１５、１６を用いて、対応付けの具体例について説明する。

【００５５】

図１４は、第４の実施例における文字認識結果（ａ）と音声認識結果（ｂ）のスコア情報（尤度や確率等で表された認識結果）を保持した複数候補の例を示す図である。図１４に示す例では、Ｎ＝４，Ｍ＝４，Ｋ＝３，Ｌ＝３である。ここで、第１の実施例で説明したように、文字認識結果と音声認識結果の文字列を直接比較することによって、静止画と音声の対応付けを行うことにする。例えば、図１４に示すように、Ｉ１は「春」、Ｓ１は「冬」、ＰＩ１１＝０．７，ＰＳ４３＝０．１，ＲＩ１３は「空」、ＲＳ３２は「足」等となる。

【００５６】

このとき、ｎ＝１、すなわち「春」、「香」、「空」と文字認識された文字領域に対する音声区間は、ｉ＝１，ｍ＝２，ｊ＝１の場合、ＰＩ１１＝０．７、ＰＳ２１＝０．７、ＲＩ１１は「春」、ＲＳ３２は「春」でδＲＩ１１，ＲＳ２１＝１となり、上記ａｒｇｍａｘの中が最大０．４９（＝０．７×０．７×１）となる。尚、その他の場合は、いずれもδＲＩｎｉ，ＲＳｍｊ＝０となるため、ａｒｇｍａｘの中は０となる。よって、Ｃ１＝（Ｉ１，Ｓ２）と決定される。同様の計算を行うことによって、Ｃ２＝（Ｉ２，Ｓ３）、Ｃ３＝（Ｉ３，Ｓ４）、Ｃ４＝（Ｉ４，Ｓ１）と対応付けがなされる。

【００５７】

次に、第２の実施例で説明したように、文字認識結果を発音列に変換し、これと音声認識結果の発音列を比較することによって静止画と音声の対応付けを行う際に、スコア付きの複数候補を用いる例について説明する。

【００５８】

図１５は、第４の実施例における文字認識結果を発音列に変換した結果（ａ）と音声認

識結果から得られる発音列（ｂ）のスコア情報を保持した複数候補の例を示す図である。この場合、文字認識結果のスコア情報をそのまま発音列のスコア情報とする。また、１つの文字認識結果から複数の発音列が得られる場合には、それぞれの発音列に対して同じスコア情報を用いる。

【００５９】

例えば、ｎ＝１の場合は、ｉ＝１で「ハル」と「シュン」の２通り、ｉ＝２で「カ」と「コウ」の２通り、ｉ＝３で「ソラ」と「アキ」と「クウ」の３通りの発音列に対して、図１４に示した例の場合と同様の計算を行う。この結果、例えば、ｎ＝１，ｉ＝１の「ハル」とｍ＝２，ｊ＝１の「ハル」のａｒｇｍａｘの中は０．４９（＝０．７×０．７×１）であり、ｎ＝１，ｉ＝３の「アキ」とｍ＝３，ｊ＝１の「アキ」のａｒｇｍａｘの中は０．０６（＝０．１×０．６×１）であり、Ｃ１＝（Ｉ１，Ｓ２）と対応付けられる。また、ｎ＝４，ｉ＝２の「フユ」とｍ＝１，ｊ＝１の「フユ」は０．１５（＝０．３×０．５×１）であり、ｎ＝４，ｉ＝３の「ツ」とｍ＝４，ｊ＝２の「ツ」は０．０２（＝０．２×０．１×１）であり、Ｃ４＝（Ｉ４，Ｓ１）と対応付けられる。同様に、Ｃ２＝（Ｉ２，Ｓ３）、Ｃ３＝（Ｉ３，Ｓ４）と対応付けがなされる。

【００６０】

次に、第３の実施例で説明したように、音声認識結果を文字列に変換し、これと文字認識結果の文字列を比較することによって静止画と音声の対応付けを行う際に、スコア付きの複数候補を用いる例について説明する。

【００６１】

図１６は、第４の実施例における文字認識結果から得られる文字列（ａ）と音声認識結果を文字列に変換した結果（ｂ）のスコア情報を保持した複数候補の例を示す図である。この場合も、図１５で示した発音列の対応付けと同様であり、例えば、ｎ＝１，ｉ＝１の「春」とｍ＝２，ｊ＝１の「春」は０．４９（＝０．７×０．７×１）であり、ｎ＝１，ｉ＝３の「空」とｍ＝３，ｊ＝１の「空」は０．０６（＝０．１×０．６×１）であり、Ｃ１＝（Ｉ１，Ｓ２）と対応付けられる。

【００６２】

尚、上述したように本実施例では、δは完全に一致する場合に１、一致しない場合は０という２値の値のいずれかをとる関数を用いていたが、これに限らず、例えば一致の度合いに応じた値とする等、別の定義でもよい。また、文字認識結果のスコアと音声認識結果のスコアは同等に扱っているが、例えば、文字認識のスコアを音声認識のスコアよりも重視する等、これらのスコアに重みをつけてもよい。

【００６３】

以上の説明から明らかなように、本実施例によれば、文字認識結果と音声認識結果をスコア付きで複数候補出力することで、１位の候補に正解の認識結果が含まれない場合でも、より正確に静止画像と音声の対応付けを行うことが可能となる。

【実施例５】

【００６４】

上述した第２～第４の実施例では、静止画音声対応部２０５において、発音列又は文字列のどちらかに変換された結果に基づいて静止画像と音声の対応付けを行っていたが、これらの両方を用いて対応付けを行うこともできる。すなわち、文字認識結果を読みに変換した文字認識結果発音列と音声認識結果として得られる音声認識結果発音列のマッチングと、文字認識結果として得られる文字認識結果文字列と音声認識結果を文字列に変換した音声認識結果文字列のマッチングの両方を用いる。これは、図１０と図１２のそれぞれで示されるモジュール構成を併用することによって実現することができる。

【実施例６】

【００６５】

上述した実施例では、文字認識に関する処理と音声認識に関する処理は、それぞれ独立に行われていたが、文字認識の結果を音声認識で利用することも可能である。この際、以下に説明するように様々な利用の仕方が考えられる。

【００６６】
まず、文字認識結果を音声認識結果情報出力部で利用する場合について説明する。図１７は、本発明の第６の実施例における静止画像・音声認識装置のモジュール構成を示すブロック図である。図１７において、文字認識部７０１は文字認識部２０２と、また、音声分析部７０２、探索部７０３、音声認識用音響モデル７０４及び音声認識用言語モデル７０５は、それぞれ音声分析部４０１、探索部４０２、音声認識用音響モデル４０４、音声認識部言語モデル４０５と、さらに静止画音声対応部７０７は静止画音声対応部２０５と同じであるため説明は省略する。
【００６７】
７０６は音声認識結果情報出力部であり、探索部７０３の探索結果に加えて、文字認識部７０１の文字認識で得られる結果も利用する。例えば、図１４に示す場合、図１４（ｂ）に示される音声認識結果に対して、図１４（ａ）の結果に含まれない「古」、「露」、「樽」、「白」、「足」、「薪」、「松」、「津」の８種類の文字列は音声認識結果候補としない。この結果、これらの８種類の文字列に対しては、第４の実施例で説明した計算を行う必要がなくなり、処理の効率化が図れる。
【００６８】
次に、文字認識結果を音声認識の探索部で利用する場合について説明する。図１８は、本発明の第６の実施例における文字認識結果を音声認識に利用する静止画像・音声認識装置のモジュール構成を示すブロック図である。
【００６９】
図１８において、文字認識部８０１は文字認識部２０２と、また音声分析部８０２及び音声認識用音響モデル８０４から音声認識結果情報出力部８０６はそれぞれ音声分析部４０１及び音声認識用音響モデル４０４、音声認識用言語モデル４０５、音声認識結果情報出力部４０３と、さらに静止画音声対応部８０７は静止画音声対応部２０５と同じであるため説明は省略する。
【００７０】
探索部８０３は、音声認識用音響モデル８０４と音声認識用言語モデル８０５の２つのモデルを用いて音声認識を行う際に、文字認識部８０１で得られる結果を利用する。例えば、図１４（ａ）に示された結果が文字認識の結果として得られた場合、探索部８０３は、これらの１２種類の文字列（単語）のみを用いた探索処理を行う。すなわち、探索部８０３は、音声認識用言語モデル８０５に含まれる音声認識対象語としてこれらの１２種類のみを用いて音声認識を行う。この結果、探索部８０３の計算が大幅に低減され、文字認識の結果候補に正解が含まれている場合、文字認識と音声認識を独立に行うものと比較して、音声認識の性能も一般に向上させることができる。
【００７１】
次に、文字認識結果を発音列に変換し、これを音声認識結果情報出力部で利用する場合について説明する。図１９は、第６の実施例における文字認識結果を発音列に変換して利用する静止画像・音声認識装置のモジュール構成を示すブロック図である。
【００７２】
図１９において、文字認識部９０１は文字認識部２０２と、また文字認識結果発音列変換部９０２は文字認識結果発音列変換部５０１と、さらに音声分析部９０３から音声認識用言語モデル９０６はそれぞれ音声分析部４０１、探索部４０２、音声認識用音響モデル４０４、音声認識用言語モデル４０５と、さらにまた静止画音声対応部９０８は静止画音声対応部２０５と同じであるため説明は省略する。尚、図１９では、文字認識結果発音列変換部９０２の処理を行う際に必要な発音変換辞書５０２は省略している。
【００７３】
図１９において、音声認識結果情報出力部９０７は、探索部９０３の結果に加えて、文字認識部９０１の文字認識結果を発音列に変換した結果も利用する。例えば、図１５に示す例の場合、図１５（ｂ）に示される音声認識結果に対して、図１５（ａ）の結果に含まれない「フル」、「ツユ」、「タル」、「ハク」、「アシ」、「マキ」、「マツ」の７種

類の発音列は音声認識結果候補としない。この結果、これらの７種類の文字列に対しては、第４の実施例で説明した計算を行う必要がなくなる。

【００７４】

次に、文字認識結果から得られる発音列を音声認識の探索部１００４で利用する場合について説明する。図２０は、第６の実施例における文字認識結果を発音列に変換して探索部で利用する静止画像・音声認識装置のモジュール構成を示すブロック図である。

【００７５】

図２０において、文字認識部１００１は文字認識部２０２と、また文字列認識結果発音列変換部１００２は文字列認識結果発音列変換部５０１と、さらに音声分析部１００３及び音声認識用音響モデル１００５から音声認識結果情報出力部１００７はそれぞれ音声分析部４０１及び音声認識用音響モデル４０４、音声認識用言語モデル４０５、音声認識結果情報出力部４０３と、さらにまた静止画音声対応部１００８は静止画音声対応部２０５と同じであるため説明は省略する。尚、図２０では、文字認識結果発音列変換部１００２の処理を行う際に必要な発音変換辞書５０２は省略している。

【００７６】

図２０において、探索部１００４は、音声認識用音響モデル１００５と音声認識用言語モデル１００６の２つのモデルを用いて音声認識を行う際に、文字認識結果発音列変換部１００２で文字認識結果を発音列に変換した結果も利用する。例えば、図１５（ａ）に示された結果が文字認識の結果から得られる発音列であるとき、探索部１００４は、これらの２５種類の発音列のみを用いた探索処理を行う、すなわち、探索部１００４は、音声認識用言語モデル１００６に含まれる音声認識対象語として、これらの２５種類のみを用いて音声認識を行う。

【００７７】

この結果、探索部１００４の計算が大幅に低減され、文字認識の結果から得られる発音列候補に正解が含まれている場合、文字認識と音声認識を独立に行うものと比較して、音声認識の性能も一般に向上させることができる。

【００７８】

次に、文字認識結果から得られる文字列を音声認識結果から文字列に変換する際に利用する静止画音声対応処理について説明する。

【００７９】

図２１は、第６の実施例における文字認識結果の文字列を音声認識結果を文字列に変換する際に利用する静止画像・音声認識装置のモジュール構成を示すブロック図である。図２１において、文字認識結果文字列抽出部１１０１は文字認識結果文字列抽出部６０１と、文字変換辞書１１０３から静止画音声対応情報出力部１１０５はそれぞれ文字変換辞書６０３から静止画音声対応情報出力部６０５と同じであるため説明は省略する。

【００８０】

図２１において、１１０２は音声認識結果文字列変換部であり、音声認識結果を文字列に変換する際に、文字認識結果文字列抽出部１１０１による文字認識結果から抽出される文字列も利用する。例えば、図１６（ａ）に示された結果が文字認識の結果から抽出される文字列であるとき、音声認識結果文字列変換部１１０２の音声認識結果を文字列に変換する際に、これらの１６種類の文字列に変換しうる音声認識結果のみを文字列変換候補として選択する。

【００８１】

以上の説明から明らかなように、本実施例によれば、文字認識で得られる結果を音声認識において利用することで、計算量の低減や、音声認識性能を向上させることが可能となる。

【実施例７】

【００８２】

前述した実施例における、文字認識の結果を音声認識の探索部で利用する処理は、文字認識の結果の文字列をそのまま用いることによって行われていたが、一般に文字認識の結

果通りに音声が発声されるとは限らないため、文字認識の結果から音声として発声されると予想される重要語を抽出し、これを音声認識の探索部で利用することが好ましい。

【００８３】

図２２は、第７の実施例における文字認識結果から重要語を抽出して探索部で利用する静止画像・音声認識装置のモジュール構成を示すブロック図である。図２２において、文字認識部１２０１は文字認識部２０２と、また音声分析部１２０３及び音声認識用音響モデル１２０５から音声認識結果情報出力部１２０７はそれぞれ音声分析部４０１及び音声認識用音響モデル４０４、音声認識用言語モデル４０５、音声認識結果情報出力部４０３と、さらに静止画音声対応部１２０８は静止画音声対応部２０５と同じであるため説明は省略する。

【００８４】

図２２において、１２０２は重要語抽出部であり、文字認識の結果から重要語を抽出する。例えば、文字認識の結果が、「提案法は、統計的言語モデルを用いたアプローチです。」といった文字列であり、重要語の抽出方法が文字列を形態素解析し、この結果から自立語を抽出するものであるとすると、この結果から、「提案法」、「統計的」、「言語」、「モデル」、「アプローチ」の５単語が重要語として抽出される。

【００８５】

また、１２０４は探索部であり、音声認識用音響モデル１２０５と音声認識用言語モデル１２０６の２つのモデルを用いて音声認識を行う際に、重要語抽出部１２０２で抽出された単語を利用する。具体的には、上述した５単語をキーワードとしたキーワードスポッティングによる音声認識を行う、または、大語彙連続音声認識の場合には、音声認識結果からこれらの５単語が含まれる部分を抽出する、若しくは、上記５単語に関する音声認識用言語モデルの確率値を増加させて音声認識を行う。尚、重要語の抽出規則は、ここでは自立語の抽出としたが、これに限らず他の規則や手法を適用してもよい。また、１２０９は重要語を抽出ための規則やデータ（単語辞書）である。

【００８６】

以上の説明から明らかなように、本実施例によれば、文字認識の結果通りの音声でない場合においても、静止画像と音声の対応付けを好適行うことが可能となる。

【実施例８】

【００８７】

一般に、静止画像に含まれる文字に関する情報は、単なる文字列のみではなく、フォントサイズ、文字種、色、斜体やアンダーライン等スタイルや文字飾りに関する情報も含まれているため、これらのフォント情報を抽出し、これを音声認識で利用することによって、より正確に静止画像と音声を対応付けることができる。

【００８８】

そこで、例えば、図２３に示されるような静止画像からフォント情報を抽出し、これを音声認識で利用する実施例を考える。図２３は、第８の実施例における種々のフォント情報をもった静止画像の例である。また、図２４は、第８の実施例における文字領域からフォント情報を抽出して文字認識結果情報として出力する静止画像・音声認識装置のモジュール構成を示すブロック図である。

【００８９】

図２４において、１３０１はフォント情報抽出部であり、文字領域に対して、フォントサイズ、文字種、色、斜体やアンダーラインの有無等のフォント情報を抽出する。また、他のモジュールは、図４に示す例と同じであるため省略する。

【００９０】

図２５は、図２３に示す静止画像からの文字認識結果と各文字領域のフォント情報を示す図である。次に、図２５に示されるフォント情報を音声認識で利用する。尚、このときのモジュール構成は、図１８に示す装置と同様である。但し、図１８の文字認識部８０１は、図２４に示した構成となる点で異なる。

【００９１】

ここで、フォント情報の音声認識での利用の仕方は様々であるが、例えば、フォントサイズが大きい文字列や斜体やアンダーラインが施されている文字列は、キーワードスポッティングの対象とする、又は統計的言語モデルの確率値を増加させて音声認識を行う。他にも、黒以外の色については、色の情報を音声認識の対象語彙に追加するといったことができる。

【００９２】

以上の説明から明らかなように、本実施例によれば、静止画像に含まれる文字領域のフォント情報を音声認識で利用することによって、より正確に静止画像と音声を対応付けることが可能となる。

【実施例９】

【００９３】

上述した第６の実施例では、文字認識の結果を音声認識で利用する場合について説明したが、これとは逆に、音声認識の結果を文字認識で利用することもできる。この際、以下に説明するように様々な利用の仕方が考えられる。

【００９４】

まず、音声認識結果を文字認識結果情報出力部で利用する場合について説明する。図２６は、第９の実施例における文字認識結果情報出力部の細部モジュール構成を示すブロック図である。図２６において、音声認識部１４０１は音声認識部２０４と、また前処理部１４０２から文字認識用言語モデル１４０６はそれぞれ前処理部３０１、特徴抽出部３０２、識別部３０３、文字認識用テンプレート３０５、文字認識用言語モデル３０６と、さらに静止画音声対応部１４０８は静止画音声対応部２０５と同じであるため説明は省略する。

【００９５】

１４０７は文字認識結果情報出力部であり、識別部１４０４の識別結果に加えて、音声認識部１４０１の音声認識で得られる結果も利用する。例えば、図１４の場合、図１４（ａ）に示される文字認識結果に対して、図１４（ｂ）の結果に含まれない「香」、「空」、「科」、「和」、「新」、「厚」、「各」、「尽」の８種類の文字列は文字認識結果候補としない。この結果、これらの８種類の文字列に対しては、第４の実施例で説明した計算を行う必要がなくなる。

【００９６】

次に、音声認識結果を文字認識の識別部で利用する場合について説明する。図２７は、第９の実施例における静止画像・音声認識装置のモジュール構成を示すブロック図である。図２７において、音声認識部１５０１は音声認識部２０４と、また前処理部１５０２、特徴抽出部１５０３及び文字認識用テンプレート１５０５から文字認識結果情報出力部１５０７はそれぞれ前処理部３０１、特徴抽出部３０２及び文字認識用テンプレート３０５、文字認識用言語モデル３０６、文字認識結果情報出力部３０４と、さらに静止画音声対応部１５０８は静止画音声対応部２０５と同じであるため説明は省略する。

【００９７】

識別部１５０４は、文字認識用テンプレート１５０５と文字認識用言語モデル１５０６の２つのモデルを用いて文字認識を行う際に、音声認識部１５０１の音声認識で得られる結果を利用する。例えば、図１４（ｂ）に示された結果が音声認識の結果として得られた場合、識別部１５０４は、これらの１６種類の文字列のみを用いた識別処理を行う。すなわち、識別部１５０４に含まれる文字認識対象語としてこれらの１６種類のみを用いて文字認識を行う。この結果、識別部の計算が大幅に低減され、音声認識の結果候補に正解が含まれている場合、文字認識と音声認識を独立に行うものと比較して、文字認識の性能も一般に向上させることができる。

【００９８】

次に、音声認識結果を文字列に変換し、これを文字認識結果情報出力部で利用する場合について説明する。図２８は、第９の実施例における音声認識結果を文字列に変換して利用する静止画像・音声認識装置のモジュール構成を示すブロック図である。図２８におい

て、音声認識部１６０１は音声認識部２０４と、また音声認識結果文字列変換部１６０２は音声認識結果文字列変換部６０２と、さらに前処理部１６０３から文字認識用言語モデル１６０７はそれぞれ図４に示す前処理部３０１、特徴抽出部３０２、識別部３０３、文字認識用テンプレート３０５、文字認識用言語モデル３０６と、さらにまた静止画音声対応部１６０９は図３に示す静止画音声対応部２０５と同じであるため説明は省略する。尚、図２８では、音声認識結果文字列変換部１６０２の処理を行う際に必要な文字変換辞書６０２は省略している。

【００９９】

図２８において、１６０８は文字認識結果情報出力部であり、識別部１６０５の識別結果に加えて、音声認識部１６０２の音声認識結果を文字列に変換した結果も利用する。例えば、図１６に示す例の場合、図１６（ａ）に示される文字認識結果に対して、図１６（ｂ）の結果に含まれない「香」、「科」、「和」、「真」、「厚」、「各」、「尽」の７種類の文字列は文字認識結果候補としない。この結果、これらの７種類の文字列に対しては、第４の実施例で説明した計算を行う必要がなくなる。

【０１００】

次に、音声認識結果から得られる文字列を文字認識の識別部で利用する場合について説明する。図２９は、第９の実施例における音声認識結果から得られる文字列を文字認識で利用する静止画像・音声認識装置のモジュール構成を示すブロック図である。図２９において、音声認識部１７０１は図３に示す音声認識部２０４と、また音声認識結果文字列変換部１７０２は図１２に示す音声認識結果文字列変換部６０２と、さらに前処理部１７０３、特徴抽出部１７０４及び文字認識用モデル１７０６から文字認識結果情報出力部１７０８はそれぞれ図４に示す前処理部３０１、特徴抽出部３０２及び文字認識用テンプレート３０５、文字認識用言語モデル３０６、文字認識結果情報出力部３０４と、さらにまた静止画音声対応部１７０９は図３に示す静止画音声対応部２０５と同じであるため説明は省略する。尚、図２９では、音声認識結果文字列変換部１７０２の処理を行う際に必要な図１２に示す文字変換辞書６０３は省略している。

【０１０１】

識別部１７０５は、文字認識用モデル１７０６と文字認識用言語モデル１７０７の２つのモデルを用いて文字認識を行う際に、音声認識結果文字列変換部１７０２の音声認識結果を文字列に変換した結果も利用する。例えば、図１６（ｂ）に示された結果が音声認識の結果から得られる文字列であるとき、識別部１７０５は、これらの３２種類の文字列のみを用いた識別処理を行う。すなわち、識別部１７０５は、文字認識用モデル１７０６や文字認識用言語モデル１７０７に含まれる文字認識対象語としてこれらの３２種類のみを用いて文字認識を行う。

【０１０２】

この結果、識別部の計算が大幅に低減され、音声認識の結果から得られる文字列候補に正解が含まれている場合、文字認識と音声認識を独立に行うものと比較して、文字認識の性能も一般に向上させることができる。

【０１０３】

次に、音声認識結果から得られる発音列を文字認識結果の発音列の変換で利用する際の静止画音声対応手段について説明する。図３０は、第９の実施例における音声認識結果から得られる発音列を文字認識結果の発音列の変換に利用する静止画像・音声認識装置のモジュール構成を示すブロック図である。図３０において、音声認識結果発音列抽出部１８０１は図１０に示す音声認識結果発音列抽出部５０３と、発音変換辞書１８０３から静止画音声対応情報出力部１８０５はそれぞれ発音変換辞書５０２、発音列マッチング部５０４、静止画音声対応情報出力部５０５と同じであるため説明は省略する。

【０１０４】

図３０において、１８０２は文字認識結果発音列変換部であり、文字認識結果を発音列に変換する際に、音声認識結果発音列抽出部１８０１の音声認識結果から抽出される発音列も利用する。例えば、図１５（ｂ）に示された結果が音声認識の結果から抽出される発

音列であるとき、文字認識結果発音列変換部１８０２の文字認識結果を発音列に変換する際に、これらの１６種類の発音列に変換しうる文字認識結果のみを発音列変換候補として選択する。

【０１０５】

以上の説明から明らかなように、本実施例によれば、音声認識で得られる結果を文字認識において利用することで、計算量の低減や、文字認識性能を向上させることが可能となる。

【実施例１０】

【０１０６】

前述した実施例で示した図２（ａ）や図２３に示す静止画像の例は、非常に単純なものであったが、本発明をより複雑な静止画像に対して音声との対応付けを行うためには、静止画像から単純に文字領域を見つけるだけでなく、静止画像の構造を正確に捉える必要がある。すなわち、まず大局的に静止画像を複数の領域に分割し、次に個々の分割静止画像領域に対して文字領域を抽出し、文字認識を行う。

【０１０７】

図３１は、より複雑な静止画像（ａ）とそれに関連付けられた音声（ｂ）の一例を示す図である。また、図３２は、図３１に示すような複雑な静止画像を分割する機能を有する静止画像・音声認識装置のモジュール構成を示すブロック図である。また、図６２は、第１０の実施例に係る静止画像・音声認識装置の処理動作を説明するためのフローチャートである。

【０１０８】

図３２に示すように、静止画像分割部１９０１は、１つの静止画像を複数の静止画像領域に分割する（ステップＳ１１）。尚、この分割方法としては、既存の技術を利用すればよい。また、文字検出部１９０２から静止画音声対応部１９０６は、図３に示す文字検出部２０１から静止画音声対応部２０５と同様であるため説明は省略する。但し、図３の文字検出部２０１は静止画像全体が入力であったが、本実施例では静画像分割部１９０１で分割された個々の静止画像が入力となる点で異なる。

【０１０９】

図３３は、図３１（ａ）の静止画像を静止画像分割部１９０１によって５つの領域に分割された結果を示す図である。また、図３４は、図３３に示す各分割領域に対して領域ＩＤを付与して図３２（ａ）の座標系における座標情報を表した図である。尚、図３５は、図３４と図３２（ａ）の対応関係について示す図である。

【０１１０】

次に、各分割画像に対して文字検出部１９０２によって文字領域の検出処理を施す（ステップＳ１２）。図３６は、文字検出部１９０２による文字領域の検出結果例を示す図である。さらに、図３６に示す文字領域に対して文字認識部１９０３による文字認識を行うとともに（ステップＳ１３）、図３１（ｂ）の音声に対して音声検出部１９０４による音声検出（ステップＳ１４）及び音声認識部１９０５による音声認識を行う（ステップＳ１５）。尚、ステップＳ１１～Ｓ１３の文字認識、ステップＳ１４～Ｓ１５の音声認識は、両認識処理を同時にしてもよく、どちらの処理を先にしてもよい。

【０１１１】

図３７は、文字認識及び音声認識の結果得られる文字認識結果情報（ａ）及び音声認識結果情報を（ｂ）示す図である。尚、文字認識結果情報の座標情報は、図３８に示すような矩形領域として２点の座標で表している。すなわち、図３８は、図３６に示す文字領域の検出結果に文字認識結果情報を対応させた結果を示す図である。そして、静止画音声対応部１９０６で、前述した実施例と同様の方法によって、図３７（ａ）の文字認識結果と図３７（ｂ）の音声認識結果を対応付け、静止画音声対応情報を得る（ステップＳ１６）。

【０１１３】

以上の説明から明らかなように、本実施例によれば、静止画像領域を大局的に分割する

ことにより、複雑な静止画像に対しても文字領域と音声との対応付けを行うことが可能となる。

【実施例１１】

【０１１４】

前述した実施例で示した図２（ｂ）や図３１（ｂ）の音声の例は、各発声区間の間には十分な無音区間が含まれ、また、発声される内容は静止画像の文字領域のいずれかと全く同じであるという非常に単純なものであった。しかしながら、実際の音声は、文字領域の内容と同じ発声をするとは限らず、さらに、ある文字領域の内容に関する発声は全くされていなかったり、どの文字領域とも関係のない発声が含まれていたりする場合がある。また、複数の文字領域に関する音声が十分な無音区間なしに連続的に発声される場合や、雑音や音楽といった音声以外のものも含まれている場合もある。よって、本発明をより一般的な音声であっても静止画像との対応付けが行えるようにするためには、音声区間の正確な抽出、音声認識結果と文字認識結果の柔軟なマッチングを行う必要がある。

【０１１５】

そこで、音声区間の正確な抽出に関しては、まず、雑音や音楽といった音声以外のものが入力音声に含まれている場合の対処について説明する。このような音声が入力される場合には、最初に音声を複数のセグメントに大局的に分割し、次に個々の音声セグメントに対して、音声／非音声の判定や音声区間の検出を行うことが望ましい。

【０１１６】

図４０は、第１１の実施例に係る静止画像・音声認識装置のモジュール構成を示すブロック図である。図４０において、文字検出部２００１、文字認識部２００２及び音声認識部２００５、静止画音声対応部２００６は、それぞれ図３に示す文字検出部２０１、文字認識部２０２、音声認識部２０４、静止画音声対応部２０５と同じであるため説明は省略する。

【０１１７】

図４０において、２００３は音声分割部であり、音声を大局的に捉え、複数のセグメントに分割する。具体的には、音声信号をフレーム処理し、スペクトル情報を求め、複数フレーム間のスペクトルの類似性から着目しているフレームをセグメント境界とするか否かを判定する等の方法によって分割することができる。

【０１１８】

次に、音声検出部２００４では、音声分割部２００３で分割された各セグメントに音声が含まれるか否かを判定し、音声が含まれる場合には音声区間を検出する。具体的には、音声及び非音声のそれぞれに対して事前にＧＭＭ（Gaussian Mixture Model）を作成し、入力音声をフレーム処理することによって得られるスペクトル情報とこれらのＧＭＭを用いて、当該セグメントに音声が含まれるか否かを判定する。そして、音声が含まれていないと判定された場合は音声認識の対象とせず、音声が含まれていると判定された場合は、２００４における次の処理として音声区間を検出し、検出された音声区間を２００５の音声認識部に入力する。

【０１１９】

ここで、セグメント数は、セグメント間又はセグメント境界における音声スペクトルに関する尤度基準を用いて音声から決定する方法が考えられるが、これに限らず、静止画像分割、文字領域、文字認識結果によって得られる情報を用いて決定することもできる。具体的には、静止画像分割および文字領域の情報としては、分割数又は領域数に応じてセグメント数を変更する。文字認識結果の情報としては、文字認識結果全体の確からしさが高い場合にはセグメント数を増やすといった方法である。

【０１２０】

次に、文字領域の内容と同じ発声がされていない場合、一部の文字領域の内容に関する発声は全くされていない場合、どの文字領域とも関係のない発声が含まれていたりする場合、複数の文字領域に関する音声が十分な無音区間なしに連続的に発声される場合について説明する。

【０１２１】
図３９は、図３１（ａ）に示す静止画像に関連する音声を説明するための図である。この例では、図３６の文字領域の内容と同じ発声がなされておらず、また、図３９の３番目の音声区間である「これまでの研究では、...」の部分は静止画像のどの文字領域とも関係のない発声であるとする。さらに、図３９に示されるように、２番目から４番目の音声には十分な無音区間が存在しないものとする。
【０１２２】
図３９に示すような発声に対しては、音声分割部２００３又は音声検出部２００４が、正確に静止画像の文字領域に対応した音声分割又は音声区間を検出することは困難である。そこで、音声検出部２００４で検出された音声区間に対して音声認識部２００５で音声認識を行い、音声認識の結果から音声検出部２００４で決定された音声区間を必要に応じてさらに分割するようにする。
【０１２３】
具体的には、無音区間が十分に存在しない音声に対しては、音声認識部２００５による音声認識として大語彙連続音声認識に基づく方法を用いれば、句点を推定することによって文の区切りが分かるため、図４１に示すように、この情報を用いて音声区間を分割することができる。ここで、図４１は、図３１の例に対する文字認識結果情報と音声認識結果情報の一例を示す図である。また、文字領域の内容に関する発声がない場合、又はどの文字領域とも関係のない発声がなされている音声に対しては、音声認識結果と文字認識の結果をそれぞれ部分マッチングすることによって対応付けを行うことが可能である。
【０１２４】
また、第７の実施例で説明したように、文字認識の結果から重要語を検出すれば、この重要語をキーワードとしたワードスポッティングに基づく方法を音声認識部２００５による音声認識とすれば、より直接的に文字認識の結果と音声認識の結果を対応付けることが可能となる。図４２は、重要語抽出によるワードスポッティングを用いた場合の音声認識結果情報の一例を示す図である。図４２に示す例では、文字認識結果から重要語として抽出された「音声認識」、「文字認識」、「統計的言語モデル」、「目的」等の言葉を音声認識のワードスポッティングとしている。なお、図４２における「＊」は、これらのキーワード以外の音声区間を表し、また、「NO_RESUTLS」は、この音声区間に対してはどのキーワードもマッチングしなかったことを表している。このワードスポッティング結果と文字認識結果から得られる重要語をマッチングさせることによって、文字領域と音声の対応付けを行うことができる。
【０１２５】
以上の説明から明らかなように、本実施例によれば、音声に雑音や音楽といった音声以外のものが含まれている場合や、無音区間が十分に存在しない場合、文字領域の内容に関する発声がない場合、どの文字領域とも関係のない発声がなされている場合の音声であっても文字領域と音声との対応付けを行うことが可能となる。
【実施例１２】
【０１２６】
上記第１０の実施例では、複雑な静止画像に対しても文字領域と音声との対応付けを行えるようにするために、静止画像領域を大局的に分割する方法について説明した。本実施例では、この静止画像分割処理を分割数の異なる分割静止画像を階層的な構造として得ることによって、より柔軟な対応付けを行うことができることを説明する。
【０１２７】
図４３は、図３３で示した静止画像の分割をさらに行った場合の分割結果（ａ）（一点破線）、（ａ）をさらに分割した場合の結果（ｂ）（二点破線）を示す図である。尚、分割数の増減は、分割するか否かの基準（例えば、尤度基準に対する閾値）を変化させることによって制御することができる。ここで、図４３（ａ）は図３３の結果を元に分割されており、また、図４３（ｂ）は図４３（ａ）の結果を元に分割されているため、分割は階層的に行われている。

【０１２８】
図４４は、階層的な静止画像の分割を木構造で表現した例を示す図である。図４４において、黒丸はルートノードであって静止画像全体を表している。また、Ｉ１～Ｉ５の５個のノードは、図３３の分割領域に対する静止画像であり、Ｉ１は、図３３の分割領域の「音声認識・文字認識のための統計的言語モデルの利用」を含む画像領域、Ｉ２は、「目的」、「音声認識性能の向上」、「文字認識性能の向上」を含む画像領域、Ｉ３は、「提案法」、「統計的言語モデルの利用」、「単語間、文字間の...可能となる」を含む画像領域、Ｉ４は、「実験結果」、「認識率」、「音声認識」、「文字認識」を含む画像領域、Ｉ５は、「結論」、「統計的言語モデルは、...分かった。」を含む画像領域である。
【０１２９】
また、次の階層のＩ２１～Ｉ５２の１１個のノードは、図４３（ａ）の分割領域に対する静止画像であり、Ｉ２１は、「目的」を含む画像領域、Ｉ２２は、「音声認識性能の向上」及び「文字認識性能の向上」を含む画像領域、Ｉ３１は、「提案法」を含む画像領域、Ｉ３２は、「統計的言語モデルの利用」を含む画像領域、Ｉ３３は、下矢印記号を含む画像領域である。尚、図４３（ａ）の分割時にはＩ１の画像領域分割が施されていないため、Ｉ１のノード分割はない。
【０１３０】
同様に、最下階層のＩ２２１～Ｉ４３２の４個のノードは、図４３（ｂ）の分割領域に対する静止画像であり、Ｉ２２１は、「音声認識性能の向上」を含む画像領域、Ｉ２２２は、「文字認識性能の向上」を含む画像領域、Ｉ４３１は、「音声認識」を含む画像領域、Ｉ４３２は、「文字認識」を含む画像領域である。
【０１３１】
本実施例では、音声のセグメント分割又は音声区間検出は必ずしも階層的に行う必要はないが、ここでは階層的に行った場合の例を示す。図４５は、階層的に音声分割を行った場合の例を示す図である。また、図４６は、図４５で階層的に分割された音声を木構造で表現した例である。
【０１３２】
次に、前述した実施例で説明したいずれかの方法によって、図４４に示す各ノードに対応する画像領域から文字領域を抽出し、文字認識を施すことによって、文字認識結果情報を得ることができる。同様に、前述した実施例で説明したいずれかの方法によって、図４６に示す各ノードに対応する音声セグメントから音声区間を検出し、音声認識を施すことによって、音声認識結果情報を得ることができる。
【０１３３】
そして、これらの文字認識結果情報に音声認識結果情報を対応付ける。対応付けの方法は、前述した実施例で説明したいずれの方法を用いればよい。また、木構造の特徴を生かした対応付けの方法として、静止画像の上位ノードから下位ノードの順に対応付けを行い、その際に、上位ノードの対応付けの結果を下位ノードの対応付けにおいて制約として利用することができる。例えば、下位ノードの音声を対応付ける際に、上位ノードで対応付けられた音声区間に含まれる音声を優先的にもしくは限定的に選択する。他にも、上位ノードほど時間的に長い音声区間を優先的に選択し、下位ノードほど時間的に短い音声区間を優先的に選択する等の方法を用いることができる。
【０１３４】
図４７は、静止画像の木構造ノードに複数候補の分割音声を対応付けた結果の一例を示す図である。図４７において、「NULL」は音声区間の候補がなかった場合を示しており、特にＩ３３に対しては、どの音声区間にも対応付けられなかったことを表している。図４８は、図３１の例に対する静止画像と音声の対応結果を用いたアプリケーションの一例を示す図である。図４８に示す例では、静止画像の文字の場所にマウスカーソル（矢印マーク）を持っていくと、この文字に対応した音声データが再生され、スピーカー等の音声出力装置から出力される。
【０１３５】

また、図４８とは逆に、音声を先頭から、或いはマウス等で任意の時間を指定することによって音声を再生し、再生されている音声区間に対応する静止画像に枠を付与して表示することも可能である。図６１は、図４３に示す静止画像と音声との対応結果を用いた別のアプリケーションに基づく表示例を示す図である。この例では、利用者が「そこで、本研究では、統計的...」と音声認識された音声区間（ｓ４からｅ４）にマウスカーソル（矢印マーク）を持っていくと、この音声区間に対応した文字領域の座標に文字領域分の外枠が生成・表示される。この結果、出力されている音声が静止画像のどの部分に対応しているかを理解することができる。

【０１３６】

本実施例で説明した静止画像を木構造表現すること、また、複数候補の音声を対応付けることは、静止画像と音声の対応付けに誤りを含む場合に特に有効である。図４９は、静止画像の木構造の結果及び複数候補音声を利用する際のユーザインタフェースの一例を示す図ある。図４９では、上位候補の音声出力に左矢印キー「←」を、下位候補の音声出力に右矢印キー「→」を、静止画像の親ノードへ移動して１位候補の音声出力をするために上矢印キー「↑」を、静止画像の子ノードへ移動して１位候補の音声出力をするために下矢印キー「↓」をそれぞれ割り当てている。そして、利用者がマウス等によって所望の画像領域を選択（クリック等）すると、選択領域に含まれる画像領域の最下位ノードに対応する文字領域を枠で囲み画面上に表示し、さらに１位候補の音声を出力する。この際、音声又は画像領域が所望のものでない場合には、これら４つのキーのみを用いて他を選択する簡単な操作によって、他の候補を効率よく探すことが可能となる。

【実施例１３】

【０１３７】

前述した実施例では、文字認識の結果又はこれから抽出された重要語と音声認識の結果をマッチングしていたため、文字認識から得られる文字列と音声認識結果から得られる文字列が少なくとも部分的には同じである必要があった。すなわち、例えば、「題目」という文字認識結果に対して「タイトル」という発声がなされたり、「夏」に対して「暑い」という発声がなされた場合には対応付けを行うことはできない。そこで、本実施例は、このような場合においても静止画像と音声を対応付けることが可能となる方法を提供する。

【０１３８】

図５０は、第１３の実施例における静止画像と音声の例を示す図である。図５０より明らかなように、静止画像に含まれる「春」、「夏」、「秋」、「冬」という単語列は、音声の中に一切含まれていない。この場合、文字認識の結果と、音声認識の結果をそれぞれ抽象化、すなわち概念に変換し、それぞれの概念レベルでマッチングを行うことによって図５０のような場合であっても静止画像と音声を対応付けることが可能となる。

【０１３９】

図５１は、第１３の実施例における文字概念変換機能及び音声概念変換機能を有する静止画像・音声認識装置のモジュール構成を示すブロック図である。図５１において、文字検出部２１０１、文字認識部２１０２、音声検出部２１０４、音声認識部２１０５は、それぞれ図３に示す静止画像・音声認識装置のモジュールと同様であるため説明は省略する。図５１において、２１０３は文字概念変換部であり、文字認識部２１０２で得られる文字認識の結果を予め定められた概念に抽象化する。

【０１４０】

また、２０１６は音声概念変換部であり、音声認識部２１０５で得られる音声認識の結果を予め定められた概念に抽象化する。２１０７は概念対応部であり、文字概念変換部２１０３と音声概念変換部２１０６で得られる結果に対して概念レベルでマッチングを行う。静止画音声対応部２１０８は、概念対応部２１０７で対応付けられた概念に対して静止画像と音声を対応付ける。

【０１４１】

例えば、$SPRING、$SUMMER、$AUTUMN、$WINTERという４つの概念が定義されており、各概念に含まれる文字列として、S$SPRING＝｛春、spring、桜、入学式、...｝、$SUMMER＝

{夏、summer、hot、暑、...}、$AUTUMN＝{秋、autumn、fall、紅葉、...}、$WINTER＝{冬、winter、cold、寒、...}が定義されているとする。図５２は、文字概念変換結果と静止画像の座標情報、及び音声概念変換結果と音声の時間情報の一例を示す図である。そこで、図５０における静止画像及び音声に対して、図５２に示すような関係があるとする。尚、この例の場合は、音声認識として英語が認識できるものを用いているとする。

【０１４２】

そこで、この結果を概念対応部２１０７で対応付けることによって、$SPRING同士、$SUMMER同士等がそれぞれ対応付けられ、静止画音声対応部２１０８では、「春」の画像領域に対して「入学式の...」の音声が、「夏」の画像領域に対して「暑くなって...」の音声が、「秋」の画像領域に対して「紅葉狩りに...」の音声が、「冬」の画像領域に対して「Winter is a...」の音声がそれぞれ対応付けられる。

【０１４３】

以上の説明から明らかなように、本実施例によれば、文字列ではなく概念レベルでマッチングを行うことによって、文字認識から得られる文字列と音声認識結果から得られる文字列が全く一致しない場合であっても文字領域と音声との対応付けを好適に行うことが可能となる。

【実施例１４】

【０１４４】

前述した実施例では、静止画像の文字領域の部分に対してのみ音声と対応付けることが可能であり、静止画像中の文字以外の、例えば円や三角形等の図形や、人、車等のオブジェクトに対しては音声を対応付けることはできなかった。そこで、本実施例では、このような場合においても静止画像と音声を対応付けることが可能な方法を提供する。

【０１４５】

図５３は、第１４の実施例において用いる静止画像とそれに対応付けられる音声の例を示す図である。図５３より明らかなように、静止画像には文字列が一切含まれていない。この場合、前述した実施例における文字認識の代わりに、オブジェクト認識を行い、その認識結果と音声認識の結果をマッチングすることによって図５３のような場合であっても静止画像と音声を対応付けることが可能となる。

【０１４６】

図５４は、本発明の第１４の実施例に係るオブジェクト認識処理機能を有する静止画像・音声処理装置のモジュール構成を示すブロック図である。図５４において、音声検出部２２０３及び音声認識部２２０４は、図３に示すそれぞれのモジュールと同様であるため説明は省略する。図５４において、２２０１はオブジェクト検出部であり、静止画像からオブジェクト領域を抽出する。また、２２０２はオブジェクト認識部であり、オブジェクト検出部２２０１で抽出されたオブジェクトを認識する。尚、オブジェクト検出処理及びオブジェクト認識処理については、既存の技術を用いることができる。

【０１４７】

本実施例では、例えば、円、三角形、長方形、正方形等の図形の形状、棒グラフ、折れ線グラフ、円グラフ等のグラフの形状、およびそれぞれの形状に対する代表的な色の抽出が可能なオブジェクト検出処理及びオブジェクト認識処理が実施できるとする。この場合、図５４（ａ）の静止画像に対して、図５５（ａ）に示されるようなオブジェクト認識結果情報が得られる。

【０１４８】

図５５は、オブジェクト認識結果情報の例（ａ）とオブジェクト認識結果情報から得られる画像領域の例（ｂ）を示す図である。 図５５に示すように、オブジェクト認識結果として得られる「長方形」、「黒」、「正方形」、「白」といったオブジェクトの形状や色を表す言葉を文字列とし、この文字列と音声認識結果を２２０５で比較することによって、静止画像と音声を対応付けることができる。この結果、図５５（ｂ）で示されるように、静止画像のオブジェクトと音声が対応付けられる。

【０１４９】

以上の説明から明らかなように、本実施例によれば、オブジェクトを検出・認識する機能を備えることによって、静止画像に文字列が含まれない場合であっても音声との対応付けを好適に行うことが可能となる。

【実施例１５】

【０１５０】

前述した実施例では、静止画像と音声を対応付ける場合に、音声は音声認識を行っていたが、静止画像に人物が含まれ、この人物もしくは人物のクラスが特定でき、さらに、音声は、静止画像の人物もしくは人物クラスに関連している場合には、音声認識を行う代わりに、話者もしくは話者クラスの識別を行うことによって、静止画像と音声を対応付けることが可能となる。

【０１５１】

図５６は、第１５の実施例において用いる静止画像とそれに対応付けられる音声の例を示す図である。図５６より明らかなように、静止画像には文字列が一切含まれていない。また、音声は、高齢者・男性音声で「戦争の頃は...」、成人・男性音声で「僕は来年受験が...」、子供・女性音声で「今日の給食は...」、成人・女性音声で「今夜のドラマは...」という発声がなされているものとする。

【０１５２】

図５７は、本発明の第１５の実施例に係る人物認識機能及び話者認識機能を有する静止画像・音声認識装置のモジュール構成を示すブロック図である。図５７において、２３０１は人物検出部であり、静止画像から人物に関する画像領域を検出する。２３０２は人物認識部であり、人物検出部２３０１で検出された画像領域に対して、人物又は人物クラスの認識を行う。２３０３は音声検出部であり、音声区間の検出を行う。２３０４は話者認識部であり、音声検出部２３０３で検出された音声区間に対して、話者又は話者クラスの認識を行う。

【０１５３】

いま、人物認識部２３０２が、男性／女性の性別、及び子供／成人／高齢者の年代からなる人物クラスが認識できるとし、話者認識部２３０４も同様に男性／女性の性別、及び子供／成人／高齢者の年代からなる話者クラスが認識できるものとする。図５８は、第１５の実施例における人物認識結果情報及び話者認識結果情報の一例を示す図である。ここで、静止画音声対応部２３０５は、人物クラスと話者クラスのマッチングをとることによって、図５９に示すように静止画像と音声の対応付けをすることができる。すなわち、図５９は、人物認識結果情報から得られる画像領域を示す図である。

【０１５４】

以上の説明から明らかなように、本実施例によれば、人物又は人物クラスを検出・認識する機能と話者又は話者クラスを認識する機能を備えることによって、静止画像に文字列が含まれない場合に、音声認識を行うことなく音声との対応付けを行うことが可能となる。

【実施例１６】

【０１５５】

前述した実施例では、静止画像と音声がそれぞれ１つずつ存在する場合の対応方法について説明したが、本発明の適用はこれだけに限られることなく、例えば静止画像２つと音声３つを対応付ける等、任意の数の静止画像と音声を対応付けるようにしてもよい。

【０１５６】

尚、上述した第１～第１５の実施例では静止画像を対象として説明したが、動画像が例えば複数のカテゴリ等に分割されており、各カテゴリの代表的なフレーム（静止画像）に対して本発明を適用することで、所望の動画像を検索することも可能である。

【実施例１７】

【０１５７】

以上、実施形態例を詳述したが、本発明は、例えば、システム、装置、方法、プログラムもしくは記憶媒体等としての実施態様をとることが可能であり、具体的には、複数の機

器から構成されるシステムに適用しても良いし、また、一つの機器からなる装置に適用しても良い。

【０１５８】

尚、本発明は、前述した実施形態の機能を実現するソフトウェアのプログラム（実施形態では図に示すフローチャートに対応したプログラム）を、システムあるいは装置に直接あるいは遠隔から供給し、そのシステムあるいは装置のコンピュータが該供給されたプログラムコードを読み出して実行することによっても達成される場合を含む。

【０１５９】

従って、本発明の機能処理をコンピュータで実現するために、該コンピュータにインストールされるプログラムコード自体も本発明を実現するものである。つまり、本発明は、本発明の機能処理を実現するためのコンピュータプログラム自体も含まれる。

【０１６０】

その場合、プログラムの機能を有していれば、オブジェクトコード、インタプリタにより実行されるプログラム、ＯＳに供給するスクリプトデータ等の形態であっても良い。

【０１６１】

プログラムを供給するための記録媒体としては、例えば、フロッピー（登録商標）ディスク、ハードディスク、光ディスク、光磁気ディスク、ＭＯ、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、ＣＤ－ＲＷ、磁気テープ、不揮発性のメモリカード、ＲＯＭ、ＤＶＤ（ＤＶＤ－ＲＯＭ，ＤＶＤ－Ｒ）などがある。

【０１６２】

その他、プログラムの供給方法としては、クライアントコンピュータのブラウザを用いてインターネットのホームページに接続し、該ホームページから本発明のコンピュータプログラムそのもの、もしくは圧縮され自動インストール機能を含むファイルをハードディスク等の記録媒体にダウンロードすることによっても供給できる。また、本発明のプログラムを構成するプログラムコードを複数のファイルに分割し、それぞれのファイルを異なるホームページからダウンロードすることによっても実現可能である。つまり、本発明の機能処理をコンピュータで実現するためのプログラムファイルを複数のユーザに対してダウンロードさせるＷＷＷサーバも、本発明に含まれるものである。

【０１６３】

また、本発明のプログラムを暗号化してＣＤ－ＲＯＭ等の記憶媒体に格納してユーザに配布し、所定の条件をクリアしたユーザに対し、インターネットを介してホームページから暗号化を解く鍵情報をダウンロードさせ、その鍵情報を使用することにより暗号化されたプログラムを実行してコンピュータにインストールさせて実現することも可能である。

【０１６４】

また、コンピュータが、読み出したプログラムを実行することによって、前述した実施形態の機能が実現される他、そのプログラムの指示に基づき、コンピュータ上で稼動しているＯＳなどが、実際の処理の一部または全部を行ない、その処理によっても前述した実施形態の機能が実現され得る。

【０１６５】

さらに、記録媒体から読み出されたプログラムが、コンピュータに挿入された機能拡張ボードやコンピュータに接続された機能拡張ユニットに備わるメモリに書き込まれた後、そのプログラムの指示に基づき、その機能拡張ボードや機能拡張ユニットに備わるＣＰＵなどが実際の処理の一部または全部を行ない、その処理によっても前述した実施形態の機能が実現される。

【図面の簡単な説明】

【０１６６】

【図１】本発明の第１の実施例に係る画像データと音声データの部分データ同士を対応付ける静止画像・音声処理装置の構成を示すブロック図である。

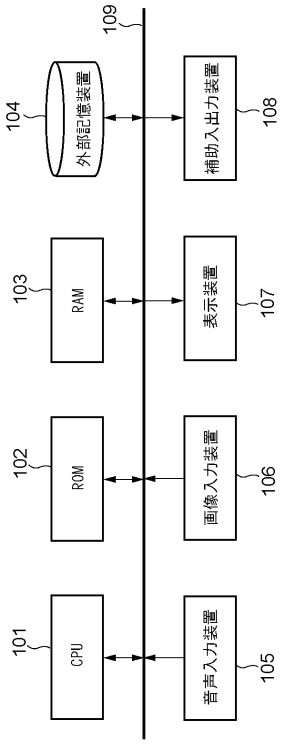
【図２】第１の実施例で互いに部分データの対応付け処理の対象となる静止画像（ａ）と当該静止画像に関連する音声（ｂ）の一例について示す図である。

【図３】本発明の第１の実施例において静止画像と音声を入力して静止画像と音声との対応関係（画像音声対応情報）を求める際のモジュール構成を示すブロック図である。
【図４】第１の実施例における文字認識部２０２の細部モジュール構成を示すブロック図である。
【図５】第１の実施例における音声認識部２０４の細部モジュール構成を示すブロック図である。
【図６】図７に示す文字認識結果情報と音声認識結果情報の例を示す図である。
【図７】図２に示す静止画像と音声の例に対する文字認識結果情報と音声認識結果情報を対応させた結果を示す図である。
【図８】第１の実施例における静止画像と音声との対応付けの一例を示す図である。
【図９】静止画像と音声との対応結果を用いたアプリケーションの例である。
【図１０】本発明の第２の実施例における発音列マッチングによる静止画音声対応部２０５の細部モジュール構成を示すブロック図である。
【図１１】第２の実施例における文字認識結果と音声認識結果に対する発音列の例を示す図である。
【図１２】本発明の第３の実施例における文字列マッチングを行う静止画音声対応部２０５の細部モジュール構成を示すブロック図である。
【図１３】第３の実施例における文字認識結果と音声認識結果に対する文字列の例である。
【図１４】第４の実施例における文字認識結果（ａ）と音声認識結果（ｂ）のスコア情報（尤度や確率等で表された認識結果）を保持した複数候補の例を示す図である。
【図１５】第４の実施例における文字認識結果を発音列に変換した結果（ａ）と音声認識結果から得られる発音列（ｂ）のスコア情報を保持した複数候補の例を示す図である。
【図１６】第４の実施例における文字認識結果から得られる文字列（ａ）と音声認識結果を文字列に変換した結果（ｂ）のスコア情報を保持した複数候補の例を示す図である。
【図１７】本発明の第６の実施例における静止画像・音声認識装置のモジュール構成を示すブロック図である。
【図１８】本発明の第６の実施例における文字認識結果を音声認識に利用する静止画像・音声認識装置のモジュール構成を示すブロック図である。
【図１９】第６の実施例における文字認識結果を発音列に変換して利用する静止画像・音声認識装置のモジュール構成を示すブロック図である。
【図２０】第６の実施例における文字認識結果を発音列に変換して探索部で利用する静止画像・音声認識装置のモジュール構成を示すブロック図である。
【図２１】第６の実施例における文字認識結果の文字列を音声認識結果を文字列に変換する際に利用する静止画像・音声認識装置のモジュール構成を示すブロック図である。
【図２２】第７の実施例における文字認識結果から重要語を抽出して探索部で利用する静止画像・音声認識装置のモジュール構成を示すブロック図である。
【図２３】第８の実施例における種々のフォント情報をもった静止画像の例である。
【図２４】第８の実施例における文字領域からフォント情報を抽出して文字認識結果情報として出力する静止画像・音声認識装置のモジュール構成を示すブロック図である。
【図２５】図２３に示す静止画像からの文字認識結果と各文字領域のフォント情報を示す図である。
【図２６】第９の実施例における文字認識結果情報出力部の細部モジュール構成を示すブロック図である。
【図２７】第９の実施例における静止画像・音声認識装置のモジュール構成を示すブロック図である。
【図２８】第９の実施例における音声認識結果を文字列に変換して利用する静止画像・音声認識装置のモジュール構成を示すブロック図である。
【図２９】第９の実施例における音声認識結果から得られる文字列を文字認識で利用する静止画像・音声認識装置のモジュール構成を示すブロック図である。

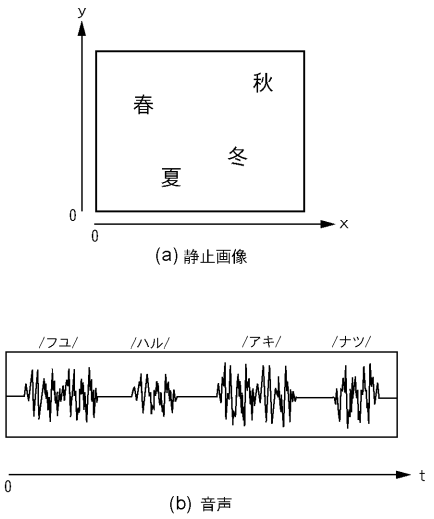
【図３０】第９の実施例における音声認識結果から得られる発音列を文字認識結果の発音列の変換に利用する静止画像・音声認識装置のモジュール構成を示すブロック図である。
【図３１】より複雑な静止画像（ａ）とそれに関連付けられた音声（ｂ）の一例を示す図である。
【図３２】図３１に示すような複雑な静止画像を分割する機能を有する静止画像・音声認識装置のモジュール構成を示すブロック図である。
【図３３】図３１（ａ）の静止画像を静止画像分割部１９０１によって５つの領域に分割された結果を示す図である。
【図３４】図３３に示す各分割領域に対して領域ＩＤを付与して図３２（ａ）の座標系における座標情報を表した図である。
【図３５】図３４と図３２（ａ）の対応関係について示す図である。
【図３６】文字検出部１９０２による文字領域の検出結果例を示す図である。
【図３７】文字認識及び音声認識の結果得られる文字認識結果情報（ａ）及び音声認識結果情報を（ｂ）示す図である。
【図３８】図３６に示す文字領域の検出結果に文字認識結果情報を対応させた結果を示す図である。
【図３９】図３１（ａ）に示す静止画像に関連する音声を説明するための図である。
【図４０】第１１の実施例に係る静止画像・音声認識装置のモジュール構成を示すブロック図である。
【図４１】図３１の例に対する文字認識結果情報と音声認識結果情報の一例を示す図である。
【図４２】重要語抽出によるワードスポッティングを用いた場合の音声認識結果情報の一例を示す図である。
【図４３】図３３で示した静止画像の分割をさらに行った場合の分割結果（ａ）（一点破線）、（ａ）をさらに分割した場合の結果（ｂ）（二点破線）を示す図である。
【図４４】階層的な静止画像の分割を木構造で表現した例を示す図である。
【図４５】階層的に音声分割を行った場合の例を示す図である。
【図４６】図４５で階層的に分割された音声を木構造で表現した例である。
【図４７】静止画像の木構造ノードに複数候補の分割音声を対応付けた結果の一例を示す図である。
【図４８】図３１の例に対する静止画像と音声の対応結果を用いたアプリケーションの一例を示す図である。
【図４９】静止画像の木構造の結果及び複数候補音声を利用する際のユーザインタフェースの一例を示す図ある。
【図５０】第１３の実施例における静止画像と音声の例を示す図である。
【図５１】第１３の実施例における文字概念変換機能及び音声概念変換機能を有する静止画像・音声認識装置のモジュール構成を示すブロック図である。
【図５２】文字概念変換結果と静止画像の座標情報、及び音声概念変換結果と音声の時間情報の一例を示す図である。
【図５３】第１４の実施例において用いる静止画像とそれに対応付けられる音声の例を示す図である。
【図５４】本発明の第１４の実施例に係るオブジェクト認識処理機能を有する静止画像・音声処理装置のモジュール構成を示すブロック図である。
【図５５】オブジェクト認識結果情報の例（ａ）とオブジェクト認識結果情報から得られる画像領域の例（ｂ）を示す図である。
【図５６】第１５の実施例において用いる静止画像とそれに対応付けられる音声の例を示す図である。
【図５７】本発明の第１５の実施例に係る人物認識機能及び話者認識機能を有する静止画像・音声認識装置のモジュール構成を示すブロック図である。
【図５８】第１５の実施例における人物認識結果情報及び話者認識結果情報の一例を示す

図である。
【図５９】人物認識結果情報から得られる画像領域を示す図である。
【図６０】図２に示す静止画像と音声との対応結果を用いた別のアプリケーションに基づく表示例を示す図である。
【図６１】図４３に示す静止画像と音声との対応結果を用いた別のアプリケーションに基づく表示例を示す図である。
【図６２】第１０の実施例に係る静止画像・音声認識装置の処理動作を説明するためのフローチャートである。
【符号の説明】
【０１６７】
２０１　文字検出部
２０２　文字認識部
２０３　音声検出部
２０４　音声認識部
２０５　静止画音声対応部
３０１　前処理部
３０２　特徴抽出部
３０３　識別部
３０４　文字認識結果情報出力部
３０５　文字認識用テンプレート
３０６　文字認識用言語モデル
４０１　音声分析部
４０２　探索部
４０３　音声認識結果情報出力部
４０４　音声認識用音響モデル
４０５　音声認識用言語モデル

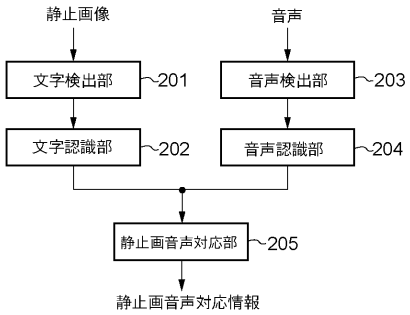
【図１】



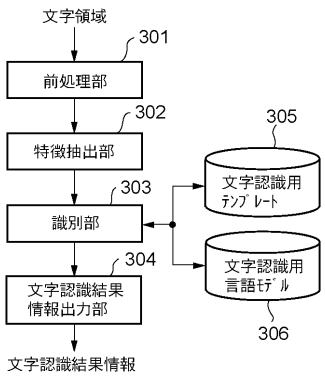
【図２】



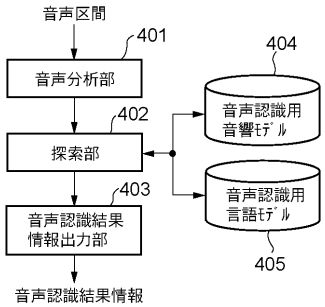
【図３】



【図４】



【図５】



【図６】

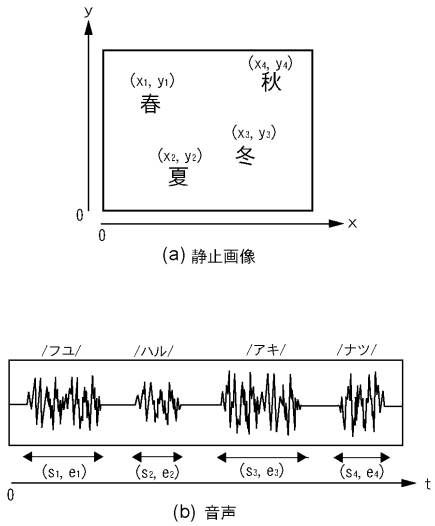
文字認識結果	座標情報
春	(x_1, y_1)
夏	(x_2, y_2)
秋	(x_3, y_3)
冬	(x_4, y_4)

(a) 文字認識結果情報の列

音声認識結果	時間情報
冬(フユ)	(s_1, e_1)
春(ハル)	(s_2, e_2)
秋(アキ)	(s_3, e_3)
夏(ナツ)	(s_4, e_4)

(b)音声認識結果情報の列

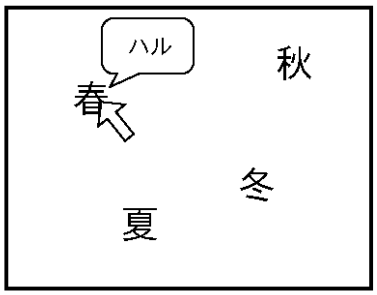
【図７】



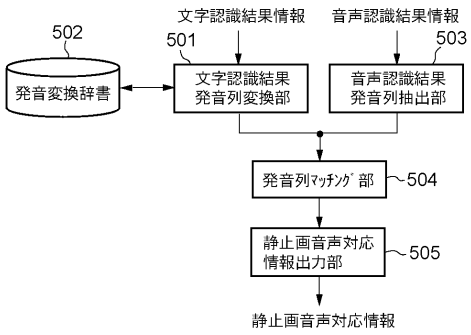
【図８】

静止画像	音声
(x_1, y_1)	(s_2, e_2)
(x_2, y_2)	(s_4, e_4)
(x_3, y_3)	(s_3, e_3)
(x_4, y_4)	(s_1, e_1)

【図９】



【図１０】



【図１１】

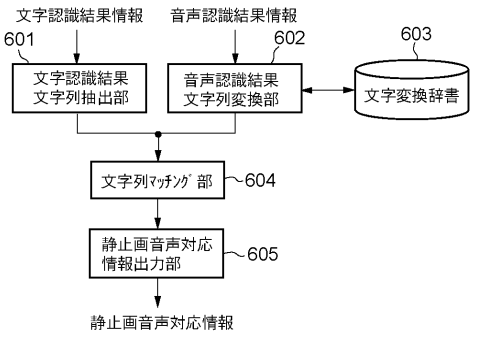
発音候補		座標情報
ハル	シュン	(x_1, y_1)
ナツ	カ	(x_2, y_2)
アキ	シュウ	(x_3, y_3)
フユ	トウ	(x_4, y_4)

(a) 文字認識結果に対する発音列候補

発音候補	時間情報
フユ	(s_1, e_1)
ハル	(s_2, e_2)
アキ	(s_3, e_3)
ナツ	(s_4, e_4)

(b) 音声認識結果に対する発音列

【図１２】



【図１３】

文字候補	座標情報
春	(x_1, y_1)
夏	(x_2, y_2)
秋	(x_3, y_3)
冬	(x_4, y_4)

(a) 文字認識結果に対する文字列

文字候補			時間情報
冬	不輸	一	(s_1, e_1)
春	張る	貼る	(s_2, e_2)
空	飽き	秋	(s_3, e_3)
夏	奈津	捺	(s_4, e_4)

(b) 音声認識結果に対する文字列候補

【図１４】

文字認識結果候補			
n/i	1	2	3
1	春(0.7)	香(0.2)	空(0.1)
2	科(0.5)	秋(0.3)	和(0.2)
3	夏(0.6)	真(0.3)	厚(0.1)
4	各(0.5)	冬(0.3)	尽(0.2)

(a) 文字認識結果候補の例

音声認識結果候補			
m/j	1	2	3
1	フユ(冬)(0.5)	フル(古)(0.4)	ツユ(露)(0.1)
2	ハル(春)(0.7)	タル(樽)(0.2)	ハク(白)(0.1)
3	アキ(秋)(0.6)	アシ(足)(0.2)	マキ(薪)(0.2)
4	マツ(松)(0.8)	ツ(津)(0.1)	ナツ(夏)(0.1)

(b) 音声認識結果候補の例

【図１５】

文字認識結果候補			
n/i	1	2	3
1	ハル(0.7) シュン(0.7)	カ(0.2) コウ(0.2)	ソラ(0.1) アキ(0.1) クウ(0.1)
2	カ(0.5) シナ(0.5)	アキ(0.3) シュウ(0.3)	ワ(0.2) ナゴ(0.2)
3	ナツ(0.6) カ(0.6)	シン(0.3) マ(0.3)	コウ(0.1) アツ(0.1)
4	カワ(0.5) オノ(0.5)	フユ(0.3) トウ(0.3)	ジン(0.2) ツ(0.2)

(a) 文字認識結果候補に対する発音列候補

音声認識結果候補			
m/j	1	2	3
1	フユ(0.5)	フル(0.4)	ツユ(0.1)
2	ハル(0.7)	タル(0.2)	ハク(0.1)
3	アキ(0.6)	アシ(0.2)	マキ(0.2)
4	マツ(0.8)	ツ(0.1)	ナツ(0.1)

(b) 音声認識結果候補に対する発音列

【図１６】

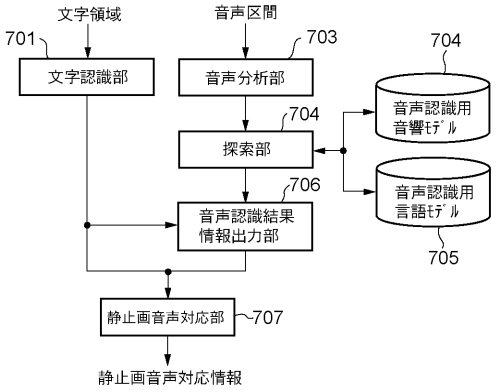
文字認識結果候補			
n/i	1	2	3
1	春(0.7)	香(0.2)	空(0.1)
2	科(0.5)	秋(0.3)	和(0.2)
3	夏(0.6)	真(0.3)	厚(0.1)
4	各(0.5)	冬(0.3)	尽(0.2)

(a) 文字認識結果候補に対す文字列

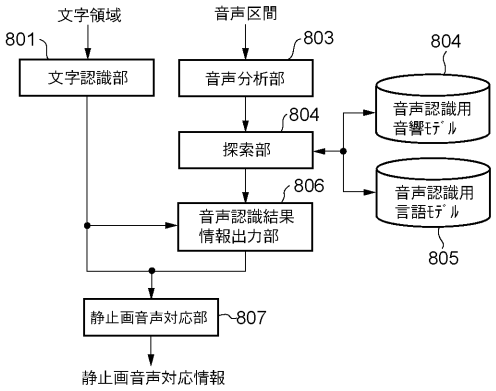
音声認識結果候補			
m/j	1	2	3
1	冬(0.5) 不輸(0.5)	古(0.4) 振る(0.4) 降る(0.4)	露(0.1) 梅雨(0.1) 汁(0.1)
2	春(0.7) 張る(0.7) 貼る(0.7)	樽(0.2) 足る(0.2)	白(0.1) 吐く(0.1) 履く(0.1)
3	空(0.6) 飽き(0.6) 秋(0.6)	足(0.2) 脚(0.2) 芦(0.2)	薪(0.2) 巻き(0.2) 真希(0.2)
4	松(0.8) 末(0.8) 待つ(0.8)	津(0.1)	夏(0.1) 奈津(0.1) 擦(0.1)

(b) 音声認識結果候補に対する文字列

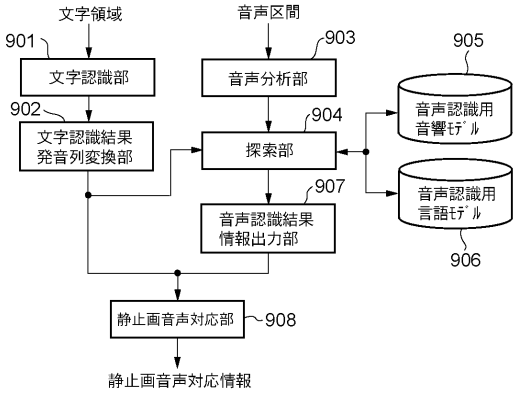
【図１７】



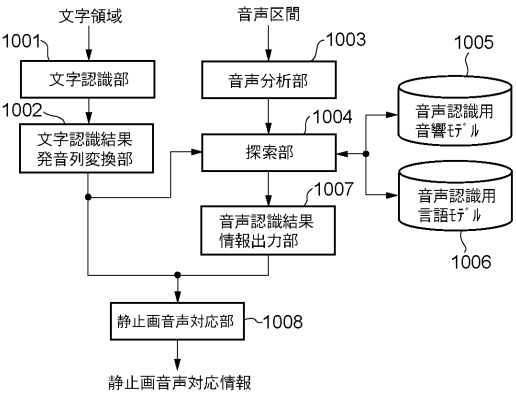
【図１８】



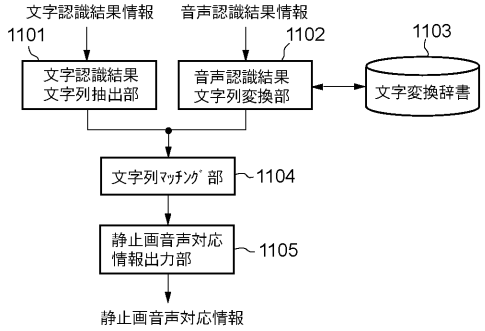
【図１９】



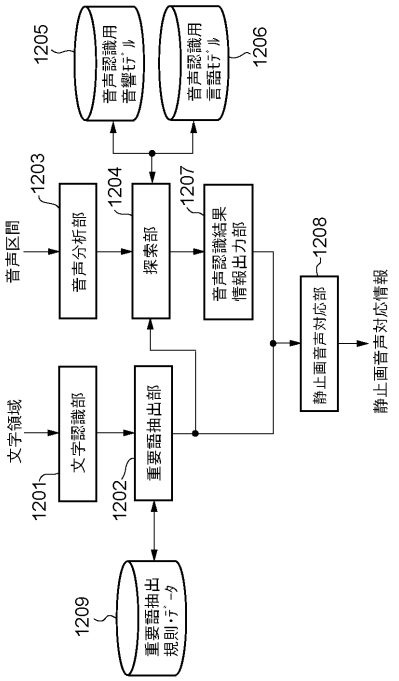
【図２０】



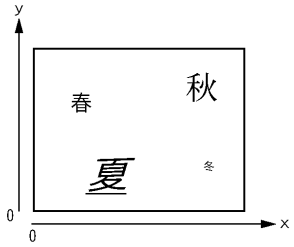
【図２１】



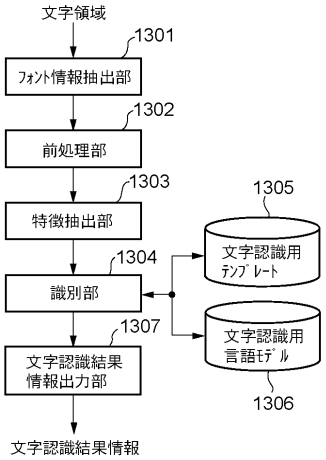
【図２２】



【図２３】



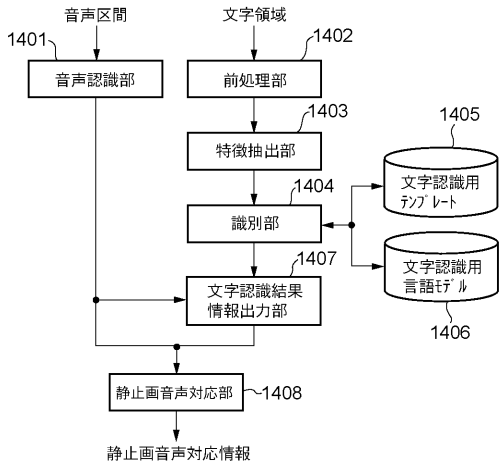
【図２４】



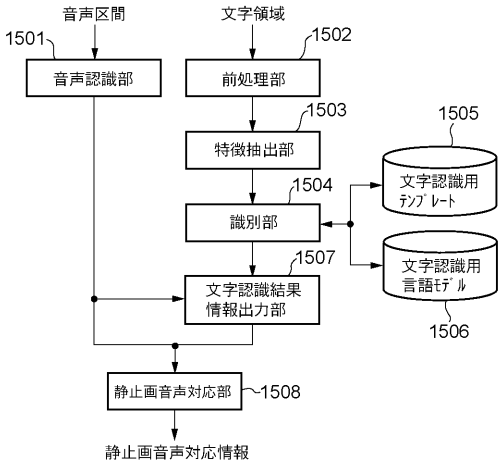
【図２５】

文字認識結果	フォント情報					座標情報
	サイズ	文字種	色	斜体	アンダーライン	
春	24	ゴシック	黒	N	N	(x_1, y_1)
夏	44	ゴシック	黒	Y	Y	(x_2, y_2)
秋	36	明朝	赤	N	N	(x_3, y_3)
冬	12	ゴシック	黒	N	N	(x_4, y_4)

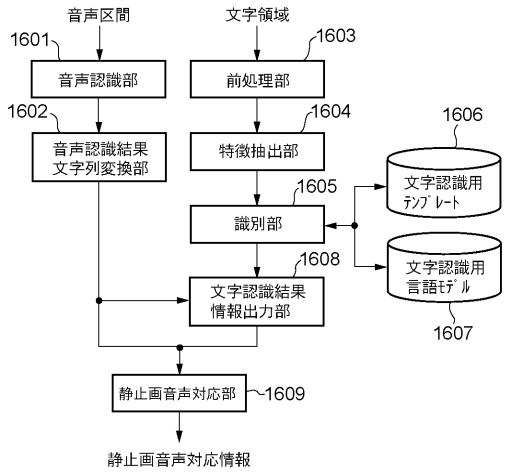
【図２６】



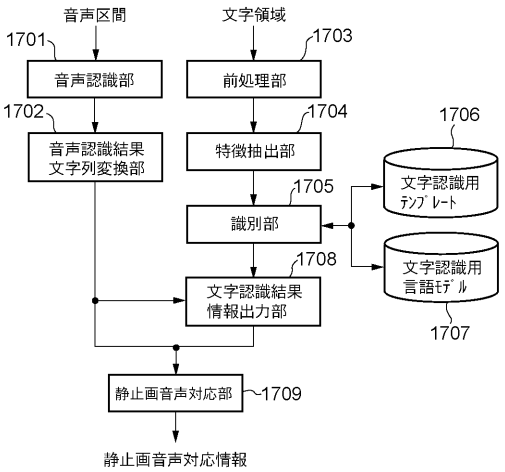
【図２７】



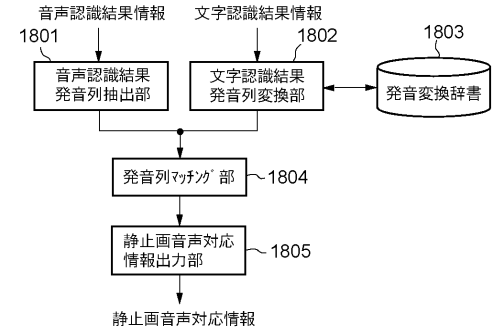
【図２８】



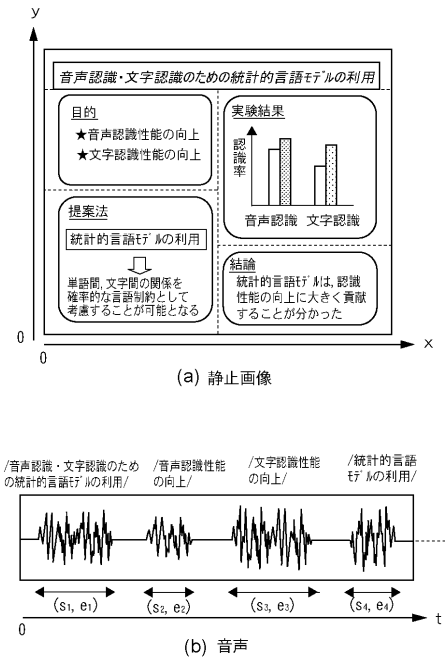
【図２９】



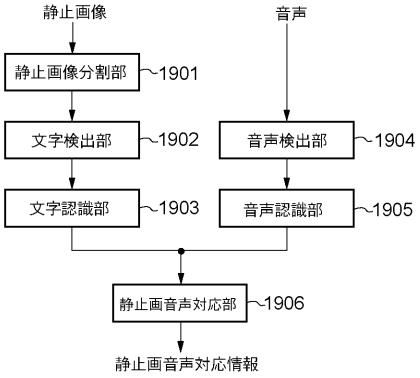
【図３０】



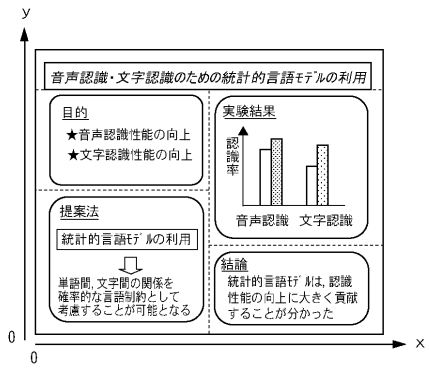
【図３１】



【図３２】



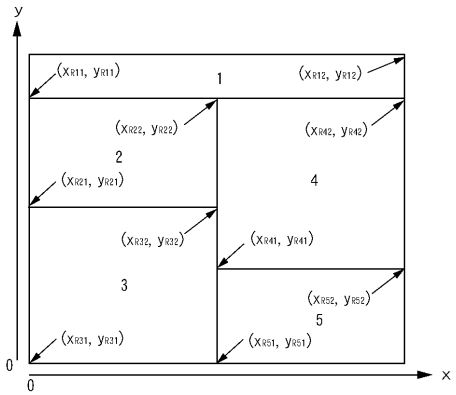
【図３３】



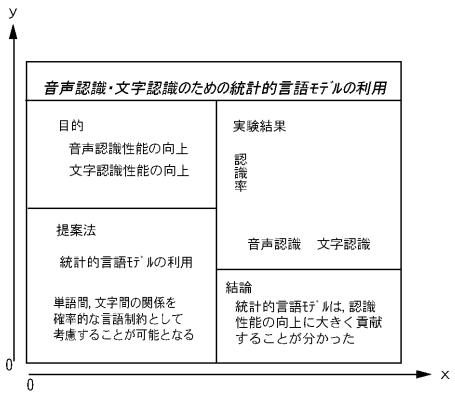
【図３４】

静止画分割結果(領域 D)	座標情報
1	$(x_{R11}, y_{R11}, x_{R12}, y_{R12})$
2	$(x_{R21}, y_{R21}, x_{R22}, y_{R22})$
3	$(x_{R31}, y_{R31}, x_{R32}, y_{R32})$
4	$(x_{R41}, y_{R41}, x_{R42}, y_{R42})$
5	$(x_{R51}, y_{R51}, x_{R52}, y_{R52})$

【図３５】



【図３６】



【図３７】

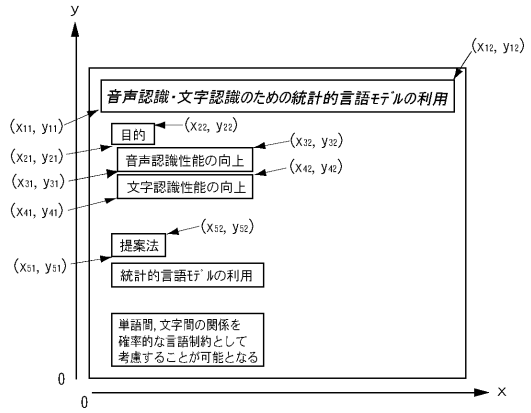
文字認識結果	座標情報
音声認識・文字認識のための統計的言語モデルの利用	$(x_{11}, y_{11}, x_{12}, y_{12})$
目的	$(x_{21}, y_{21}, x_{22}, y_{22})$
音声認識性能の向上	$(x_{31}, y_{31}, x_{32}, y_{32})$
文字認識性能の向上	$(x_{41}, y_{41}, x_{42}, y_{42})$
提案法	$(x_{51}, y_{51}, x_{52}, y_{52})$
・・・	・・・

(a)文字認識結果情報の列

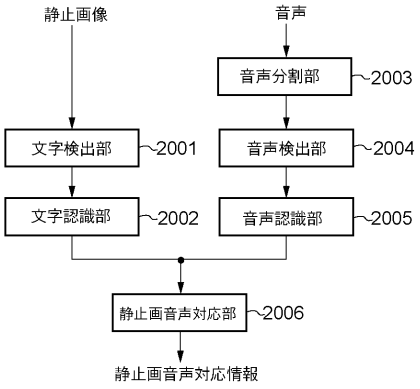
音声認識結果	時間情報
音声認識・文字認識のための統計的言語モデルの利用	(s_1, e_1)
目的	(s_2, e_2)
音声認識性能の向上	(s_3, e_3)
文字認識性能の向上	(s_4, e_4)
提案法	(s_5, e_5)
・・・	・・・

(b) 音声認識結果情報の列

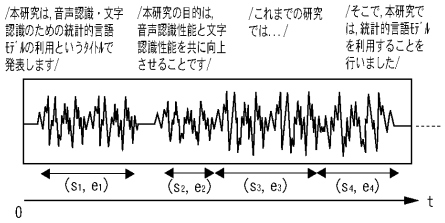
【図３８】



【図４０】



【図３９】



【図４１】

文字認識結果	座標情報
音声認識・文字認識のための統計的言語ﾓﾃﾞﾙの利用	(x11, y11, x12, y12)
目的	(x21, y21, x22, y22)
音声認識性能の向上	(x31, y31, x32, y32)
文字認識性能の向上	(x41, y41, x42, y42)
提案法	(x51, y51, x52, y52)
…	…

(a)文字認識結果情報の列

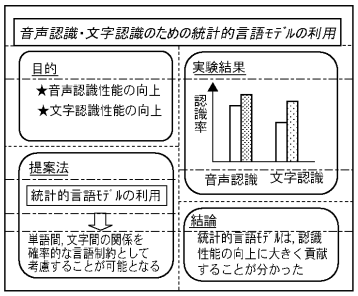
音声認識結果	時間情報
本研究，音声認識・文字認識の…	(s1, e1)
本研究の目的は，音声認識性能と…	(s2, e2)
これまでの研究では，…	(s3, e3)
そこで，本研究では，統計的言語…	(s4, e4)
…	…

(b) 音声認識結果情報の列

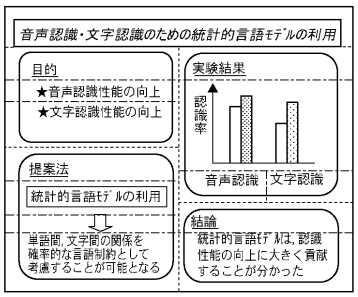
【図４２】

音声認識結果	時間情報
*音声認識*文字認識*統計的言語ﾓﾃﾞﾙ*	(s1, e1)
*目的*音声認識*文字認識*	(s2, e2)
NO_RESULTS	(s3, e3)
統計的言語ﾓﾃﾞﾙ	(s4, e4)
…	…

【図４３】

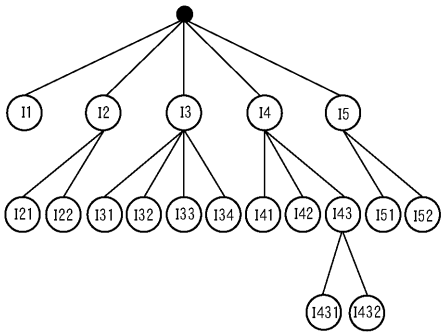


(a) 更に画像分割を行った結果

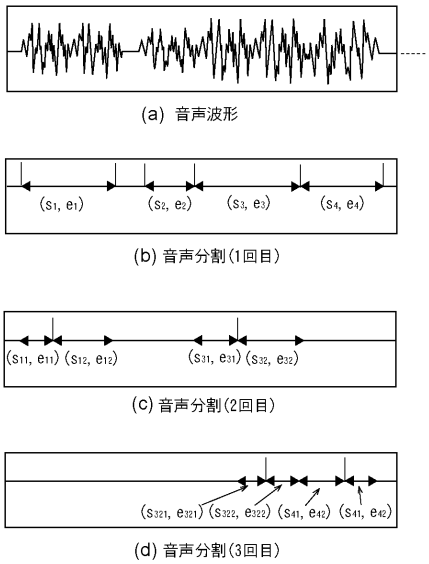


(b) (a)を更に画像分割した結果

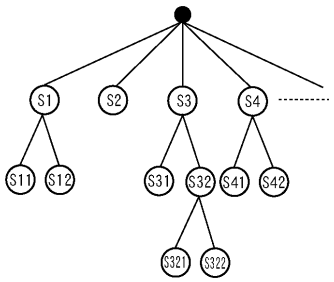
【図４４】



【図４５】



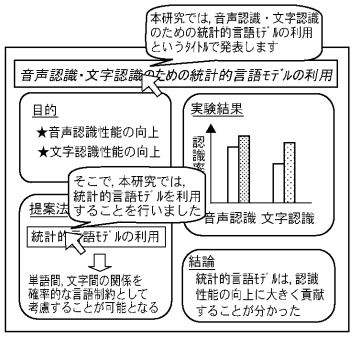
【図４６】



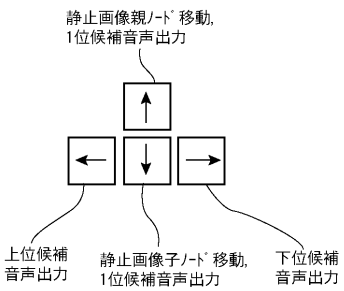
【図４７】

静止画像	音声		
	第1候補	第2候補	第3候補
I1	S2	S31	NULL
I2	S3	S31	S321
I3	S4	S41	S42
I4	S3	S321	S31
I5	S1	S11	S12
I21	S31	NULL	NULL
I22	S3	S2	S42
I31	S12	S1	S2
I32	S322	S321	NULL
I33	NULL	NULL	NULL
...	...		...

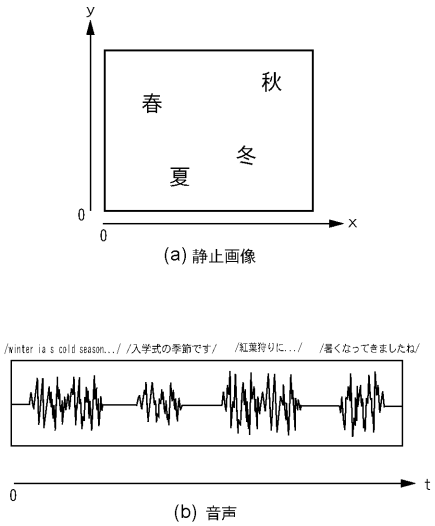
【図４８】



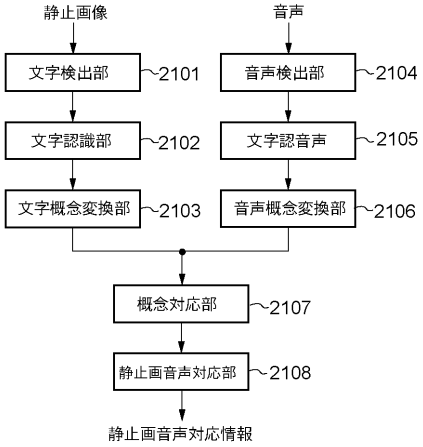
【図４９】



【図５０】



【図５１】



【図５２】

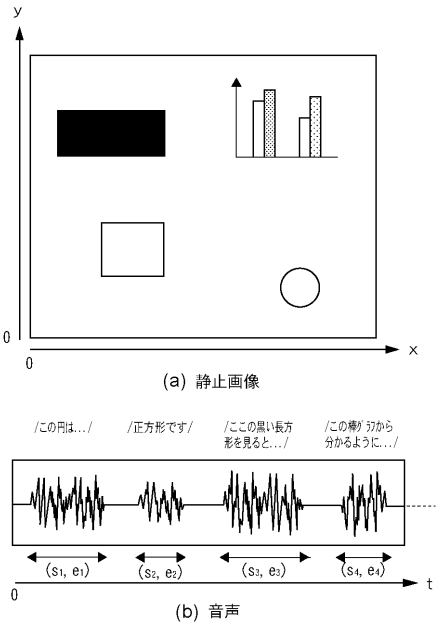
文字概念変換結果（文字認識結果）	座標情報
$ SPRING(春)	(x_1, y_1)
$ SUMMER(夏)	(x_2, y_2)
$ AUTUMN(秋)	(x_3, y_3)
$ WINTER(冬)	(x_4, y_4)

(a) 文字概念変換結果と座標情報の列

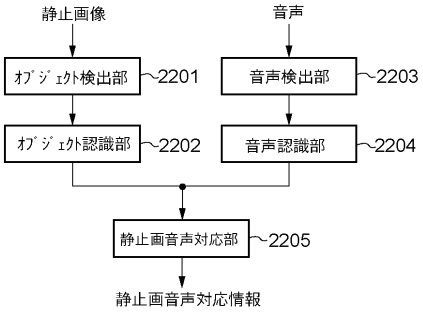
音声概念変換結果（音声認識結果）	時間情報
$ WINTER(Winter is a ...)	(s_1, e_1)
$ SPRING(入学式の...)	(s_2, e_2)
$ AUTUMN(紅葉狩りに...)	(s_3, e_3)
$ SUMMER(暑くなって...)	(s_4, e_4)

(b) 音声概念変換結果と時間情報の列

【図５３】



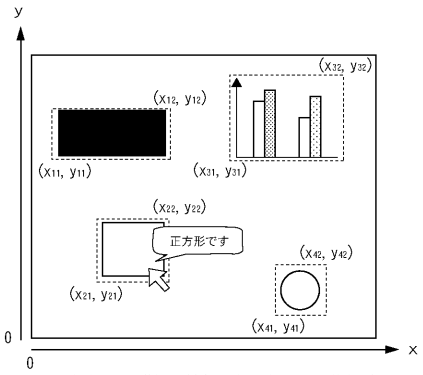
【図５４】



【図５５】

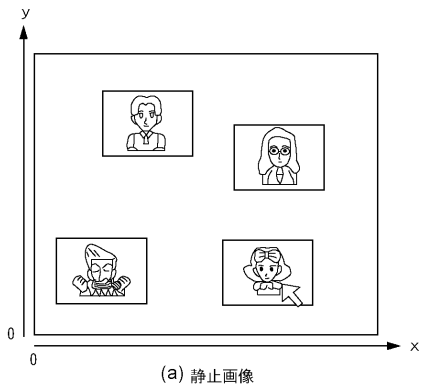
ｵﾌﾞｼﾞｪｸﾄ認識結果	座標情報
長方形, 黒	(x_{11}, y_{11}, x_{12}, y_{12})
正方形, 白	(x_{21}, y_{21}, x_{22}, y_{22})
棒ｸﾞﾗﾌ, 灰	(x_{31}, y_{31}, x_{32}, y_{32})
円, 白	(x_{41}, y_{41}, x_{42}, y_{42})

(a) ｵﾌﾞｼﾞｪｸﾄ認識結果情報の例

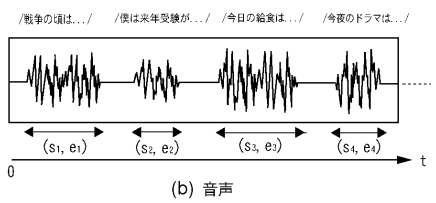


(b)ｵﾌﾞｼﾞｪｸﾄ認識結果情報から得られる画像領域

【図５６】

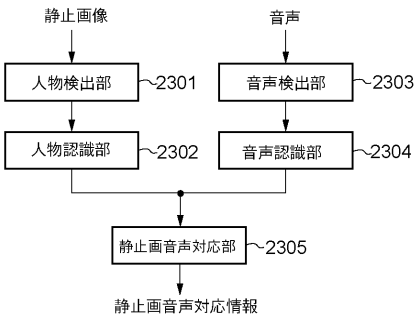


(a) 静止画像



(b) 音声

【図５７】



【図５８】

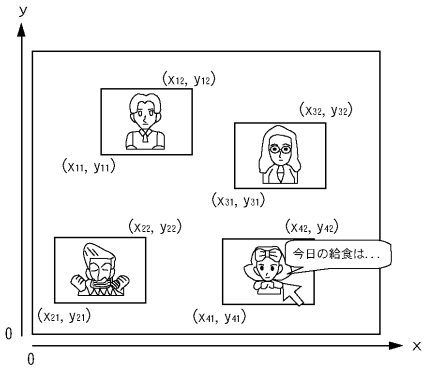
人物認識結果	座標情報
成人, 男性	(x_{11}, y_{11}, x_{12}, y_{12})
高齢者, 男性	(x_{21}, y_{21}, x_{22}, y_{22})
成人, 女性	(x_{31}, y_{31}, x_{32}, y_{32})
子供, 女性	(x_{41}, y_{41}, x_{42}, y_{42})

(a) 話者認識結果情報の列

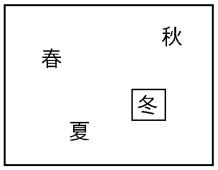
話者認識結果	時間情報
高齢者, 男性	(s_1, e_1)
成人, 男性	(s_2, e_2)
子供, 女性	(s_3, e_3)
成人, 女性	(s_4, e_4)

(b) 話者認識結果情報の列

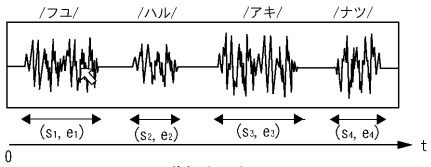
【図５９】



【図６０】

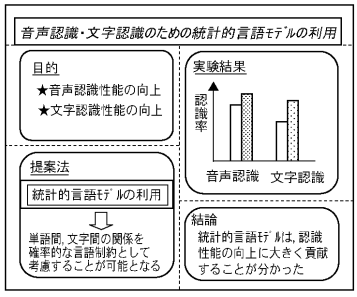


(a) 静止画像表示

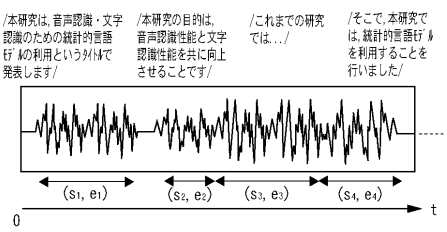


(b) 音声表示

【図６１】

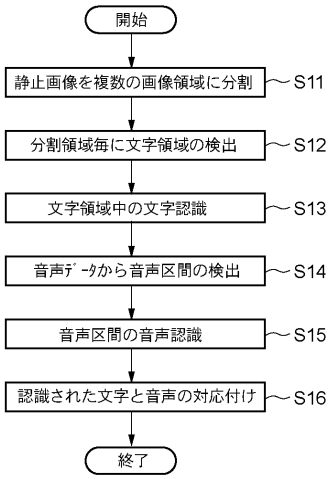


(a) 静止画像表示



(b) 音声表示

【図６２】



フロントページの続き

審査官　加藤　恵一

(56)参考文献　特開平０９－２１８９５５（ＪＰ，Ａ）
特開平０９－２３３４４２（ＪＰ，Ａ）
特開２００１－０３４１５１（ＪＰ，Ａ）
特開２００２－１９７１０３（ＪＰ，Ａ）
特開２００３－０８５５７２（ＪＰ，Ａ）
特開２００３－２４２１５０（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｈ０４Ｎ　５／７６－５／９５６