(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第5351958号
(P5351958)

(45)発行日　平成25年11月27日(2013.11.27)　(24)登録日　平成25年8月30日(2013.8.30)

(51)Int.Cl.		F I		
G06T 1/00	(2006.01)	G06T	1/00	200D
G06T 7/00	(2006.01)	G06T	7/00	300F
G06F 17/30	(2006.01)	G06F	17/30	170B
H04N 5/76	(2006.01)	G06F	17/30	210A
		H04N	5/76	B

請求項の数 2　(全 17 頁)

(21)出願番号	特願2011-512451 (P2011-512451)
(86)(22)出願日	平成21年5月22日 (2009.5.22)
(65)公表番号	特表2011-525012 (P2011-525012A)
(43)公表日	平成23年9月8日 (2011.9.8)
(86)国際出願番号	PCT/US2009/003160
(87)国際公開番号	WO2009/148518
(87)国際公開日	平成21年12月10日 (2009.12.10)
審査請求日	平成24年3月22日 (2012.3.22)
(31)優先権主張番号	61/058,201
(32)優先日	平成20年6月2日 (2008.6.2)
(33)優先権主張国	米国(US)
(31)優先権主張番号	12/331,927
(32)優先日	平成20年12月10日 (2008.12.10)
(33)優先権主張国	米国(US)

(73)特許権者　513077243
インテレクチュアル　ベンチャーズ　ファンド　８３　エルエルシー
アメリカ合衆国、８９１２８　ネバダ州、ラスベガス、ウエスト　レイク　ミード　ブールバード　７２５１、スイート　３００
(74)代理人　100107766
弁理士　伊東　忠重
(74)代理人　100070150
弁理士　伊東　忠彦
(74)代理人　100091214
弁理士　大貫　進介

最終頁に続く

(54)【発明の名称】デジタルコンテンツ記録のための意味論的イベント検出

(57)【特許請求の範囲】

【請求項１】

イベントに関連する画像記録のグループの意味論的イベント分類を容易にする方法であって、

前記画像記録の各々から複数の視覚的特徴を抽出するステップと、

前記視覚的特徴を使用して前記画像記録の各々に対する複数の概念スコアを生成するステップと、

前記画像記録の前記概念スコアに基づいて、各イベントを記述するためのＢＯＦ特徴ベクトルを、意味論的イベントに対応する予め定められたコードブックに前記イベントの前記画像記録の前記概念スコアをマッピングすることにより生成するステップと、

前記イベントに意味論的イベントが現れる確率の指標である検出スコアを生成する意味論的イベント分類器に前記マッピングされた特徴ベクトルを供給するステップと、

を包含し、

各前記概念スコアは、視覚的概念に対応し、前記画像記録が前記視覚的概念を含む確率の指標であることを特徴とする方法。

【請求項２】

請求項１に記載のイベントに関連する画像記録のグループの意味論的イベント分類を容易にする方法であって、

前記画像記録の対の間のペアワイズ類似性を決定するステップと、

スペクトルクラスタ化を適用して、前記決定されたペアワイズ類似性に基づいて、前記

意味論的イベントの訓練画像記録を各クラスタが一つのコードワードに対応する異なるクラスタにグループ化することによって各前記意味論的イベントの前記コードブックを生成するステップと、

前記訓練イベントの前記画像記録の前記概念スコアを意味論的イベントに対応する前記コードブックにマッピングして、各前記訓練イベントを記述するためのＢＯＦ特徴ベクトルを生成するステップと、

前記イベント分類器を前記訓練イベントに対応する前記ＢＯＦ特徴ベクトルに基づいて訓練するステップと、

を包含する訓練プロセスを有することを特徴とする方法。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、デジタル静止画像又はビデオのようなデジタルコンテンツ記録のカテゴリ化に関している。特に、本発明は、意味論的イベント（ｓｅｍａｎｔｉｃ　ｅｖｅｎｔｓ）の検出に基づいたデジタルコンテンツ記録のカテゴリ化に関している。

【背景技術】

【０００２】

低コストの電子消費者撮像技術の出現は、平均的な消費者によって獲得されるデジタル画像の数の顕著な増加をもたらす結果となっている。実際、様々な形態の電子メモリが時間とともにますます安価になっているので、消費者は、より一層多くのデジタル静止画像及びビデオを撮影するとともに、以前には廃棄したであろうデジタル静止画像及びビデオも保持する傾向にある。結果として、平均的な消費者は、記憶及び後の検索のためにデジタル画像を適切に識別及びカタログ化するにあたって、ますます困難な問題に直面している。一般的に、そのような識別及びカタログ化は通常は手作業で実行され、これは消費者にとって極端に時間を消費するプロセスになることがある。

【０００３】

単なる一つの描写として、消費者は1回の休暇の間にいくつもの異なる場所に旅行するかもしれない。消費者は、特定の場所の各々で、ならびに他の主題カテゴリ又はイベントに関係している場所の各々で、画像を撮影し得る。例えば、消費者は、それらの場所の各々で家族メンバの画像を撮影し、それらの場所の各々で特定のイベントの画像を撮影し、それらの場所の各々で歴史的な建造物の画像を撮影し得る。旅行から戻ると、消費者は、デジタル画像を人物、誕生日、博物館などの様々なグループ分けに基づいて分類し、デジタル画像をそのグループ分けに基づいて電子アルバムに記憶したいと思うかもしれない。消費者は、現在のところ、何百というデジタル静止画像及びビデオセグメントを特定のイベントで識別するために、それらを手作業で分類するということに直面している。

【０００４】

上記のことを考慮して、最近、消費者の写真及びビデオの自動アルバム化が大きな関心を集めている。自動アルバム化に対する一つの人気のあるアプローチは、デジタル画像及びビデオを日付順及び画像コンテンツ内の視覚的な類似性によるイベントに従って組織化することである。例えば、非特許文献１には、デジタル画像のグループがどのようにして自動的にイベントにクラスタ化されることができるかが記載されている。

【先行技術文献】

【非特許文献】

【０００５】

【非特許文献１】A.C.Loui及びA.Savakis, 「Automated event clustering and quality screening of consumer pictures for digital albuming（デジタルアルバム化のための消費者写真の自動イベントクラスタ化及び質のスクリーニング）」, IEEE Trans. on Multimedia, ２００３年, Vol.5, No.3, p.390－402

【発明の概要】

【発明が解決しようとする課題】

【０００６】
画像の基本的なクラスタ化は単一のイベントに関連しているように見える画像をグループ化することができるが、自動アルバム化プロセスを改善するために、クラスタ化されたイベントに意味論的意味（ｓｅｍａｎｔｉｃ　ｍｅａｎｉｎｇｓ）のタグ付けをすることができることが望ましいであろう。しかし、意味論的イベントの検出は、基本的な問題を提示する。第１に、実用的なシステムは、デジタル静止画像及びビデオを同時に処理することができる必要がある。これは、しばしば両方が、実際の消費者画像コレクションに存在するからである。第２に、実用的なシステムは実際の消費者コレクション内の様々な意味論的コンテンツを収容し、それによって、各々の特定の意味論的イベントを検出する特定の個別の方法の代わりに、異なる意味論的イベントを検出する包括的な方法を組み込んだシステムを提供することを望ましくする必要がある。最後に、実用的なシステムは、識別及び分類における誤りを防ぐために、ロバストである必要がある。

【課題を解決するための手段】

【０００７】
本発明は、デジタル画像コンテンツ記録における意味論的イベント検出のためのシステム及び方法を提供する。特に、イベントレベルの「Ｂａｇ－ｏｆ－Ｆｅａｔｕｒｅｓ（特徴のバッグ）」（ＢＯＦ）表現がイベントをモデル化するために使用され、包括的な意味論的イベントが、ＢＯＦ表現に基づいて、元の低レベルの視覚的特徴空間の代わりに概念空間で検出される。

【０００８】
好適な実施形態では、イベントレベル表現が開発され、そこでは各イベントがＢＯＦ特徴ベクトルによってモデル化され、ＢＯＦ特徴ベクトルに基づいて意味論的イベント検出器が直接的に構築される。分類器の訓練のために画像レベル特徴ベクトルが使用される単純なアプローチに比べて、本発明は、イベント内の難しい画像又は誤って組織化された画像に対して、よりロバストである。例えば、任意の所与のイベントにおいて、いくつかの画像は分類が難しいことがある。これらの困難な画像は、通常、決定境界を複雑にし、モデル化を困難にする。イベントレベル特徴表現を適用することによって、イベントレベルの測定における困難な又は誤ったデジタル静止画像及びビデオセグメントの影響を減らすことによって、感度の問題を避けることができる。後述のように、良好な検出性能が、サポートベクトルマシン（Ｓｕｐｐｏｒｔ　Ｖｅｃｔｏｒ　Ｍａｃｈｉｎｅ，ＳＶＭ）分類器に対する少数のサポートベクトルで達成され得る。すなわち、分類の問題が、イベントレベル表現によって顕著に単純化され得る。

【０００９】
好適な実施形態では、あるイベントに関連した画像記録のグループの意味論的イベント分類を容易にする方法が提供され、その方法は、画像記録の各々から複数の視覚的特徴を抽出するステップと、それらの視覚的特徴を使用して画像記録の各々に対して複数の概念スコアを生成するステップであって、各概念スコアが視覚的概念に対応し、且つ各概念スコアは画像記録がその視覚的概念を含む確率を示す、ステップと、画像記録の概念スコアに基づいてそのイベントに対応する特徴ベクトルを生成するステップと、イベント分類器に特徴ベクトルを供給するステップであって、当該イベント分類器は、当該イベントに対応する少なくとも一つの意味論的イベント分類器を特定する、ステップと、を含む。

【００１０】
画像記録は、少なくとも一つのデジタル静止画像及び少なくとも一つのビデオセグメントを含み得る。したがって、このシステムは、通常はデジタル静止画像及びビデオセグメントの両方を含む実生活の消費者画像データセットを取り扱うことができる。

【００１１】
複数の視覚的特徴の抽出は、ビデオセグメントからのキーフレームの抽出、ならびにキーフレーム及びデジタル静止画像の両方からの複数の視覚的特徴の抽出を含む。それから初期概念スコアが、抽出された視覚的特徴の各々に対応する各キーフレーム及び各デジタル静止画像に対して生成される。それから好ましくは、アンサンブル概念スコアが、初期

概念スコアに基づいて各キーフレーム及び各デジタル静止画像に対して生成される。

【００１２】

アンサンブル概念スコアは好ましくは、所与のキーフレーム又は所与のデジタル静止画像に対する各々の抽出された視覚的特徴に対する個別の概念スコアを融合することによって生成される。

【００１３】

意味論的イベント分類器がひとたび特定されると、デジタル静止画像及びビデオセグメントは、画像及びビデオセグメントの適切な分類、記憶、及び検索を容易にするためにタグ付けされることができる。

【図面の簡単な説明】

【００１４】

【図１】本発明に従った意味論的イベント検出システムの模式的ブロック図である。

【図２】図１に描かれた意味論的イベント検出システムによって利用される処理モジュールを描いた流れ図である。

【図３】意味論的イベント検出のために図１に描かれたシステムを訓練するために利用される処理モジュールを描いた流れ図である。

【図４】図１に描かれたシステムで使用される概念検出器を訓練するために利用される処理モジュールを描いた流れ図である。

【図５】テストプロセスで使用される異なる意味論的イベントを、それらの詳細な定義を含めて描いた表である。

【図６】図１に描かれたシステムの結果と、ＢＯＦ特徴ベクトルが元の低レベルの視覚的特徴に基づいて構築される従来のアプローチとの比較を描いたグラフである。

【図７】本発明の結果を、ベースラインイベント検出器の結果、及び画像レベル概念スコア表現を直接的に使用するＳＶＭ検出器（ＳＶＭ－Ｄｉｒｅｃｔ）の結果と比較するグラフである。

【図８】ＳＶＭ－Ｄｉｒｅｃｔ方法と比較して本発明によって要求されるサポートベクトルの数を比較するグラフである。

【発明を実施するための形態】

【００１５】

本発明は、ある好適な実施形態及び添付の図面を参照して記述される。

【００１６】

複雑な意味論的イベントは通常は初歩的な視覚的概念の同時発生によって生成される。例えば、「結婚」は、「人々」「花」「公園」などのような視覚的概念に関連した意味論的イベントであり、あるパターンで進展する。一般的に、視覚的概念は、画像の画像コンテンツ特性として定義されることができ、通常は、特定のイベントを識別するために使用されるワード（ｗｏｒｄ）よりも広いワードによって意味論的に表現される。したがって、視覚的概念は、特定のイベントに与えられることができる画像コンテンツ特性のサブセットを形成する。

【００１７】

本発明では、初歩的な視覚的概念が最初に画像から検出され、意味論的イベント検出器が、元の低レベルの特徴空間の代わりに概念空間に構築される。そのようなアプローチからの恩恵は、少なくとも２つの局面を含む。第１に、視覚的概念は、元の低レベルの特徴よりも高レベルであり、且つより直観的である。S.EbadollahiらのIEEE ICME（2006）「Visual event detection using multi dimensional concept dynamics（多次元概念ダイナミクスを使用する視覚的イベント検出）」に記述されているように、概念スコアは、意味論的イベントをモデル化するために強力である。第２に、本発明における概念空間は、好ましくは、例えばS.F.ChangらのACM MIR（2007）「Multimodal semantic concept detection for consumer video benchmark（消費者ビデオベンチマークのための多モード意味論的概念検出）」に記述されているように、意味論的概念検出器によって形成され、例えばA.C.LouiらのACM MIR（2007）「Kodak consumer video benchmark data set: concept de

finition and annotation（コダック消費者ビデオベンチマークデータセット：概念の定義と注釈付け）」に記述されたタイプの既知の消費者電子画像データセットで訓練される。これらの意味論的概念検出器は、以前の画像データセットから付加的な情報を組み込む重要な役割を果たし、現在の画像データセットにおける意味論的イベントを検出する手助けをする。

【００１８】

例えば、上述のデータセットが実際の消費者からのデジタル静止画像及びビデオセグメントを含むと仮定すると、データセット全体が最初にマクロイベントのセットに区分され、各マクロイベントがさらにイベントのセットに区分されることが望ましい。この区分は、好ましくは、上述の以前に開発されたイベントクラスタ化アルゴリズムを使用することによって、ビデオセグメントの各デジタル静止画像の撮影時間及びそれらの間の色の類似性に基づく。例えば、E_tがｔ番目のイベントを指し、$m_p{}^t$個の写真及び$m_v{}^t$個のビデオを含むとする。$I_i{}^t$及び$V_j{}^t$は、E_tにおけるｉ番目の写真及びｊ番目のビデオを指す。画像はこのアルゴリズムを使用してイベントにグループ化又はクラスタ化されることができるが、イベント自身は、意味論的な意味で識別されたり関連付けられたりはしない。したがって、本発明のゴールは、特定の意味論的意味、すなわち、特定のイベントE_t及びそのイベントに対応する画像記録に対して、「結婚」及び「誕生日」のような意味論的イベントS_Eのタグ付けをすることである。

【００１９】

「人々」「公園」及び「花」のような、同時に発生する視覚的概念によって意味論的イベントが生成されることが仮定されるであろう。ここでC_1、…、C_NはＮ個の視覚的概念を示す。上述の視覚的概念検出器を使用して、２１個（Ｎ＝２１）のＳＶＭベースの概念検出器が、適用されたデータセットに対して、好ましくは低レベルの色、テクスチャ、及びエッジの視覚的特徴を使用して形成される。これらの意味論的概念検出器は、各画像$I_i{}^t$に対する２１個の個別の概念スコア$p(C_1, I_i{}^t)$，…，$p(C_N, I_i{}^t)$を生成するために適用されることができる。これらの概念スコアはそれから、以下により詳細に記述されるように、概念空間における画像$I_i{}^t$を

【数１】

$$\mathbf{f}(I_i^t) = [\, p(C_1, I_i^t), \ldots, p(C_N, I_i^t)]^T$$

と表現するための特徴ベクトルを形成するために使用される。

【００２０】

実際の消費者からのビデオセグメントは、通常は一つの長い撮影映像（ロングショット）からの様々な視覚的コンテンツを有するので、各ビデオ$V_j{}^t$は、好ましくはセグメント$V^t_{j,1}$，…，$V^t_{j,mj}$のセットに区分される。各セグメントは所与の長さ（例えば５秒）を有する。それからキーフレームが、ビデオセグメントから一様に周期的にサンプリングされる（例えば０．５秒ごとに）。例えば、$I^t_{j,k,l}$がｋ番目のセグメント$V^t_{j,k}$のｌ番目のキーフレームであるとすると、そのときには、$I^t_{j,k,l}$はまた、デジタル静止画像と同じように概念空間内の特徴ベクトル

【数２】

$$\mathbf{f}(I^t_{j,k,l})$$

によっても表現されることができる。上述されたものとは異なるサンプリングレートが容易に使用され得ることが理解されるであろう。

【００２１】

デジタル静止画像及びビデオセグメントの両方が、ｘによって表されるデータポイントとして定義されることができる。例えば、イベントE_tは合計で、

【数３】

$$m^t = m_p^t + \tilde{m}_v^t$$

のデータ点を含み、

【数４】

$$\tilde{m}_v^t$$

はE_tにおける$m_v{}^t$個のビデオクリップからのビデオセグメントの全数である。意味論的イベント検出器がそれから、これらのデータ点、及び概念スコアから開発された対応する特徴ベクトルに基づいて、実行される。

【００２２】

ＢＯＦ表現は、画像に対する包括的概念を検出するために有効であることが証明されてきている。例えば、J.Sivic及びA.Zisserman, 「Video google: a text retrieval approach to object matching in videos（ビデオグーグル：ビデオにおけるオブジェクトマッチングに対するテキスト検索アプローチ）」, ICCV, pp.1470-1477（2003）を参照のこと。ＢＯＦでは、画像は、順序なしのローカル記述子のセットによって表現される。クラスタ化技術を通して、中レベルの視覚的語彙が構築され、そこでは各々の視覚的ワードがローカル記述子のグループによって形成される。各々の視覚的ワードは、画像を記述するためのロバストで且つノイズが除去された視覚的用語であるとみなされる。

【００２３】

例えば、S_Eが意味論的イベント、例えば「結婚」を指し、E_1, …, E_Mがこの意味論的イベントを含むＭ個のイベントを指すとする。各E_tは、$m^t{}_p$個の写真及び

【数５】

$$\tilde{m}_v^t$$

個のビデオセグメントによって形成されている。視覚的語彙と同様に、概念語彙が、これら

【数６】

$$\sum_{t=1}^{M} m^t$$

のデータ点（ここで、

【数７】

$$m^t = m_p^t + \tilde{m}_v^t$$

）をｎ個の概念ワードにクラスタ化することによって構築されることができる。各概念ワードは、S_Eを含む全てのイベントを記述するための共通の特性である概念同時発生のパターンとして取り扱われることができる。特に、静止ビデオ画像及びビデオデータ点の両方を収納するために、スペクトルクラスタ化アルゴリズム（例えばA.Y.Ng, M.Jordan及びY.Weiss, 「On spectral clustering: analysis and an algorithm（スペクトルクラスタ化について：分析及びアルゴリズム）」, Advances in NIPS（2001）を参照のこと）が適用されて、アースムーバーの距離（Ｅａｒｔｈ　Ｍｏｖｅｒ’ｓ　Ｄｉｓｔａｎｃｅ，ＥＭＤ）によって測定されたペアワイズ類似性（ｐａｉｒｗｉｓｅ　ｓｉｍｉｌａｒｉｔｙ）に基づいて概念語彙を構築する。ＥＭＤは、Y.Rubner, C.Tomasi及びL.Guibas, 「The earth mover's distance as a metric for image retrieval（画像検索のための指標としてのアースムーバーの距離）」, IJCV（2000）に記述されている。

【００２４】

各データ点は画像のセットとして取り扱われる。すなわち、静止ビデオ画像に対して一つの画像、及びビデオセグメントに対して複数の画像である。それから、２つのデータ点

（画像セット）の間の類似性を測定するためにＥＭＤが使用される。２つの画像セットの間の距離を計算するためには多くの方法があり、例えば、これら２つのセットにおける画像の間の最大／最小／平均距離がある。これらの方法は雑音の多い異常画像（ｏｕｔｌｉｅｒ　ｉｍａｇｅｓ）によって容易に影響されるが、ＥＭＤは、よりロバストな距離指標を提供する。ＥＭＤは、重み正規化された制約の対象となる２つの画像セットの間の対距離（ｐａｉｒｗｉｓｅ　ｄｉｓｔａｎｃｅ）の全てにおいて最小の重み付け距離を見出し、データ点の間の部分的なマッチングを許容し、異常画像の影響を低減することができる。

【００２５】

２つのデータ点の間のＥＭＤは、以下のようにして計算される。データ点x_1及びx_2にそれぞれn_1個及びn_2個の画像があるとする。x_1及びx_2の間のＥＭＤは、任意の２つの画像$I^1_p \in x_1$及び$I^2_q \in x_2$の間のフロー$f(I_p^1, I_q^2)$によって重み付けされたグラウンド距離$d(I_p^1, I_q^2)$の線形組み合わせである。

【数８】

$$D(x_1,x_2)=\frac{\sum_{p=1}^{n_1}\sum_{q=1}^{n_2}d(I_p^1,I_q^2)f(I_p^1,I_q^2)}{\sum_{p=1}^{n_1}\sum_{q=1}^{n_2}f(I_p^1,I_q^2)} \qquad (1)$$

ここで、最適フローマトリクス$f(I_p^1, I_q^2)$は、以下の線形プログラムから得られる。

【数９】

$$\min\sum_{p=1}^{n_1}\sum_{q=1}^{n_2}d(I_p^1,I_q^2)f(I_p^1,I_q^2)$$
$$\text{w.r.t } f(I_p^1,I_q^2),1\le p\le n_1,1\le q\le n_2$$
$$\text{s.t. } f(I_p^1,I_q^2)\ge 0,\sum_{q=1}^{n_2}f(I_p^1,I_q^2)\le w_p^1,\sum_{p=1}^{n_1}f(I_p^1,I_q^2)\le w_q^2$$
$$\sum_{p=1}^{n_1}\sum_{q=1}^{n_2}f(I_p^1,I_q^2)=\min\left\{\sum_{p=1}^{n_1}w_p^1,\sum_{q=1}^{n_2}w_q^2\right\}$$

ここで、w^1_p及びw^2_qはそれぞれデータ点x_1及びx_2における画像I_p^1及びI_q^2の重みである。ここで、等しい重み$w^1_p = 1/n_1$及び$w^2_q = 1/n_2$を取る。概念スコア特徴に対するユークリッド距離が、距離$d(I_p^1, I_p^2)$として使用される。式（１）より、ＥＭＤは、２つのデータ点における最もマッチする画像の対を見出す。重み正規化制約は、各画像が他のセットで十分なマッチングを有することを確実にする。x_1及びx_2の両方が写真であるときには、ＥＭＤは単にユークリッド距離である。この対のＥＭＤはそれから、ガウス関数$S(x_1, x_2) = \exp(-D(x_1, x_2)/r)$によって、ペアワイズ類似性に変換される。ここで、ｒは全訓練データ点の間の全ての対の距離の平均である。

【００２６】

上記で言及したスペクトルクラスタ化は、データ点の対の間の類似性からなるデータセットにおけるグループを見出すための技術である。ここで、エング（Ng）らにより開発されたアルゴリズムが適用され、以下のように記述されることができる。類似性マトリクス$S(x_i, x_j)$が与えられると、

・アフィンマトリクス$A_{ij} = S(x_i, x_j)$　（ｉ≠ｊの場合），かつ$A_{ii} = 0$を得る。

・対角線マトリクス$D_{ij} = \Sigma_j A_{ij}$を定義する。$L = D^{-1/2} A D^{-1/2}$を得る。

・最大のものから順にｎ個の固有値に対応するＬの固有ベクトル

【数１０】

$$u_1,\ldots,u_n$$

を見出し、

【数１１】

$$U = [u_1,\ldots,u_n]$$

を得る。ここで、ｎは、保持すべき固有値のエネルギー比によって決定される。
・Ｕの行が単位長さを有するように再正規化することによって、ＵからマトリクスＶを得る。
・Ｖにおける各行をR^n（元のｉ番目のデータ点に対応するｉ番目の行）における点として取扱い、Ｋ－ｍｅａｎｓアルゴリズムを介して全ての点をｎ個のクラスタにクラスタ化する。

【００２７】

スペクトルクラスタ化アルゴリズムによって得られる各データクラスタは概念ワードと呼ばれ、全てのクラスタは、意味論的イベントを表し且つ検出するための概念語彙を形成する。W_j^iが意味論的イベントS_{Ei}に対して学習されたｊ番目のワードを表し、$S(x, W_j^i)$が、ｘとW_j^iにおけるメンバ・データ点（ｍｅｍｂｅｒ　ｄａｔａ　ｐｏｉｎｔｓ）との間の最大類似性として計算されたワードW_j^iに対するデータｘの類似性を指すとする。

【数１２】

$$S(x, W_j^i) = \max_{x_k \in W_j^i} S(x_k, x)$$

であり、ここで、$S(x_k, x)$は上記と同じように定義される。各データｘに対して、ベクトル$[S(x, W_1^i), \cdots, S(x, W_n^i)]^T$が、ｘに対するＢＯＦ特徴ベクトルとして取り扱われることができる。イベントE_tがm^t個のデータ点を含むとし、上記のＢＯＦ特徴ベクトルに基づいて、イベントE_tはまた、ＢＯＦ特徴ベクトル

【数１３】

$$f_{bof}(E_t)$$

によって、

【数１４】

$$f_{bof}(E_t)=[\max_{x \in E_t} S(x, W_1^i),\ldots,\max_{x \in E_t} S(x, W_n^i)]^T$$

として表されることもできる。最後に、ＢＯＦ特徴

【数１５】

$$f_{bof}$$

を使用すると、二値化された一対全ＳＶＭ分類器が、意味論的イベントS_{Ei}を検出するために学習されることができる。

【００２８】

ここで図１を参照すると、本発明のある実施形態に従ったデジタルコンテンツ記録のための意味論的イベント検出のためのシステム１００が描かれている。このシステム１００は、データ処理ユニット１１０、周辺ユニット１２０、ユーザインターフェースユニット１３０、及びメモリユニット１４０を含む。メモリユニット１４０、周辺ユニット１２０、及びユーザインターフェースユニット１３０は、データ処理システム１１０に通信的に接続されている。

【００２９】

データ処理システム１１０は一以上のデータ処理装置を含み、このデータ処理装置が、ここで記述される図２～４の例示的なプロセスを含む本発明の様々な実施形態のプロセスを実現する。「データ処理装置」又は「データプロセッサ」という表現は、中央処理装置

（ＣＰＵ）、デスクトップコンピュータ、ラップトップコンピュータ、メインフレームコンピュータ、個人デジタル端末、ブラックベリー™、デジタルカメラ、携帯電話、あるいは、電気的、磁気的、光学的、生物学的構成要素とともに実現されるか又はその他の方法で実現された任意の他のデータ処理、データ管理、又はデータ取扱いのための装置のような任意のタイプのデータ処理装置を含むことが意図されている。

【００３０】

メモリユニット１４０は情報を記憶するように構成された一以上のメモリ装置を含み、記憶される情報は、ここで記述される図２～４の例示的なプロセスを含む本発明の様々な実施形態のプロセスを実行するために必要とされる情報を含む。メモリユニット１４０は分散プロセッサアクセス可能メモリシステムであってもよく、これは、複数のコンピュータ及び／又は装置を介してデータ処理システム１１０に通信的に接続された複数プロセッサアクセス可能メモリを含む。一方、メモリユニット１４０は、分散プロセッサアクセス可能メモリシステムである必要は無く、したがって単一のデータプロセッサ又は装置内に配置された一以上のプロセッサがアクセス可能なメモリを含み得る。さらに、「メモリユニット」という表現は、揮発性又は不揮発性、電子的、磁気的、光学的、又はその他のものであってもよい任意のプロセッサアクセス可能データ記憶装置を含むことが意図されており、これは、フロッピー（登録商標）ディスク、ハードディスク、コンパクトディスク、ＤＶＤ、フラッシュメモリ、ＲＯＭ、及びＲＡＭ、又は任意の他のデジタル記憶媒体を含むが、これらに限定されるものではない。

【００３１】

「通信的に接続された」という表現は、有線であっても無線であっても、データが通信され得る装置、データプロセッサ、又はプログラムの間の任意のタイプの接続を含むことが意図されている。さらに、「通信的に接続された」という表現は、単一のデータプロセッサ内の装置又はプログラム間の接続、異なるデータプロセッサに配置された装置又はプログラム間の接続、及びデータプロセッサには全く配置されていない装置間の接続を含むことが意図されている。この点に関して、メモリユニット１４０はデータ処理システム１１０から離れて示されているが、当業者は、メモリユニット１４０がデータ処理システム１１０内に完全に又は部分的に実現され得ることを理解するであろう。さらに、この点に関して、周辺システム１２０及びユーザインターフェースシステム１３０がデータ処理システム１１０から離れて示されているが、当業者は、それらのシステムの一方又は両方がデータ処理システム１１０内に完全に又は部分的に実現され得ることを理解するであろう。

【００３２】

周辺システム１２０は、データ処理システム１１０にデジタルコンテンツ記録を提供するように構成された一以上の装置を含み得る。例えば、周辺システム１２０はデジタルビデオカメラ、携帯電話、通常のデジタルカメラ、又はその他のデータプロセッサを含み得る。加えて、周辺システム１２０は、データ処理システム１１０を離れたデータ源に接続するために必要な機器、装置、回路などを含み得る。例えば、システム１００は、インターネットを介して、データセットが記憶されるサーバーにリンクされ得る。データセットは、システム１００を訓練するために使用されるデジタルコンテンツ記録のデータセット、あるいは、システム１００によって分析されるべきデジタルコンテンツ記録を含むデータセットを含み得る。データ処理システム１１０は、周辺システム１２０内の装置からデジタルコンテンツ記録を受領すると、そのようなデジタルコンテンツ記録を、さらなる処理のためにプロセッサアクセス可能メモリシステム１４０に記憶し得て、あるいは、十分な処理パワーが利用可能であれば、受領したデータストリームとしてリアルタイムでデジタルコンテンツ記録を分析し得る。

【００３３】

ユーザインターフェースシステム１３０は、マウス、キーボード、他のコンピュータ、又はデータがそこからデータ処理システム１１０に入力される任意の装置又は複数の装置の組み合わせを含み得る。これに関して、周辺システム１２０がユーザインターフェース

システム１３０から離れて示されているが、当業者は、周辺システム１２０がユーザインターフェースシステム１３０の一部として含まれ得ることを理解するであろう。

【００３４】

ユーザインターフェースシステム１３０はまた、ディスプレー装置、プロセッサアクセス可能メモリ、又はデータ処理システム１１０によってデータがそこに出力される任意の装置又は複数の装置の組み合わせを含み得る。これに関して、ユーザインターフェースシステム１３０がプロセッサアクセス可能メモリを含むならば、そのようなメモリは、ユーザインターフェースシステム１３０及びメモリユニット１４０が図１では離れて示されているが、メモリユニット１４０の一部であり得る。

【００３５】

システムの基本的な動作がここで図２を参照して記述される。これは図１に描かれたユニットの一以上によって実現される処理モジュールを描く流れ図である。処理モジュールがシステム１００に設けられたユニットの一以上によって実行される指示を含むことが理解されるべきである。図示されている例では新しいイベント（E_0）がデータエントリモジュール２００を介してシステム１００に与えられる。ここではE_0が特定の意味論的イベントに属する確率が決定されることが望ましい。例えば、ユーザインターフェースユニット１３０を介して受領されたオペレータ指示に基づいて、データ処理ユニット１１０は、E_0に対応するデータをメモリユニット１４０にダウンロードするように周辺ユニット１２０の動作を制御する。各イベントは、複数のデジタルコンテンツ記録を含み、図示されている例ではデジタル静止画像$m_{0,p}$及びビデオコンテンツ$m_{0,v}$を含む。これらのデジタルコンテンツ記録は、撮影時刻及び色の類似性にしたがって、先に記述されたクラスタ化方法を利用して一緒にグループ化される。クラスタ化方法は、システム１００への提出に先立って、静止デジタル画像及びビデオセグメントのデータセットに適用されることができる。あるいは、データセットがシステム１００に与えられて、データエントリモジュール２００が、E_0を生成するためにデータ処理ユニット１１０の一つの動作要素としてクラスタ化動作を実行してもよい。

【００３６】

例えば、消費者は電子カメラを使用して、複数の異なるイベントの１００個のデジタル静止画像及びビデオからなるデータセットを撮影し得る。電子カメラからのメモリカードが、周辺ユニット１２０の一部としてのカードリーダーユニットに提供される。ユーザインターフェースユニット１３０を介してユーザによって入力された制御指示に反応して、データ処理ユニット１１０は、データセットをメモリカードからメモリユニット１４０にダウンロードするように、周辺ユニット１２０の動作を制御する。データ処理ユニット１１０はそれから先に進んで、デジタル静止画像及びビデオを複数のイベントに対応する複数のクラスタにグループ化するために、データセットに対してクラスタ化アルゴリズムを実行する。これにより、データエントリモジュール２００内に提供された指示の機能が完了し、ある数のデジタル静止画像及びビデオ（例えば元の１００個のうちの１０個）が、E_0に関連しているとして識別される。この時点で、１０個のデジタル静止画像及びビデオがE_0に関連付けられるが、E_0は、「結婚」のような特定の意味論的イベントにはまだ関連付けられていない。

【００３７】

視覚的特徴抽出モジュール２１０がそれから使用されて、E_0内のビデオセグメントからキーフレームを獲得し、キーフレーム及びE_0内に含まれるデジタル静止画像の両方から視覚的特徴が抽出される。図示されている例では、視覚的特徴抽出モジュール２１０は、格子ベースの色モーメント、ガボール・テクスチャ（Ｇａｂｏｒ　ｔｅｘｔｕｒｅ）、及びエッジ方向性ヒストグラムを、デジタル静止画像及びビデオの各々に対して決定する。しかし、図示されている例で使用されているもの以外の視覚的特徴が容易に利用され得ることが、理解されるであろう。

【００３８】

データ処理ユニット１１０は、視覚的特徴抽出モジュール２１０内に提供された指示に

したがって、E_0とともに含まれるデジタル静止画像及びビデオの各々に対して、従来の技術を利用して必要なキーフレーム及び視覚的特徴の抽出を実行する。したがって、E_0に対応する１０個のデジタル静止画像及びビデオの各々についての３つの視覚的特徴表現が、さらなる分析のためにここでは利用可能である。

【００３９】

特徴抽出モジュール２１０によって抽出された３つの視覚的特徴は、概念検出モジュール２２０によって使用されて、特定のキーフレーム又は静止デジタル画像が特定の意味論的イベントに関係している確率を反映した概念スコアを生成する。概念検出モジュール２２０は、好ましくは、２ステップのプロセスを使用して概念スコアを決定する。第１に、概念スコア検出モジュール２２２が設けられ、これは、２１個の上記のＳＶＭ意味論的概念決定子（データ処理ユニット１１０によって実現される）を利用して、各デジタル静止画像及びキーフレームに対する各視覚的特徴空間における各々の個別の分類器に基づいて、概念スコアを生成する。第２に、個々の概念スコアがそれから融合モジュール２２４（データ処理ユニット１１０によって実現される）によって融合され、特定のデジタル静止画像及びキーフレームに対するアンサンブル概念検出スコアを生成し、それによって、さらに処理されるべきデータ量を低減する。

【００４０】

好適な実施形態では、融合モジュール２２４は最初に、異なる特徴からの異なる分類出力を、シグモイド関数１／（１＋ｅｘｐ（－Ｄ））によって正規化する。ここで、Ｄは、決定境界までの距離を表すＳＶＭ分類器の出力である。融合は、２１個の概念の各々に対する異なる視覚的特徴からの分類出力の平均を取ることによって完了されて、アンサンブル概念検出スコアを生成する。

【００４１】

簡単化された例では、３つの概念「人々」「公園」及び「花」が議論される。「人々」「公園」及び「花」に対する概念スコアが、E_0の１０個の画像の各々の３つの視覚的特徴表現の各々について生成される。例えば、１０個の画像のグループの最初の画像の色の特徴表現は、人々を含む確率が９０％、公園を含む確率が５％、及び花を含む確率が５％であり得て、最初の画像のテクスチャの特徴表現は、人々を含む確率が５％、公園を含む確率が８０％、及び花を含む確率が１５％であり得て、最初の画像のエッジ方向の特徴表現は、人々を含む確率が１０％、公園を含む確率が５０％、及び花を含む確率が４０％であり得る。

【００４２】

１０個の画像の３つの視覚的特徴表現が与えられると、３０セットの概念スコアが生成され（各視覚的特徴表現に対して一つ）、各セットは３つの個別の概念スコア（「人々」に対して一つ、「公園」に対して一つ、及び「花」に対して一つ）を含む。最初の画像に対するアンサンブル概念スコアを生成するために、視覚的表現の各々に対する概念の各々についての確率が平均され、第１の画像のアンサンブル概念スコアは、人々を含む確率が３５％（人々の確率として色９０％、テクスチャ５％、エッジ５％の平均）、公園を含む確率が３０％（公園の確率として色５％、テクスチャ８０％、エッジ５％の平均）、及び花を含む確率が２０％（花の確率として色５％、テクスチャ１５％、エッジ４０％の平均）となる。

【００４３】

アンサンブル概念スコアは引き続いてＢＯＦモジュール２３０に与えられ、これがE_0に対するＢＯＦベクトルを決定する。E_0に対するＢＯＦ特徴ベクトルは、最初に、各々の各デジタル静止画像及びビデオセグメントのアンサンブル概念スコアを使用してE_0内に含まれるデジタル静止画像及びビデオセグメントの各々に対する個別の特徴ベクトルを決定することによって得られる。好適な実施形態では、各デジタル静止画像又はビデオセグメントはデータ点として扱われて、E_0内の各データ点のアンサンブル概念スコアの間のペアワイズ類似性、及び所与の意味論的イベント（ＳＥ）、例えば「結婚」に対する各々の予め定められた正の訓練データ点のアンサンブル概念スコアが、それからＥＭＤを使

用して類似性検出モジュール２３２によって計算される。効果的には、個別の特徴ベクトルは、E_0内に含まれるデジタル静止画像及びビデオセグメントの各々に対して得られる。マッピングモジュール２３４がそれから、E_0の個別の特徴ベクトルの各々を意味論的イベントのコードブック（以下により詳細に記述される訓練プロセスの間に先に開発されている）にマッピングするために使用され、E_0に対するイベント特徴ベクトルが、マッピングされた類似性に基づいて生成される。

【００４４】

イベント特徴ベクトルが、ここで分類器モジュール２４０に供給されることができる。描かれている例では、分類器モジュール２４０はＳＶＭ分類器を使用して、E_0に対するイベント検出スコアを生成する。イベント検出スコアは、新しいイベントE_0が「結婚」のような所与の意味論的イベントに対応する最終的な確率を表す。イベント検出スコアはそれから、好ましくは予め定められた閾値と比較され、E_0が結婚イベントとしてカテゴリ化されるべきかどうかが決定される。予め定められた閾値は、所与のアプリケーションにてシステム１００によって要求される正確さのレベルに依存して変化してもよい。

【００４５】

ひとたびE_0が適切にカテゴリ化されると、E_0に対応する静止デジタル画像及びビデオセグメントが、適切な意味論的イベント分類器でタグ付けされ、適切なアルバムフォルダ又はファイルに分類されて、後の検索のためにメモリユニット１４０内に記憶されることができる。あるいは、タグ付けされた静止デジタル画像及びビデオセグメントは、周辺ユニット１２０を介して画像記憶媒体に書き込まれることができる。静止デジタル画像及びビデオセグメントの意味論的イベント分類器によるタグ付けは、画像及びビデオセグメントがサーチエンジンによって容易に検索されることを可能にするという付加的な効果を提供する。

【００４６】

システム１００の訓練が、ここで図３を参照して記述される。最初に、Ｔ個の正の訓練イベントE_1，…，E_Tが、データエントリモジュール２００を使用して入力される。各イベントE_tは、$m_{t,p}$個の写真及び$m_{t,v}$個のビデオを含み、これらは、先に記述されたクラスタ化方法によって、撮影時間及び色の類似性にしたがってグループ化されている。視覚的抽出モジュール２１０がそれから使用されて、ビデオセグメントからキーフレームを抽出し、キーフレーム及びデジタル静止画像の両方から視覚的特徴が抽出される。上述の動作の場合と同様に、視覚的特徴は、格子ベースの色モーメント、ガボール・テクスチャ、及びエッジ方向性ヒストグラムを含む。概念検出モジュール２２０がそれから使用されて、上述のようにキーフレーム及びデジタル画像に対するアンサンブル概念スコアを生成する。

【００４７】

ＢＯＦ学習モジュール２５０がそれから使用されて、システム１００を訓練する。最初に、各デジタル画像又はビデオセグメントがデータ点として取り扱われて、データ点の各々の対の間のペアワイズ類似性が、先に記述された類似性検出モジュール２３２を使用してＥＭＤによって計算される。ペアワイズ類似性マトリクスに基づいて、スペクトルクラスタ化モジュール２５２が使用されてスペクトルクラスタ化を適用し、データ点を異なるクラスタにグループ化する。ここで、各クラスタは一つのコードワードに対応する。意味論的イベントＳＥを検出するために分類器を訓練するために、全ての訓練イベントE_i（E_iはＳＥに対する正の訓練イベント及び負の訓練イベントの両方を含む）が上述のコードブックにマッピングされて、マッピングモジュール２５４によって各訓練イベントに対するＢＯＦ特徴ベクトルが生成される。ＢＯＦ特徴ベクトルに基づいて、分類器訓練モジュール２６０が使用されて、特定の意味論的イベントＳＥを検出するように二値化ＳＶＭ分類器を訓練する。

【００４８】

図４は、概念スコア検出モジュール２２２で使用されるビデオ概念検出器のための訓練プロセスの詳細を描いている。概念Ｃに対して、ベンチマーク消費者ビデオデータセット

からＮ個の正の訓練ビデオがデータエントリモジュール２００を介して提供される。キーフレームがビデオから得られて、視覚的特徴が、先の例におけるように視覚的特徴抽出モジュール２１０を使用してキーフレームから得られる。視覚的特徴は、格子ベースの色モーメント、ガボール・テクスチャ、及びエッジ方向性ヒストグラムを含む。概念訓練モジュール２７０がそれから使用されて、概念決定子を訓練する。すなわち、視覚的特徴の各タイプに基づいて、各キーフレームは特徴ベクトルとして表されて、二値化ＳＶＭ分類器が概念Ｃを検出するように訓練される。特徴の個々のタイプに対するこれらの分類器の判別機能が一緒に平均されて、概念Ｃに対するアンサンブル概念検出器を生成する。

【００４９】

上述された意味論的決定システム及び方法のテストが、コダックの消費者データセットから１９７２個の消費者イベントを評価することによって実行された。イベントは１０個の異なる意味論的イベントにラベルされ、その詳細な定義が、図５に与えられた表に示されている。合計１２６１個のイベントが訓練のためにランダムに選択され、残りはテストのために使用された。訓練及びテストデータはマクロイベントレベルで区分された。すなわち、同じマクロイベントからのイベントが、訓練又はテストデータとして一緒に扱われた。これは、同じマクロイベントからの類似のイベントが分離されることを避けるためであり、分類問題を単純化する。

【００５０】

平均精度（ａｖｅｒａｇｅ　ｐｒｅｃｉｓｉｏｎ，ＡＰ）が性能の指標として使用されたが、これは、ビデオ概念検出のための公式指標として使用されている。例えば、ニスト（Ｎｉｓｔ）の「ツリービデオ検索評価（Tree video retrieval evaluation (treevid)）」，２００１－２００６，http://www--nlpir.nist.gov/projects/treevidを参照のこと。これは、精度－再生曲線における異なる再生点での精度値の平均を計算し、これにより特定の意味論的イベントを検出する際の分類器の有効性を評価する。複数の意味論的イベントを考慮するときには、ＡＰの平均（ｍｅａｎ　ｏｆ　ＡＰｓ，ＭＡＰ）が使用される。

【００５１】

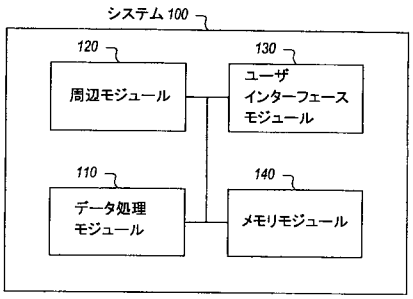
意味論的イベント検出アルゴリズムにおける概念スコア表現の有効性を示すために、本発明の方法とＢＯＦ特徴ベクトルが元の低レベルの視覚的特徴に基づいて構築されるアプローチとを比較する実験が行われた。具体的には、上述のS.F.Changらに記述されたものと同じ低レベルの視覚的特徴が使用された。図６は、性能の比較を与える。図６に示されるように、両方の方法は、一致してランダム推測よりも性能がよい。しかし、概念スコアを伴うＳＥ検出は、大抵の概念について、ＡＰに関して低レベル特徴を伴うＳＥ検出よりも性能がよく、ＭＡＰに関しては２０．７％よい。この結果は、意味論的イベントの検出を手助けするために先の概念検出モデルを使用するパワーを確かめるものである。

【００５２】

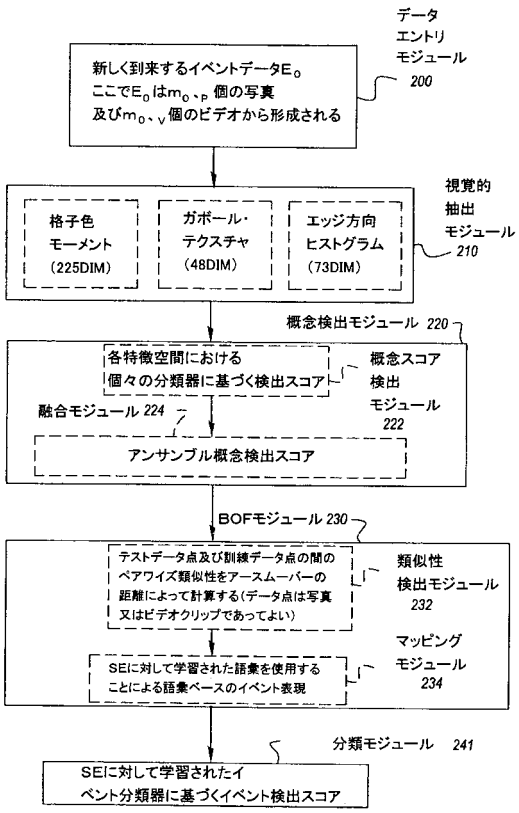
第２の実験が行われ、イベントレベル表現対画像レベル表現の比較を行った。この実験では、本発明の意味論的検出方法（ＳＥ検出）と２つの他の検出器、すなわち（１）ベースラインイベント検出器（ベースライン）及び（２）画像レベル概念スコア表現を直接的に使用するＳＶＭ検出器（ＳＶＭダイレクト）との間の比較が行われた。図７は、異なる方法のＡＰ比較を与える。示されるように、提案されるＳＥ検出は意味論的イベントの大抵に対して最も良く機能し、「結婚」「クリスマス」及び「学校活動」のような多くの意味論的イベントに対して、２番目に良い方法に比べて２０％より多くの顕著な性能の改善を得ている。この結果は、イベントレベルＢＯＦ表現の成功を確かめるものである。加えて図８は、異なるアルゴリズムからのサポートベクトルの数の比較を与える。一般的に、サポートベクトルが少ないほど、決定境界が単純になる。図より、決定境界は、イベントレベル表現によって顕著に単純化され、ＳＶＭ分類器は意味論的イベントを非常によく分離することができている。さらに、ベースライン検出器及びＳＥ検出による「動物」に対するトップ５個の検出イベントの比較は、ＳＥ検出方法が１００％の精度を達成することができるのに対して、画像ベースのＳＶＭダイレクト方法は２０％の精度しか得ることが

できないことを示した。

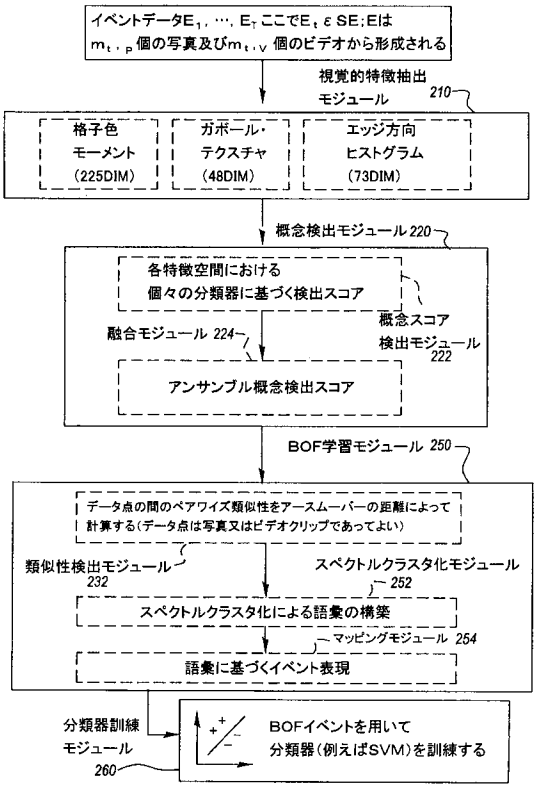
【図１】



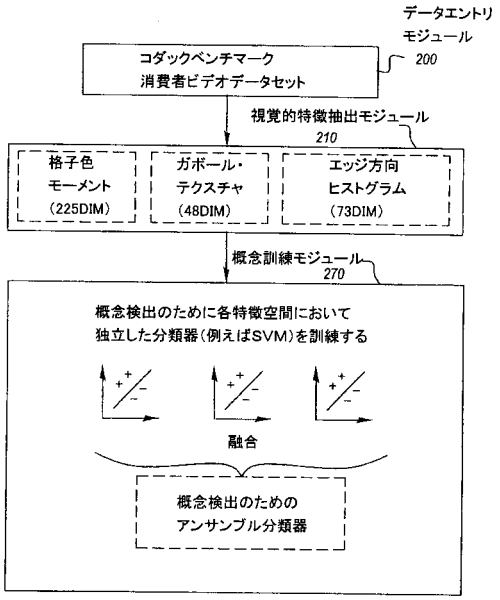
【図２】



【図３】



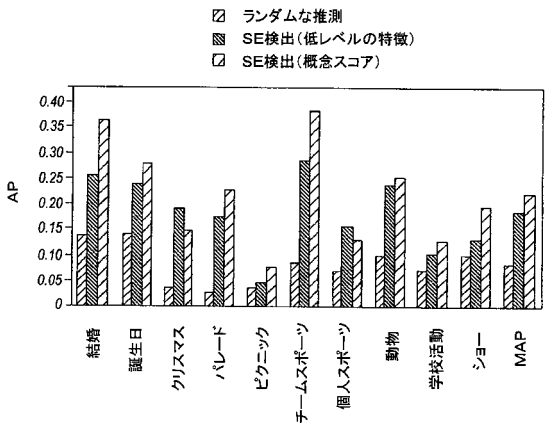
【図４】



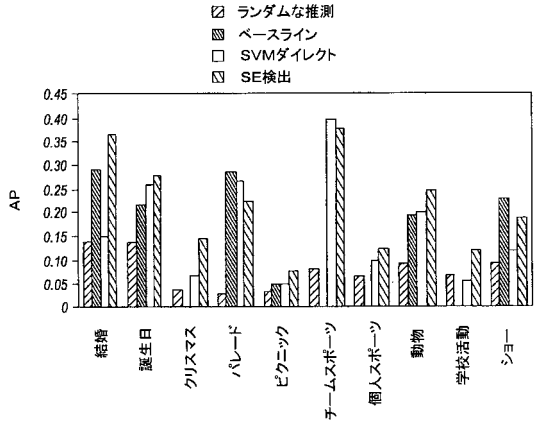
【図５】

意味論的イベント	定義
結婚	花嫁及び花婿、ケーキ、装飾された車、レセプション、結婚パーティー、あるいは結婚の日に関連した何か
誕生日	誕生日ケーキ、風船、包装されたプレゼント、及び誕生日の帽子 通常は有名な誕生日の歌を伴う
クリスマス	クリスマスツリー及び通常のクリスマス装飾、 必ずしもクリスマスの日に撮影されない
パレード	公共の場を動いていく人々又は車両の進行
ピクニック	屋外、ピクニックテーブルがある又はない、日よけがある又はない、 視野の中の人々及び食べ物
チームスポーツ	バスケットボール、野球、フットボール、ホッケー、及び他のチームスポーツ
個人スポーツ	テニス、水泳、ボーリング、及び他の個人スポーツ
動物	ペット(例えば、犬、猫、馬、魚、鳥、ハムスター)、野生の動物 動物園、及び動物ショー
学校活動	学校の卒業、学校の日々(学校の初日又は最終日)、及び学校に関する他のイベント
ショー	ショー及びコンサート、リサイタル、演劇、及び他のイベント

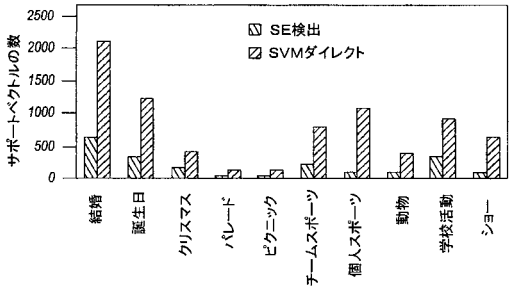
【図６】



【図７】


ランダムな推測
ベースライン
SVMダイレクト
SE検出
AP
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0
結婚
誕生日
クリスマス
パレード
ピクニック
チームスポーツ
個人スポーツ
動物
学校活動
ショー

【図８】


SE検出
SVMダイレクト
サポートベクトルの数
2500
2000
1500
1000
500
0
結婚
誕生日
クリスマス
パレード
ピクニック
チームスポーツ
個人スポーツ
動物
学校活動
ショー

フロントページの続き

(72)発明者　ルイ　アレクサンダー　シー
アメリカ合衆国　ニューヨーク　ロチェスター　ステート　ストリート　３４３
(72)発明者　ジアン　ウェイ
アメリカ合衆国　ニューヨーク　ニューヨーク　ウェスト　１２０　ティーエイチ　ストリート
マッド　５００　エス　ダブリュー　１３００

審査官　真木　健彦

(56)参考文献　特開２００７－３１７０７７（ＪＰ，Ａ）
特開２００３－１５３００７（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｔ　　１／００
Ｇ０６Ｔ　　７／００
Ｇ０６Ｆ　１７／３０