データベースシステム　127－17
情　報　学　基　礎　　67－17
（２００２．５．２２）

画像情報とテキスト情報を統合的に利用したインタラクティブな映像検索システム

椎谷秀一，馬場孝之，遠藤進，上原祐介，増本大器，長田茂美

(株)富士通研究所　ＩＴメディア研究所

{shiitani,baba-t,endou.susumu-02,yuehara,masumoto.daiki,nagata.shigemi}@jp.fujitsu.com

計算機性能の向上とブロードバンドの普及により，映像コンテンツの量が急速に増え続けている．本稿では，これら大量の映像コンテンツから目的の映像やシーンを効率的に検索するための手法について説明する．本手法では，映像やシーンの内容を表す画像を，その画像から抽出した色や形状などの特徴量の似ているものが近くに集まるように配置する．これにより，ユーザは目的の映像やシーンに似ているものが集まっている付近を重点的に探索し，目的の映像やシーンを直感的かつ効率的に探すことができる．テレビ映像を対象にしたシーンの検索実験を行った結果，早送り再生で検索する場合に比べ短時間で目的のシーンを探し出すことができ，本手法の有効性が確認できた．

Interactive video retrieval system integrating visual search with textual search

Shuichi Shiitani, Takayuki Baba, Susumu Endo, Yusuke Uehara,

Daiki Masumoto and Shigemi Nagata

IT MEDIA LABORATORIES, FUJITSU LABORATORIES LTD.

The performance of a computer improves and the broadband has spread recently. Therefore, the quantity of video contents is continuing increasing quickly. In this paper, we explain the technique for searching the target video or scene efficiently from a lot of video contents. This technique arrange the image showing the contents of a video so that similar images get closer. A user looks around the area in which the similar images has gathered, and can search the target video intuitively and efficiently. We verified the effect of the method by the experiment of the scene retrieval for a television.

1. はじめに

昨今，我々はデジタルビデオカメラや DVD などによりデジタル映像を利用する機会が増えてきている. 映像の流通･放送においても, CS, BS デジタル放送やインターネットにおけるストリーミング配信などのようにデジタル化されつつある．さらに映像の制作現場でも，素材映像をデジタルで保存し，ノンリニア映像編集をすることは必要不可欠になっている．このように，利用・流通・制作すべての場面においてデジタル映像が普及している.

また，デジタル映像編集やテレビ映像のキャプチャなども，一般のユーザがパソコンで簡単に利用できるようになっており，パーソナルなデジタル映像の量も増えている.

このように大量のデジタル映像が存在するようになると，それらの中から目的の映像を検索する必要がでてくる.

検索方法としては，関連するキーワードで絞り込むのが一般的である．そのためにはあらかじめデジタル映像に検索キーとなるメタ情報を付与しておく必要があるが，このメタ情報は人手で入力しなければならない．検索時に入力するキーワードを想定して映像に合ったメタ情報を入力することは，家庭で撮影したデジタルビデオのように数が少なく利用者も限られている場合には可能だが，映像制作用の素材などの数が多く適切な情報が求められる場合はたいへん手間がかかるとともに，重要なキーワードの記述漏れなど情報の精度に問題が生じる.

このメタ情報入力に関する問題を解決するため，映像内の音声やテロップ文字などを自動的に認識し，メタ情報として利用する研究も行われている[1]. しかし音声認識は未だ研究レベルであり，さまざまな環境において実用に耐える精度は得られていない．テロップ文字は一部のショットにのみ記述されており，かつ簡潔に書かれているため，検索キーワードに対応するのに充分なメタ情報を得ることはできない.

一方, 我々はこれまでに, 画像を一覧表示し，それをユーザが眺めて目的の画像を探し出すマルチメディア情報検索システム MIRACLES (Multimedia Information RetrievAl, CLassification, and Exploration System)の研究開発を行ってきた[2][3]. MIRACLES では，画像から色や形状といった画像特徴量を抽出し，その特徴量が似ている画像が近くに集まるように仮想三次元空間に配置する．ユーザはその三次元空間を動き回り，目的の画像に似ているものが集まっている付近を重点的に探索することで，目的の画像を直感的かつ効率的に探すことができる.

この方法では，人間の検索能力を利用し，計算機はそのサポートをすることで，計算機だけあるいは人間だけでは困難な検索を可能としている．本稿では，この検索方法を映像に対応させた，映像を直感的かつ効率的に検索できるシステムについて説明する.

2. MIRACLES

まず，今回の映像検索システムの基となる，MIRACLES について説明する.

MIRACLES はテキストによる意味的検索と画像による視覚的検索とを兼ね備えたクロスメディア検索技術をベースにしており，クローラによる情報収集，情報の類似性に基づく配置，インタラクティブな情報検索といった機能を持つ.

ここでは MIRACLES の応用の一つである Web 検索を例に，それぞれの機能について説明する.

2.1. クローラによる情報収集

まず始めに，ユーザは欲しい情報が掲載されている Web ページの URL，あるいは欲しい情報に関するキーワードを入力する．キーワードが指定された場合はそのキーワードを外部のテキスト検索エンジンに渡し，結果として得られた URL を起点とする．クローラは起点ページからページ内に埋め込まれたアンカーを辿ることによってページを巡回する.

巡回した各ページにおいて，クローラはそのページにある画像と，その画像の周辺にあるテキストをペアにして収集する．一般に画像の近

くにあるテキストはその画像に関連しているテキストであると考えられるため，これらを画像の関連テキストとして収集し，画像を意味的に検索する際の情報として利用する．

収集した画像・関連テキストからは，各種特徴量を抽出する．画像特徴の代表的なものとして，画像の各画素を HSI 色座標に変換し HSI 空間を格子状にブロックに分割して各ブロックに含まれる画素数をカウントした HSI ヒストグラム特徴や，画像の輝度値を Wavelet 変換して画像の大まかな形状成分や細かな模様成分に分離する Wavelet 特徴などがある[4]. 関連テキストからは単語の出現頻度をベースにした単語頻度特徴を抽出する[3].

2.2. 情報の類似性に基づく配置

MIRACLES は，収集した画像を特徴量が似ている画像が近くに集まるように平面に配置する．

この配置には自己組織化マップを用いている[5]．自己組織化マップでは，データの分布を把握できるように高次元の特徴ベクトル空間を低次元空間に写像する．このとき，高次元空間における分布の状態を低次元空間においても保存するように配置する．

図 1 に HSI ヒストグラム特徴を利用して配置した画面例を示す．配置の結果，黄色いバッグや赤いバッグなど同じ色のバッグが集まって配置されている．もしユーザが赤いバッグを探しているならば，表示されている画像すべてをチェックしなくても，赤いバッグが集まっている付近だけに着目すればよく，直感的かつ効率的に探し出すことができる．

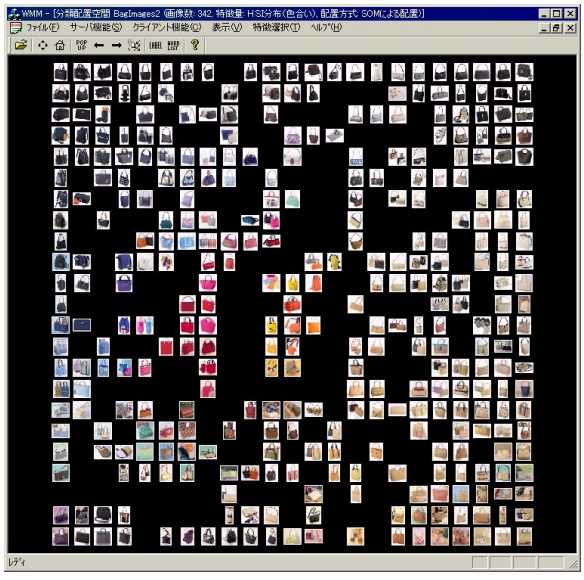

図 1 HSI ヒストグラム特徴による配置例

2.3. インタラクティブな情報検索

画像群は，仮想的な三次元空間に配置されており，最初の視点は図 1 のように画像群全体を見渡せる位置に設定されている．ユーザはこの三次元空間内をフライスルーすることによって画像に近づいてより詳細な内容を確認できる．さらには，配置の基準となる特徴を変更して，目的に合った配置を選ぶことも可能である．

また，ユーザはこれらの画像からキーワードを入力することで検索対象を絞り込むこともできる．キーワードを入力すると，収集時に画像とペアで収集した関連テキスト内にそのキーワードを含む画像だけが図 2 のように手前に浮き出てポップアップされる．ユーザはポップアップされた画像だけを対象に検索すればよいので効率的である．

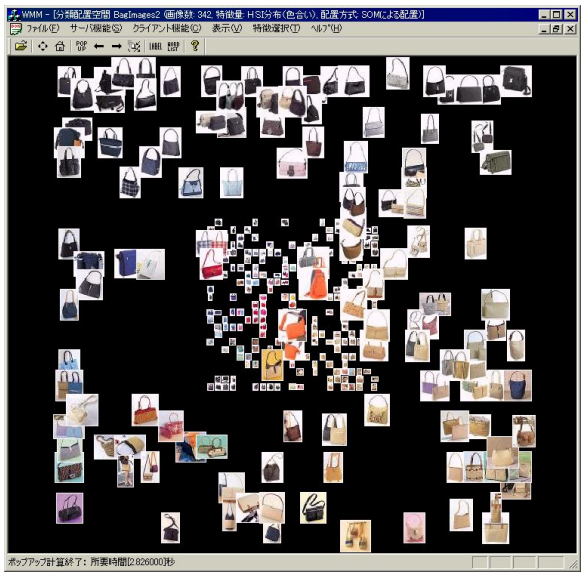

図 2 ポップアップした例

以上のような操作によって探し出した画像をユーザが選択すると，その画像が掲載されていた Web ページを表示することができる．

このように MIRACLES では，テキストと画像をそれぞれが補完しあうような形で利用できるので，ユーザはそれぞれの情報を使い分けながら，目的の情報を探し出していくことができる．

3. 映像検索

前章で述べたような，大量の情報をユーザにわかりやすく提示するという手法は，そのまま映像の検索にも適用できる．映像の検索は大きく，個々の映像コンテンツを検索する場合と，映像コンテンツ内の特定のシーンを検索する場合とに分けられる．ユーザは大量の映像コンテンツの中から目的の映像コンテンツを検索し，その後にその映像コンテンツから見たいシーンを検索する．以下では，それぞれの検索について前章の検索手法を適用する方法を説明する．

3.1. 映像コンテンツの検索

MIRACLES では，ユーザが目的の画像を探し出せるように大量の画像を同時にユーザに提示する．映像の場合も同様に大量の映像を同時にユーザに提示する必要がある．

我々は大量の映像を同時にユーザに提示するために，二つの方法を検討している．一つ目は複数の映像を再生しながら同時にユーザに提示する方法であり，二つ目は映像をその映像の内容を表す画像と関連づけ，その画像をユーザに提示する方法である．

それぞれについて以下に詳しく説明する．

3.1.1. 映像の同時再生

第一の方法は複数の映像を再生しながらユーザに提示するものである．図 3 に縦横に映像を並べて表示し，同時再生している画面を示す．このように複数の映像を同時に表示することで，ユーザはこれらの映像を見比べ，どれが自分が探している映像なのかを把握し，その映像を選択する．

この方法は一本一本を再生して内容を確認していくのに比べ，同時に多くの映像の内容を把握できるので，検索時間が少なくて済む．そのため，CS 放送やケーブルテレビのチャンネル選択などによく使われている．

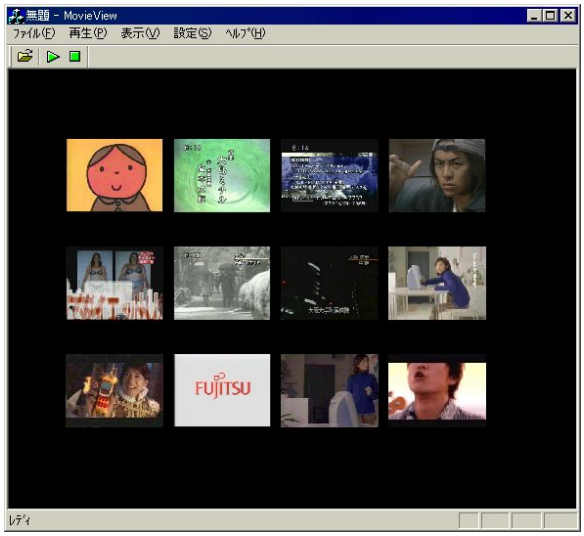

図 3 映像同時再生画面

しかしこの方法でも映像の数が増えてくると，どこにどのような映像があるかを瞬時に判断することができず，ひとつひとつの映像を順に確認していかなくてはならない．

そこで，図 4 のようにらせん状に映像を配置し，映像が再生しながら流れていくような表示方法を実現した．映像の位置が自動的に動いていくので，ユーザはらせん中央近くに来た映像だけに着目すれば，楽に順次映像の内容を確認することができる．

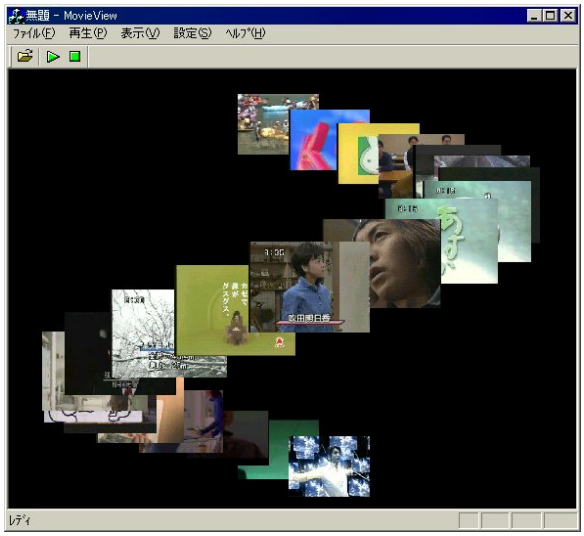

図 4 らせん状映像同時再生画面

また，映像の時系列を踏まえた特徴量を抽出し，その特徴量にしたがって自己組織化マップで画像の場合と同様に配置し，同時再生する方法が考えられる．しかし今のところ映像をユーザにわかりやすいように配置するための特徴は定義できていない．今後，配置に適した映像特徴を検討していく必要がある．

3.1.2. 代表画像の一覧表示

第二の方法は映像の内容を表す画像をユーザに提示するものである．映像は再生しなければその内容を把握できない．一方で画像は一瞬で内容を識別できるという特徴を持つため，二次元に配置してユーザに提示する際の視認性に優れていると考えられる．

そこで映像の代わりにその映像の内容を表す画像をユーザに提示すれば，ユーザは瞬時に複数の映像を把握できる．このとき映像の代わりに表示する画像を映像の代表画像と呼ぶ．代表画像は特徴によって配置されるため，ユーザが検索するのに適した画像である必要がある．また，用意した代表画像が映像の内容をうまく反映しているほど，検索効率の向上が期待できる．

代表画像は映像に関連する別の画像である場合も，映像内の任意のフレーム画像である場合も考えられる．例えば映画の検索を考えた場合，画像はパッケージやポスターのような，ユーザがよく目にして記憶に残っているものが望ましい．一方，映像制作現場での素材映像の検索の場合は，その素材映像内のフレーム画像を見ることでその映像の内容を詳しく把握することができ，欲しい素材を効率的に検索できる．

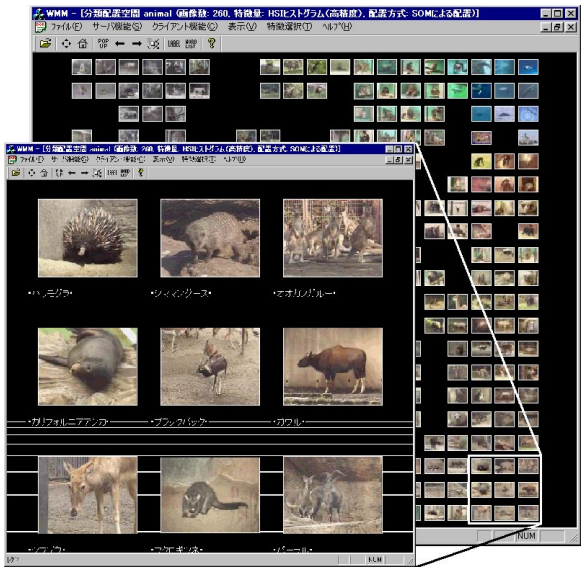

図 5 動物映像の配置例

図 5 は動物の映像を代表画像で配置した例である．映像は素材集のものであるが，映像の先頭フレームは映像の内容を表しておらず，代表画像として利用するには問題がある．またメタ情報は付与されていない．そこで MIRACLES を利用して動物の画像を Web ページから収集し，それぞれの画像に映像を人手で対応付けた．MIRACLES で収集したため，各画像には関連テキストが付与されており，それらの関連テキストで配置することも可能である．例えば関連テキスト内の動物の生息地情報を利用して配置すると，図 6 のようになる．

このようにして映像そのものではなく，映像に関連する画像を一覧表示することにより，容易に検索することができる．

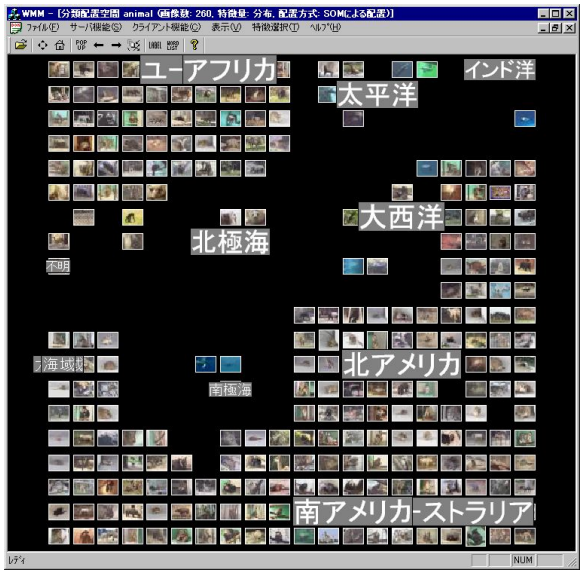

図 6 生息地情報による動物映像の配置例

3.2. 映像シーンの検索

映像内のシーンを検索するためには，映像の内容をユーザに提示する必要がある．我々は映像の内容の提示方法として，映像からカットを検出し，そのカット画像を提示する方法と，一

定時間ごとに抽出したフレーム画像を提示する方法を検討した．

図 7 は映像から検出したカット画像を時間順に一覧表示したものである．フレーム画像を４×４に領域分割したそれぞれの部分画像の色ヒストグラム特徴量の差を計算し，そのうち値の小さい８つの総和を評価値とする，分割 χ自乗検定法によってカットを検出した[6].

図 7 カット画像の時間順表示

このようにカット画像を並べると，その映像がどのようなシーンから構成されているか，どのような順序でシーンが並んでいるかが一目で把握できる．そのために，映像を再生あるいは早送りして映像の内容を閲覧しなくても，映像の内容を検索することが可能となる．

また，画像検索と同様にカット画像の特徴によって配置することも可能である（図 8）．このように配置することでさらに目的のシーンが含まれる部分を映像全体から絞り込むことができ，効率的に検索することができる．

図 9 は一定時間ごとに抽出したフレーム画像を一覧表示した画面例である．ここではフレーム画像をフィルムのイメージで配置している．ユーザはこの空間内を動き回り，映像の内容を把握することができる．このように表示することで映像の長い区間を同時に見て把握できるので，映像を早送りするのに比べてより簡単に内容を理解できる．

また，この場合にもカット画像と同様に時間順に配置したり，特徴によって配置することもできる．ユーザはそのときの検索に合った方法で配置することで，効率的に検索できる．

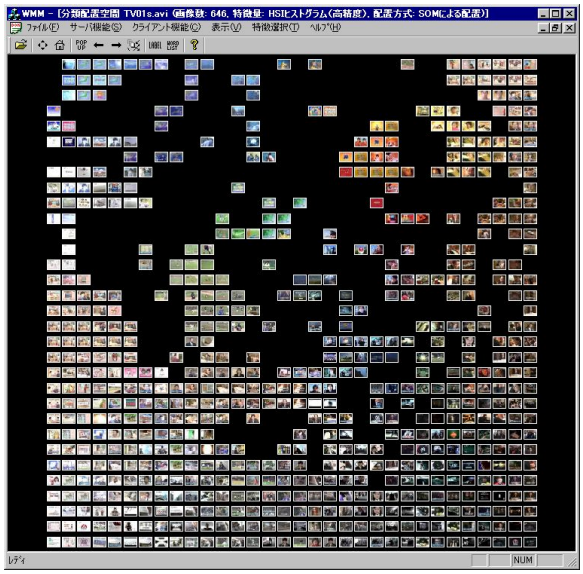

図 8 カット画像の HSI ヒストグラム配置

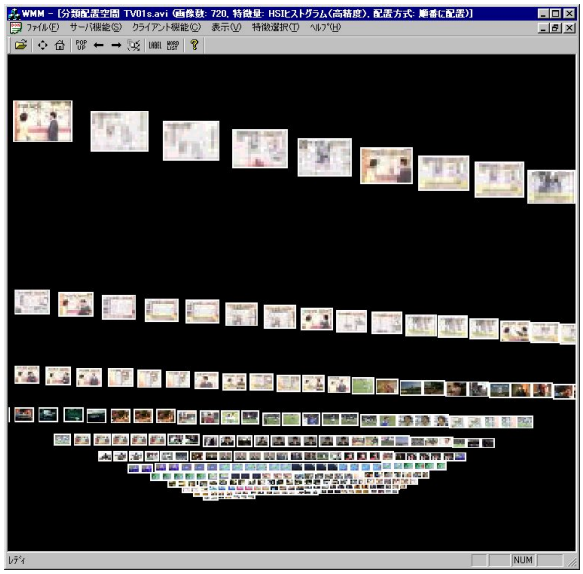

図 9 一定間隔フレーム画像のフィルム状表示

4. 実験

4.1. 実験データと方式

前章で述べた映像検索手法の有効性を確かめるために実験を行った．テレビを録画した 1 時間の映像を準備し，その映像の中から指定したシーンを探し出すときの所要時間を測定した．

実験に用いた映像は日本テレビの朝 5 時 30 分からの 1 時間で，天気予報やニュースの番組が続き，比較的多くの CM が流れているものである．この映像をカット検出した結果の全 646 カット画像を色ヒストグラム特徴で配置した．PentiumIII700MHz の計算機でカット検出にかかった時間は約 14 分，自己組織化マップによる配置計算時間は約 1 分であった．

この映像から，次の三つのシーンを被験者に検索してもらった．ここではキーワードは使わず，画像をキーとして検索することとした．

A．コーヒーの CM

B．広末涼子が映っているシーン

C．イチローが映っているシーン

コーヒーの CM は，あらかじめキー映像として CM 映像を被験者に見せ，それと同じ CM を検索してもらった．検索前にはっきりとイメージを作ることで，画像をキーとして検索することが容易になると考えられる．

広末涼子のシーンは映像中のとある CM に 1 回のみ登場する．カット検出の結果，いくつかのカット画像にアップで映っている．この場合はコーヒーの CM とは異なり，どのような画像のシーンかはユーザにあらかじめ指示しない．

イチローのシーンはスポーツニュースの 1 シーンに登場している．イチローの場合には野球のシーンを想像すれば検索しやすいものの，カット画像には小さく映っているだけでそれを見ただけではイチローと判断することは難しい．

4.2. 実験結果と考察

表 1 に各被験者が検索に要した時間を示した．

1 時間の映像から任意のシーンを検索する場合，例えばシーンをはっきり視認できる 3 倍速で早送りしながら検索すると，シーンの映像内の位置によるが，平均で 10 分かかる計算になる．実験の結果を見ると遅くても 5 分で検索できており，このようにカット画像を一覧表示する検索方法は早送りによる検索に比べて検索効率が高いことがわかる．

表 1 実験結果

被験者	A	B	C
	時間	時間	時間
1	0’51”	1’15”	1’28”
2	0’30”	3’24”	4’19”
3	0’44”	2’51”	2’03”
4	0’35”	1’35”	4’08”

以下，それぞれの場合について，被験者の行動について考察する．

コーヒーのＣＭの検索の場合，被験者はあらかじめ見た映像の中で印象に残ったシーンと似た色が集まっている部分に着目し，その部分を中心に動き回って探すという行動が見られた．その結果，646 のカット画像の中から目的の CM のカットを 1 分以内で探し出している．

広末涼子のシーンの場合は，全体を見渡して人物が映っている画像を探し，そこに近づくという動作をくり返して探したり，各カット画像の内容が認識できる大きさになるまで接近し，顔の映っているものだけを順に見ていく方法で探していた．646 カットを順に見ていくのは骨の折れる作業であるが，同時に多くの画像を把握できるので早送りで検索するのに比べると短い時間で検索できていることがわかる．

イチローのシーンの場合は，野球のシーンは背景が緑である場合が多いので，全体のカット画像から野球選手が映っているカット画像を探すことは容易であった．しかしカット画像にアップで映っていなかったため，その先の探索は困難であった．このトライアンドエラーにどの被験者もかなりの時間を費やしている．被験者によっては同じ球団の選手である佐々木投手や，同じくメジャーリーグで活躍している小宮山投手のカット画像を見つけ，フレーム順に並べ替えてそのカット画像の前後を見ることで探していた．

今回の実験では，検索するシーンのイメージがはっきりしている場合にはそうでない場合に比べて短時間で検索できている．しかし HSI ヒストグラム特徴で配置しおおまかな色で検索した場合とフレーム順で一覧配置して検索した場合とでどのような違いがあったかはわからない．今後はそのような違いを確認するための比較実験を行っていく必要がある．

5. まとめ

本稿では，映像や映像の代表画像を一覧表示して，その中からユーザが目的の映像を探し出すという検索方法について述べた．また，テレビ映像から特定のシーンを検索する実験を行い，人間の持つ検索能力を十分に発揮できるよう計算機でサポートすることによって，効率的で実用的な検索を実現できることを確認した．

今後は映像の内容を自動的に解析し，ユーザがより探し出しやすいような配置を作成することを検討していく予定である．

参考文献

[1] 古山浩志，八塩仁，江村恒一，井上郁夫，遠藤充，星見昌克：“ 音声認識とメタデータを利用した映像検索システムの開発 ”，信学技報 PRMU，Vol.99，NO.181，pp.67-72，1999.

[2] 長田茂美，遠藤進，椎谷秀一，上原祐介，増本大器：”マルチメディア情報検索システム“ MIRACLES ””，人文科学とコンピュータシンポジウム 2001 論文集，pp.267-274，2001.

[3] 遠藤進，椎谷秀一，上原祐介，増本大器，田茂美：“ テキストによる意味的な検索と画像による視覚的な検索を統合したマルチメディア検索システム MIRACLES ”，DBWeb2001，IPSJ Symposium Series ，Vol.2001，No.17，pp.249-256，2001.

[4] 村尾晃平，安藤淳禎：“ 画像をキーとする類似画像検索システム ”，1998 年電子情報通信学会 情報・システムソサイエティ大会，D-11-60，p.175，1998

[5] T.コホーネン著，徳高平蔵，岸田悟，藤村喜久郎訳：“ 自己組織化マップ ”，シュプリンガー・フェアラーク東京，1996.

[6] 長坂晃朗，田中譲：“ カラービデオ映像における自動索引付け法と物体探索法 ”，情処論，Vol.33，No.4，pp.543-550，1992.

[7] Rainer Lienhart，Silvia Pfeiffer，Wolfgang Effelsberg ：“ VIDEO ABSTRACTING ”，COMMUNICATIONS OF THE ACM，Vol.40，No.12，pp.55-62，1997.

[8] J.Boreczky，A.Girgensohn，G.Golovchinsky，S.Uchihashi ：“ An Interactive Comic Book Presentation for Exploring Video ”，CHI2000 Conference Proceedings ， ACM Press ，pp.185-192，2000.