ビデオ・データベースにおけるシーン粒度の多様性を考慮した類似検索手法

A Method for Similarity-Based Retrieval of Scenes in Diverse Granularity on Video Database

牛尼 剛聡[**]　小坂 健[*†]　渡邉 豊英[*†]
USHIAMA Taketoshi[**]　KOSAKA Ken[*†]　WATANABE Toyohide[*†]

あらまし　本論文では，ビデオ・データベースにおけるシーンを対象とした類似検索手法を提案する．現在，類似検索のためには，検索対象を単一の特徴ベクトルによって表現し，ベクトル間の距離によって類似度を定義する手法が広く利用されている．シーンとはビデオ中で意味的なまとまりがあるフレーム区間である．ビデオ中には，利用者の興味に基づいて，多様な視点から，様々な粒度のシーンを考えることができる．したがって，検索対象となるシーンをあらかじめ個別化することは困難であるため，単一の特徴ベクトルを利用する一般的な類似検索手法をそのまま適用することが不適切である．本手法では，利用者によって例示されたシーンの粒度に基づいて，指定されたシーンの特徴となる特性を決定する．そして，限定された特性に関してのみマッチングを行うことにより，利用者の意図を反映した効率的なシーンの類似検索を実現する．

キーワード　ビデオ・データベース，類似検索，シーン検索，意味内容記述，内容検索

1　はじめに

ビデオは対象世界の動的な活動を人間にとって直感的に分かりやすく表現可能であるため，報道，娯楽，教育等様々な分野で対象世界の動的な側面の記録・伝達に利用されている．近年，計算機の処理速度の向上，および 2 次記憶媒体の大容量化と低価格化に伴い，ビデオをコンピュータ上で簡単に利用可能となった．こうした中で，大量のビデオを効率的に管理し，効果的に利用するためのビデオ・データベースに関する研究が活発に行われている．ビデオはフレームの時系列であり，その部分系列を考えることができるが，意味的なまとまりがある部分系列はシーンと呼ばれる．ビデオ・データを利用する際には特定のシーンのみを必要とする場合があるため，ビデオ・データベースから特定のシーンを検索する機構が重要である．

これまでに，シーン検索に関して多くの研究がなされている．従来の一般的なシーン検索手法では，利用者がキーボードから文字列を入力したり，GUI 上のメニューを選択することにより検索条件を指定する必要があった．しかし，一般の利用者が検索条件を正確に記述することは困難である．また，検索条件の指定に熟練した利用者であっても，検索条件を正確に指定するためには多大な労力が必要である．さらに，利用者がビデオ・データベース中に含まれているデータに関して正確な知識を持たない場合には，考えられる検索条件の数が膨大になり，利用者は要求する検索結果が得られるまで，様々な検索条件で検索を繰り返す必要がある．また，利用者が予想しなかったデータがデータベース中に含まれており，それが利用者の検索要求と一致する場合も考えられる．

今後，インターネットやケーブルテレビ回線を通じて一般の利用者がビデオ・データベースにアクセスする環境が整備されていくと予想される．こうした環境では，より簡単な操作でシーンを検索するための機構が必要である．こうした背景の下，我々はデータベース

[**] 九州芸術工科大学芸術工学部 Facility of Design, Kyushu Institute of Design
[*†] 名古屋大学大学院工学研究科 Graduate School of Engineering, Nagoya University

をブラウジングして利用者の検索要求に類似したシーンを指定し, 指定されたシーンを例として類似したシーンを検索するシステムの開発を行っている．我々はこれまでに概略表現の生成に関する手法 [1] やハイパービデオを用いたブラウジング手法 [2] を提案している．これらの手法を用いることにより，利用者は簡単な操作で内容を確認し，例となるシーンを発見できる．本稿では，利用者が例となるシーンが発見したことを前提として，利用者が例として指定したシーンに類似するシーンを検索する手法を提案する．

本手法は，人手によって内容記述がなされたビデオを対象とする．また，データベースに蓄えるビデオは同種の対象を記録したものとする．本論文では，説明のために野球の試合を記録したビデオを例題として利用する．

2 背景

2.1 従来の類似検索手法とシーン検索

静止画像データ，音データ，テキスト・データ等を対象とした類似検索はこれまでにも活発に研究がなされてきた．類似検索の代表的な手法は，距離アプローチと変換アプローチに大別できる [3]．現在，データベースにおける類似検索においては，距離アプローチが採用されることが多い．距離アプローチでは特徴ベクトルを利用する手法が一般的である．この手法では，検索対象となるオブジェクトを特徴ベクトルとして表現する．特徴ベクトルはオブジェクトの特徴を表す多次元ベクトルである．例えば，静止画像データに対する類似検索のために利用される特徴量としては色ヒストグラム，テクスチャ特徴量，形状特徴量などがある [4].

本論文では，ビデオのシーンを対象とした類似検索について考える．ビデオはフレームの時系列であり，シーンとはビデオ中で意味的なまとまりをもつフレーム区間である．ビデオ中にはさまざまな視点からシーンを捉えることができるため，一つのビデオの中にも多種多様なシーンが存在する．一般に，それらの中には，再生時間が数秒の短いシーンから，数分に渡る長いシーンまで，様々な粒度のシーンが含まれる．例えば，野球の試合においては，一つの打席を一つの意味的なまとまりとして捉える場合もあれば，複数のイニングを一つの意味的なまとまりとして捉える場合もある．

従来の特徴ベクトルに基づく一般的な検索手法をビデオのシーン検索に適用する場合には，検索対象となるすべてのシーンをあらかじめ個別化し，それぞれに対応する特徴ベクトルを設定しなければならない．ビデオにおいて，意味的なまとまりがあるフレーム区間は，利用者の視点によって無数に存在するため，シーンをあらかじめ個別化することは非常に困難である．

上記の問題点を解決するため，シーンを単一の特徴ベクトルではなく特徴ベクトルの系列として表現するアプローチが考えられる．このアプローチでは, フレームにおける特徴を一つのベクトルとして表現し，ベクトル系列のマッチングを行う．ベクトル系列のマッチングは，要素となる単一のベクトルのマッチングよりも処理コストが大きい．例えば，長さ n のベクトル系列に対して，長さ m の部分系列を検索するために必要なマッチング処理の一般的なオーダーは $O(nm)$ である．

ビデオの内容をベクトル系列として表現した場合，空間的な広がりと対象を捉える概念の多様性はベクトルの次元数に反映され，時間的な広がりはマッチングの対象となる部分系列の長さ（ベクトルの数）に反映される．

利用者がシーンを捉える視点の多様性を反映した検索を実現するためには，ベクトルの次元数を多くする必要がある．これは，ベクトルの 1 要素が対象を捉える 1 つの視点に対応しているからである．しかし，単純に様々な視点を考慮し，ベクトルの次元数を増加させると，マッチングの際の処理コストが増大するという問題が発生する．また，類似度を決定する際に，利用者が意図しない視点が多く含まれるため，それらがノイズとなり，利用者の意図を反映した検索結果が困難となる．

2.2 本手法のアプローチ

本手法では上記の問題を解決するために，データの適切な抽象化に基づくアプローチを用いる．具体的には，データの抽象化により内容的な特徴を失うことがないよう以下を実現する．

1. マッチングの対象となる特徴ベクトルの次元数を削減する．
2. マッチングの対象とする特徴ベクトル系列の系列

長を短くする．

フレーム毎に対応づけられているベクトルは時間の経過に従って変化するが，全ての特徴がフレームごとに変化するわけではない．そこで，本手法では，特徴が不変な区間を一つのベクトルとして表現する．これにより，マッチングする系列長を短縮できる．マッチングする系列長の短縮は，マッチング時の処理時間の短縮につながる．

さらに，指定したシーンの粒度に応じて，類似度の計算に利用する特徴を限定する．ビデオの内容の特性を表す概念図を図 1 に示す．本手法では，ビデオの内容を特徴の集合と考える．それらの特徴は様々な抽象度に位置付けることができる．抽象度が高い特徴は，抽象度が低い特徴を解釈することによって導出される．すなわち，抽象度が高い特徴は，抽象度が低い特徴に依存している．そして，特徴の変化の頻度に注目すると，一般的に抽象度が高い特徴は抽象度が低い特徴に比べて変化の頻度が少ない傾向にある．

こうした傾向に基づいて, 利用者が粒度の大きいシーンを捉えるときは抽象度が高い（変化の頻度が少ない）特徴を中心に内容を捉え，粒度の小さいシーンを捉えるときは抽象度が低い（変化の頻度が多い）特徴を中心に内容を捉えていると仮定する．例えば，一つの試合をシーンとして指定した場合には，ストライク・カウントの変化など頻繁に変化する特徴量はあまり重要でないが，一つの打席を考えた場合には，ストライク・カウントの変化は非常に重要である．本手法では，この仮定に基づいて対象とする特徴を限定する．特徴を限定することにより，特徴ベクトルの次元数を減少できる．特徴ベクトルの次元数の減少はマッチング時の処理時間の短縮につながる．また，特徴ベクトルの次元数が減少すると，特徴が不変な区間が長くなるため，マッチングする系列長も短縮できることも，マッチング時の処理時間の短縮につながる．

3　ビデオ・データベース

対象とするビデオ・データベースを定義するために，ビデオをモデル化するために利用する概念とデータ構造を導入する．

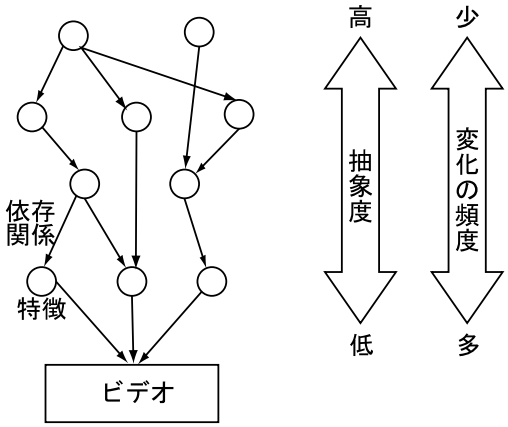

図 1: 特徴の抽象度と変化の頻度

■ビデオ　ビデオはフレームの系列である．$n+1$ フレームから構成されるビデオ v の i 番目 $(0 \leq i \leq n)$ のフレームを $\mathrm{frame}(v,i)$ と表記する．また，フレームの前後関係 $\preceq$ を $\mathrm{frame}(v,i) \preceq \mathrm{frame}(v,j)$ $(0 \leq i \leq j \leq n)$ として定義する．

■実体　ビデオの特徴を実体と呼び，ビデオの内容を実体の集合としてモデル化する．実体とは対象世界上で区別可能な物理的または概念的な「モノ」である．物理的な実体の例としては,「選手」,「ボール」などがあげられる．概念的な実体の例としては,「イニング」,「ストライク・カウント」などのような単純な概念や,「得点差」,「逆転チャンス」,「連打」などのように対象の場面を判断するための複雑な概念もある．実体は文字列で表すこととする．

■状態　実体は状態を有する．状態とは, 任意のフレームにおける実体の様相である．実体 e のフレーム f における状態を 2 項 (e,f) として表記する．実体 e の状態に割り当てることが可能な値の集合を, 実体 e のドメインと呼び D_e と表記する．このとき, 状態 $st=(e,f)$ を値を割り当てる関数を $\phi_{\mathrm{state}}(st) \in D_e$ として定義する．例えば，$\phi_{\mathrm{state}}((\text{イニング},\mathrm{frame}(v_1,50000)) = 3$ は，ビデオ v_1 の 50000 フレームにおけるイニングが 3 であることを表している．

なお，同一のドメインの任意の 2 要素 val_1, val_2 間には距離 $\mathrm{dist}_{\mathrm{state}}(val_1,val_2)$ が定義されているものとする．

実体の状態は，他の実体の状態に依存する場合がある．例えば,「得点差」を表す実体の状態は，それぞれ

のチームの得点を表す実体の状態から導出可能である．実体の状態を自動的に導出することにより，内容記述を簡素化できるが，本論文では自動的に導出する方法についての詳細は省略する．他の実体から導出された実体が表す概念は，導出元とした実体要理も抽象度が高い概念であると考えられる．

■ビデオ・データベース　ビデオ・データベースを $(V, E, D, \phi_{\text{state}})$ として定義する．ここで，V はビデオの集合，E は実体の集合，D はドメインの集合，ϕ_{state} は状態に値を割り当てる関数である．

4　類似シーン検索

前節で定義したビデオ・データベースにおいて，類似シーン検索を実現するためにいくつかの概念を導入し，それらを利用した検索処理について説明する..

4.1　検索に利用する概念

■状況　特定のフレームにおいて対象とする状態の組を状況と呼ぶ．状況を，対象とする実体のリスト $t = \langle e_1, \cdots, e_n \rangle$ とフレーム f の 2 項 (t, f) として定義する．状況 (t, f) の値はベクトル $(val_1, \cdots, val_n)$ である．ここで，$val_i = \phi_{\text{state}}(e_i, f)$ である．状況 si の値を返す関数を $\phi_{\text{situation}}(si)$ と表記する．例えば，$\phi_{\text{situation}}((\langle \text{先攻チーム得点}, \text{イニング}, \text{アウトカウント} \rangle, \text{frame}(v_1, 50000))) = (1, 3, 2)$ は，ビデオ v_1 の 50000 フレームにおいて，先攻チームの得点が 1 であり，イニングが 3 であり，アウトカウントが 2 であることを表している．

同一の実体リストを対象とする 2 つの状況 si_1, si_2 の距離を $\text{dist}_{\text{situation}}(si_1, si_2) = \sqrt{\sum_{i=1}^{n} d_i^2}$ とする．ここで，$d_i = \text{dist}_{\text{state}}(val_{1,i}, val_{2,i})$ である．

■状況区間　ある状況に関して，その値が不変であるようなフレーム区間を状況区間と呼ぶ．状況区間は対象とする実体リスト t と 2 フレーム $f_{\text{head}}, f_{\text{tail}}$ から構成される 3 項 $(t, f_{\text{head}}, f_{\text{tail}})$ として定義する．なお，ここで $f_1 \preceq f_2$ に対して，条件

$$f_{\text{head}} \preceq f_1 \preceq f_2 \preceq f_{\text{tail}}$$

ならば

$$\phi_{\text{situation}}((t, f_1)) = \phi_{\text{situation}}((t, f_2))$$

を満足する．

状況区間を構成する状況は不変であるため，状況区間の値を単一のベクトルとして表現可能である．状況区間 $in = (t, f_{\text{head}}, f_{\text{tail}})$ の値を，$\phi_{\text{interval}}(in) = \phi_{\text{situation}}((t, f))$ とする．ここで，$f_{\text{head}} \preceq f \preceq f_{\text{tail}}$ を満足するものとする．

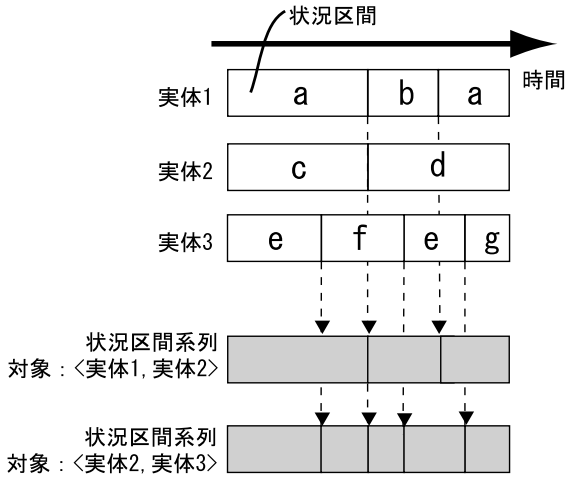

図 2: 状況区間系列の例

同一の実体リストを対象とする 2 つの状況区間の距離は，状況の距離を用いて，$\text{dist}_{\text{interval}}(in_1, in_2) = \text{dist}_{\text{situation}}(si_1, si_2)$ と定義する．ここで，si_i は状況区間 in_i に含まれる任意の状況である．

■状況区間系列　ビデオ中のシーンをフレーム系列の先頭フレームと終了フレームの 2 項 $(f_{\text{head}}, f_{\text{tail}})$ と表記する．シーンは，対象とする実体を指定することにより，状況区間の系列に一意に分解することができる．この系列を状況区間系列と呼ぶ．シーン $sc = (f_{\text{head}}, f_{\text{tail}})$ を，実体リスト t を用いて状況区間系列に分解する演算を $\text{div}(sc, t)$ と表記し，$\text{div}(sc, t) = \langle in_0, \cdots, in_n \rangle$ と定義する．ここで，$in_i (0 \leq i \leq n)$ は状況区間であり，以下の 3 条件を全て満足する．

【条件 1】　$in_i = (t, f_i, f_{i+1})$

【条件 2】　$f_i \preceq f_{i+1}$

【条件 3】　$f_0 = f_{\text{head}}$かつ $f_{n+1} = f_{\text{tail}}$

状況区間系列の例を図 2 に示す．

■シーン間の距離　いま，2 つのシーン sc_1, sc_2 に対して，実体リスト t に着目した場合の距離を $\text{dist}_{\text{scene}}(\text{sc}_1, \text{sc}_2, \text{t}) = \left(\sum_{i=1}^{n} \text{dist}_{\text{interval}}(\text{in}_{1,\text{i}}, \text{in}_{2,\text{i}})\right) / n$ として定義する．ここで，$in_{j,i}$ は状況区間系列 $\text{div}(sc_j, t)$ の i 番目

の要素（状況区間）を表す．なお，シーン間の距離が小さいほど，シーンは類似している．

4.2 検索アルゴリズム

利用者が例としてシーン sc を指定した場合の類似シーン検索を行う手続きの概要を以下に示す．

〔ステップ 1〕 利用者が指定したシーン sc に基づいて，対象とする実体リスト t を決定する．これについての具体的な方法は，次節で述べる．

〔ステップ 2〕 シーン sc に対して，実体リスト t に関する区間状況系列 $\mathrm{div}(sc,t)$ を作成する．ここで生成された状況区間系列の長さを n とする．

〔ステップ 3〕 データベース中に含まれるすべてのビデオ v_i に対して，実体リスト t に関する区間状況系列 $\mathrm{div}(v_i,t)$ を作成し，$\mathrm{div}(v_i,t)$ に含まれる長さが n である全ての部分系列と $\mathrm{div}(sc,t)$ の距離を計算する．

ここでは，比較する 2 つの状況区間系列の長さは同一であるとしている．系列の類似検索を行う際には，異なる長さの系列を時間的にロバストなマッチングを実現するために，系列データを非線形的に伸縮して対応づける DP マッチングを利用することが多い [5]. 本手法における DP マッチングの適用は今後の課題である．

4.3 対象とする実体の決定

利用者が指定した例に基づいて，対象とする実体リストを自動的に選択することについて考える．本手法では，シーンの粒度と平均周波数を利用する．平均周波数は, ある単一の実体に関する状況区間の要素がもつ平均フレーム数である．利用者が指定した例において重要な実体を決定するために, 以下のヒューリスティックを利用する．

- 平均周波数が少ない状況が変化した場合，利用者はその状況に注目している
- 平均周期（平均周波数の逆数）が例示されたシーンの長さよりも大幅に小さい実体は重要度が低い．

これらのヒューリスティックに基づいて，実体に対して重み付ける．そして，重みに基づいて対象とする実体リストを自動的に決定する．

5 関連研究

長坂ら [6] はビデオを構成するフレームの色平均を特徴量とし，特徴量系列に対してランレングス圧縮を行い，ラン単位の照合を行うことにより高速のシーン照合を実現する手法を提案している．また，片岡ら [7] は MPEG 形式における GOP 単位で特徴ベクトル列の平均値を用いた手法を提案している．この手法は特徴の不変区間に着目し，マッチングの対象となる要素を減少させる点は本論文で提案する手法と類似している．しかし，これらの研究では照合するシーンの粒度に関しては考慮していない点が本手法と異なる．

6 おわりに

本論文では，ビデオ・データベースにおけるシーン粒度の多様性に対処可能な，意味内容に基づく類似検索手法を提案した．

今後の評価実験により，本手法の有効性を評価する予定である．

参考文献

[1] 牛尼剛聡, 渡邉豊英: “ビデオ・データベース・ブラウジングのための意味内容記述に基づいたストーリーボード生成手法”, 情報処理学会論文誌：データベース, Vol. 42, No. SIG8(TOD10), pp. 124–138 (2001).

[2] 牛尼剛聡, 渡邉豊英: “ハイパービデオを利用したビデオ・データベースの意味構造に基づくブラウジング手法”, 第 12 回データ工学ワークショップ (DEWS01) (2001).

[3] V.S. Subrahmanian: “*Principles of Multimedia Database Systems*”, Morgan Kaufumann (1998).

[4] M. Flickner, et al.: “Query by Image and Video Content: The QBIC System”, *IEEE Computer*, Vol. 28, No. 9, pp. 23–32 (1995).

[5] 柳沼良知, 坂内正夫: “DP マッチングを用いたドラマ映像・音声・シナリオ文書の対応づけ手法の一提案”, 信学論 (D-II), Vol. J79-D-II, No. 5, pp. 747–755 (1996).

[6] 長坂晃朗, 宮武孝文: “時系列フレーム特徴の圧縮

符号化に基づく映像シーンの高速分類手法”, 信学論 (D-II), Vol. J81-D-II, No. 8, pp. 1831–1837 (1998).

[7] 片岡良治, 遠藤斉: “MPEG 符号化情報に基づく類似シーン検出方式”, 情報処理学会論文誌：データベース, Vol. 41, No. SIG3(TOD6), pp. 37–47 (2000).