データベースシステム 113－29
（1997. 7. 15）

映像自己組織化機構に基づく内容記述と類似シーン検索

波多野 賢治† 亀井 俊之‡ 田中 克己†
†神戸大学大学院自然科学研究科情報メディア科学専攻
‡神戸大学大学院自然科学研究科情報知能工学専攻

本稿では、Kohonenの自己組織化マップを用いたビデオ映像カットへの内容記述、その記述情報とコンテンツ情報を共に取り入れたハイブリッドな自己組織化システム、類似シーンの検索機構について述べる。従来までのコンテンツ情報のみによる分類精度には限界が生じるため、映像に対する内容記述情報とDCT情報を併用して特徴ベクトルを構成することを試みた。また、コンテンツのみによる分類結果をもとに内容記述を行うことで、一般に手作業となる動画像データへの記述をできるだけ効率的に行える。また、シーン特徴ベクトルを生成し、そのベクトル値を用いた分類を行い評価を行なった。

Authoring and Retrieval of Scenes supported by SOM-based Video-Clustering

Kenji Hatano† Toshiyuki Kamei‡ Katsumi Tanaka†
†Division of Media and Computer Sciences,
Graduate School of Science and Technology, Kobe University
‡Division of Computer and Systems Engineering,
Graduate School of Science and Technology, Kobe University

In this paper, we discuss a video shot authoring based on a hybrid self-organizing system which uses both of video-content information and metadata (description data) as input for Kohonen's Self-Organizing Map.

Conventional methods relaying on contents description only has its limits when considering the accuracy of classification. We sought to improve its accuracy by introducing image description as a feature vector. By using classification by contents only, the task of describing animated picture data which is done manually can be carried out more effectively. We also produced scene feature vector to use its values for classification, and we have made evaluations.

1 はじめに

WWW(World Wide Web)の急速な普及と発展、および、計算機性能の向上、2次記憶媒体の大容量化、さらにはヒューマンインターフェイスの重要性に伴い、文書データのみならず、映像や音声などのマルチメディアデータを取り扱うことが可能かつ、必要になってきている。しかし、それら多種多様なマルチメディアデータはインターネットなどの通信ネットワーク上に分散しているため、それらのデータを動的・自己組織的に構造化することが必要不可欠になってくる。中でも、動画像データは意味情報が複雑かつ内容や構成に関する情報が明示的には含まれていないことから、動画像データに対するデータベース化は非常に困難である。

この動画像データベースを構築する際には、ユーザからの問い合わせの方法を考慮することが重要になってくるものと考えられる。動画像データの場合、ユーザの問い合わせというものは、動画像のコンテンツ情報を基にしたものというものよりも、むしろ、その内容を基にしたものであろうと予測される。このことからも、データベースを構築する際には、動画像データに対する内容記述を行なう必要性がでてくるものと考えられる。また、我々はこれまでに画像のコンテンツ情報のみを用いた自己組織化マップによる自動分類を行なってきた[1]。過去の結果を見てみると精度はそれほど高くないが、ある程度の自動分類が可能であることが判明している。

このような背景から、本研究では、自己組織化マップを用いた動画像データのコンテンツ情報をもとにしたクラスタリング結果を映像の内容記述を行なう際の補助的な手段として用いることにし、実際に動画像データのカットに対し記述を行い、その記述情報と動画像のコンテンツ情報を両方取り入れたハイブリッド型の特徴ベクトルを生成し、それをもとに自己組織化マップによる学習およびクラスタリングを行なった。また記述によって与えられた記述情報をシーンへ継承することで、記述情報のみを含んだシーン特徴ベクトルを生成し、これを入力として自己組織化マップの学習を行ない、シーンの自動クラスタリングを試みた。

2 基本的事項

2.1 自己組織化マップ (Self-Organizing Map)

ニューラルネットワークの一種であるSOMは、1990年にT.Kohonenによって提案された教師なし競合学習モデルである[2]。出力層の各ユニットが層の中で位置を持つという点が他の学習モデルと異なる。このモデルの特徴はデータに隠されているトポロジカルな構造を学習アルゴリズムにより発見し、通常2次元空間で表示するというものである。

具体的には、入力データを通常高次元の特徴(feature)ベクトルxにパターン化し、出力層にある各ユニットiが入力パターンxと同次元のベクトルm_iを持っており、2次元平面上に配置される。学習はこれらのユニットが入力パターンに選択的に近付けることによって進行する。競合というのはSOM法が入力パターンに一番近いパターンを持つ出力ユニットcおよびその近傍のユニットの集合N_cのみが入力パターンに近付けることができるようなアルゴリズムをとっている。また、統計的に正確な学習効果を得るため、一定の学習回数Tをとらなければならない。

2.2 離散コサイン変換 (Discrete Cosine Transformation)

DCTとは、JPEG(Joint Photographic Expert Group)とよばれる静止画像圧縮技術で用いられている画像の変換符合化方式である。1枚の自然画像をN×N画素の正方形の領域(ブロック)に分割し、各ブロックに対して変換処理を行うと、領域内の平均的な画像(領域全体が一様)に始まり、徐々に精細さを表現する画像へと段階的な画像に分解することができる。この分解操作を直交変換といい、精細さが高いことを別のいい方では、周波数が高いという。自然画像は、第1低周波項(平均値画像)から順に、高周波項へと分解した画像の重ね合わせの表現になる。

DCTのメリットは、変換前にランダムに分布していた画素値(輝度など)が、変換後には低周波項に大きな値が集中する性質がある。したがって、高周波項を落とす操作(量子化)をすれば情報圧縮を行なうことができる。

3 カットに対するハイブリッドシステムの適用

我々は、これまでコンテンツ情報のみを用いた画像のクラスタリングを行ない[1]、その結果を評価してきた。しかし、一般にコンテンツ情報のみを用いると、動画像の内容の情報が欠落しているため、意味的な類似性が低いデータの集合にクラスタリングされてしまっているという結果になることが多々ある(図1参

照)。そのため、コンテンツ情報のみを用いたクラスタリングには限界があることが判明している。

よって、更なるクラスタリング精度の向上のためには映像に対する記述情報が必要になってくるものと考えられるため、我々は、コンテンツ情報およびキーワードによって与えられる映像の内容記述情報を共に含んだハイブリッド型の特徴ベクトルを生成し、これを入力値としてSOMによる学習を行なった。

3.1 ビデオデータの選択

従来の我々の研究ではニュース映像などの実世界の映像を主に使用してきたが、我々が採用したDCTによる画像情報の取得では、フルカラーの映像に対しては、カットベースの分類でさえも精度の向上が困難な例が多々見られた。

そこで、入力データとなるビデオデータをある種のもの、例えばアニメーション映像や劇場映画の予告編などビデオの種類を限定して、どの種類の映像を用いた場合にこのDCTによるカット分類がうまくいくかどうかを確かめた。この結果、色構成や画面の構図がほぼ決まっているアニメーション映像が比較的良好にカット分類ができることから、実験対象とする対象とするデータにアニメーション映像を選択した。

3.2 ハイブリッド型特徴ベクトルの生成

カットに対するハイブリッド型特徴ベクトルの生成のために以下のような処理を行なう。

1. **コンテンツ情報をもとにした、カットのクラスタリング**
   ビデオデータを有木らのカット検出プログラム[3]を用いて、カット点を検出する。次に、このカットの特徴ベクトルの生成方法として、カット検出で分割された各カットに含まれる全てのフレームのフレーム特徴ベクトルの重みつき平均値を用いる方式を採用した。この方法では、DCTを行なうと低い周波数にその画像のパワー(特長)が集中するという特徴を利用することにより、情報が膨大な動画像の特徴をできるだけ低次元のベクトルに反映させることができる。こうして求めたカット特徴ベクトルを入力としてSOMに学習させ、その結果をマップ表示させる。

2. **キーワード付与による内容記述**
   コンテンツ情報の類似度が高い動画像どうしでは、記述すべき情報も同じようなものであろうという考えから、本研究ではこのコンテンツ情報のみを用いたクラスタリングの結果を画像に対する内容記述の補助的手段として利用することを考えた。具体的には、マップ中の同じセルに配置されたカットには同じ内容の記述を行ない、すべてのカットに対して人間が記述を行なうという手間を省くものである。
   また、記述によって生じるであろう各セルに配置されたデータのノイズであるが、これに対しては手動で取り除くことにした。

3. **記述情報のみを含む特徴ベクトルの生成**
   記述されたキーワードにより、まず、映像の内容の記述情報を含んだ特徴ベクトルを生成する。

   第iカットの記述情報特徴ベクトルを式で表すと、次のようになる。

$$\boldsymbol{V}(cut_i) = w\{k_1, k_2, \ldots, k_n\}$$

$$k_i = \begin{cases} 0 \\ 1 \end{cases}$$

   ただし、全てのカットへの記述の際に用いたキーワード集合を$\{k_1, k_2, \ldots, k_n\}$、$cut_i$は第iカット、$k_j$は第j番キーワード、$w$はコンテンツ情報に対する内容記述への重みを表す。ここでwはコンテンツのDC成分の大きさに合わせ100とした。

4. **コンテンツ情報と記述情報を統合させる**
   コンテンツ特徴ベクトル、記述情報特徴ベクトルをそれぞれ

$$\boldsymbol{V}(cut_i) = \{c_1, c_2, \ldots, c_m\}$$

$$\boldsymbol{V}(cut_i) = \{wk_1, wk_2, \ldots, wk_n\}$$

   とすると、ハイブリッド型の特徴ベクトル$\boldsymbol{V}(cut_i)$は、次のようになる。

$$\boldsymbol{V}(cut_i) = \{c_1, c_2, \ldots, c_m, wk_1, wk_2, \ldots, wk_n\}$$

3.3 実行例

前述した方法により生成されたハイブリッド型特徴ベクトルを入力として、SOMを学習させ、その結果をVRMLによって出力する。実際に、15分のアニメーションをカット分割し、15×15のマップを用いてマップ生成を行なった。この時、学習回数は12,000回、データとなったカットの総数は125個で、ユニッ

図 1: A Contents based 3D-SOM for Video Cuts ©Sun Television Corporation Limited

ト数225個のマップは約1分で生成された。分類結果を図2に示す。

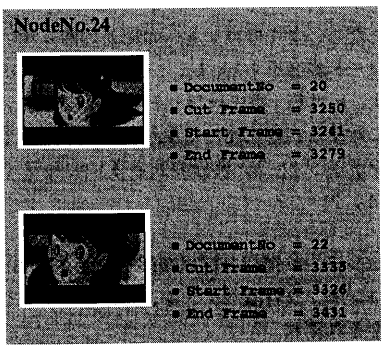

図 2: A Hybrid-type 3D-SOM for Video Cuts ©Toei Corporation Limited

3.4 評価

実際、行なった実験がどれほどの結果なのかを判断するために「精度(適合率)」と「再現率」を測定する。従来から良く知られている「精度(適合率)」と「再現率」は、検索されなかった適合情報を A、検索された適合情報を B、検索された不適合情報を C とした場合、精度は $B/(B+C)$、再現率は $B/(A+B)$ で表されるものであるが、本研究で利用したものは少し定義が異なる。ここで用いた「精度」と「再現率」はSOMであることを考慮して次のように定義する。

すべてのカットの中から、マップ上の円筒の上に張りつけられたあるカット v に似ているカットの数を $similar(v)$ とし、カット v が張りつけられている円筒付近にあるカットの数を $neighbour(v)$ で表わすとする。このもとで精度は、

$$\frac{|neighbour(v) \cap similar(v)|}{|neighbour(v)|}$$

で表わされ、また再現率は、

$$\frac{|neighbour(v) \cap similar(v)|}{|similar(v)|}$$

で表わされる。

表1では画像のコンテンツ情報のみ、表2ではハイブリッド情報による場合の精度と適合率を、それぞれ測定範囲が中心の円筒から近傍距離が1, 2の場合を示す。

distance	precision ratio(%)	recall ratio(%)
1	64.59	45.08
2	35.54	55.94

表 1: Precision ratio and Recall ratio of contents based retrieval

本研究で用いた動画像データはアニメーション画像であるため、映像に用いられている色、構図などがほぼ決まっていることから、記述の前段階でも普通のテレビ映像を用いた場合に比べかなり高いクラスタリ

distance	precision ratio(%)	recall ratio(%)
1	65.03	47.67
2	37.95	60.69

表 2: Precision ratio and Recall ratio of hybrid type retrieval

ング精度が得られている[1]。このクラスタリング結果を基に人がキーワードを与えることで内容の記述を行なうと、その与えられたキーワードをも考慮された入力により SOM に学習されるので、コンテンツ情報のみの場合と比較すればわずかながらであるが良好な結果が得られ、同一の映像でも異なる内容記述を行ったカットは異なるセルに配置されるという結果が得られた。ただし、この方法では同一のキーワードを与えると不都合なカットも当然存在するので、今回はこのようなノイズは手動で除外している。

しかし、カットが類似しているか否かという判断は我々人間の目で行っており、また、コンテンツ情報のみで類似している、コンテンツ情報と映像の内容情報で類似しているという異なる 2 つの基準で精度、再現率を計算しているため、この数値を単純に比較してどちらが良いと判断するのは安易であると思われる。

4 シーンへの記述および類似シーン検索

一般に動画像データはシーンと呼ばれる映像を編成する 1 つ 1 つの記事から構成されている。またこのシーンはカットと呼ばれる映像の中でカメラの切り替えや素早いカメラワークなどによって分割される 1 つ 1 つの場面から構成されている。シーンの場合はカットとは異なり、含まれている情報量が多いためカットとは異なる特徴ベクトルを生成する必要があると考えられる。以下にシーンに対する特徴ベクトルの生成法を示す。

4.1 シーンの特徴ベクトルの生成

シーンはカットとは異なり、その映像における話の流れ、場面の展開などから決定されるものであるため、カットのようにコンテンツ情報の変化から自動的に検出することが困難である。従って、本研究におけるシーン検出は手動で行なうことにした。また抽出した各シーンに対して直接キーワードを付与することも考えられるが、一般に 1 つのシーンには複数の人物、物体やその動作の情報が含まれており、これらの情報をその場で理解し、記述を行なうことは難しいと考えられる。それに対して、カットに描写されている対象物や動作は比較的少ないことから、シーンに対する記述を行なう代わりに、カットに対して記述を行なった後に、その記述情報をシーンに継承させる。シーンの特徴ベクトルを生成するには以下のような処理を行なう。

1. カットへの内容記述を行なう
   3.2 で前述した方法によりカットに対する内容記述を行なう。

2. シーンに含まれるカットを求める

3. カットの記述情報をシーンに継承させる
   各カットに付与された、映像の内容記述情報をもとにシーンの特徴ベクトルを生成する。以下に本研究で用いた計算式を示す。

$$\boldsymbol{V}(S) = \sum_{cut_i \in S} \frac{n(cut_i)}{N} \boldsymbol{V}(cut_i)$$

ただし、シーン S は

$$S = \{\, cut_1,\ cut_2,\ \ldots,\ cut_n \}$$

$\boldsymbol{V}(S)$ は記述情報のみを含むシーンの特徴ベクトル、$\boldsymbol{V}(cut_i)$ はカットの記述情報特徴ベクトル、シーンに含まれるフレーム数を N、カット cut_i に含まれるフレーム数を $n(cut_i)$ とする。
ここでは、まずシーン中に含まれるフレーム数に対するカット中に含まれるフレーム数の割合を算出し、その値をキーワード特徴ベクトルに重みとして加えることでシーンの記述情報のみを用いた。

4.2 実行例

生成されたシーン特徴ベクトルを入力として SOM による学習を行い、その結果を VRML によって出力する。出力結果を表示する際には、6 × 6 のマップを用いた。また、シーン分割を手作業で行なった結果、検出されたシーンの総数は 11 個で、ユニット数 36 個のマップは学習回数 15,000 回の場合、約 5 秒で生成された。分類結果を図 3、図 4 に示す。

[1] テレビ映像を用いた場合の精度および再現率は次の通り。精度 53.71%、再現率 43.52%

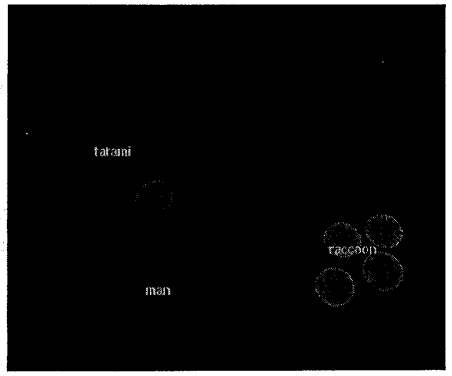

図 3: A 3D-SOM for Video Scene

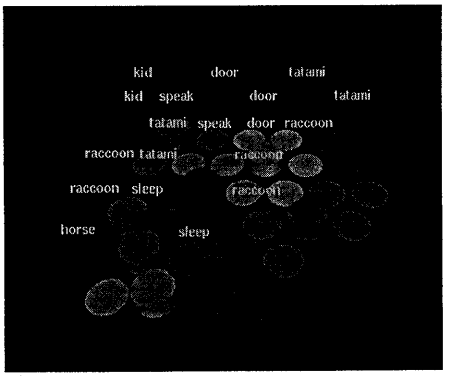

図 4: A Hierarchical 3D-SOM and Zoom-in Operation

4.3 評価

映像データに対する内容記述にあたっては、映像に登場してくる人物の名前、物体、およびそれらのとりうる動作や、状態を表わす単語を付与することにした。結果を見る限りでは、我々が行なった内容記述情報を反映してある程度クラスタリングされていることが分かった。しかし、シーンの総数が 11 個と極端に少ないとや、記述者の主観が多少影響していることなどから、中にはそれほど類似性の高くないデータ集合にクラスタリングされているものも見受けられた。

5 おわりに

本研究では、自己組織化マップを用いて動画像データに対する内容記述をカットに対して行ない、その記述情報および動画像データのコンテンツ情報を共に用いたカットのクラスタリング、およびカットの記述情報をシーンの特徴ベクトル生成に適用させ、その特徴ベクトルを用いたシーンのクラスタリングを行なった。

本研究の結果として、利点を挙げると次のものが挙げられる。

- コンテンツ情報のみを用いたカットのクラスタリング結果を利用することによる動画像データに対する内容記述の手軽さ
- カットへの記述情報をシーンに対して継承させることによるシーンに対する記述の省略化

今後の課題として、

- 複数の人が記述を行なう際に生じる記述情報間の矛盾をどう解決し、記述内容の統合をどのようにして行なうか
- シーン分割の自動化
- 記述をより効率的に行なうための工夫や、ツールの開発
- 数多くのデータに対して同様の実験を行ない、この方法の有効性の実証

が挙げられる。

謝辞

本研究において、貴重な映像資料の学術利用を許可して下さった、東映株式会社およびサンテレビジョンに感謝致します。また、本研究の一部は、日本学術振興会未来開拓学術研究推進事業における研究プロジェクト「マルチメディア・コンテンツの高次処理の研究」および文部省科学研究費重点領域研究 (課題番号 08244103) による。ここに記して謝意を表します。

参考文献

[1] K.Hatano, Q.Qing and K.Tanaka. *A SOM-Based Information Organizer for Text and Video Data.* Proceedings of the Fifth International Conference on Databese Systems for Advanced Applications (DASFAA'97), pages 205–214, Apr. 1997.

[2] T.Kohonen. *The Self-Organizing Map,* Proc. of the IEEE, Vol.78, No.9, pp.1464-1480, 1990

[3] 岩成 英一, 有木 康雄, 「DCT 成分を用いた動画シーンのクラスタリングとカット検出」, 電子情報通信学会, パターン認識と理解研究会, PRU93-119, pp.23-30, 1994