ヒューマンインタフェース 67－8
（1996．7．11）

複数映像の空間統合による映像アクセスインタフェース

阿久津 明人，外村 佳伸，浜田 洋
NTTヒューマンインタフェース研究所

あらまし

複数映像の空間統合技術を用いた映像の撮影された空間の再合成による複数映像のアクセスインタフェースを提案し，その具体的な実現について述べる．本報告で実現している映像のユーザインタフェースは，映像が撮影された空間に関する構造情報を用いたコンテンツ，コンテクスト構造の可視化と映像の空間を用いた時間情報への直感的なアクセスを可能としたものである．

アクセスインタフェース上に実現した映像の表現は，撮影空間の広がりの表現と被写体の動き等の時間の経過・軌跡の表現である．また，再合成した撮影空間上での再生であるパノラマ再生，空間による指定と再生を行う空間領域再生，撮影空間にストロボ表現された被写体の指定と再生を行う被写体ランダム再生等によりユーザの映像とのインタラクションを実現した．

提案するユーザインタフェースにより，フレームを超えた広い撮影空間，ショット毎に分割・分離された被写体の撮影空間に対する絶対的な構造（配置，大きさ等），撮影空間に対する絶対的な動きの直感的な把握や理解支援を実現した．

Video Interface with video shot synthesis

Akihito AKUTSU, Yoshinobu TONOMURA and Hiroshi HAMADA
NTT Human Interface Laboratories

ABSTRACT

We propose a new video user interface based on synthesizing video shots and describe concrete implementation techniques. The video interface is realized using video contents and context structure extracted automatically. The video interface comprehensibly visualizes the video contents and context structure to allow us to access video information intuitively.

In this paper, camera operation information and the spatial relationship among video shot, are used to realize the new video interface. We visualize each video shot as a panoramic space and the trajectory object motion. We implement several space-based play back modes, 1) panoramic play back, 2)area play back and 3) object play back, for the structured videos in the video user interface. The proposed video interface can support the synthesis several video shots, and allows us to grasp the scenes intuitively.

1. はじめに

最近のコンピュータ技術の進展と高速広帯域ネットワークの発展により，コンピュータはもとよりネットワーク上で映像を処理し，伝送することが可能になってきた．このような高度にネットワーク化された情報社会における多量の映像情報に，インタラクティブ性を与え，ハンドリング可能にすることで，効率的且つ効果的に映像を利用する環境の実現が期待されている．ここで言うハンドリングとは，映像情報の効率的な検索，加工，編集であり，また個人の目的に応じた様々なスタイルでの映像情報の効果的な利用・鑑賞である．

映像情報をインタラクティブにハンドリングする場合，映像情報をいかに簡素化し構造的に扱うか，すなわち映像をいかに単位化しインデクスを付与して記述することで効率的に管理するかが課題である。また，構造化された映像情報をいかに人がハンドリングし易い形で表現し，効果的にインタラクションさせるのかも重要な課題である．

この課題に対して，映像の構造化に基づく映像のハンドリングに関する報告がなされてきた[1][2][3]．我々も映像解析（処理）技術を用いて自動抽出した物理的な映像情報（カット点，カメラ操作，音楽，音声情報等）に基づいて映像を構造的に扱い，ユーザの目的に応じたスタイルで映像に対してアクセス可能なインタフェースを提案してきた[4]．図1に映像の構造化による映像インタフェースの概念を示す．

映像解析により映像に含まれる物理的な情報を自動抽出する．自動抽出された情報を用いることで，シーケンシャルな映像を複数のセグメントで表現することが可能となる．さらにこのセグメントの内容を分析することでコンテンツ構造の抽出とその構造へのインデクス付けを行う．さらにセグメント間の関係（リンク関係）を分析することで映像のコンテキスト構造を表すリンク構造を抽出する．これらコンテンツ構造，コンテキスト構造に知識等を適用することで映像の意味的な構造を表現する可能性が広がる．最後にコンテンツ構造，コンテキスト構造をユーザの目的に応じて可視化し，ユーザに呈示することで，直感的かつ効率的な映像へのインタラクションを実現する．

本報告では，セグメント化された複数の映像から空間に関する情報を抽出し，複数のセグメント間の空間的なリンク関係を自動的に抽出する．次に，抽出した空間的なリンク関係を用いて映像を空間的に統合する．最後に映像の空間的統合を用いた，効率的且つ効果的な映像のアクセスインタフェースを提案しその具体的な実現方法を述べ，検討結果を示す．

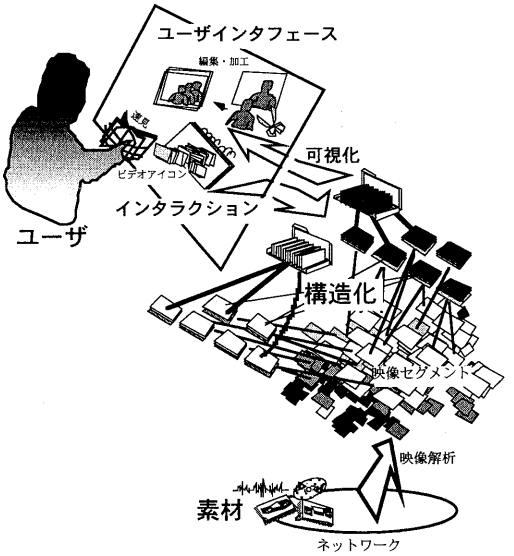

図1　映像の構造化と映像インタフェース

2. 映像のユーザインタフェース

映像のユーザインタフェースで用いるコンテンツ構造，コンテキスト構造は，ユーザの目的や対象とする映像の種類によって異なる．映像の構造化は，ドラマや映画のような制作者によって編集されたシーケンシャルな映像（セグメント列）を扱う構造化と素材映像のような断片映像の集合群（セグメント群）を扱う構造化とに大別できる．

2-1 で述べるビデオブラウザで用いる映像の構造化は，シーケンシャルな映像（セグメント列）を扱う構造化であり，ストーリーが持つ制作者の意図を反映した階層構造に基づいた構造化である．ユーザの目的の明確さに応じてそのスタイルが容易に選択・変化可能な個人のスタイルを反映した効率的な映像のブラウズの支援を目指している．

一方，2-2 で述べる映像空間統合型ユーザイン

タフェースで用いる映像の構造は，断片映像の集合群（セグメント群）を扱う構造化であり，セグメント間の関係を空間ベースでリンク付けする構造化である．複数のセグメント間にまたがる空間情報（被写体の配置，大きさ，絶対的な動き）の直感的な把握や理解の支援を目指している．

2-1 ビデオブラウザ

ビデオブラウザは、映像のコンテンツとコンテキスト構造を直感的にユーザへ呈示することで映像を効率よくブラウズさせることを支援するものである[5]．映像から自動抽出したショット情報を用いた映像コンテンツ構造の可視化は，各ショットを代表するキーフレームを選択しその画面を表示することや３次元（画像軸２次元+時間軸１次元）のビデオアイコンで表現する等で行われている．また，キーフレーム画面，ビデオアイコン等を２次元空間（ディスプレー）へ時間軸に沿って表示することでコンテンツ構造の可視化と共にコンテンツ構造の時間的な変化によるコンテクスト構造の表現も試みられている．我々が実現してきた幾つかの構造化に基づくインタラクティブな映像利用インタフェースを紹介する．

ペーパービデオは，映像のコンテント，コンテキスト情報を紙へ定着することを試みたビデオブラウザである．ビデオ中のシーンの変わり目を自動的に捕らえ，その情報を基にサンプルした画像の一覧を自動的に作成し紙へ出力するシステムである．紙メディアへ映像を変換することで，何時でも，何処でも手軽に映像の内容の把握を容易にしている[6]．

また，階層的な映像のランダムアクセスを可能にすることで大量の映像から情報を効率よく直感的に取得可能なインタフェースも実現した[7]．粒度の粗いアクセスは映像の持つ属性情報（放送時間，作成場所等）を用いて，細かい粒度のアクセスには自動抽出した構造情報（カット点，音楽，音声情報等）を用いて構造化した様々な観点でのコンテント，コンテキスト構造を可視化することで実現している．ユーザの目的の明確さに応じてそのスタイルが容易に選択・変化可能な個人のスタイルを反映したユーザインタフェースの実現である．

ビデオブラウザでは，ユーザの目的の明確さに応じて映像を効率良くブラウジングすることは可能であるが，各セグメントに分割・分離された映像情報を直感的に且つ統一的に把握・理解することは難しい．

2-2 映像空間統合型ユーザインタフェース

映像空間統合型ユーザインタフェースは，複数の映像に対して，空間ベースで映像を統合することにより，映像の内容や映像間の関係が直感的に把握可能なユーザインタフェースである．

映像空間統合型ユーザインタフェースで実現すべき映像の表現とインタラクションは，

1)撮影空間の広がりの表現，
2)被写体の動き等の時間の経過・軌跡の表現，
3)空間をインデクスとした時間情報へのランダムアクセス，

である．

1)では，映像を撮影空間をフレームの枠を超えた広い空間として表現すること，また，複数の映像に対してそれぞれの撮影空間を統合し，より広い空間として表現することで，ユーザの空間に対する把握や複数の映像間の空間位置関係を直感的に理解させることが実現できる．

また，2)では，被写体の動きを撮影空間（背景）に対する絶対的な動きとして撮影空間に定着表現することで，ユーザの時間情報に対する直感的把握や理解を支援することが実現できる．

映像へのインタラクションでは，再合成された空間や時間情報が定着された空間をインデクスとして用い，時間情報への直接的なランダムアクセスを実現することで，直感的に時間情報の選択と取得を支援することが可能である．

本報告で実現する映像空間統合型ユーザインタフェースの表現は，自動抽出したカメラ操作情報とショット間の空間位置関係情報を用いた，画像の再合成による撮影空間のパノラマ表現と，被写体の動きを撮影空間に定着させたストロボ表現である．また，ランダムアクセスでは，パノラマ表現，ストロボ表現された画像をインデクスとして用い，ユーザが所望する映像の時間位置の直感的且つ直接的な指定と再生による映像情報の取得を実現する．

3.空間構造情報の抽出法

本報告で実現するインタフェースで用いる映像の構造情報は，映像処理により自動抽出可能な映像の空間に関する情報である．具体的には，ショット内フレーム間の関係を表すカメラ操作情

報と複数のショット間の空間的な位置関係を表す情報である.

3-1 カメラ操作情報抽出法

映像から自動的に抽出するカメラ操作は, パンニング, チルティング, ズーミングとこれらの組

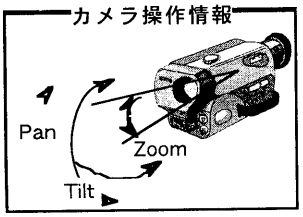

図2　抽出カメラ操作

み合わせ操作である（図2).

カメラ操作情報抽出法は, 時空間投影法とHough変換をベースとした映像解析である[8]. 映像を時空間画像（2次元画像軸+時間軸の3次元データ）として扱い, 時空間画像の投影方向への積分処理することで, 映像中の動き情報を強調表現する. 動きが強調された時空間画像が時空間投影画像である.

図3に時空間投影画像の例を示す. 上図がy軸方向を投影方向としたx-t時空間投影画像であり, 下図がx軸方向を投影方向としたy-t時空間投影画像である. 時空間投影画像上では, 動きは時間軸に沿った流れとして表現される. カメラ操作による動きは, フレーム全体（グローバル）に及ぶ流れのパターンを作る.

この時間に沿った流れを自動追跡しパラメータ化することでカメラ操作を自動的に抽出する.

カメラ操作による時空間投影画像上の流れをアフィン変換でモデル化し, 流れの時間変化の相関関係をHough変換しHough空間での最大投票値をグローバルな動き（カメラ操作による動き）として捉えカメラ操作をパラメータ化する.

抽出したカメラ操作パラメータは, パンニング, チルティング成分に対してはフレーム間での変位画素数, ズーミング成分に対しては, フレーム間でのフレームの大きさ（面積）の変化率である.

抽出誤差は, パンニング, チルティング成分に対しては数画素以下の誤差を含み, ズーミング成分に対しては, 1/10以下の誤差を含む. この誤差は, 後でフレームを空間ベースで統合する場合に大きな影響を及ぼさない程度のものである.

3-2 ショット間の空間位置関係情報抽出法

ショット間の空間位置関係を抽出するために画像間で空間的な対応関係を算出する. 対応をとる画像を P_1, P_2 で表し, 画像間の関係はアフィン変換でモデル化する. このモデルは, 焦点距離を変えずに撮影した映像から作成したパノラマ画像間で近似できる. 図4に示す様に P_1 上を P_2 の位置を変化させて次式で表す相互相関値を算出し, 最大の値 ρ_{max} を持つ大きさと位置を二つの空間的な関係パラメータとする[9].

$$\rho=\{\Sigma[P_1(x,y)-\overline{P_1}(x,y)]\Sigma(P_2(x+b,y+c)-\overline{P_2}(x+b,y+c))/\Sigma[(P_1-\overline{P_1})(P_2x+b,y+c-\overline{P_2}\ x+b,y+c)]\}\cdot A,$$

ここで, $\overline{P_1}$, $\overline{P_2}$ はそれぞれ P_1, P_2 の平均値を表す. (b, c)が空間位置を表すパラメータである. また, A は画像間での重なり領域の面積（画素数）である.

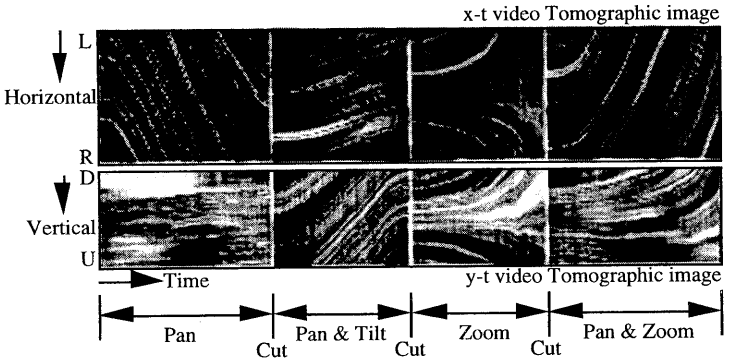

図3　時空間投影画像

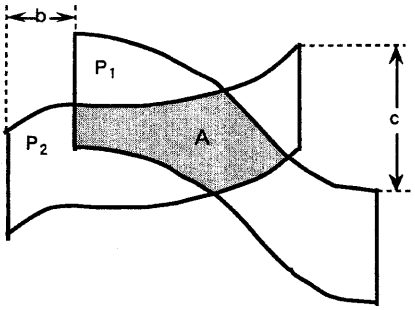

図4　パノラマ画像間の空間位置関係

3つ以上の画像間の空間関係の算出は，基準になる画像を一つ選び，基準画像にた対して逐次上記の方法で，空間関係を決定する．

本報告では，次章で述べるショット内フレーム間統合された撮影空間（パノラマ画像）間で対応をとった．一般の映像はカメラのレンズ系による中心射影変換であり，各フレーム画像は個々に変換による歪みを有する．空間的に離れた画像に対してアフィン変換モデルは成り立たないが，近接する画像に対してはその仮定が成立する．

一方，パノラマ画像は，円柱変換[10]された画像でありそれに伴う画像の歪みを個々のパノラマ画像で有する．しかし，個々のパノラマ画像が，大きなパノラマ画像の部分的な切り出し画像として仮定できるため，固定位置の一台のカメラで撮影された複数の映像から作成した個々のパノラマ画像に対しては，空間的に離れた画像に対してもアフィン変換モデルで近似可能である．

4.複数映像の空間統合によるユーザインタフェース

本報告で対象としている複数映像とは，固定位置の一台のカメラで撮影された複数の映像である．このカメラで撮影された複数の映像は空間的に共通する情報（領域）を持つが，撮影時間が異なるため被写体等の情報は異なる．この複数の映像を用いて空間ベースで統合した映像のユーザインタフェースの実現を試みた．

次に実現したインタフェースの具体的な検討結果を示す．

4-1 撮影空間の広がりの表現

映像から抽出したカメラ操作を用いてフレームの枠を越える撮影空間を再合成する方法を図5に示す．カメラ操作に応じてフレームの位置

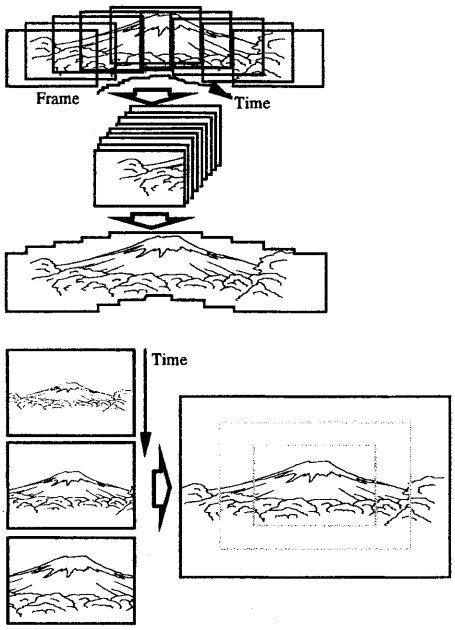

図5　撮影空間再合成方法

（上図）や大きさを変化（下図）させて撮影空間を再合成する．再合成された各々のフレームは，空間的に統合されており，我々は，このショット内フレーム間で空間統合された映像をパノラマビデオと呼んでいる．パノラマ写真は，空間のみの2次元データであるのに対して，パノラマビデオは空間と時間の情報を持つ3次元データである．

パンニング，チルティング操作で撮影された映像からカメラ操作を自動抽出し，カメラ操作情報を用いてフレーム間を再合成して作成した撮影空間を図6に示す．

フレーム間統合により時間情報に埋もれていた広い空間情報を明示することで，ショットを再生せずに直感的な撮影空間の把握や理解が可能になった．

次にカメラ操作情報に加えてショット間の空間的関係情報を用いることでパノラマビデオを越える撮影空間の再合成を試みる．

用いた複数の映像は，スキーのジャンプ台を固定位置の一台のカメラで撮影した映像である．9個のショットから再合成した撮影空間を図7に示す．各ショット毎に再生したのでは把握しずらい撮影空間の被写体間の絶対的な構造（ノーマルヒル，ラージヒル，ジャッジタワー等の位置関係や大きさ等）が直感的に把握可能になった．

図６　撮影空間

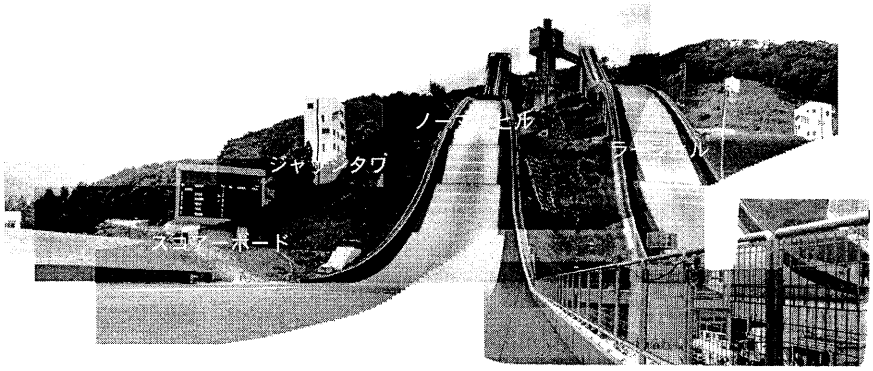

図７　ショット間の空間配置情報から再合成された撮影空間

4-2 被写体の動き等の時間の経過・軌跡の表現

次に撮影空間に被写体の動き等の時間情報を定着表現することを試みた．定着方法として自動的に抽出した注目する被写体の領域，位置情報を用いて，被写体をストロボ的に再合成された撮影空間上へ張り合わせることで実現した．図８に被写体のストロボ表現を示す．

この表現によりスキージャンプの選手動きが撮影空間（背景）に定着でき，撮影空間に対する絶対的な動きとして捉えることが可能になった．この表現によりスキージャンプの選手の動きと共にフォームの時間経過に対しても直感的に把握，理解することができる．

また，ショット間（パノラマビデオ間）の空間的関係情報を用い，同一空間で撮影時間の異なる複数の被写体のストロボ表現をオーバーラップさせることで，複数の被写体間の動きを比較表現することを試みた．２人の選手のストロボ表現をショット間の空間位置関係情報を用いてオーバーラップさせた画像を図９に示す．

この表現により，マルチモニタ等で比較したのでは得られない，他選手との空間内の絶対的な動きや連続的に変化するフォームの違い等の情報が直感的に把握可能になった．

図８　被写体のストロボ表現

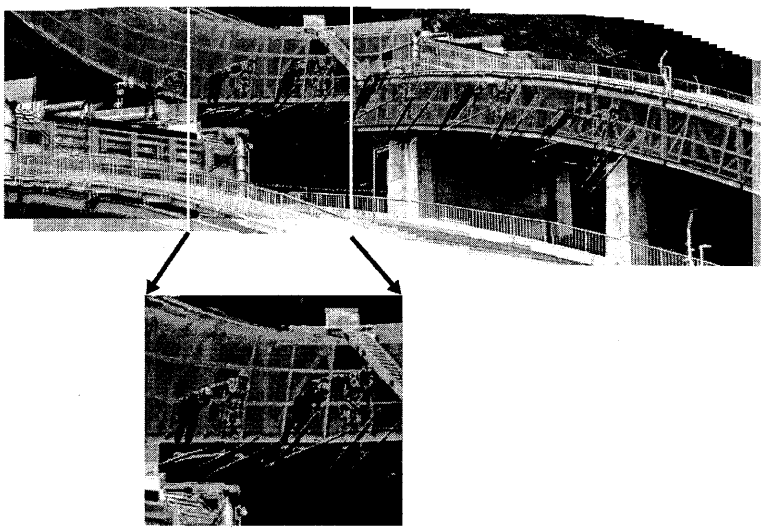

図9　空間へ定着された時間情報のオーバーラップ表現

4-3 空間統合された撮影空間を用いた時間情報へのインタラクション

映像の時間情報へのインタラクションでは，パノラマ表現，ストロボ表現，オーバーラップ表現をインデクスとして空間を指定することでタイムコード等を介さない直感的で直接的な映像へのアクセスを試みた．

ユーザのアクセスモードについて以下の 3 つのモードをインタフェースに設けた．空間統合された映像へのアクセスの様子を図 10 に示す．図 10 は，複数の映像が空間統合された 3 次元空間（2 次元の撮影空間軸+時間軸）の XT 平面である．パンニング操作で撮影された 2 つの映像を例にしている．

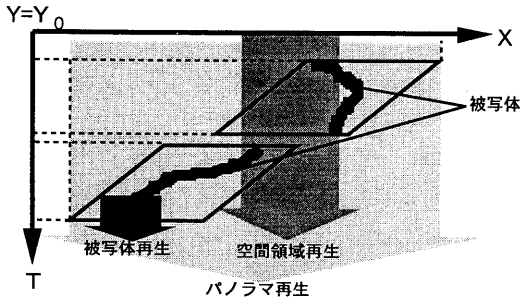

図10　時間情報へのインタラクションモード

実現した再生モードは，撮影空間全体を把握しながら再生可能なパノラマ再生，ユーザの見たい撮影空間の一部領域によるランダム指定とその領域のみの再生（空間領域再生），ストロボ表現を手がかりとした見たい被写体によるランダム指定とその被写体の再生（被写体再生）をそれぞれ実現した．

これらのアクセスにより，時間情報へのランダムアクセスがタイムコード等を介さない空間的にシームレスな操作になる．撮影空間を把握しながら，ユーザの所望する被写体の動き等に対する映像の再生時間位置が直接的に指定可能となることで，効率の良い，直感的な映像情報の取得が実現できた．

5. まとめ

複数映像の空間統合技術を用いた映像の撮影された空間の再合成による複数映像のアクセスインタフェースを提案し，その具体的な実現について述べた．本報告で実現した映像のユーザインタフェースにより，映像が撮影された空間に関する構造情報を用いたコンテンツ，コンテクスト構造の可視化と映像の空間を用いた映像の時間情報への直感的なアクセスを可能とした．

アクセスインタフェース上に実現した映像の表現は，撮影空間の広がりと被写体の動き等の時間の経過・軌跡の表現である．また，再合成した

撮影空間上での映像の再生であるパノラマ再生，空間を介した時間の指定と映像の再生である空間領域再生，撮影空間にストロボ表現された被写体を介した時間の指定と映像の再生である被写体ランダム再生によりユーザの映像との直感的なインタラクションを実現した．

提案するユーザインタフェースにより，フレームを超えた広い撮影空間，ショット毎に分割・分離された被写体の撮影空間に対する絶対的な構造（配置，大きさ等），撮影空間に対する絶対的な動きの直感的な把握や理解の支援が可能になったと考える．

今後の課題として，カメラの並進移動による撮影やマルチカメラ等で撮影された 3 次元空間の情報を含む映像に対する空間統合によるユーザインタフェースへの発展とデジタルライブラリ等の多量の映像情報に対する新たな映像の管理概念への応用があげられる．

謝辞

本研究を進めるにあたり，有益な御助言を頂きましたNTTヒューマンインタフェース研究所映像処理研究部徳永幸生部長，御指導，御討論頂いた映像処理研究部の各位に深く感謝致します．

参考文献

[1]外村，安部：”動画像データベースハンドリングに関する検討”，信学技報，IE89-33（1989）．
[2]H. Ueda, T. Miyatake and S. Yoshizawa :"IMPACT: An Interactive Natural-Motion-Picture Dedicated Multimedia Authoring System", Proceedings of CHI'91, pp.343-350 (1991).
[3]S. W. Smoliar and H.J. Zhang :"Content-Based Video Indexing and Retrieval", IEEE MultiMedia Magazine, summer, pp. 62-72 (1994).
[4]Y. Tonomura, A. Akutsu, Y. taniguchi and G. Suzuki:"Structured Video Computing", IEEE Multimedia, Vol. 1, No. 3, pp. 34-43(1994).
[5]浜田，阿久津：“ビデオの高次利用を実現するユーザインタフェース”，1996TV 学会年次大会論文集 (1996).
[6]外村，谷口，阿久津：“PaperVideo:紙を用いた新しい映像インタフェース”，信学技報，IE94-59, pp.15-20(1994-09).
[7]Y. Taniguchi, A. Akutsu, Y. Tonomura and H. Hamada:"An intuitive and efficient access interface to real-time incoming video based on automatic indexing", Proc. of ACM Multimedia '95, pp. 25-33(1995).
[8]A. Akutsu, Y. Tonomura, Video Tomography: An efficient method for Camerawork Extraction and Motion Analysis, Proceedings of ACM Multimedia 94, pp. 349-3566(1994).
[9]阿久津，浜田“VideoJigsaw:映像群からのシーン再構築方法”，1995 年信学ソサイエティ大会論文集（1995）．
[10]J. Y. Zheng, M. Asada and S. Tsuji :"Color-based panoramic representation of outdoor environment for a mobile robot", In Proc. 9th ICPR, pp.801-803, Rome, Italy (1988).