様式 C-19

科学研究費補助金研究成果報告書

平成２２年　５月２８日現在

研究種目：基盤研究（B）
研究期間：2007～2009
課題番号：19300028
研究課題名（和文）　ビデオオントロジーの導入による動画像検索エンジンの開発

研究課題名（英文）　Video retrieval by using video ontology

研究代表者
上原　邦昭（Kuniaki Uehara）
神戸大学・工学研究科・教授
研究者番号：60160206

研究成果の概要（和文）：テレビ放送のデジタル化、高速ネットワークを介した映像提供サービス、あるいは映像検索サービスなどの普及により、一般ユーザが大量の映像情報を気軽に楽しむ環境が構築されてきている。このような大量の映像を効率よく扱うために、高度な映像検索技術の開発が必要となってきている。本研究課題では、映像から意味的な高次特徴を抽出し、その結果を用いてユーザから与えられた問い合わせに合致する映像ショットを検索する、「クエリーベースの手法」と呼ぶ映像検索技術を開発した。本技術は、統計的手法や機械学習などを用いて、大規模な実データの持つ多様性、曖昧性、多義性などに柔軟に対応する手法である。特に多義性や多様性については「ラフ集合理論」、曖昧性については「部分空間クラスタリング」、大規模性については「部分教師つき学習」を拡張した検索技術となっている。

研究成果の概要（英文）：In videos, the same event can be taken by different camera techniques and in different situations. Therefore, shots of the same event may contain significantly different features. In order to retrieve such diverse sets of shots for a given event (query), we propose a method which defines an event based on the rough set theory. First, given subsets of shots for an event as positive examples, we represent the event as the union of the subsets. Then, we adopt a partially supervised learning approach to obtain negative examples from a large amount of unlabeled data. To be precise, we identify ``likely'' negative examples from the unlabeled data based on their dissimilarities to the given positive examples. In calculating dissimilarities, we take advantage of subspace clustering to find clusters in different subspaces of the high-dimensional feature space.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	5,400,000	1,620,000	7,020,000
２００８年度	4,900,000	1,470,000	6,370,000
２００９年度	4,200,000	1,260,000	5,460,000
年度			
年度			
総　計	14,500,000	4,350,000	18,850,000

研究分野：総合領域
科研費の分科・細目：情報学・メディア情報学・データベース
キーワード：マルチメディア情報処理

１．研究開始当初の背景

現在、アメリカでは「通信」の本来的な技術である双方向のやりとりが行われるとともに、コンテンツのデジタル化と蓄積容量の拡大を反映して、「共有」モデルの開発が盛んに行われている。これらは，通常「動画共有サイト」と言われている。すなわち、視聴者がインターネットのサイトに自身の動画コンテンツを蓄積するとともに、番組制作者、放送局、企業も同サイトに動画コンテンツを蓄積するものである。通信回線の進化に伴って生まれた「動画共有サイト」は、日本でも流行の兆しを見せ、現在の放送局やネット企業は大きな転換を迫られている。ただ、ここでの技術的課題は、インターネットで席巻している検索エンジン技術が有効に機能しないことにある。言い換えると、通常のインターネットで用いられる検索エンジンにも画像や動画を対象としたものがあるが、極めて精度が悪い。これは、従来の検索エンジンがテキストを対象として進化した技術であり、画像，動画、音楽などのマルチメディアコンテンツに関しては、全く異なるアプローチを開発しなければならないからである。「動画共有」が、今後、大量に行われるようになっても，現在の検索技術レベルのままであれば、利用者が望むコンテンツは探すことができないという事態が生じうる。このため、アメリカでは動画検索エンジンの研究開発が盛んに行われており、日本においても積極的に開発を進めないと、かつてのテキスト検索のようにアメリカのネット企業に対して大きく立ち遅れる可能性がある。

２．研究の目的

動画像の検索は従来から研究されており，いくつかの成果も報告されている．たとえば，画像処理のアプローチを用いた「内容に基づく動画像検索」では，低レベルの特徴ベクトルから高レベルの意味的概念への写像を仮定し，まず概念ごとに単一の概念モデルを構築し，つぎに「オントロジーに基づく概念間の関係」を考慮して，各ショットをいくつかの概念からなる組み合わせとして注釈付けするという研究がある．この手法では，「オントロジーの階層構造」が概念間の意味的制約条件として働くため，妥当な注釈付け手法として注目されている．しかしながら，動画像をキーフレーム（静止画）の集合として考え，静止画ごとに画像処理するというアプローチであり，動画像が時系列データであるという特色を無視しているために，動きのある典型的なショットの注釈付けに失敗してしまうという問題がある．

一方，映像のようなマルチメディア処理の本質は，内容を理解することによるメディア統合にある．すなわち，音声，動画像などによって表現されている知識（内容）をメディア間で変換，要約，統合するための基盤技術を開発することが喫緊の課題である．ここでマルチメディア処理には少なくとも２種類のオントロジーが必要となる．すなわち，メディア自体のオントロジー（ビデオオントロジーと呼ぶ）と「内容」に踏み込んだオントロジー（ドメインオントロジーと呼ぶ）を用いた「理解によるメディア統合」である．もちろん，日本でもオントロジーに関する研究は盛んに行われているが，マルチメディアオントロジーに関する研究は皆無である．

このような状況に対して，我々はデータマイニング技術を用いて映像から有用な知識を発見する「ビデオマイニング」という考え方を提唱している．データマイニングは，膨大なデータから過去には知られていなかった興味深いパターンを発見する技術である．これに倣うと，ビデオマイニングは，映像からイベントを表現するための「意味的なパターン」を発見し，映像検索のための有用な知識を発見することになる．たとえば，「車が駆ける」シーンを「意味的なパターン」で記述すると，「映像中で直線の数が少なく，かつ動きが激しいショット」となる．特にカーアクションの場合は BGM が多用されるため，さらに「音楽を含むショットの連続」という条件がパターンに付加される．このように，「意味的なパターン」は低レベルの特徴ベクトルを組み合わせたものに過ぎないが，具体的に対象物を認識することなく，映像特有の現象を記述できるため，有効なアプローチであることが分かっている．

本研究課題で「内容」に踏み込んだマルチメディアオントロジーを新規に検討し，カテゴリーごとに定義されたドメインオントロジーを動画像検索に適用することが第一の目的である．また，ビデオマイニングの成果として得られるパターンから，ビデオオントロジーを定義付けることが第二の目的である．最終的に，上記の枠組みを統合し，動的な映像注釈技術を開発し，革新的な動画検索エンジンを開発する．

３．研究の方法

映像アーカイブから所望のイベントを効率的に検索するためには，いくつかの重要な技術的課題が残されている．まず，ユーザが興味をもつイベントは多種多様であり，これら全てを事前には列挙できないという点が挙げられる．そのため，事前にイベントの検索モデルを構築しておく“モデルベースの手法”や，事前にイベントに関連する概念を定義しておく“概念ベースの手法”では，ユーザからの多種多様な検索要求に対応できない．一方，イベントが表現されたサンプル映像と，特徴量に関して類似した映像を検索する“類似度ベースの手法”がある．この手法

は，サンプル映像が与えられれば，任意のイベントを検索できるという利点がある．しかしながら，類似度だけでは高精度にイベントを検索できない．これは，類似度ベースの手法では，所望のイベントが表現されたサンプル映像という“正例”しか使用していないことによる．

本研究では，正例に加えて，所望のイベントが表現されていない“負例”を使用した“クエリーベースの手法”を提案する．これにより，正例と負例の比較から，適切な特徴量と不適切な特徴量の区別が可能になり，類似度ベースの手法よりも高精度にイベントを検索可能になる．以下では，４つのクエリーベース検索における課題について述べ，さらに検討した手法について述べる．

(1) 同一イベントにおける特徴量の多様性：

同一イベントのショットでも，撮影技法，撮影状況，オブジェクトの動きといった様々な要因によって，特徴量が大きく異なってくる．図１に，「車が街を走っている」イベントのショットを３つ示す．

図１

ここで，郊外で撮影された Shot 1 では，空に対応して，画面上部からほとんどエッジが検出されない.都市部で撮影された Shot 2 では，ビルに対応して，画面上部から多数のエッジが検出される．また，車をタイトショットでとらえた Shot 3 では，画面全体から大きな動きが検出される．一方，車をロングショットでとらえた Shot 1 や Shot 2 では，画面中央部のみから動きが検出される．この考察から，同一イベントのショットは，特徴空間中で部分集合に分かれて分布していると仮定できる．

上記のような部分集合性を考慮するために，“ラフ集合理論”という，集合論に基づいて事例間の識別可能性を検証する分類手法を導入する．具体的には，まず，どの特徴量を使用すれば，正例と負例が識別可能か検証する．そして，正例（もしくは，負例）を正確に分類可能な部分集合を特定するための“決定ルール”を抽出する．最終的に，このような部分集合の和集合として，正例という１つのクラスを外延的に定義する．

しかしながら，従来のラフ集合理論が主に質的データを対象としていたのに対して，ショットから抽出される特徴量は，ショット長などの連続値，色などのヒストグラム，音声信号などの時系列というように，様々な形式で記述される．このような特徴量を少数のカテゴリーに離散化すれば必然的にエラーを伴うことになる．すなわち，意味的に無関係なショットが，同一のカテゴリーに割り当てられてしまうという問題が生じる．そこで，本研究では，近年提案された連続データに対するラフ集合理論を拡張して，任意形式の特徴量に対応可能なラフ集合理論を提案する．具体的には，事例間の識別可能性を，特徴量ごとの類似度に基づいて定義している．

(2) 負例選択の困難さ：

ユーザは，興味のあるイベントに関して，少数であれば正例を用意することができる．しかしながら，適切な負例を用意することは非常に困難である．これは，正例の補集合である負例には，多種多様なショットが該当するからである．すなわち，イベントごとに，大量のショットを人手で検証することは困難である．また，選択余地が広範囲になるため，選択された負例には，ユーザの主観が多分に含まれる可能性がある．このため，少数の正例のみが利用可能であり，その他の大量のショットはラベルなし事例であるという状況下でクエリーベース検索を行われなければならない．

そこで，正例とラベルなし事例から分類器を学習する“部分教師付き学習”として，クエリーベース検索を定式化する．したがって，負例はラベルなし事例から選択されることになる．通常，部分教師付き学習では正例の分布に基づいて負例が選択される．例えば，SVM やナイーブベイズを用いて，正例の分布を推定する手法がある．このような手法では，多くの正例がなければ正確に分布を推定できず，負例選択の精度が低下するという問題がある．これに対して，Fung らは正例とラベルなし事例の類似度のみから負例を選択する手法を提案している．Fung らの手法は，ごく少数の正例しか利用できない状況下でも，際立って優位であることが実証されている．そこで，この手法を拡張してクエリーベース検索に導入する．

(3) 特徴量の高次元性：

色，エッジ，動きといった様々な特徴量で記述されたショット（事例）は，非常に高次元のデータである．こうした高次元空間では，“次元の呪い”と呼ばれる問題が生じる．そこで，正例とラベルなし事例の類似度を正確に測るためには，適切な次元と不適切な次元を区別する必要がある．

一方，様々なイベントが表現されているラベルなし事例は，それぞれ異なった特徴量の組み合わせによって特徴づけられると考えられる．そこで，部分空間クラスタリングを用いて，高次元空間における部分空間で類似したラベルなし事例のクラスターを抽出する．すなわち，各クラスターには，特定の特徴量の組み合わせが関連づけられ，その特徴

量に関してのみ類似したラベルなし事例が含まれるようになる，最終的に，ラベルなし事例が属するクラスターに関連づけられた特徴量のみを用いて，正例との類似度を測るようにしている．

(4) ビデオオントロジーの利用

検索精度の向上のために，イベントを独立に検索するのではなく，過去に検索したイベントとの関係を考慮した検索手法を開発する．たとえば，「車が街を走る」イベントを検索するとき，「水上に船が浮かんでいる」，「人が室内で会話している」といった無関係なイベントに関する決定ルールにマッチするようなショットは検索すべきでないことは明らかである．これらの問題を解決するために，本研究では，「内容」に踏み込んだビデオオントロジーについて検討し，概念ごとに定義されたオントロジーを用いて，動画像検索に適用する手法を開発する．特に，Large Scale Concept Ontology for Multimedia (LSCOM) で既に定義されている 374 個の概念をオントロジーとして組織化する．さらに，映像検索の精度向上のため，ビデオオントロジーを用いて，ユーザの検索要求を精緻化して不要な映像は排除し，曖昧な要求は詳細化する手法を開発する．

４．研究成果

本研究では，438 本の映像，71,872 ショットからなる TRECVID 2008 の映像データを実験データとして使用した．特に，以下の３つのイベントに関して，提案手法の性能評価を行った．

Event 1: 人がドアを開ける
Event 2: 人が路上でカメラに向かって話している
Event 3: 車が街を走る

表１に，上記のイベントに関する実験結果の概要を示す．まず，２列目から分かるように，提案した部分教師付き学習による負例選択 (PSL) を，手動 (Manual)，ランダム (Random) での負例選択と比較している．５列目は，検索された 300 ショットに関する精度を表しており，正解ショット数を括弧に記述している．さらに６列では， SVM による検索結果との比較を行っている．

表１

	n-example selection	# of p-examples	# of n-examples	P@300 (# of rel.)	P@300 by SVM
Event 1	*Manual*	9	16	0.070 (21)	0.060 (18)
	Random	9	50	0.087 (26)	–
	PSL	9	50	0.070 (21)	–
Event 2	*Manual*	11	16	0.087 (26)	0.060 (18)
	Random	11	50	0.050 (15)	–
	PSL	11	50	0.050 (15)	–
Event 3	*Manual*	9	14	0.217 (65)	0.223 (67)
	Random	9	50	0.127 (38)	–
	PSL	9	50	0.170 (51)	–

表１から，選択された負例によって，検索性能が大きく変化していることが分かる．特に，Event 1 を除いて，Manual の精度が PSL と Random に比べて際立って高いことが分かる．これは，機械的に選択された負例よりも，手動で選択された負例の方が質の高いことを表している．一方，Event 3 に関しては Random より PSL の方が，Event 1 に関しては PSL より Random の方が，精度が高くなるという結果が得られた．この結果から，PSL の性能は，イベントに対する正解ショット数に依存するということが分かった．

具体的には，Event 3 の「車が街を走っている」ショットは実験データ中に比較的多く存在し，Event1 の「人がドアを開ける」ショットはほとんど存在しない．さらに，Event 3 の場合，特徴量を解析する PSL の方が，特徴量を全く解析しない Random よりも正確に負例を選択できるため，検索精度が高くなっている．一方，Event1 の場合，負例拡張を行っても PSL では正例と類似しているような質の高い負例が選択されなかった．これに対して，Random では，ランダム性と正解ショットが少ないことから，誤って正解ショットを負例とすることなく，正例と類似したショットを負例として選択することに成功し，高い検索精度が得られている．このため，正解ショットが少ない状況下における PSL の性能を向上させるためには，例えば遺伝アルゴリズムのようなランダム性を伴って負例を選択するメカニズムを導入することが有効であると考えられる．

次に，右端の列から，本手法による精度は SVM と比べて相対的に高くなっていることが分かる．特に，分類アルゴリズム以外に関して，本手法と SVM で全く同一事例を使用しているため，上記の結果は，ラフ集合理論の有効性を実証していると言える．また，両手法により検索されたショットを比較したところ，大きな違いがあることが分かった．

図２に，本手法，もしくは SVM により検索された３つのショットを示す．まず，図２(a) から分かるように，本手法によって，Shot 1 のようなタイトショット，Shot 2 のようなミドルショット，Shot 3 のようなロングショットなど，様々なショットサイズで撮影された Event 3 を検索できていることが分かる．また，Shot 1 と Shot 3 の上部には空が映っており，Shot 2 と Shot 3 には建物が画面中の大きな領域を占めている．このことから，本手法は，同一イベントにおける特徴量の多様性にうまく対応できていると言える．

図２

一方，SVM では，ショットが正例と類似した特徴量を少しでも含んでいれば，検索されるという傾向がある．例えば，図２(b) において，Shot 4 は正解であるが，Shot 5 と Shot 6 は明らかに不正解である．これらのショットは，「下部に灰色が多い」という特徴量（本来は，道路を特徴づける）を含んでいる理由のみで検索されている．このように SVM から得られたイベント検索モデルは一般的すぎるのに対して，本手法では多くの決定ルールを含んでいるかどうかという基準でショットを検索しているため，上記のような不正解のショットは検索されることがない．

最後に，図３を用いて，抽出された決定ルールについて詳しく考察する．

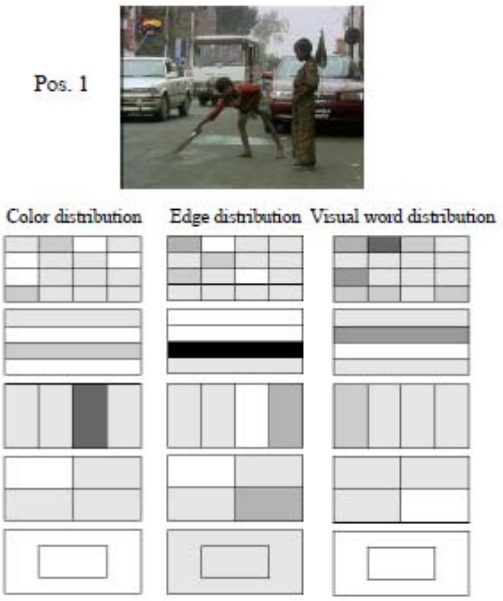

図３

図３は，Event 3 に対する１つの正例 Pos1 と抽出された決定ルールとの関係を表している．特徴量が多くの決定ルールに含まれているほど，対応する領域が黒くなっている．たとえば，19 番目の領域のエッジ特徴は非常に多くの決定ルールに含まれていることが分かる．そして，この特徴量は「道路からはほとんどエッジが検出されない」ことを表している．また，18 番目の領域の SIFT 特徴も多くのルールに含まれており，車を特徴づけていることが分かる．さらに，1，2 番目の領域の SIFT 特徴も多くのルールに含まれており，街路樹や建物を特徴づけている．このように，本手法によって，イベントに関連するオブジェクトを特徴づける有用な決定ルールが抽出されていることが分かる．

しかしながら，上記の有用な決定ルールに加えて，正例に過度に特化した意味のないルールも抽出されている．たとえば，図３の Pos 1 から，23 番目の領域の色特徴が多くのルールに含まれている．ただし，これらのルールは，「車の前に立っている人物の服の色」を特徴づけており，多くの不正解ショットが検索される要因になっている．このような過度に特化したルールが抽出される最大の理由として，1 つの正例と全ての負例の識別可能性からルールを抽出している点が挙げられる．その結果，ある正例にたまたま映っていたレアなオブジェクトを特徴づけるルールが抽出されやすくなっている．そこで，いくつかの正例の集合と全ての負例の識別可能性からルールを抽出すれば，ある正例にだけ出現しているレアなオブジェクトを特徴づけるルールは抽出されなくなると考えられる．

５．主な発表論文等

〔雑誌論文〕（計 4 件）

1) K. Shirahama and K. Uehara, A Novel Topic Extraction Method based on Bursts in Video Streams, International Journal of Hybrid Information Technology, Vol. 1, No. 3, pp. 21-32, (2008).
2) 野宮浩揮，上原邦昭，相補的な視覚的学習による複数の認識手法の統合，電子情報通信学会論文誌，Vol.J90-D, No.11, pp.3043-3054（2007）.
3) 熊野雅仁，有木康雄，上原邦昭，輝度投影相関と二分化テンソルヒストグラムを併用したオンライン処理向けカメラワーク解析法の精度向上 ～訓練指向型オンライン映像撮影ナビゲーションシステム～，映像情報メディア学会誌，Vol.61, No.8, pp.1159-1167（2007）.
4) 熊野雅仁，有木康雄，上原邦昭，実時間カメラワーク評価に基づく単一ショット訓練指向型オンライン映像処理ナビゲーションシステム ～映像文法を背景とした映像撮影学習システムに向けて～映像情報メディア学会誌，Vol.61, No.8, pp.1150-1158（2007）.

〔学会発表〕（計 13 件）

1) K. Shirahama, C. Sugihara and K. Uehara, Query-based Video Event Definition Using

Rough Set Theory and High-dimensional Representation, Proc. of the 16th International Conference on Multimedia Modeling (MMM 2010), pp.358-369 (2010).
2) K. Shirahama, C. Sugihara, Y. Matsuoka and K. Uehara, Query-based Video Event Definition Using Rough Set Theory and Video Prototypes, Proc. of IS&T/SPIE Electronic Imaging Multimedia Content Access: Algorithms and Systems IV, 7540B-41 (2010).
3) K. Shirahama, C. Sugihara, K. Matsumura, Y. Matsuoka and K. Uehara, Mining Event Definitions from Queries for Video Retrieval on the Internet, Proc. of the 1st International Workshop on Internet Multimedia Mining in conjunction with IEEE ICDM 2009 (IMM 2009), pp.176-183 (2010).
4) Kimiaki Shirahama, Chieri Sugihara, Yuta Matsuoka, Kana Matsumura and Kuniaki Uehara. Kobe University at TRECVID 2009 Search Task, Proc. of TREC Video Retrieval Evaluation (TRECVID) 2009 Workshop, pp. 76-84, 2009
5) K. Shirahama, C. Sugihara, Y. Matsuoka and K. Uehara, Query-based Video Event Definition Using Rough Set Theory, Proc. of the 1st ACM International Workshop on Events in Multimedia (EiMM 2009), pp.9-15 (2010).
6) Kimiaki Shirahama, Akihito Mizui and Kuniaki Uehara, Characteristics of Textual Information in Video Data from the Perspective of Natural Language Processing, Proc. of Semantic Knowledge Discovery, Organization and Use, 2008
7) Akihito Mizui, Kimiaki Shirahama and Kuniaki Uehara, TRECVID 2008 NOTEBOOK PAPER: Interactive Search Using Multiple Queries and Rough Set Theory, Proc. of TREC Video Retrieval Evaluation (TRECVID) 2008 Workshop, pp. 123-132, 2008
8) K. Shirahama and K. Uehara, A Novel Topic Extraction Method based on Bursts in Video Streams, Proc. of the 2nd International Conference on Multimedia and Ubiquitous Engineering, pp.249-252, (2008).
9) K. Shirahama and K. Uehara, Query by Shots: Retrieving Meaningful Events Using Multiple Queries and Rough Set Theory, Proc. of the 9th International Workshop on Multimedia Data Mining, pp. 43-52, (2008).
10) H. Nomiya and K. Uehara, Multistrategical Image Classification for Image Data Mining, Proc. of International Workshop on Multimedia Data Mining, pp.22-30 (2007).
11) K. Shirahama and K. Uehara, Video Data Mining: Discovering Topics by Burst Detection in Video Streams, Proc. of ICDM 2007 Workshop on Knowledge Discovery and Data Mining from Multimedia Data and Multimedia Applications, pp.57-62 (2007).
12) H. Nomiya and K. Uehara, Multistrategical Approach in Visual Learning, Proc. of 8th Asian Conference on Computer Vision, pp.502-511 (2007).
13) K. Shirahama, K. Otaka and K. Uehara, Content-Based Video Retrieval Using Video Ontology, Proc. of the 3rd IEEE International Workshop on Multimedia Information Processing and Retrieval, pp.283-288 (2007).

〔図書〕(計１件)

(1) Hiroki Nomiya and Kuniaki Uehara, Content-based Image Classification via Ensemble Visual Learning, pp.141-166, Julio Ponce and Adem Karahoca (eds.) Data Mining and Knowledge Discovery in Real Life Applications, intechweb.org

６．研究組織

(1)研究代表者

上原　邦昭 (Kuniaki Uehara)
神戸大学・工学研究科・教授
研究者番号：60160206