多段階自己組織化マップによるビデオ映像記述支援と類似シーン検索

波多野 賢治† 亀井 俊之†† 田中 克己†

意味情報に基づいてビデオ情報の分類や検索を行うためには，映像に対して何らかの意味記述を行う必要がある．しかし，ビデオ映像に対して意味記述を行う作業は，通常，人手によるものであるため，作業自体に多くの手間を要する．本論文では，カット割りされた映像群をその内容によって自動分類するシステムをKohonenの自己組織化マップを用いて構築し，これをカット映像に対する意味記述支援機構としてシーン映像の意味記述や検索を効果的に行う手法を提案する．本手法では，カット映像に対する記述情報をシーン映像の記述情報として継承させ，その記述情報を基に再び自己組織化マップに学習させる多段階自己組織化マップを生成し，これにより意味情報に基づいたシーン映像の分類と検索が可能となっている．

Authoring and Retrieval of Video Scenes by Multi-level Self-organizing Maps

KENJI HATANO,† TOSHIYUKI KAMEI†† and KATSUMI TANAKA†

In order to classify and retrieve video data based on their semantics, we need to add indexing information to video data, which describe their content. Conventionally, this authoring work is done manually, and it is a very tedious work. In this paper, we propose a content-based clustering system of video shots using Kohonen's Self-Organizing Map (SOM), which works as an authoring support system. Since the SOM for video shots enables clustered video shots to share the same description information, it is useful to save our labor in video authoring. Most of the description information for video shots is used to compute feature vectors of each video scene, which consists of several video shots. By using the computed feature vectors, we again make a SOM for video scenes, in which 'similar' video scenes are clustered on the SOM. In our method, SOMs are used twice to classify video scenes and so, we call it a "multi-level self-organizing map".

1. はじめに

WWW（World Wide Web）の急速な普及と発展，計算機性能の向上，2次記憶媒体の大容量化などにともない，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機会が増加している．そのため，爆発的に増加するマルチメディアデータをデータベース化する技術[1]はますます重要になってきている．しかし，マルチメディアデータ，特に映像データは，その意味情報表現が複雑であり，かつ意味情報や構成に関する情報が明示的には含まれていないなどの理由から，映像データの現在の分類方法や自己組織化の技術ではまだまだ不十分な点が残されている．従来より映像データベースの構築には，映像データにキーワードや説明文など索引となる情報を付加し，この2次情報を利用して映像の分類や検索を行うといった方法がとられている[2)~4)]．この場合，映像の分類体系が静的に規定されていることが多く，これらの分類や2次情報の付加は手作業であるため，データベースの構築と維持に巨大なコストがかかるという問題や，分類作業の効果的な支援機能が不十分であるという問題が生じている．

最近，データベース全体の把握や曖昧検索および分類作業の支援などの目的を達成するため，ニューラルネットワーク技術を用いた研究が行われている．我々も，ニューラルネットワークの一種である自己組織化マップ（Self-Organizing Map: SOM）[5)~7)]を用いて文書群の動的な分類，検索システムの開発[8),9)]や

† 神戸大学大学院自然科学研究科情報メディア科学専攻
Division of Information and Media Sciences, Graduate School of Science and Technology, Kobe University
†† 神戸大学大学院自然科学研究科情報知能工学専攻
Division of Computer and Systems Engineering, Graduate School of Science and Technology, Kobe University

カット映像のコンテンツ情報☆による分類システムの開発[10]を行ってきている．しかしこのような従来のコンテンツ情報によるカット映像の分類システムは，映像の意味的な情報が無視されるという欠点を持っている．本論文では，映像のコンテンツ情報による自己組織化マップの機能拡張として，映像データに意味的な情報を効果的に付与できる機能をあわせ持った類似シーン検索システムの試作を行ったのでそれについて報告する．

本論文で提案する手法の要点を以下にあげる．

- **カット映像単位の内容記述**
  一般に，映像データは，シーンと呼ばれる映像単位（たとえば，ニュース映像でいえば記事と呼ばれるもの）で構成されている．カット映像とは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つの映像のことをいうが，シーンは複数のカット映像から構成されている．本論文では各カット映像ごとに意味記述を行う方法を採用する．なぜなら，シーンに対して意味記述を行う場合，1つのシーンに多くのカット映像が含まれているため記述すべき情報が多く，記述者が映像の内容を容易に記述することができないと考えられるからである．また，逆にフレーム画像に対して意味記述を行うという方法は非現実的である．このような理由から，記述する単位はカット映像と定めた．
- **SOMによるカット映像分類マップによる記述情報の共有**
  我々が映像データに対して意味情報を付加する場合は，すべてのカット映像に対してその内容の記述を施さなくてはならない．しかし，これらの作業は人手によるところが大きく，依然，記述作業に手間を要するという大きな問題を生じる．そこで，映像データを映像そのものの色情報や色合いの情報であるコンテンツ情報（ここではDiscrete Cosine Transform情報，以下コンテンツ情報（DCT）と記す）で分類した自己組織化マップをあらかじめ用意しておき，そのマップ中の同じセルに分類されたカット映像に対しては同じ記述を施すという方法をとることで，ユーザがすべてのカット映像に対して意味記述するという労力を省き，効率的に記述が行えるように工夫した．
- **カット映像の意味記述からのシーン映像の意味記述の生成**
  一般ユーザが映像データに対して検索を行うことを想定した場合，考えられる検索データの単位はシーンである場合が多い．しかし，シーン映像はカット映像に比べ含まれている情報量が多く，それに対して記述を行うには，映像データの内容の知識がある程度必要であるなど，記述者が簡単に意味内容を記述するのは困難である．そこで，描写されている対象物やその動作内容などが少なく比較的記述の行いやすいカット映像に対し意味記述を行い，その記述内容をそのカット映像から構成されているシーン映像に継承させることで，シーンの記述情報が得られる．
- **多段階自己組織化マップ**
  本研究においてSOMは，カット映像のコンテンツ情報（DCT）によるカット映像の分類と，シーンに対して与えられた意味記述情報によるシーンの分類の2段階に用いられている．このように，SOMを多段階で使用するという研究は他にも始まっている．Kohonenの研究グループでは，文書データに対しこの多段階SOMを用いており，出現した単語のカテゴリ分類[11),12)]に第1段階のSOM，カテゴリ分類された単語を基にした文書データの分類に第2段階のSOM[13),14)]を用いている．扱うデータの違いがあるにせよ，この多段階SOMの利用は有効であるものと考えられる．

以下，2章では基本的事項としてKohonenの自己組織化マップの原理とDCTについて，3章ではカット映像に対する記述の方法について，4章ではカット映像に与えられた記述情報をシーンに継承させる方法について述べ，5章で結論および問題点について述べる．

2. 基本的事項

2.1 自己組織化マップ（Self-Organizing Map: SOM）

ニューラルネットワークの一種である自己組織化マップ（SOM）[5)~7)]は教師なし競合強化学習モデルである．出力層の各セル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．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特徴を持っているため，特徴のよく似たデータどうしは出力マップ上の近い位置に配置されるようになっている．生成されたマップはそれぞれのデータの位置関係によって，類似しているデータかどうかが直観的に理解しやすいという点からシステムの視覚化に利用できる．SOMには様々な

☆ 本論文では，コンテンツ情報はビデオ映像を構成する画像フレーム系列情報のことを指し，これをJPEGやMPEG形式に変換して得られるものもこれに含むものとする．

種類があるが，ここでは最も基本的なものについて述べる．

SOM で用いられるネットワークは，セルを 2 次元に六角格子状に配置したものである．それぞれのセル i はセルの特徴ベクトル $\boldsymbol{m}_i(t) \in R^n$（$R$ は実数）を持っており（t は時間を表し，$\boldsymbol{m}_i(0)$ は適切な方法で初期化されている），これらのセルの特徴ベクトルを，入力である特徴ベクトル $\boldsymbol{x}_j \in R^n (j = 1, 2, \cdots, d)$ に選択的に近づけることによって学習は進行する．このとき，SOM では入力となる特徴ベクトルに一番近いパターンを持つ出力セルおよびその近傍のセルの集合のみが入力ベクトルに近づくことができるようなアルゴリズムをとる．

SOM のアルゴリズムを以下に示す．

(1) 各入力特徴ベクトルを生成し，その集合を X とする

$$X = \{\boldsymbol{x}_j \mid \boldsymbol{x}_j \in R^n, j = 1, 2, \cdots, d\}$$

(2) 出力層にある各ユニットの持つパターンを初期化する

$$M = \{\boldsymbol{m}_i \mid \boldsymbol{m}_i \in R^n, i = 1, 2, \cdots, k\}$$

（ただし，$\boldsymbol{m}_i(0) = [0, 0, \cdots, 0]$ とした）

(3) T をあらかじめ設定された学習回数とする．このとき，$t = 0, 1, \ldots, T$ について以下を繰り返す

(i) x_j に最も近いセル c を探す．つまり，$\|\boldsymbol{x}_j - \boldsymbol{m}_c(t)\|$ を最小にするセル c を求める

(ii) 探し出したセル c の特徴ベクトル $\boldsymbol{m}_c$ を更新し，さらにその近傍のセルの集合 N_c も入力パターンに近づける

$$\boldsymbol{m}_c(t+1) = \begin{cases} \boldsymbol{m}_c(t) + \alpha(t)[\boldsymbol{x}_j(t) - \boldsymbol{m}_c(t)] \\ \qquad (i \in N_c(t)) \\ \boldsymbol{m}_c(t) \quad (i \notin N_c(t)) \end{cases}$$

N_c の中央はセル c である．N_c の半径は，学習の初期段階ではたいてい大きく，学習を繰り返していくうちに単調に減少させる．また，$\alpha(t) \in (0,1)$ は「学習率」を表し，これもまた時間とともに単調に減少させる．

$$\alpha(t) = \alpha_0(t) exp\left(-\frac{\|\boldsymbol{r}_c - \boldsymbol{r}_i\|^2}{\sigma(t)^2}\right)$$

ただし，$\boldsymbol{r}_c$ と $\boldsymbol{r}_i$ はそれぞれセル c とセル i の持つベクトルを表す．$\alpha_0(t)$ や $\sigma(t)$ には単調減少の一次関数や指数関数がよく用いられる．

(iii) $j = 1, 2, \ldots, d$ について (i)(ii) を繰り返す

2.2 離散コサイン変換（Discrete Cosine Transform: DCT）

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[15),16)]．1 枚の自然画像を $N \times N$ 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ということである．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うことができる．

1 枚の画像から分割された画素ブロックの大きさが $N \times N$ 画素のとき，画素信号を $f(x,y)$，変換によって得られる係数（DCT 係数）を $F(u,v)$ とすると，$F(u,v)$ は次のように求まる．

$$F(u,v) = \frac{2}{N}\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} C(u)C(v) \cdot f(x,y) \cdot \cos\frac{(2x+1)u\pi}{2N}\cos\frac{(2y+1)v\pi}{2N}$$

ただし，

$$C(u), C(v) = \begin{cases} \frac{1}{\sqrt{2}} & \text{for } u, v = 0 \\ 1 & \text{otherwise} \end{cases}$$

このようにして得られた DCT 係数のうち $F(0,0)$ を DC（Direct Current，直流）係数といい，それ以外の DCT 係数を AC（Alternate Current，交流）係数と呼ぶ．DCT 係数は，画素数と同じ $N \times N$ 個求まり，低周波成分に集中する．DC 係数はブロック内の画素値の平均値を表し，AC 係数はその周波数の活性度を示す．また次式に示す逆変換でブロック画像の再生画素値が求まる．

$$f(x,y) = \frac{2}{N}\sum_{u=0}^{N-1}\sum_{v=0}^{N-1} C(u)C(v) \cdot F(u,v) \cdot \cos\frac{(2x+1)u\pi}{2N}\cos\frac{(2y+1)v\pi}{2N}$$

ただし，

$$C(u), C(v) = \begin{cases} \frac{1}{\sqrt{2}} & \text{for } u, v = 0 \\ 1 & \text{otherwise} \end{cases}$$

3. カット映像に対する記述方法

映像データに対して映像の意味情報を付加する場合，多かれ少なかれ，我々はすべてのカット映像に対してその内容の記述を施さなくてはならない．しかし，これらの作業は人手によるところが大きいため，ビデオデータが増えれば増えるほど，記述を施す手間がかかるという大きな問題がある．そこで我々は，ビデオデータをカット映像のDCT情報などのコンテンツ情報や，コンテンツ情報（DCT）と意味情報をともに用いたハイブリッド情報によりカット映像を分類したマップをあらかじめ生成しておき，それを用いることでそのような手間を省き，効率的に記述を行うことができるように工夫した．

以下に，その手順を示す．

3.1 カット映像への分割

DCTを利用した，映像データのカット映像分割プログラムは，有木らが開発したソフトウェアを利用している[17]．カット映像分割とは，映像のフレームのDCT値が大きく変化する部分，つまり，場面と場面の切り替わる時点で行われるため，カット映像は1つの意味の持った単位であると考えることができる．こうして，1つの映像データは，多くのカット映像に分割される．

3.2 DCT値を用いたフレーム特徴ベクトルの生成

映像を構成している各フレーム画像をいくつかのブロック（たいてい 16×16 または 8×8）に分割し，そのブロックを2次元DCTにより周波数成分に変換する．このうち，直流成分である DC および，水平，垂直周波数成分である交流成分のうち，値の大きなもの上位2つの AC_1，AC_2 の3つの成分を用いることとする（**図1** 参照）．これから1つのフレームに対するDCT成分を要素とするベクトル

$$(DC(1), AC_1(1), AC_2(1), \ldots, \ldots, DC(n), AC_1(n), AC_2(n))$$

を構成し，これをフレーム特徴ベクトルと呼ぶ．ここで $DC(k), AC_1(k), AC_2(k)$ $(k = 1, 2, \ldots, n)$ は分割された第 k 番目ブロックにおける DC 成分と AC 成分である．

3.3 DCTによるカット映像特徴ベクトルの生成

1つのカット映像として，フレーム画像系列 $f_1, f_2, \ldots, f_m$ $(m \geq 1)$ が得られたとする．ここで，それぞれの f_i $(i = 1, 2, \ldots, m)$ は映像のフレーム画像を表す．上記3.2節に従って得られたフレーム画像 f_i のフレーム特徴ベクトルを以下のように表すことにする．

$$(DC^i(1), AC_1^i(1), AC_2^i(1), \ldots, \ldots, DC^i(n), AC_1^i(n), AC_2^i(n))$$

ここで，あるカット映像 $F_l = \{f_1^l, f_2^l, \ldots, f_m^l\}$ のカット映像DCT特徴ベクトル $\boldsymbol{Vector}_{DCT}(F_l)$ を，そのカット映像を構成するすべてのフレーム画像のフレーム特徴ベクトルの平均の各成分に重みをつけたものとして定義する．すなわち，カット映像DCT特徴ベクトルは，

$$\begin{aligned} &\boldsymbol{Vector}_{DCT}(F_l) \\ &= \left(\frac{w_1}{m} \cdot \sum_{i=1}^{m} DC^i(1), \frac{w_2}{m} \cdot \sum_{i=1}^{m} AC_1^i(1), \right. \\ &\frac{w_3}{m} \cdot \sum_{i=1}^{m} AC_2^i(1), \ldots, \frac{w_1}{m} \cdot \sum_{i=1}^{m} DC^i(n), \\ &\left. \frac{w_2}{m} \cdot \sum_{i=1}^{m} AC_1^i(n), \frac{w_3}{m} \cdot \sum_{i=1}^{m} AC_2^i(n) \right) \end{aligned}$$

として表される．ここでいう重みとは DC 成分に対する w_1，AC_1 成分に対する w_2，AC_2 成分に対する w_3 のことである．これらの重みを用いた理由は，重みを考慮しなかった場合の映像の分類傾向が DC 成分の影響が強すぎて映像の色で分類されるのではなく，むしろ映像の色の明暗のみで分類されるという結果が得られたからである．また，実験を繰り返した結果，$w_1 < w_2 < w_3$ という関係が満たされていれば，おおむね良好な結果が得られることが判明している[18]．

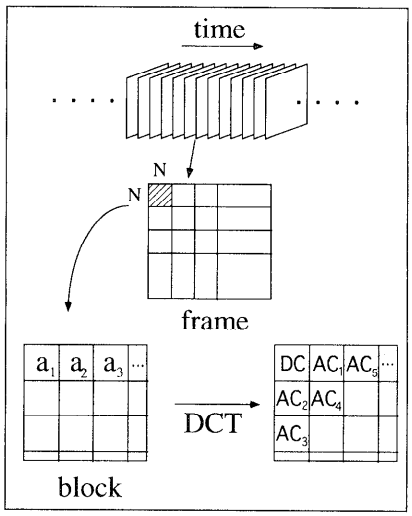


図1 DCT成分の抽出
Fig. 1 DCT value extraction from frame images.

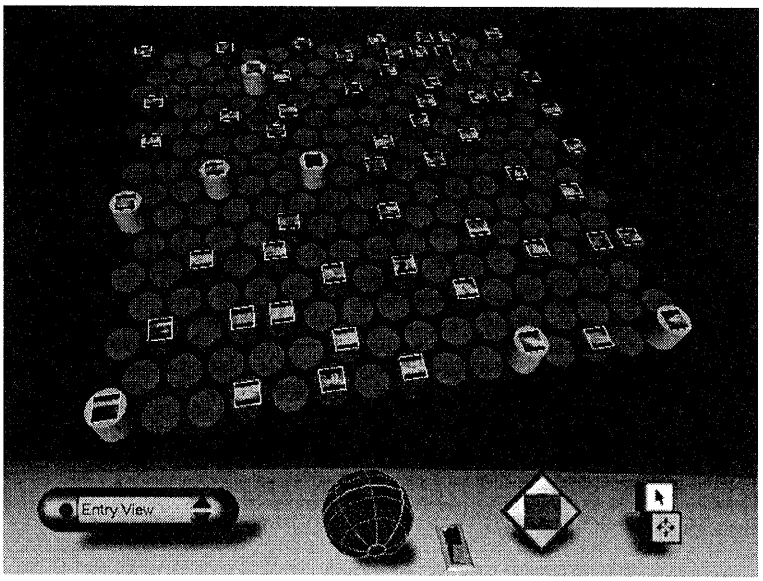


図2 カット映像群に対する3次元SOM
Fig. 2 Contents-based 3D-SOM for video shots.

3.4 カット映像群の3次元SOMの生成

カット映像DCT特徴ベクトルをSOMの入力として学習を行い，SOMマップの生成を行う．SOMの学習部分はKohonenの研究グループが開発したSOM_PAK[19]を用い，その出力結果をVRML形式に変換をすることでカット映像の分類およびブラウジングができるようになっている．これらのソフトウェアはすべてシリコングラフィックス社のIndy，Indigo2ワークステーション上で実装されている．図2はカット映像をカット映像DCT特徴ベクトルのみで分類した3次元SOM（これをコンテンツ情報3D-SOMと呼ぶ）により分類した例であり，VRMLブラウザにより表示・操作することができる．図において，円筒の高さは該当セルに写像されたカット映像の数を表しており，また各円筒上部には，そのセルの持つ特徴ベクトルの値に最も近いカット映像が表示されている．

3.5 SOM上でのカット映像群に対する内容記述

映像データに対してその内容の記述を施すという作業は，自動化が困難であるため記述者に非常に大きな手間を要するという問題が生じる．そこで，その手間を最小限におさえるために，我々は次のような2つの方式を提案する．

3.5.1 カット映像のDCT情報に基づくSOM上での記述

カット映像とは，場面の変わり目で切られる映像の1つの単位であるので，映像に映っているものに動きがあったとしても，映っている物体はカット映像内ではさほど変化しないと考えられる．つまり，コンテンツ情報（DCT）の類似度が高ければ，記述する情報も類似性があると考えることができるため，コンテンツ情報3D-SOMをあらかじめ生成しておき，そのマップ上において同じセルに分類されたカット映像には同じ記述を施すという方法を採用することにした．これによって，記述者がすべてのカット映像に対してその映像の意味内容を記述しなければならないという労力を省くことができると思われる．また，コンテンツ情報（DCT）では類似しているが，意味記述が異なるカット映像も存在するため，意味記述が異なるカット映像は手動で取り除くことにした☆．

内容記述はキーワードによって行うが，これをシーンへ継承する際に適切なキーワードとして，登場人物や場所などがあげられる．

3.5.2 ハイブリッド型SOM上での記述

我々が映像データの分類システムを構築する場合に考慮した点は，ユーザが映像に対する問合せを行う場合は，データのコンテンツ情報（DCT）を基にした分類よりも，その映像の意味内容を基にした分類が必要であるという点である．こうして，開発されたカット映像の分類システムがハイブリッド型SOMである[20]．

☆ ただし，扱うデータによっては取り除くべきカット映像の数も多くなると考えれるが，第1段階のSOMによるカット映像の分類の時点でクラスタリングの精度を上げれば対処できると考えられる．たとえば，DCT情報だけではなく他のコンテンツ情報（カラーヒストグラムなど）を組み合わせるなど．

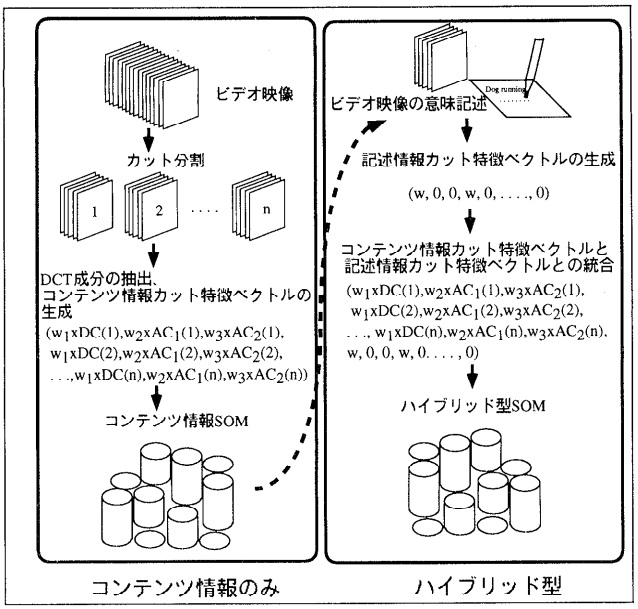


図3 ハイブリッド型SOMの生成過程
Fig. 3 Generation of hybrid-type SOM.

3.3節の式より得られた各カット映像 F_l のカット映像DCT特徴ベクトルを

$$\boldsymbol{Vector}_{DCT}(F_l) = (c_1^l,\ c_2^l,\ \ldots,\ c_k^l,\ \ldots,\ c_n^l)$$

(ただし, $c_k^l = (DC^l(k), AC_1^l(k), AC_2^l(k))$, $(k = 1, 2, \ldots, n)$) とする. 3.5.1項で述べた方法により与えられたカット映像の意味記述情報から生成された, カット映像キーワード特徴ベクトルを

$$\boldsymbol{Vector}_{keyword}(F_l) = (k_1^l, k_2^l, \ldots, k_j^l, \ldots, k_h^l)$$

(ただし, すべてのカット映像への記述の際に用いたキーワード集合を $\{K_1, K_2, \ldots, K_j, \ldots, K_h\}(h \geq 1)$ とし, k_j^l はカット映像 F_l に対しキーワード K_j が与えられているとき1をとり, そうでないときは0をとる) とする. これらの2つのベクトルを結合したハイブリッド型カット特徴ベクトル

$$\boldsymbol{Vector}_{hybrid}(F_l) = (c_1^l,\ \ldots,\ c_n^l,\ w \times k_1^l,\ \ldots,\ w \times k_h^l)$$

を生成 (ただし, w は正規化のために用いられる重み) し, これをSOMに学習させることで生成されるマップがハイブリッド型SOMである (生成手順は**図3**を参照). このマップは, カット映像のコンテンツ情報 (DCT) と人手により記述されたキーワードによる意味情報を用いてカット映像を分類しており, 分類精度はコンテンツ情報 (DCT) のみの分類システムと比較しても, 比較的良い結果が得られている[21].

このハイブリッド型SOM自身もまた多段階SOMと位置づけられ, コンテンツ (DCT) 情報によるカット映像の分類を第1段階に, コンテンツ情報 (DCT) と意味情報によるカット映像の分類を第2段階にSOMを利用している. こうしてできたハイブリッド型SOMを用いて, 3.5.1項に述べたような方法で, カット映像に対して記述を行えば, 計3段階のSOMの学習を行うことになる.

4. シーンへの記述および類似シーン検索

シーンの場合はカット映像とは異なり, 登場人物や場面のストーリーなど含まれている情報量が多いため, 記述する際には何らかの工夫を施す必要があると考えられる.

そこで, 3.5節に示したように, カット映像を分類したマップをあらかじめ生成しておき, カット映像に対して与えられたキーワード情報を利用することにする. 具体的には, シーンに含まれているカット映像を求め, そのカット映像に対して与えられている記述情報をそのままシーン映像の記述として継承する方法である. このとき, シーンの長さに対するカット映像の長さを考慮した重み付けを行う.

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

シーンはカット映像とは異なり, 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.

したがって，カット映像のようにコンテンツ情報の変化から自動的にシーンを検出するということは困難である．よって，本研究ではシーン分割は手動で行うことにした．

また分割され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，記述を行うことは困難であると考えられる．これに対して，カット映像に描写されている対象物や動作はシーンの場合と比較すると比較的少ないため，シーンに対して直接記述を行う代わりにカット映像に対して記述を行うことにした．そして，そのカット映像に与えられた記述情報をそのカット映像が属しているシーンに継承させる．

シーンの特徴ベクトルを生成する際の手順を以下に示す．

(1) **カット映像の3D-SOM生成**（3.4節参照）

(2) **カット映像へのキーワード付与による内容記述**（3.5節参照）

(3) **シーンに含まれるカット映像を求める**
話の流れや場面展開，登場人物の変化などを考慮し手作業でシーンの分割を行う．また，のちにカット映像の記述情報をシーンに継承させるために，こうして分割された各シーンの中に含まれている複数のカット映像もあらかじめ求めておく．

(4) **カット映像の記述情報をシーンに継承させる**
各カット映像に付与された映像の意味記述情報を基に，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．

$$\boldsymbol{Vector}_{scene}(S) = \sum_{F_l \in S} \frac{n(F_l)}{N} \boldsymbol{Vector}_{keyword}(F_l)$$

ただし，シーン $S = \{F_1, F_2, \ldots, F_l, \ldots, F_r\}$ $(r \geq 1)$，$\boldsymbol{Vector}_{scene}(S)$ はシーン特徴ベクトル，$\boldsymbol{Vector}_{keyword}(F_l)$ はカット映像キーワード特徴ベクトル，N はそのシーンの全フレーム数，$n(F_l)$ はカット映像 F_l に含まれるフレーム数とする．

ここでは，シーンのフレーム数に対するカット映像のフレーム数の割合を算出し，その値をカット映像キーワード特徴ベクトルへの重みとして加えている．なぜなら，各カット映像の時間的な長さが互いに異なることから，カット映像の持つ記述情報をシーンに継承させる際に，キーワード情報に対してこの映像の長さに比例した重みを与え，正規化処理を行うためである．このように，シーンに含まれている各カット映像に対応する重み付け済みの各カット映像キーワード特徴ベクトルの和をとることでそのシーンの特徴ベクトルの生成を行っている．

このようにして得られたシーン特徴ベクトルを入力ベクトルとして再度SOMに学習させ，この結果をマップ表示させる．ここでは，カット映像のコンテンツ情報（DCT）によるSOMの学習を第1段階，シーンの記述情報によるSOMの学習を第2段階の計2段階で使用している．

4.2 実験・評価

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，今回，実験において用いた映像データは，人間に助けられた動物が恩返しをするというアニメーション映像（約30分）であったが，その映像のカット映像の総数は250個，さらにシーン分割を手作業で行った結果，検出されたシーンの総数は22個であった．また，1シーンあたりの平均カット数は11.36であり，意味記述が異なるという理由でノイズとされたカット映像は6個であった．セル数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4**，**図5**に示す．このマップ生成においては，我々がこれまでに行ったテキスト文書分類・検索システム[10]で用いたマップ生成プログラムを使用した．

図4上で“man”という領域内の高さを持ったあるセルを選択すれば，そのセル内に分類されているシーンがブラウザ上に現れるようになっている．また，“man”自身を選択すれば図5のように，1つの領域がさらに細かい複数の領域に分けられた詳細図が現れる（3D-SOMの詳細化機能）．

この映像から得られたシーンの分類状況を見てみると，たとえば，人および動物が一緒に登場しているシーンが多かったが，そういったシーンはマップ上で互いに近くに集まっている様子が目についた．

そこで，本システムがどれほどの分類精度を持っているかを定量的に判断するために，「精度（適合率）」と「再現率」を測定することで判断する．世に知られている「精度」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 $\frac{B}{B+C}$，再現率は $\frac{B}{A+B}$ 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

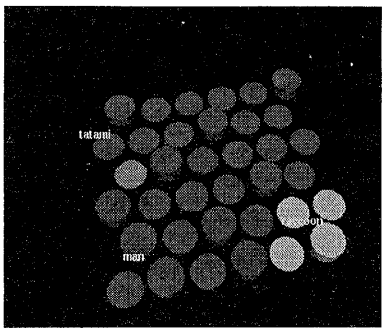


図4 ビデオシーン3D-SOM
Fig. 4 A 3D-SOM for video scene.

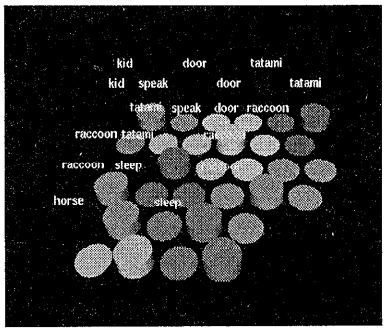


図5 3D-SOMの詳細化機能
Fig. 5 A hierarchical 3D-SOM and zoom-in operation.

現率」というのは次のように定義できる．

すべてのカット映像の中から，マップ上の円筒の上に張りつけられたあるカット映像 v に似ているカット映像の数を $similar(v)$ とし，カット映像 v が張りつけられているセル付近にあるカット映像の数を $neighbour(v)$ で表すとする．このもとで精度は，

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

で表され，また再現率は，

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

で表されたものである．

表1は，$neighbour(v)$ に含まれる範囲が中心の円筒から近傍距離1の場合と近傍距離2の場合の精度と再現率である．ただし，ここでは w_1，w_2，w_3 の値はそれぞれ1，10，20を用いている[18]．またこの時，22個あるシーンは8つの類似シーンとして分類されている．

4.3 考　　察

表1を見ると精度，再現率ともに40～80%前後と分類結果としては十分満足のいく結果であると思われる．しかし，カット映像およびシーンの類似判定は，

表1 「精度」と「再現率」
Table 1 Precision ratio and recall ratio.

Distance	Precision Ratio (%)	Recall Ratio (%)
1	85.00	37.50
2	66.50	62.50

分類を行う人物により異なり，そのために評価も異なるという問題点を考える必要があるため，この実験結果から一概に本システムが有効であるということはいえない．また，システムの詳細部分である特徴ベクトル生成の際の重み付けや複数のコンテンツ情報を組み合わせるなどのチューニングができれば，本システムのパフォーマンスが上昇すると考えられるため，さらにシステムの実験・評価を行う必要がある．

映像データに対する意味記述にあたっては，今回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，状態を表す単語を付与することにした．この分類結果を見る限りでは，我々が行った意味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（視覚的にも，また実験評価表1を見ても）分かった．しかし，シーンの総数が22個と極端に少ないことや，記述者の主観が多少影響していることなどから，なかにはそれほど類似性の高くないデータが分類されている例も見受けられた．

今回，シーン分割は話の流れや場面の展開をもとに我々の判断によって行ったため，各シーンを構成するカット数は当然ながらまちまちになった．シーンの意味記述をどの程度のカット数から構成するかという問題は非常に重要であるが，もしシーンを構成するカット数に制限を設けてしまえば意味的なつながりを持つにもかかわらず異なるシーンに分類されてしまうカット映像が生じる可能性も考えられるため，今後は，このような不都合をいかにして解消していくのか検討していく必要があると思われる．

また，今回シーンの特徴ベクトルを生成する際に，重み付けされたカット映像キーワード特徴ベクトルの和を単純に求めただけであるため，今後何らかの改良について考えていかねばならない．

5. おわりに

我々は，Kohonenの自己組織化マップ（SOM）を用いたカット映像のコンテンツ情報（DCT）またはハイブリッド情報による分類システムをシーンの意味記述支援ツールとして応用する手法を提案した．また，カット映像に与えられた記述情報を継承させたシーン特徴ベクトルを生成し，これを再度SOMに学習させるという2段階SOMを用いたシーンの分類システム

の提案およびそのプロトタイプシステムの試作を行い，これを用いたシーン分類の実験，評価を行った．

本論文のまとめを以下にあげる．

- **シーンの意味記述支援ツールとしてカット映像の分類結果を利用**
  すべてのカット映像に対して，人が手作業で記述する手間を軽減するためにKohonenの自己組織化マップ（SOM）を，カット映像のコンテンツ情報またはハイブリッド情報を基に生成したマップをシーンの意味記述支援ツールとして用いた．ここでは，生成したマップにおいて同じセルに分類されたカット映像に対しては同じ記述を与えることにした
- **カットに対する記述情報のシーンへの継承**
  一般にシーンには複数の人物，物，出来事等が描写されているため，これらの情報を一度に把握，記述することはかなり困難な作業となり，ユーザへの負担も大きくなる．そこで意味記述はカット映像単位で行い，その記述情報をそれらカット映像により構成されているシーンに継承させることで，シーンへの意味記述を行った
- **VRMLによるSOMの実装**
  VRMLによりシステムを実装しているので，Webツールとして利用できるという汎用性の確保を行った
- **実験・評価および考察**
  今回試作したシステムによって，実際にカット映像への内容記述とそれを利用したシーン映像に対する内容記述およびその記述情報を基にしたシーン分類を行い，考察を行った

今後の課題としては以下の問題があげられる．

- 与えられたキーワード間の相関関係，共起関係を考慮した，カット映像キーワード特徴ベクトルの生成
- 手動で行っているシーン分割の自動化
- 記述をより効率的に行うためのインタフェースの開発や記述方法の提案
- 1つのカット映像に対して複数の人間が記述した場合，その記述情報を1つに統合する方法の提案

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，東映株式会社に感謝いたします．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) Elmagarmid, A., Jiang, H., Helal, A., Joshi, A. and Ahmed, M.: *VIDEO DATABASE SYSTEMS: Issues, Products and Applications*, Kluwer Academic Publishers (1997).
2) Kim, Y. and Shibata, M.: Content-Based Video Indexing and Retrieval – A Natural Language Approach, *IEICE Trans. Inf. and Syst.*, Vol.E79-D, No.6, pp.695–704 (1996).
3) Uehara, K., Oe, M. and Maehara, K.: Knowledge Representation, Concept Acquisition and Retrieval of Video Data, *International Symposium on Cooperative Database Systems for Advanced Applications* (*CODAS'96*), pp.527–534 (1996).
4) Zettsu, K., Uehara, K. and Tanaka, K.: A time-stamped authoring graph for video databases, *Proc. 8th International Conference on Database and Expert Systems Applications* (*DEXA'97*), pp.192–201 (1997).
5) Kohonen, T.: Self-Organized formation fo topologically correct feature maps, *Biological Cybernetics*, No.43, pp.59–69 (1982).
6) Kohonen, T.: The Self-Organizing Map, *Proc. IEEE*, Vol.78, No.9, pp.1464–1480 (1990).
7) Kohonen, T.: *Self-Organizing Maps*, Springer, Berlin (1995).
8) 仁木和久，田中克己：ニューラルネットワーク技術の情報検索への応用，人工知能学会誌，Vol.10, No.1, pp.1–7 (1995).
9) Qing, Q., Shi, X. and Tanaka, K.: Document Browsing and Retrieving based on 3D Self-Organizing Map, *Proc. Workshop on New Paradigms in Information Visualization and Manipulation in Conjunction with CIKM'95* (1995).
10) Hatano, K., Qian, Q. and Tanaka, K.: A SOM-Based Information Organizer for Text and Video Data, *Proc. 5th International Conference on Database Systems for Advanced Applications* (*DASFAA'97*), pp.205–214 (1997).
11) Ritter, H. and Kohonen, T.: Self-organizing semantic maps, *Biological Cybernetics*, No.61, pp.241–254 (1989).
12) Ritter, H. and Kohonen, T.: Learning 'semantotopic maps' from context, *Proc. Int. Joint Conference on Newral Networks* (*IJCNN'90*), Vol.I, pp.23–26 (1990).
13) Honkela, T., Kaski, S., Lagus, K. and Kohonen, T.: WEBSOM – Self-Organizing Maps of Document Collections, *Proc. Workshop on Self-Organizing Maps* (*WSOM'97*) (1997).
14) Lagus, K., Honkela, T., Kaski, S. and

Kohonen, T.: Self-organizing Maps of document collections: A new approach to interactive exploration, *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining* (*KDD'96*), pp.238–243 (1996).
15) 加藤茂夫：画像データ圧縮の基礎知識，インタフェース，No.175, pp.132–159 (1991).
16) 藤原　洋：最新MPEG教科書，アスキー (1994).
17) 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).
18) 波多野賢治，田中克己：映像データベースの動的クラスタリングと素材検索機構について，情報処理学会データベースシステム研究会，96-DBS-109, pp.105–110 (1996).
19) Kohonen, T., Hynninen, J., Kangas, J. and Laaksonen, J.: SOM_PAK: The self-organizing map program package, Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science (1996).
20) 波多野賢治，亀井俊之，田中克己：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，情報処理学会データベースシステム研究会，97-DBS-113, pp.173–178 (1997).
21) Hatano, K., Kamei, T. and Tanaka, K.: Clustering and Authoring of Video Shots Using Hybrid-type Self-Organizing Maps, *Proc. International Symposium on Digital Media Information Base* (*DMIB'97*), pp.150–158 (1997).

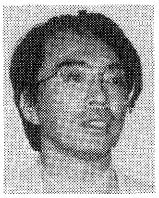
(平成 9 年 9 月 1 日受付)
(平成 10 年 2 月 2 日採録)

波多野賢治（学生会員）

1971 年生．1995 年神戸大学工学部計測工学科卒業．1997 年同大学大学院自然科学研究科情報知能工学専攻修了．同年同大学大学院自然科学研究科情報メディア科学専攻入学，現在に至る．マルチメディアデータベースの研究に従事．

亀井　俊之（学生会員）

1974 年生．1997 年神戸大学工学部情報知能工学科卒業．同年同大学大学院自然科学研究科情報知能工学専攻入学，現在に至る．マルチメディアデータベースの研究に従事．



田中　克己（正会員）

1951 年生．1974 年京都大学工学部情報工学科卒業．1976 年同大学大学院修士課程修了．1979 年神戸大学教養部助手．1986 年同大学工学部助教授．1994 年同大学工学部教授（情報知能工学科）．1995 年同大学大学院自然科学研究科（知能科学専攻）専任教授，現在に至る．工学博士．主にデータベースの研究に従事．現在本会データベースシステム研究会主査．96 年度より通信・放送機構「次世代デジタル映像通信の研究開発」の研究総括責任者，文部省科研費重点領域研究「分散発展型データベースシステム技術の研究」の研究代表者．神戸マルチメディアインターネット協議会会長．人工知能学会，ACM 等各会員．