多視点映像データの概念モデリングと代表映像検索

中西吉洋[†] 廣瀬竜男[†,☆] 田中克己[††]

現在，デジタルビデオ技術の開発により，多数のカメラで撮影され，相互に同期した多視点映像が一般にも普及し，実際ビル監視システム等は大きなアプリケーションとなっている．ここで，多視点映像すべてにキーワードの注釈をつけることは大変困難であり，また画像処理技術を用いて多視点映像の内容を判断し，注釈付けを行うことには多大なコストがかかる．よってカメラのメタデータを効率良く獲得したり，注釈情報を共有したりする効果的な方法が必要となる．本論文では，まず以前の研究である“Query by Camera”について触れる．これは多数のカメラの映像から，キーワードの注釈や画像処理技術を用いることなく，映像を検索する手法である．各カメラの撮影範囲の時系列データを蓄積し，そのデータを比較することで，多視点映像の中から「より良い」映像を検索する手法を提案している．この研究を前提に，我々は多視点映像の概念的なモデリングを行う．また，多視点映像カメラの階層構造に注目し，カメラやセンサの物理データを利用することで，より良さを表す「捕捉状態度」に基づいた映像間の捕捉度関連をはじめとしたいくつかの意味的な関連を定義する．そして，これらの定義に基づいた注釈情報の共有について提案する．また，カメラ間の切替え情報およびキーワードの出現密度に注目した多視点映像の代表映像検索について述べる．

Conceptual Modeling and Representative-video Retrieval for Multiple Perspective Video

Yoshihiro Nakanishi,[†] Tatsuo Hirose[†,☆] and Katsumi Tanaka[††]

Recently, MPV becomes popular because of recent advancement in digital video technologies. Actually, video surveillance is currently a big application of MPV. In the case of the usage of a lot of video cameras, human's keyword annotation is a tedious task, and automatic image recognition techniques takes an expensive cost to find what are taken in the MPV. In this paper, we first mention about our previous work called the “Query by camera” metaphor. The “Query by camera” metaphor means a way to search relevant video data from a lot of MPV cameras with neither keyword annotation nor image recognition techniques. Actually, we assume that each camera stores a time series data of focused area, and by comparing those time series of focus area data, we can retrieve a “better” video scene from MPV video data. Based on this previous work, we first propose a conceptual modeling for MPV data. When modeling MPV data, we consider a hierarchical setting of MPV cameras and the “better capture-ness” relationships among video scenes taken by MPV cameras. Then, we propose a way to compose representative-video of MPV data using camera switching information and keyword density. This is used for browsing or skimming a vast volume of MPV data taken by multiple cameras.

1. はじめに

多視点映像データとは，相互に時間同期した複数のカメラで撮影されたビデオデータである．現在ビル監視システムやスポーツ，コンサートのテレビ中継のように，多視点映像に関するアプリケーションが多数現れている．近い将来，各イベントにおいて一般の人々が撮影した映像を含む多視点映像を扱うことも考えられる．

ここで多視点映像の特徴をまとめると以下のようになる．

- 連続メディア
  ビデオは時間的に連続したデータである．すなわち，ビデオデータのあらゆる区間が注釈情報の表す対象となり，検索結果の答えとなりうる．
- 同期メディア
  多視点映像は，同じ被写体や場所を同時に撮影した複数の映像の集合であるといえる．多視点映像

† 神戸大学大学院自然科学研究科
Graduate School of Science and Technology, Kobe University
†† 京都大学大学院情報学研究科
Graduate School of Infomatics, Kyoto University
☆ 現在，松下電器産業株式会社
Presently with Matsushita Electric Industrial Co., Ltd.

の規模はアプリケーションにもよるがスポーツやコンサートのテレビ中継の映像の場合は数十台程度であり，ビルや高速道路の監視システムではさらに多くのカメラが使われている．

- 不特定カメラのデータ処理
  最近，DV 技術は急速な進歩を続けており，一般の人がビデオカメラを使う機会が増えてきた．イベントが行われると，多くの一般人がデジタルカメラを使用し，映像データを記録している．この場合，各カメラには独自の情報が蓄積されている．そこで，もしこれらの多視点映像データを扱おうと考えた場合，不特定のカメラのデータについても考慮する必要がある．

我々はこのような特徴を持つ多視点映像における問題点として次の 2 点に注目し，解決のアプローチを提案する．

- データ量があまりにも膨大なため，多視点映像すべてを閲覧することは困難．
- 多視点映像全体に内容を示す注釈情報を記述することは多大な労力が必要．

本論文では上記の問題を解決するために，多視点映像の概念的なモデリングとそれに基づく注釈情報の継承方式，および多視点映像から代表的な映像を検索する手法について述べる．我々は撮影するカメラより得られるパン・チルト，ズームのデータや空間センサより得られる被写体までの距離データを多視点映像データに統合させた．そして，これらのデータを用いて多視点映像の基本構造を定義し，映像間の意味的な関連を提案している．さらに，これらの意味的関連に基づいて多視点映像に注釈情報を継承する手法を提案した．また，多視点映像に注釈情報が付加されたことを前提に，ユーザは問合せとしてキーワード群を指定することで，それらのキーワードが少なくとも 1 度は現れる映像を多視点映像の代表映像として検索する手法についても提案している．代表映像を検索する際に，カメラの切替えやキーワードの出現密度を考慮することで，ユーザにとって見やすい映像を呈示する．

以下，本論文の構成を示す．2 章において本研究に関連する研究について述べ，3 章で我々の以前の研究である “Query by Camera” について述べる．4 章では多視点映像の基本的な構造を定義し，4 種類の意味的な関連（同期関連，区間包含関連，撮影範囲包含関連，捕捉度関連）を定義することで多視点映像データに相互関連を持たせるアプローチを述べる．また，これらの意味的関連を基に，多視点映像に注釈情報の継承を行う手法を提案する．5 章では上記の意味的関連と，カメラの切替え度，キーワードの出現密度という測度に基づいた多視点映像の代表映像検索について述べる．最後に 6 章で結論と今後の課題について述べる．

2. 関 連 研 究

広域インターネット技術の急速な発達にともない，膨大なデジタルビデオデータに対して問合せを行ったり，高速インターネットを利用して配信を行ったりすることへの関心が高まってきている．

ビデオデータのモデリングや構造化に対する先進的な取組みがいくつか行われているが，Allen の時区間の研究[1] は時区間に関連する多くの研究の基本となっている[2),3)]．Allen は 2 つの時区間の間には 13 の時間関連があると示している．ビデオデータベースに基づくいくつかの研究は Allen の時区間モデルに大きく影響を受けている．たとえば文献 2) では時区間に基づくモデルがビデオデータのような，時間に依存したマルチメディアデータに用いられている．

また，過去数年間にわたって，データベースに対してビデオデータを確立するシステムの構築を世界中の研究者が行っている[4)~6)]．オブジェクトビデオデータベース（OVID）[5] はインスタンス主導のビデオデータモデルである．OVID において，ある 1 組の近接する区間は意味のあるものとして定義されており，これをビデオオブジェクトと呼んでいる．継承可能な属性情報と継承不可能な属性情報がそれらのビデオオブジェクトに割り当てられ，継承可能な属性情報は区間包含関連にあるビデオオブジェクト間で共有される．この基本的な考えはビデオオブジェクトにその親となるビデオオブジェクトの継承可能な属性情報を共有させることで，注釈付けの手作業の繰返しをなくすというものである．また我々の最近の研究[7)~9)] は OVID とは異なっており，細分化されたビデオデータから指定されたすべてのキーワードを含む，すべてのビデオデータを動的に統合する *glue* 演算を提案している．また，代数演算モデル[6] は階層的なアプローチに基づいている．単一の階層とは違って，代数演算モデルでは同じビデオデータに関連している記述間に階層構造を定義することが可能であり，階層における親ノードが子ノードの前後関連を表す．この階層構造を使うことによって，同じビデオデータに異なった内容のデータを付加することができる．

Yeung ら[10] はビデオから得るストーリーデータのまとまりを拡張させることで，ビデオの要約を行う方式を提案している．彼らはビデオデータやそのメタデータから，ストーリーの構造を発見する方式を使用

している．また，是津ら[11]は時刻印付オーサリンググラフという新しい映像記述方式を提案し，映像に対する断片的な内容記述と記述間の意味的な関連性に基づいた映像記述を行っている．

上記にあげた研究において，ビデオデータに対するモデリングや要約については述べられているが，多視点映像については述べられていない．多視点映像は同じ空間を同期して撮影した映像の集合であり，データ量が膨大なために，たとえばこれらの膨大な映像データすべてに目を通し，注釈情報を付加するといった作業を行う場合や，多視点映像全体からユーザの好みとなる映像データを検索する場合，非常な労力をともなう．よってこれらの多視点映像の特徴を考慮したモデリングが必要である．近年，Bhonsle ら[12]は多視点映像データの意味的データモデルを提案し，特に時間的側面に注目しているが，上記の問題点の解決に関しては述べられていない．我々はこれらの多視点映像に関する問題点を解決するアプローチとして，多視点映像の概念的なモデリングを導入し，これに基づいた注釈情報の継承方式と多視点映像の代表映像を検索する手法を提案する．

3. カメラメタファによる多視点映像検索

この章では，本研究の前提となる，以前に我々が提案した “Query by Camera” と呼ばれる多視点映像を検索する手法[13)~15)]について述べる．

3.1 Query by Camera

ビデオカメラで被写体を撮影する場合，撮影者はファインダを覗きながら，注目する被写体を探し，カメラをフォーカスして，記録ボタンを押す．このようなカメラの操作は一般ユーザにとって，慣れ親しんだものである．我々が提案した Query by Camera において，多視点映像に対する各問合せは，カメラで被写体に焦点を合わせ，スタート・エンドボタンを押すといった慣れ親しんだ操作で行う．そしてシステムは多視点映像データから，問合せに用いたカメラよりも “より良く” 同じ被写体を写しているカメラの映像を探す．このアプローチにおいて，多視点映像データは時系列の撮影範囲データと注釈情報をともなっており，システムはより良い映像区間を得るために，問合せとなるカメラとその他のカメラの時系列な撮影範囲データを比較する．

Query by Camera により，多視点映像に対して問合せを行うアプリケーションの例（オンライン）として，バスケットボールの試合を撮影している場合を想定する（**図 1**）．あるユーザが試合の映像を撮影し終わった後，他の人が撮影した映像を要求する．ユーザが撮影していた操作自体が問合せとなるので，簡単に映像を問い合わせることができる．その結果，様々な視点から撮影された被写体を写している映像区間が検索される．さらにユーザは被写体がどのように撮影されている映像を望むかを指定し，より好ましい映像を選び出す．

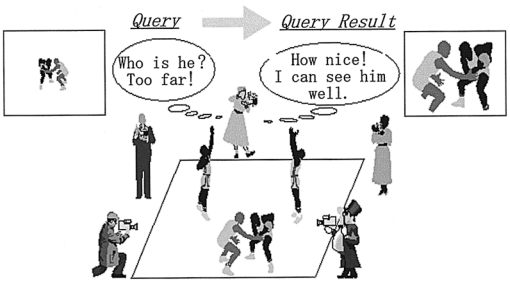
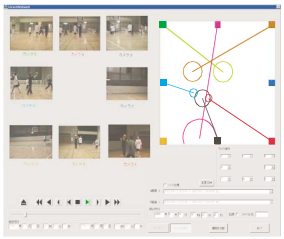


図 1 多視点映像のカメラメタファ検索
Fig. 1 Querying multiple perspective video by camera metaphor.



プロトタイプシステム

カメラ＆センサ
・パン・チルト・ズーム
・被写体までの距離

図 2 プロトタイプシステムと検索用カメラ
Fig. 2 Prototype system & Query Camera.

このような検索で用いるカメラは，カメラの位置，撮影方向，被写体までの距離等を計測するセンサや，映像データベースにアクセスするためのネットワーク機能等を実際のカメラに付加することにより実現できる．特に，検索に必要な撮影範囲や問合せ範囲に関するデータはセンサから自動的に実時間で得られるため，画像特徴量を用いたコンテンツ検索に比べ，実用時の利点が大きい．この研究ではコミュニケーションカメラと空間センサを組み合わせることで，バスケットボールゲームの映像および，カメラの撮影範囲データを時系列データとして取得し，オフライン作業により細かく同期付けを行った．この研究において我々は，ユーザが映像を見て，その映像よりもユーザの好みを反映している映像を検索するといったプロトタイプシステムを実装している．この検索はオフライン環境でのカメラメタファ検索と考えられ，評価実験を行うことで提案する検索手法の有用性を示すことができた．**図 2** に実際に撮影に使用したカメラとセンサ，および

映像データと撮影範囲を同期して表示するプロトタイプの画面を示す.

3.2 より良さの定義

“Query by Camera” において，まず解の候補となるビデオデータが撮影範囲データの時空間比較によって検索される．次にそれらのビデオデータの中から被写体がどの程度良く撮影されているかを示す尺度に基づいて，ビデオデータが検索される．“より良い” 映像とは，たとえばユーザが撮影している映像よりもより大きく被写体をとらえている映像や，ユーザの望む方向から撮影している映像を指す.

2 つのビデオデータのより良さを判断するために，我々は**捕捉状態度**という測度を提案した.

捕捉状態度の計算は我々のプロトタイプシステムで実装されており，以下の 2 段階の処理から成る.

まず，2 つのビデオデータのメタデータである撮影範囲の時系列データを比較する．各時刻において撮影範囲データを求め，比較する．現在の実装では，もし撮影範囲の中央に写っている被写体が他の撮影範囲に入っていれば，この 2 つのビデオフレームは同じ被写体をとらえていると定義している.

次に注目している被写体がどの程度良く撮影されているかを示す捕捉状態度の計算がされる．捕捉状態度を構成する要素には次の 3 種類がある.

- **大きさ**
  ビデオフレームに写る被写体の大きさはカメラの位置やズームによって変化する．ユーザは被写体を大きく写している映像や広い範囲を撮影している映像を望む場合がある.
- **角度**
  被写体に対するカメラの角度はカメラの位置や方向に依存する．ユーザが自分とは逆の方向から撮影された映像や様々な方向から撮影された映像を望む場合が考えられる.
- **時間**
  ユーザは複数の短いビデオデータの集合をイベントのハイライトとして望むかもしれない．また，1 つの被写体を長い時間追っている映像を望むかもしれない.

上記の 3 つの要素について，ビデオデータの各時刻における捕捉状態度を計算し，その値の総和を重み計算した値を各ビデオデータの捕捉状態度と定義している．ユーザは解の候補が表示されたときに，自分の好みに従って，捕捉状態を指定する．それにより捕捉状態度が高いものが選ばれ表示される.

4. 多視点映像データの概念モデリング

3 章において，多視点映像のメタデータとして得られる物理データからカメラの撮影範囲を計算し，それに基づいた映像の検索と “より良い” 映像の定義について述べた．この章では，そのメタデータとより良さを示す捕捉状態度を用いて，多視点映像の基本構造とビデオデータ間の意味的な関連を提案する．さらにこれらの関連に基づいて，ビデオデータ間で注釈情報を継承することで，多視点映像における問題であった注釈情報の付加作業の労力を軽減する手法について提案する.

4.1 基本的データ構造

我々が提案する多視点映像データの基本的な構造はビデオデータとそのメタデータ（撮影範囲，注釈情報）を含んだ**ビデオオブジェクト**とビデオオブジェクト間における複数の**意味的な関連**である．まとめると以下のようになる.

- ビデオオブジェクト
  - 連続するビデオフレーム
  - 撮影範囲情報
  - 注釈情報（キーワード群）
- 意味的関連
  - 同期関連
  - 区間包含関連
  - 撮影範囲包含関連
  - 捕捉度関連

ビデオオブジェクトはあるカメラによって撮影されたビデオデータとそのメタデータであると直感的に考えられる．ビデオオブジェクトはこのモデルの基本的実体であり，ビデオオブジェクト O_i は次の要素から成る $O_i = (F_i, A_i, K_i)$ で表される.

- 連続するビデオフレーム
  $F_i = f_{i_1} f_{i_2} \dots f_{i_p}$
  ここで，f はそれぞれカメラで撮影されたビデオフレームを示している．時間関数（$timecode(f)$ で表す）とカメラ番号関数（$camera(f)$ で表す）はそれぞれの f に対して定義される．たとえば，$timecode(f) =' 19:05:35'$ という記述はフレーム f が時刻’19:05:35 に撮影されたことを示す．また，$camera(f) =' 3'$ という記述はフレーム f がカメラ番号 3 のカメラで撮影されたことを示す.
- 撮影範囲情報
  $A_i = a_{i_1} a_{i_2} \dots a_{i_p}$
  ここで，a はそれぞれフレーム f に対応する

撮影範囲情報を意味する．撮影範囲情報は基本的に注目している被写体の位置と撮影範囲（円）の半径から成る．

- 注釈情報（キーワード群）
  $K_i = \{k_{i_1}, k_{i_2}, \ldots, k_{i_q}\}$
  ここで，K_i はキーワードを示し，それぞれのキーワードはシーン S_i の全体の内容を表すものである．

ビデオオブジェクト O_i と O_j は次のように表すことができる．$O_i = (F_i, A_i, K_i)$，$O_j = (F_j, A_j, K_j)$
ここで，各要素は以下のとおりである．

$$F_i = f_{i_1} f_{i_2} \ldots f_{i_p},$$
$$A_i = a_{i_1} a_{i_2} \ldots a_{i_p},$$
$$K_i = \{k_{i_1}, k_{i_2}, \ldots, k_{i_q}\}, \text{ and}$$
$$F_j = f_{j_1} f_{j_2} \ldots f_{j_{p'}},$$
$$A_j = a_{j_1} a_{j_2} \ldots a_{j_{p'}},$$
$$K_j = \{k_{j_1}, k_{j_2}, \ldots, k_{j_{q'}}\}.$$

4.2 意味的関連

多視点映像のビデオオブジェクトに注目してみると，いくつかの意味的関連を定義することができる．この章では 4 つの意味的関連について説明する．

4.2.1 同 期 関 連

多視点映像では，複数のカメラが同期してビデオ撮影に用いられる．よって，いくつかの被写体やイベントは複数のカメラで同時に撮られている可能性がある．もし 2 つのビデオが同じ時間撮影されたとすると，それらのビデオは同期しているといえる．もし $p = p'$ と $timecode(f_{i_r}) = timecode(f_{j_r})$ がそれぞれの $r \in \{1, 2, \ldots, p\}$ に対していえるならば，これらの 2 つのビデオオブジェクトは**同期している**と定義できる．この関連を次のように表す．

$$synchronized(O_i, O_j)$$

4.2.2 区間包含関連

あるシーンがあるシーンに完全に含まれている場合，そのシーンはもう一方のシーンによって含まれるといえる．我々はこの包含関連の概念を多視点映像に拡張させ，もし $p > p'$ で F_j が F_i の一部分であるならば，F_j は F_i に**含まれる**といえる．すなわち，シーン F_j はシーン F_i の一部分である．この**包含関連**を次のように表す．

$$include(O_i, O_j)$$

4.2.3 撮影範囲包含関連

多視点映像では，多くのカメラが他のカメラよりも広い範囲を撮影していることが考えられる．このとき，カメラが他のカメラに対して空間的により大きな範囲を撮影しているといえる．この概念を形式化すると次のように表せる．もし $synchronized(O_i, O_j)$ がいえて，$a_{i_r} \supseteq a_{j_r}$ がそれぞれの $r \in \{1, 2, \ldots, p\}$ に対していえるとき，この 2 つのビデオオブジェクトは**撮影範囲における包含関連**があると分かる．この関連を次のように表す．

$$FOA - include(O_i, O_j)$$

4.2.4 捕捉度関連

多視点映像では同じ被写体もしくは同じイベントが複数のカメラで撮影されている場合がある．このとき，あるカメラが別のカメラよりも被写体をより良く撮影している場合がある．もし $syncronized(O_i, O_j)$ でビデオオブジェクト O_i がビデオオブジェクト O_j よりも被写体をより良く写していると判断されたとき，ビデオオブジェクト O_i と O_j の間には**捕捉度関連**があると定義し，次のように表す．

$$CaptureBetter(O_i, O_j)$$

4.3 注釈情報の継承

多視点映像データに関する 1 つの大きな問題は，複数の映像ストリームに対してどのように情報を付加するかという問題である．以前我々が行った研究に OVID [5] があるが，その中でこの問題を処理する手法を提案した．1 つのビデオデータに付加されている情報を他のサブビデオデータに自動的に継承するシステムである．しかし，OVID ではシングルストリームのビデオデータに関してのみ扱っている．本研究ではこの OVID の考えを拡張させ，上記の意味的関連を用いることで注釈情報の共有を多視点映像データに対して行う手法を提案する．

4.3.1 区間包含継承

1 つのビデオオブジェクトが他のビデオオブジェクトを含んでいる場合（$include(O_i, O_j)$），注釈情報 K_i は K_j に継承される．すなわち新しい K_j は $K_i \cup K_j$ になる．確かにこれは OVID の区間包含継承のようにとらえることができるが，本研究ではシングルストリームだけでなくマルチストリームに対しての継承も行っている．

4.3.2 撮影範囲包含継承

撮影範囲包含継承は注釈情報のボトムアップ的な継承である．2 つのビデオオブジェクト O_i と O_j が撮影範囲包含関連（$FOA - include(O_i, O_j)$）にあるとき，注釈情報 K_j は K_i に継承される．これは多視点映像の環境において起こる継承である．

4.3.3 捕捉度継承

これは捕捉度関連を持つビデオオブジェクト間に起

こるキーワードの生成であり，相互に対して起こる．2つのビデオオブジェクトは同じ被写体を撮影していると考えられ，相互のビデオオブジェクトに対して，自分が被写体を「より良く」または「より悪く」撮影しているという注釈を生成する．また，撮影範囲等を考慮することで，お互いが持つ注釈情報を継承することも考えられる．

以上のような意味的関連を利用して注釈情報を自動的に継承することは，次のようなアプリケーションに対して有効であると考えている．

- 人がキーワードを1つのビデオデータに付加するとき，注釈情報の継承を行うことで，自動的に他のビデオデータにキーワードが継承され，注釈情報を付加する作業の軽減につながる
- キーワードを用いて映像検索を行うとき，システムが注釈情報の継承を自動的に行うことで，継承前にはキーワードが注釈情報として付加されていなかったビデオインターバルを解として選ぶことができる

また，キーワードをビデオデータに付加する手法としては，オフライン環境において手作業で記述する方法が考えられるが，音声認識技術の急速な発展を考えると，映像を撮影している人が被写体やイベントに関するコメントを撮影しながら発することで，キーワードを記述することも近い将来可能になると考えられる．

4.3.4 実験と評価

提案した注釈の継承手法の有効性を示すために，評価実験を行った．以下に実験の流れを示す．

(1) 8台のカメラとセンサで撮影した3分間のバスケットボール映像とカメラメタデータを使用する．

(2) このデータに対して，カメラメタファ検索を行う．このとき，ユーザは問合せカメラの映像よりも被写体を大きく撮影している映像を望んでいるとする．

(3) 8人の被験者に，1台の問合せカメラの映像と検索により得られた7台のカメラからの映像区間にキーワードを記述してもらう．キーワードの種類はあらかじめ定義した10種類（Taeko, Gaku, Chikashi, Takeshi, Tetsuo, Hiroaki, Ball, Shoot, Pass, Basketball Game）とする．

(4) 本研究で提案している注釈継承アルゴリズムに従って，キーワードの継承を各映像区間の間で行う．

(5) 継承されたキーワードと映像を同期して表示させ，映像をより詳しく示す注釈になっているかどうかを判断することで注釈継承の手法の有用性を評価する．**図3**は作成したプロトタイプシステムである．

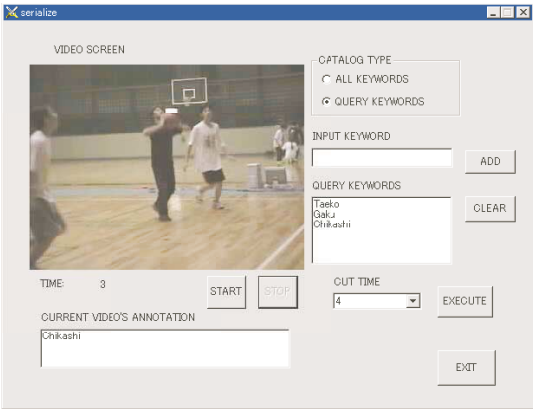


図3 注釈情報の継承
Fig. 3 Annotation inheritance.

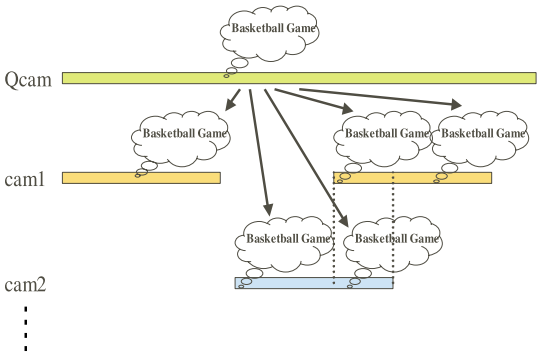


図4 区間包含継承
Fig. 4 Interval-inclusion inheritance.

各継承に関しての評価を以下に示す．

- 区間包含継承
  問合せ映像と検索結果として得られる映像区間は，すべて区間包含関係にある．ここで，付加されているキーワードをトップダウンで継承すると，すべての映像区間に問合せ映像のキーワードが継承される（**図4**）．この実験では問合せ映像を示す注釈として「Basketball Game」というキーワードを用いた．よって，各映像区間に継承されたとしても，違和感のない継承となったが，実際問題として，キーワードの種類やキーワードが示す映像の内容は瞬間なのか，区間なのか等を考慮して，継承を行う必要がある．
- 撮影範囲包含継承
  この実験では，被写体をより大きく撮影している

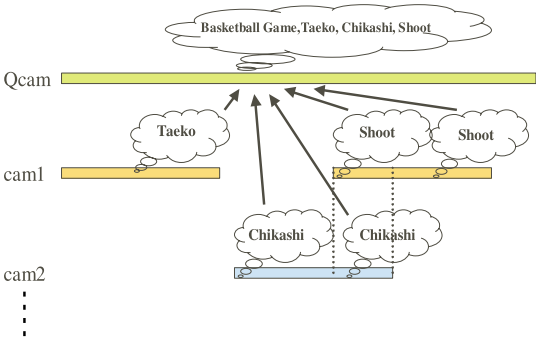


図 5 撮影範囲包含継承
Fig. 5 FOA-inclusion inheritance.

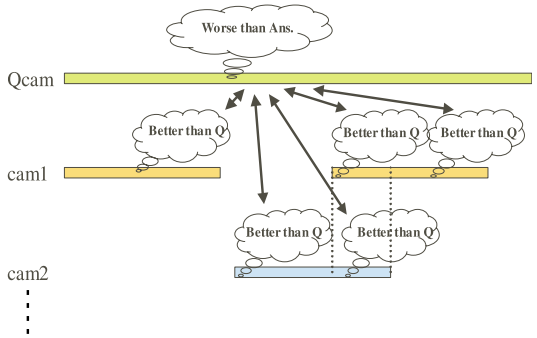


図 6 捕捉度継承
Fig. 6 Better-capturedness inheritance.

映像区間（撮影範囲がより小さい映像区間）を検索しているため，問合せ映像の各区間と得られた映像区間でその区間に同期している区間は，撮影範囲包含関係にある．ここでのキーワード継承はボトムアップ継承であり，得られた映像区間のキーワードが問合せの映像区間に継承される（**図 5**）．この継承によって，映像区間を示す詳しい注釈を得ることができ，キーワードの継承が有効であると分かる．

- 捕捉度継承
  この実験では,「被写体をより大きく写している映像」を検索している．よって，得られた映像区間と問合せ映像は，捕捉度関連にある．よって新たなキーワード「Better」「Worse」が付加されている（**図 6**）．この結果が即座に有効性を示すものではないが，これらのキーワードが付加されることで，ユーザの好みを反映した新たな検索が期待できる．

5. 多視点映像データの代表映像検索

4 章で述べたように，多視点映像の概念的モデリングを導入し，注釈情報の継承を行うことで注釈情報を付加する作業の軽減が期待できる．この章では，多視点映像にすでに注釈情報となるキーワード群が継承され，多視点映像全体に注釈情報が付加されていると考える．そして，多視点映像データが膨大であるため，全体の概要を短時間で見ることができないという問題点の解決法として上記の多視点映像データに対して，キーワードを問合せとした多視点映像の代表映像を検索する手法を提案する．また，この代表映像がユーザにとって見やすい映像となるように，ビデオオブジェクトに含まれるカメラの切替え情報とキーワードの出現密度を導入する．

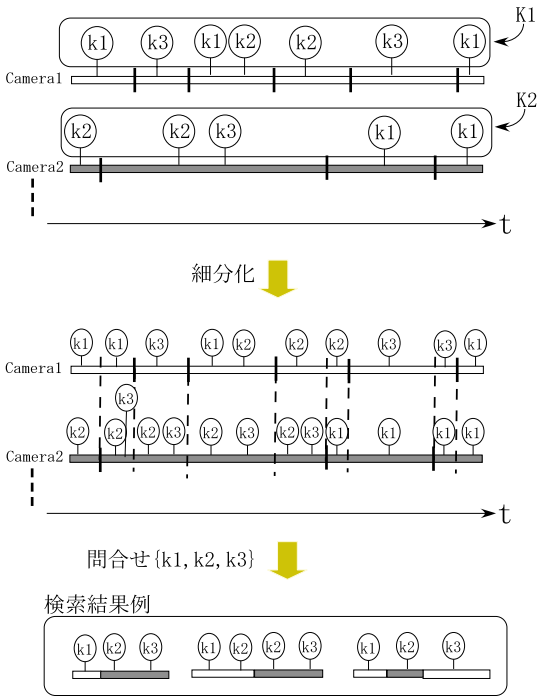


図 7 多視点映像の代表映像検索
Fig. 7 Representative-video retrieval for multiple perspective video.

まず，代表映像を検索するアルゴリズムの概要を述べた後，カメラ間の切替えに基づく代表映像の検索について述べる．最後にキーワード密度を考慮した代表映像検索について述べる．

5.1 多視点映像データの代表映像検索アルゴリズム

ここで扱う多視点映像は，撮影されるときにカメラの記録ボタンを操作する等で複数のビデオオブジェクト（ビデオデータ，撮影範囲データ，注釈情報）に分割されている（**図 7**）．多視点映像に対する問合せとしてキーワードの集合を考えるが，ユーザが複数のキーワードを指定し，そのキーワードに対する映像を検索結果として望む場合（部分代表検索）と，多視点映像

全体に付加されているキーワードに対する映像すべてを検索結果として望む場合（全体代表検索）がある．

- 部分代表検索
  問合せとして，キーワードの集合
  $K = \{k_1, k_2, \ldots, k_n\}$
  を与える．システムは各 $k \in K$ が少なくとも 1 回出現するようなビデオオブジェクトの集合を検索結果として返す．
- 全体代表検索
  問合せとして，キーワードの集合
  $K = K_1 \cup K_2 \cup K_3 \cup \ldots \cup K_n$
  $(K_i = k_{i1}, k_{i2}, \ldots, k_{in})$
  を与える．システムはこれらのキーワードが少なくとも 1 回は出現するようなビデオオブジェクトの集合を検索結果として返す．

代表映像を作成する概念的な流れを以下に示す．

(1) ユーザは多視点映像に問合せを行うために，興味あるキーワードの集合を指定する．

(2) システムは，ユーザが指定したキーワードがすべて含まれているビデオオブジェクトの集合を候補として選ぶ．たとえば，問合せが k_1, k_2, k_3 とする．あるビデオオブジェクト O_1 の注釈がキーワード k_1, k_2 を持ち，あるビデオオブジェクト O_2 の注釈がキーワード k_2, k_3 を持つとき，組合せ O_1, O_2 が問合せに対する答えの候補となる．

(3) ビデオオブジェクトの組合せの候補それぞれに対して，システムはすべてのビデオオブジェクトを各カメラについて時間順にソートする．たとえばカメラ 1，カメラ 2 に対してそれぞれビデオオブジェクトが O_{11}，O_{12}，O_{21}，O_{25} のようにソートされたとする．このときいくつかのビデオオブジェクトが時間的に重なっていることに気づく．たとえば O_{11}，O_{12}，O_{21}，O_{25} のビデオオブジェクトがそれぞれ $[10, 300]$，$[500, 800]$，$[200, 400]$，$[700, 1000]$ とすると，O_{11} と O_{21} のビデオオブジェクトと O_{12} と O_{25} のビデオオブジェクトが重なっている．

(4) 2 つのビデオオブジェクトが重なっている場合，システムはビデオオブジェクトをより小さいビデオオブジェクトに分割する．たとえば上記のビデオオブジェクトは以下のビデオブジェクトに分割される．

$O_{111} : [10, 200]$，$O_{112} : [200, 300]$，
$O_{211} : [200, 300]$，$O_{212} : [300, 400]$，
$O_{121} : [500, 700]$，$O_{122} : [700, 800]$，
$O_{251} : [700, 800]$，$O_{252} : [800, 1000]$

ここで O_{112} と O_{211} は同期しており，O_{122} と O_{251} も同期している．

(5) システムはビデオオブジェクトを時間順にソートし，解を生成する．たとえば上記のビデオオブジェクトに対する解の 1 つの例は，
$O_{111}, O_{211}, O_{212}, O_{121}, O_{122}, O_{252}$
である．また，他の解としては
$O_{111}, O_{112}, O_{212}, O_{121}, O_{251}, O_{252}$
である．

図 7 は多視点映像に対して，問合せとしてキーワードの集合 $\{k_1, k_2, k_3\}$ を与えた場合の検索結果例を示している．

5.2 カメラの切替えを考慮した検索結果の再構成

ユーザが多視点映像の代表となる映像を検索する場合，その検索結果はユーザにとって分かりやすい映像が望ましい．そこで我々はビデオオブジェクトが持つカメラの切替え情報を考慮することで，ユーザにとって見やすい映像を呈示する方法を提案する．

4 章で述べたように，分割されたビデオオブジェクト間で捕捉度関連が定義されている．上記の例を用いると，ビデオオブジェクト O_{112} と O_{211} は同期しており，これら 2 つのビデオオブジェクトが同じ被写体を写しているとすると，最終的な答えとして “より良い” 映像を選ぶべきである．しかし，もし我々が同期したビデオオブジェクト群から解を選ぶ方法として捕捉度関連のみを利用すると，最終的な映像はユーザにとって見にくい映像になる場合がある．これは最終結果は複数のカメラの映像から成るため，カメラの切替えが突然に起こった場合，視覚的混乱を招く可能性があるためである．我々が以前行った研究[16),17)] において，検索結果をカメラの切替えに基づいて呈示する手法を提案した．この手法を本研究において適用する．

たとえば人の顔を映している映像から，急に人の足を写している映像に切り替わった場合に視覚的混乱が起こると考える．“映画の文法” によると，このような視覚的混乱をさけるためにシーンの一致という規則がある．これは被写体の画面における位置と動きおよびユーザの目線を考慮することで，視覚的混乱を避けようというものである．また，ユーザは画面上の被写体の大きさが突然変化するとユーザにとって見にくい映像になると考える．しかし，撮影範囲情報だけではユーザの目線を判断することは不可能であるので，被写体の画面における位置と動きおよび大きさを考慮した検索結果の呈示方式を以下に述べる．

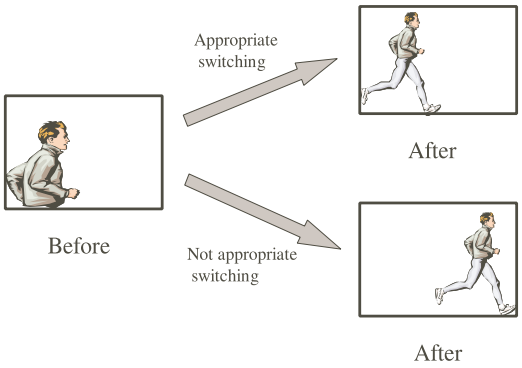


図 8 被写体の位置
Fig. 8 Object's position on screen.

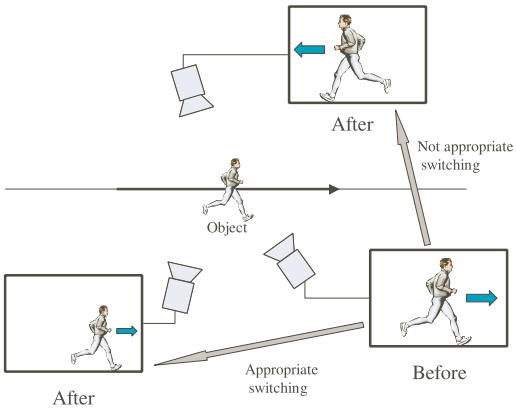


図 9 被写体の動き
Fig. 9 Object's movement on screen.

5.2.1 カメラの切替え度

カメラの切替え問題を扱うために，カメラ間の切替えの適切さ（**切替え度**）を定義する．この切替え度はカメラの切替えに対する適切さを示し，画面上の被写体の位置，動き，大きさに関する値から成る．現在のビデオオブジェクトに対する候補となるすべての同期したビデオオブジェクトに対して，切替え度を計算する．

- 位置の一致
  図 8 に示されているように，カメラが切り替わる前と後とで比較して，カメラが切り替わる前の被写体の画面上の位置と最も近い位置に被写体が存在する切替えが最も適切であるといえる．
- 動きの一致
  図 9 に示してあるように，カメラの切替え前後を比較して，カメラが切り替わる前に被写体が画面上で動いている方向とカメラの切替え後の画面に写る被写体が同じ方向に動いている場合，適切な切替えであると考える．まず，カメラの切替え前後の被写体の位置を直線で結ぶ．もしカメラが直線の同じ側にあれば切替え度を高くする．もし被写体が停止していれば，すべてのカメラに対して切替え度は等しくする．
- 大きさの一致
  画面に映っている被写体の実際の大きさは被写体とカメラの距離とカメラの画角より計算できる．カメラの切替え前後での被写体の大きさの比を比較する．この比が小さくなるカメラへの切替えが適切な切替えである．

最終的に，カメラ間の切替え度はこれら 3 つのパラメータを用いて計算される．

5.2.2 実験と評価

得られた代表映像をカメラの切替え情報を考慮することで，ユーザにとってより見やすい映像の再構成する手法の有効性の評価を行うため，以下の実験を行った．

(1) 8 台のカメラとセンサで撮影したバスケットボールのゲームの映像，および 4 人がある規則に従って動いた映像を使用する．
(2) 8 人の被験者に，8 台のカメラで撮影された映像をビデオオブジェクトに分割した後，各ビデオオブジェクトにキーワードを記述してもらう．使用するキーワードは 10 種類（Taeko, Gaku, Chikashi, Takeshi, Tetsuo, Hiroaki, Ball, Shoot, Pass, Basketball Game）とする．
(3) これらのビデオオブジェクトから映像に付加されているキーワードをすべて含むような代表映像を生成する．
(4) この代表映像に対してカメラの切替えを考慮した場合と考慮しない場合の検索結果を目視で 8 人の被験者に評価をしてもらう．

カメラの切替えを考慮しない場合，どちらの映像素材に対しても画面上での被写体の位置や動きの変化が大きい切替えが存在し，見にくい部分があると 90%以上の人が回答した．

次に，シンプルな動きをとらえた映像に対して，カメラの切替えを考慮した再構成を行うと，75%の人が見にくい部分の改善ができたと答えた．しかし，複数の被写体がまとまって素早く移動するバスケットボールの映像に対して，カメラの切替えを考慮した再構成を行った場合，見にくい切替えがいくつか存在し，被験者で見にくい部分の改善ができたと評価した人の割合は 40%であった．その原因として，興味ある被写体を写しているカメラが存在するにもかかわらず，時空

間検索で該当する映像区間を正しく，かつ十分に抽出できなかったことがあげられる．すなわち，視覚的飛躍を小さくする切替えを行うための映像区間が，うまく抽出されなかったため，仕方なく視覚的飛躍が大きくなる映像区間を選択せざるをえなかったためである．

また，バスケットボールを撮影した映像では，シュートやパス等の被写体が速く複雑な動作をしている瞬間に切替えが行われる場合があり，その場合は被写体の画面上での位置や動きが一致していても，見にくい切替えがあった．これについては，レーダで測定される被写体の位置情報だけでは不十分であり，被写体の動作に関するレベルの高い情報を画像処理や人手の注釈付けにより得る必要があると考える．

5.3 キーワードの出現密度に基づく代表映像検索

我々が提案する多視点映像の代表映像は，ユーザが指定したキーワードが少なくとも 1 回出現するような映像である．この代表映像は，ユーザが映像を見たときにその映像が何を示しているのかがはっきりと分かる映像でなければならない．そこで我々はビデオオブジェクトを生成する際に，ビデオオブジェクトに含まれるキーワードの種類ができるだけ少くなるようにビデオオブジェクトを生成する手法を提案する．

5.3.1 検索アルゴリズム

あるビデオオブジェクトの中にキーワード k_1 のみが含まれている場合と，$\{k_1, k_2, k_3\}$ がキーワードとして含まれている場合は，キーワード k_1 を表す映像としては前者を選ぶべきである．

以下にビデオオブジェクトを生成するアルゴリズムを示す．**図 10** では多視点映像に対して問合せとしてキーワードの集合 $\{k_1, k_2, k_3\}$ を与えている．

- まず可変長のウィンドウを各映像ストリームの先頭に置く．
- そのウィンドウに囲まれた区間内で他のキーワードに比べて出現回数が多いキーワード k の事象がその区間を最も象徴していると考えることができるので，その区間に付いているすべてのキーワード総数に対する，キーワード k の出現回数から，その区間におけるキーワード k の出現密度を計算する．
- ウィンドウを時間に沿って進め，上記の計算を繰り返し行う．
- すべての映像ストリームについてキーワードの出現密度を計算し，最も値の高い区間を各キーワードを示すビデオオブジェクトとする．図 10 ではキーワード k_1，k_2，k_3 それぞれに対して，出現密度 4/5，2/2，3/3 となる区間がビデオオブジェクトとして得られている．

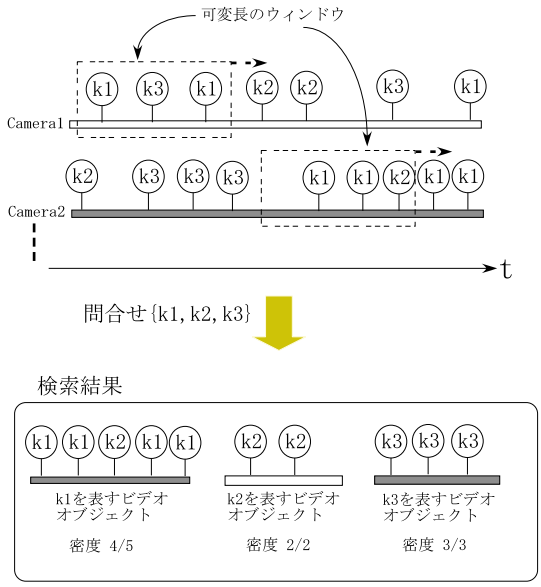


図 10 キーワードの出現密度に基づく代表映像検索
Fig. 10 Representative-video retrieval based on keyword density.

5.3.2 実験と評価

上記のアルゴリズムについて定性的な評価を行うために，次の実験を行った．使用した実験データは以下である．

- ビデオデータは，秋季関西学生競技ダンス選手権大会を 9 台のカメラで撮影した映像の中から，各カメラ約 1 分 30 秒分を使用．その映像内容は同大会のサンバ上位決勝戦である．
- 選手名（A ペア，B ペア，C ペア，D ペア，E ペア，F ペア，G ペア），ダンスの技名（スポットターン，スイング，リバースロール，プロムナードターン），その他（衝突，応援）のキーワードを用意し，映像を見ながらキーワードをテキストファイルに記述した．このとき，複数のキーワードが同じ被写体を表すことがないようにキーワードを定義し，キーワードを 1 秒ごとに記述した．

この映像に記述されたすべてのキーワードを示すビデオオブジェクトの集合から成る代表映像（**図 11**）を，10 人の被験者に見てもらった．まずキーワードの出現密度を考慮せずに得られた代表映像を見せた．このとき，各映像が何を示しているかを理解できた人の割合はキーワードの種類にもよるが，平均で 55%であった．次にキーワードの出現密度を考慮して得られた代表映像を見せたとき，各映像が何を示しているかを理解できた人の割合は平均で 77%であった．これらの結

図 11 代表映像の呈示
Fig. 11 Display represetative-video.

果から，キーワード密度を考慮することで，ユーザにとって映像の内容を理解しやすい映像を検索することが可能であると評価できる．この実験ではダンス大会の映像に対して限定したキーワードを用い，提案するアルゴリズムの定性的な評価を行った．しかし，映像コンテンツによってキーワードの種類やキーワードの示す映像の意味が変わっていると考えられ，検索アルゴリズムを改良する必要があると考えている．

6. ま と め

本論文では，多視点映像の概念的なモデリングを導入し，カメラや空間センサの物理的データ（パン・チルト，ズームデータ，カメラと被写体間の距離データ）を注釈情報と同様に多視点映像に統合した．また，多視点映像の相互関連を定義するために，4 つの意味的関連（同期関連，包含関連，撮影範囲包含関連，捕捉度関連）を提案し，これらの意味的関連に基づいて，多視点映像に対して注釈情報の継承を行う手法を提案した．最後に，キーワードの出現密度やカメラの切替えを考慮した多視点映像の代表となる映像の検索手法を提案した．

今後の課題として，多視点映像の代表映像検索において問合せとしたキーワード間に意味を持たせた問合せを考慮する必要があると考えている．たとえばキーワード k_1 とキーワード k_2 は同時に出現して，キーワード k_3 は単独で現れているようなビデオオブジェクトがほしい場合等である．また，得られる代表映像の全体の長さに関しても考慮しなければならない．本研究では注釈情報の継承アルゴリズム，カメラの切替えおよびキーワード密度を考慮した多視点映像の代表映像検索アルゴリズムの定性的な評価実験を行った．今後，様々な映像コンテンツに対応可能なアルゴリズムやキーワードの種類を考慮した検索アルゴリズムの検討を行い，統合的なプロトタイプの作成と，評価実験を行いたい．

謝辞 本研究を遂行するにあたり，懇切なるご指導をいただいた三菱電機株式会社の秦淑彦氏につつしんで感謝の意を表します．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」（プロジェクト番号 JSPS-RFTF97P00501）によっております．ここに記して謝意を表します．

参 考 文 献

1) Allen, J.F.: Maintaining Knowledge about Temporal Intervals, *Comm. ACM*, Vol.26, No.11, pp.832–843 (1983).
2) Little, T.D.C. and Ghafoor, A.: Interval-Based Conceptual Models for Time-Dependent Multimedia Data, *IEEE Trans. Knowledge and Data Engineering*, Vol.5, No.4, pp.551–563 (1993).
3) Lorentzos, N.A. and Mitsopoulos, Y.G.: SQL Extension for Intervals Data, *IEEE Trans. Knowledge and Data Engineering*, Vol.9, No.3, pp.480–499 (1997).
4) Hwang, E.J. and Subrahmanian, V.S.: Querying Video Libraries, *Journal of Visual Communications and Image Representation*, Vol.7, No.1, pp.44–60 (1996).
5) Oomoto, E. and Tanaka, K.: OVID: Design and Implementation of a Video-Object Database System, *IEEE Trans. Knowledge and Data Engineering*, Vol.5, No.4, pp.629–643 (1993).
6) Weiss, R., Duda, A. and Gifford, D.: Composition and Search with a Video Algebra, *IEEE MultiMedia*, Vol.2, No.1, pp.12–25 (1995).
7) Pradhan, S., Tajima, K. and Tanaka, K.: Interval Glue Operations and Answer Filtering for Video Retrieval, *IPSJ Trans. Databases*, Vol.40, No.SIG3(TOD1), pp.80–90 (Feb. 1999).
8) Pradhan, S., Tajima, K. and Tanaka, K.: A Query Model to Synthesize Answer Intervals from Indexed Video Units, *IEEE Trans. Knowledge and Database Systems*, Vol.13, No.6 (2001).
9) Tanaka, K., Tajima, K., Sogo, T. and Pradhan, S.: Algebraic Retrieval of Fragmentarity Index Video, *New Generation Computing*, Vol.18, pp.359–374, (2000).
10) Yeung, M., Yeo, B. and Liu, B.: Extracting Story Units from Long Programs for Video Browsing and Navigation, *International Conference on Multimedia Computing and Sys-*

tems, pp.296–305 (June 1996).
11) 是津耕司，上原邦昭，田中克己：時刻印付オーサリンググラフによるビデオ映像のシーン検索，情報処理学会論文誌，Vol.39, No.4, pp.923–932 (1998).
12) Bhonsle, S.K., Gupta, A., Santini, S. and Jain, R.: Semiorder Database for Complex Activity Recognition in Multi-Sensory Environment, *Proc. ICDE2000*, pp.689–691 (March 2000).
13) 中西吉洋，廣瀬竜男，秦　淑彦，田中克己：カメラメタファーに基づく多視点映像の検索，情報処理学会研究報告，Vol.2000, No.69, pp.215–222 (2000).
14) Hata, T., Hirose, T., Nakanishi, Y. and Tanaka, K.: Querying Multiple Perspective Video by Camera Metaphor, *Proc. 7th International Conference on Database Systems for Advanced Applications* (*DASFAA*), Hong Kong, April 18–20 (2001).
15) 秦　淑彦，廣瀬竜男，中西吉洋，田中克己：カメラメタファによる多視点映像の検索，情報処理学会論文誌：データベース，Vol.42, No.SIG4(TOD9), pp.14–26 (2000).
16) 廣瀬竜男，中西吉洋，秦　淑彦，田中克己：サンプル映像とカメラの空間情報に基づく多視点映像の検索と直列化，第 12 回データ工学ワークショップ（DEWS2000）(March 2001).
17) 廣瀬竜男，中西吉洋，秦　淑彦，田中克己：被写体の「写り具合」に基づく多視点映像の検索と表示，データベースと Web 情報システムに関する合同シンポジウム（DBWeb2000）(March 2001).

(平成 13 年 9 月 21 日受付)
(平成 14 年 1 月 31 日採録)

（担当編集委員　加藤 俊一）

中西　吉洋（学生会員）

2000 年神戸大学工学部情報知能工学科卒業．現在，同大学院自然科学研究科情報知能工学専攻博士前期課程に在籍中．マルチメディアデータベースに興味を持ち，ビデオデータベースに関する研究に従事している．

廣瀬　竜男

1999 年神戸大学工学部情報知能工学科卒業．2001 年同大学院自然科学研究科情報知能工学専攻修了．2001 年松下電器産業（株）入社，現在に至る．ビデオデータベース，コンピュータグラフィックスに興味を持つ．

田中　克己（正会員）

1974 年京都大学工学部情報工学科卒業．1976 年同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情報メディア科学専攻専任教授，2001 年京都大学大学院情報学研究科社会情報学専攻教授，現在に至る．工学博士．主にデータベースの研究に従事．人工知能学会，日本ソフトウェア科学会，IEEE Computer Society，ACM 等各会員．