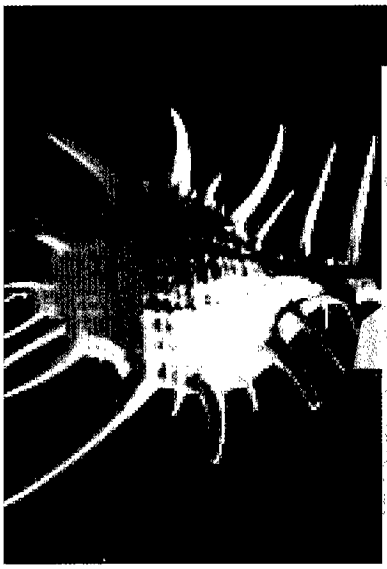


視覚的な手がかりによる動画像検索

ディジタル放送の開始とネットワークの広帯域化によりディジタル映像の流通が，数年内で急激に浸透することが予想される．そのとき，氾濫する映像データに対し，ディジタル化の利点を生かした映像の管理が重要課題である．映像を検索，再利用するには，映像の内容に基づく管理手法が必要で，テキスト中心のデータのファイル管理より，はるかに困難な状況にある．ファイルを見失ったとき，ファイルのタイトルや生成した日時をたよりに検索を行い，ファイルを開いてその内容を確認するが，映像では，内容検索とともに検索結果を有効に表示する技術が必要となる．そこで本稿では，映像管理のための情報抽出と表現，そして応用について解説する．

日本アイ・ビー・エム（株）東京基礎研究所
越後 富夫

■ 従来の画像検索手法 ——

映像や音声などのマルチメディアコンテンツは，パソコンの普及とともに一般家庭にも浸透したが，コンテンツの蓄積が高価であったため，管理方法についてこれまで大きな問題にはならなかった．しかし，圧縮技術の発展とDVD・大容量HDDなどの安価なメディアの普及，および高速なネットワークのおかげで，大量のデータが蓄積・配信可能となり，その中から特定の内容を含むコンテンツを検索し，再利用する要求が高まってきた．たとえば1台あたり75GBの容量を持つ最新のHDDでは，現行テレビ品質と同等の5MbpsでMPEG-2圧縮した2時間の映画を，16本収録することができる．また，ADSLを利用すると，MPEG-1ビデオをリアルタイムで視聴も可能である．このように，これまでテキストに対して行ってきたように，ネットワーク上でコンテンツを検索し，視聴することが実現可能になってきた．

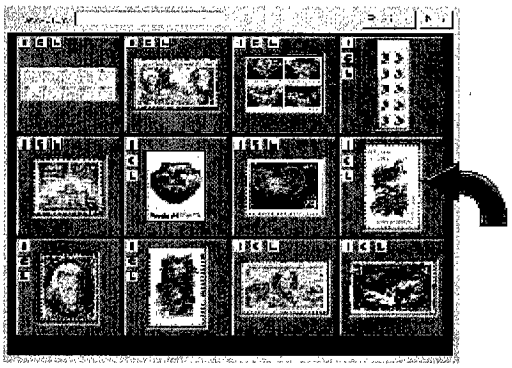
データ検索において，テキスト中心のコンテンツの場合，ユーザが検索したいキーワードはコンテンツの一部と一致することが多いが，静止画や動画像コンテンツの場合，オリジナルデータから直接的に内容を識別するのは困難で，ユーザに何をキーワード（検索条件）として問い合わせてもらえばいいのか，明確な回答はない．そのため，ユーザの求める検索・収集に必要となる，コンテンツに関する情報記述（メタデータ）が重要であると考えられている．一方，検索結果が動画像である場合，検索結果をすべて視聴しなければ結果の良否が判断できないのでは，ユーザに負担であり，結果を効果的に表現したり，短時間に要約する技術は，さらにユーザの興味をひきつけるためにも重要であると考えられる．

映像は多義の解釈が可能であるため，汎用的な処理から，計算機が内容を適切に解釈することは非常に困難である．映像は，一連の同じ場面（シーン）では，連続する画像の相関が強く，画像の変化分はわずかしかないが，シーンが変更されるとその前後で相関が急に小さくなる．動画像処理では，シーンをひとかたまりで扱うことが多く，そのシーンの代表画像である静止画を検索対象にする手法がある．静止画の検索でよく引用されるQBIC[1]は，色の分布，テクスチャ，位置による色分布，領域の輪郭形状が検索データとして利用でき，検索手法も図-1のような例示画像，スケッチなどが利用できる．たとえば，図-1（a）の矢印の画像を例示画像として入力すると，この画像の形状と構造および色の分布が類似した画像が検索され，結果として図-1（b）の画像が得

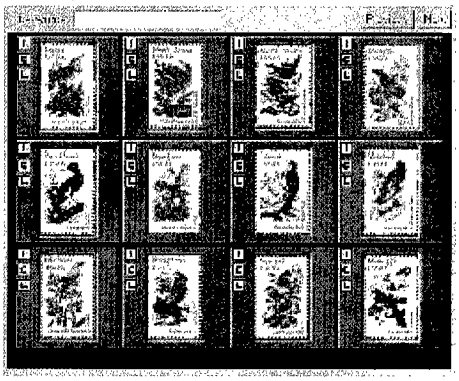
られる．QBICで扱うデータは，汎用的で，どのような映像コンテンツに対しても利用できるものであるが，検索を目的とするユーザにとって入力が容易ではなく，検索結果の良否を理解しにくいため，汎用的な処理だけで映像管理を行うには限界があった．

そこで，コンテンツの種類によって，利用するデータを使い分けることが考えられる．たとえば，映画やドラマでは，シナリオに基づいて制作されるため，画像処理の目的は映像解釈ではなく，シナリオのテキスト情報と映像をマッチングすることが主要課題である．この場合，登場人物と台詞が分かっているので，人物の顔認識，発話の有無によってシナリオと映像の同期をとることができる．

一方，蓄積する題材として膨大な量があり，かつ部分的な再利用率が非常に高く，検索ニーズの高いニュース映像では，ドラマのようにシナリオは完全ではないが，制作形式が一定しており，信頼の高い多くの種類の情報を統合することで，映像の大まかな構造を推定できる．たとえば，アナウンサーの音声，テロップ，クローズドキャプションなどの映像の注釈，映像の特徴抽出（顔の認識など）を使って検索などの応用が可能になる．Informediaプロジェクトでは，Video Skimming[2]と呼ばれる要約映像生成方法やName-Itと呼ばれる顔画像と名前の関連データベース生成がある．



(a) 矢印の画像をクリック



(b) 類似画像検索結果

図-1　QBICにおける例示画像検索

次に重要なコンテンツとしてスポーツ映像がある．スポーツ映像では利用できる注釈は非常に少なく，アナウンサーの発話や観衆の声援などの情報はあるが，コンテンツを最も正確に表現しているのは映像である．スポーツは，競技により特異な動きがあり，選手の形状変化が重要な情報となる．また，ルール，競技場の大きさ，カメラの配置，選手のユニフォーム，出場選手，競技のフォーメーションなどの事前知識が利用できる．そこで，選手をオブジェクトと定義し，映像からオブジェクトを抽出することで，1人または複数人のオブジェクトの動きを利用して，映像の解釈を行う．映像は，時間的な経過を情報として持っているので，ユーザに映像を表示するだけでなく，ユーザが映像だけでは見ることのなかった，新たなデータ解釈の一面を提示することができ，ユーザの新たな興味を引き出すことができる．

■ 動画像検索のための標準化 ——

音声や映像を含む，さまざまなタイプのマルチメディアコンテンツを，ユーザの求める検索・収集・表示・加工に必要となる，コンテンツに関するメタデータを標準化する活動がMPEG-7という名称で行われている．MPEG-7はAV情報の標準的表現であるが，符号化を含めコンテンツの表現には依存せず，応用のための参照用情報を提供することにある．映像コンテンツが持つ情報は，形・大きさ・色・位置などの特徴や，符号化形式・著作権・タイトル分類などの属性記述によって定義される．MPEG-7の具体的目標は，さまざまなタイプのマルチメディア情報を記述するのに必要な記述子の標準セットを規定することであり，特徴抽出や検索エンジンは標準化の範囲外としている．したがって標準を遵守する限り，特徴抽出を自動化もしくは人手で行ってもよい．また，コンテンツがどのような属性記述を持つかは，応用に依存し，同じコンテンツであっても応用によって，異なる属性記述が割り当てられることがある．標準化しようとしている項目は，記述子・記述スキーマ・記述定義言語と呼ばれるものである．記述子は，たとえば色や形などの属性をどのように表現するかを定義するもので，ある領域の色をRGBで表現し，R（120）G（100）B（90）と記述するように定義したものである．また，記述スキーマは，コンテンツにどのような属性が用いられ，そのときの記述子が何で，複数の記述子および属性間の構造はどのようなものかを定義するものである．たとえば「芝生の上を走る子供」の表現は，背景の芝生の属性として，色，テクスチャ，そして子供の属性として，服の色，動き属性などがあり，その属性に対する記述子の構造を定義する．記述定義言語は，記述子・記述スキー

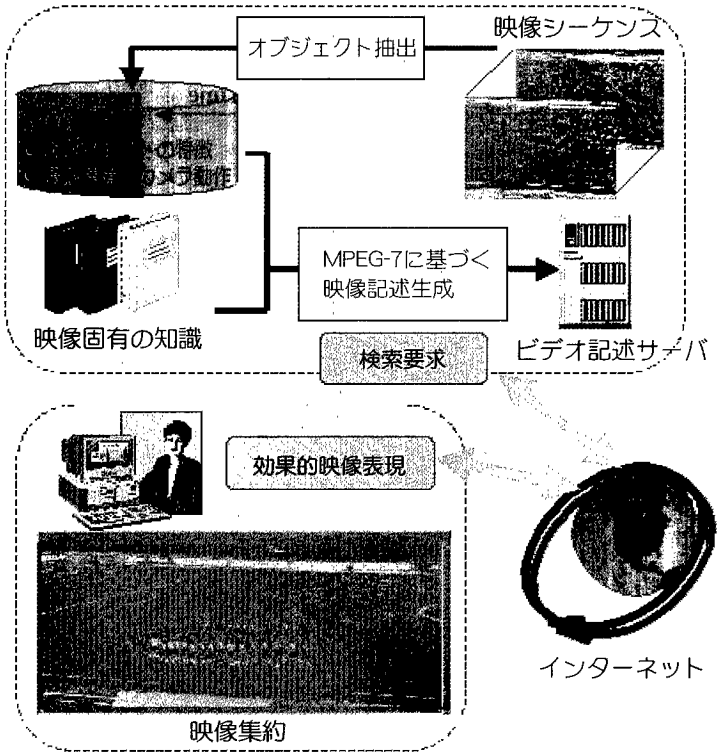


図-2　Video Enrichmentの構成

マからオブジェクト・イベントの構成を生成するための言語で，新規のオブジェクト・イベントを生成することができる．

計画では，2000年10月にCommittee Draft，2001年9月に最終的なInternational Standardが出る予定である．

なお，MPEG-7に関する，より詳細な解説については，本誌2月号[3]を参照されたい．

■ Video Enrichment

汎用的な画像処理による注釈は，コンテンツの種類に制約がない反面，応用において入力画像との類似検索が可能なだけで，限界がある．そこで，より使いやすい映像管理手法は，対象となるコンテンツの種類を明示したプロファイルを用い，対象を絞ることで有効な管理手法を実現する．日本アイ・ビー・エム（株）東京基礎研究所で開発したVideo Enrichment（http://www.trl.ibm.co.jp/projects/video/index.htm）は，スポーツなどの注釈が非常に少ない映像に有効な技術で，映像をフレーム単位でなく，オブジェクト単位で処理する技術の確立を目指している．映像をオブジェクトごとに扱うと，オブジェクトの動き・位置・速さ，および複数オブジェクト間の時空間の関係から，注釈のない映像に対し，意味的解釈が行え，映像の内容に基づく検索・要約が実現できる．特に，スポーツ映像では，ルール，競技場の大きさ，カメラの配置，選手のユニフォーム，出場選手，競技のフォーメーションなどの事前知識が利用でき，選手をオブジェクトと定義することで，1人または複数人の選手の動きを利用して，映像の解釈を行っている．

さらに，オブジェクトごとに扱うと，選手の合成や視点を変えた表示，選手の統計データ等，映像を表示するだけでなく，多様な観点から映像をデータ化して眺めることができ，ユーザの興味をさらに深く追求するように誘導することができる．映像は，時間的な経過を情報として持っているので，ユーザが見ることのない情報を提示することができ，ユーザの新たな知識発見に貢献できる．

図-2は，Video Enrichmentの構成を示す．オブジェクトを切り出し，選手の位置と動きの速さを獲得する．位置・動き情報・複数オブジェクトの関係・カメラ操作・事前知識により映像をメタデータとして記述する．ユーザは，あらかじめ提示されているキーワードを入力し，メタデータを利用して，検索・要約した結果を得る．検索結果では，オリジナル映像だけでなく，その区間のオブジェクトの動作軌跡を表現したり，上から眺めたオブジェクトの位置変化をアニメーションで表現することによって，検索結果の映像をより深く解釈できるようにユーザに提示する．

まず最初に色の類似度から画像を分割する．抽出する

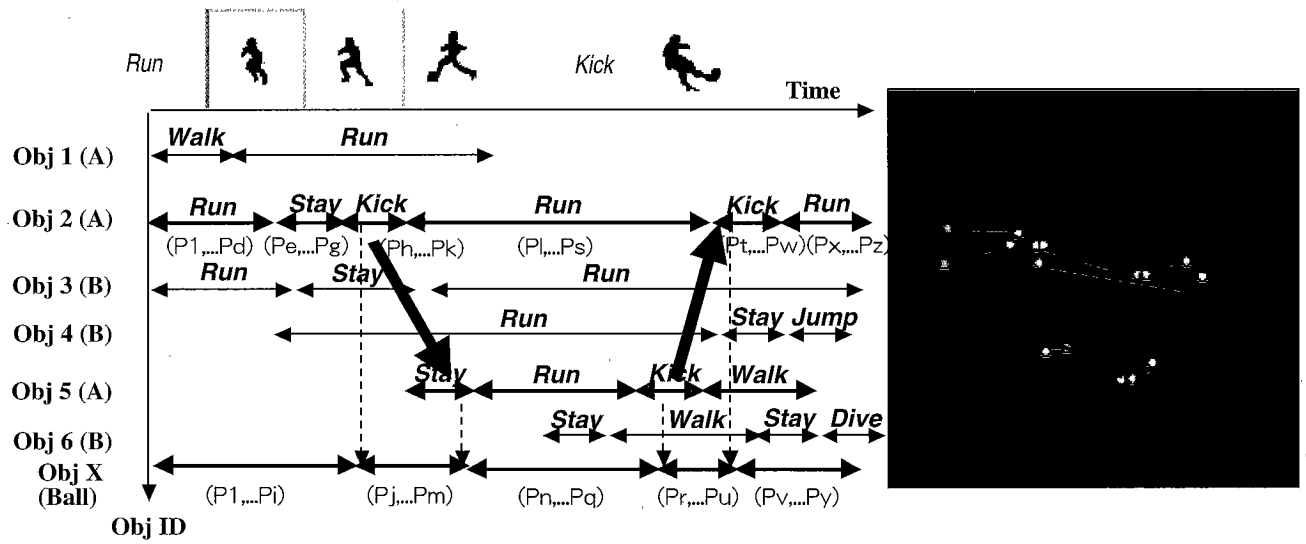


図-3　オブジェクトの表現

オブジェクトを，グランド上を移動する選手・ボールと定義すると，グランドおよびその周辺の処理だけで十分であるため，空間的な変動が小さく，面積の大きな領域をマスクし，以降はマスク領域内の画像分割を行う．オブジェクトはカメラからの奥行き距離によって大きさが変わるため，登録するオブジェクトは面積を正規化した色分布とし，登録した色分布に含まれる領域をオブジェクト領域とする．また，空間情報だけでオブジェクト領域を獲得することが難しいので，連続する画像間で確定する移動境界から領域を確定する[4]．オブジェクト領域抽出で最も困難な問題は，複数のオブジェクトの重なりによる，部分的（あるいは完全）に隠れた状態のオブジェクトの扱いにある．部分的に隠れたオブジェクト領域の分離を，誤りなく実現するには課題が多いが，隠れ状態に至る前後からオブジェクトの位置を推定することは可能であるので，この場合，オブジェクトの位置だけを有効な情報として扱う．一方，ボールの検出については，隠れや速すぎる動作による画像のブレにより自動化が困難なため，手作業で，ボールをタッチした選手に位置づけるように入力した．

また，撮影に利用するカメラは，回転とズーム以外は固定されている．この場合でも，カメラの回転移動により，わずかな平行移動が発生するが，対象となるシーンの奥行きに比べ，非常に小さいため，平行移動のないカメラと同等に扱うことができ，オブジェクトの奥行距離とは無関係に，カメラモデルを表現できる．連続する画像間の対応は，Tan[5]の手法を用い，アウトライアの除去と数回の繰返し演算によって，カメラモデルのパラメータを決定している．

オブジェクトの位置は，画像面上で表現されており，カメラの移動によって，補正が必要となる．そのため，仮想的な面を想定し，移動パラメータを復元したビデオを画像面上に射影する．これによって，固定カメラから入力したビデオにおけるオブジェクトの位置を復元するのと等価なデータが得られる．さらに，カメラは地面に対し，俯角を持つように設定しているため，仮想面を空から見下ろした，グラウンドの面にオブジェクトの位置を投影することより，画像面の距離ではなく，現実の距離として扱うことができる．切り出したオブジェクトは時間的に形状が変化することに着目し，内部の色情報を無視したシルエットを用いる．シルエットの変化は，動作によって特定の変化を示すため，あらかじめ教示した複数の動作パターンの連続シルエットを固有空間に展開し，上位の固有値から動作固有の変化を求めておく．次に入力パターンを同じ固有空間に展開したとき，どの教示パターンに最も近いかを求め，動作パターンの同定を行う．この処理では，オブジェクトの一連の動作における動作の変化点を求める必要があり，オブジェクトの移動速度の変化を動作の変化点としている．シルエットによる動作パターンの同定は，シルエットの面積が小さいと誤認識を起こしやすいので，領域の移動速度によって動作パターンを制限し，走る・歩く・立ち止まる・蹴る・倒れる，の動作IDだけを用いている．一連のオブジェクトは複数の動作によって記述が可能で，動作の変化点における動作IDとフレーム番号を登録するだけで，その間のオブジェクトはすべて同じ動作をしていたと解釈する[6]．オブジェクトは，各々が映像に存在する限り記述し，シルエットの時間的変化点を記述境界とする動作IDを最小単位として記述する．この記述をActionと

呼び，動作IDの開始・終了時の時区間と，オブジェクトの位置，およびオブジェクトの軌跡が追跡できるよう，時区間内部の複数点での位置を記述する．したがって，オブジェクトの位置は，Actionを用いてすべてのフレームにおいて近似値が計算可能となる．Actionの記述は，

Action::=<Action ID><Time Interval><Object ID><Trajectory>

である．図-3は時区間によるオブジェクトの表現で，(A)（B）はチームを表し，Obj Xはボールを表す．ボールは動作IDのないオブジェクトとして記述する．

次に，複数のオブジェクトは映像の中で複数同時に存在し，その存在期間と意味づけはそれぞれに異なるが，複数のオブジェクトの行動からなるシーンの意味づけができる．これをInteractionと呼び，

Interaction::=<Interaction ID><Time Interval><Object No>
<Object IDs><Spatial Description>

となる．Interactionは，「パス」や「ゴール」などのイベントを記述する．

イベントの検索はコンテンツに完全に依存し，コンテンツごとに異なる定義となる．ただし，記述の整合性と検索エンジンに対する適用性のため，Interactionは，他のInteractionと複数のActionによる論理演算で定義するように規定する．たとえば「スルーパス」は，複数の味方の選手間でボールのパスがあり，その間に敵チームの選手がいる場合であるため，他のInteraction"Pass"と，複数の選手のActionを定義文として

```
Begin
    Interaction Through_pass t0 O0 L0
    child_Interaction 1 Pass t1 O1 L1
    child_Action 3 Stay Walk Run t2 o2  L2
    child_Action 3 Stay Walk Run t3 o3  L3
    Where
      /* o2,and o3 are defense player*/
      get_object_from_GO o4 1 O1
      not_same_team o4 o2
      not_same_team o4 o3
      /* o2 and o3 are existed during Pass */
      temporal_overlap t2 t3
      set_temporal_overlapping_period t4 t2 t3
      temporal_overlap t1 t4
      set_temporal_overlapping_period t5 t1 t4
      ...
    Fill
      t0 t1
      O0 O1
      L0 L1
End
```

のように記述することで"Through_pass"が検索できる[7]．このように検索した結果を，Interactionとして新たに記述を加えることができ，2度目の検索からは，Interactionから対応シーケンスを引き出すことができる．図-4は，キーワードとして「ゴール」を入力したときの検索結果を示す．図-4では，4通りのゴールシーンが得られた．インタフェースはWebブラウザを用いており，映像データベースサーバに検索要求を出すと，クライアントで結果が見られるようになっている．表示には，長いMPEG-7フォーマットのシーケンスから，検索要求に合致する時区間だけを取り出し，臨時のMPEG-1シーケンスとして表示する．

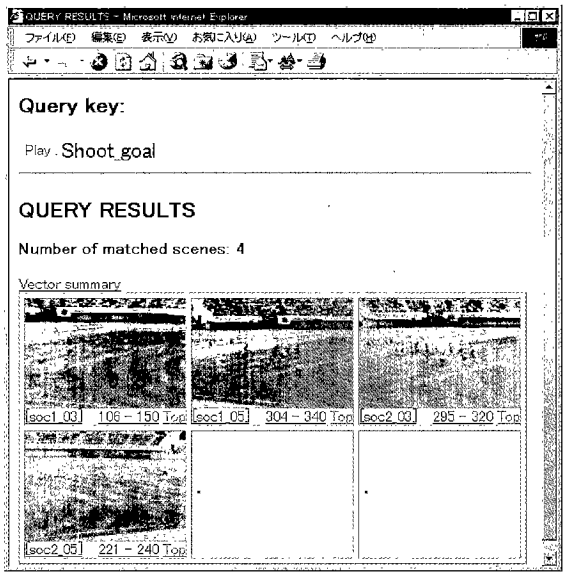


図-4　「ゴール」の検索結果

映像の特定シーケンスを検索した結果，そのシーケンスを元の映像で表示するだけでなく，短時間の映像集約と位置づけることのできるビデオモザイキングを用いる．ビデオモザイキングは，シーケンスの中で，最も大きく変化するパラメータの中央値を持つ実画像面に，画像シーケンスを平面へ射影した．背景に貼り合わせる選手は，シーケンスの長さによってサンプリングし，ストロボ撮影のように移動軌跡を表現した．図-5は，「ロングパス」に対する結果をビデオモザイキングで表示したものである．センターサークル付近の選手からペナルティーエリア左へのロングパスをカメラが撮影し，そのカメラ移動を復元すると，図-5のようなパノラマ画像の背景が得られる．その上に選手を貼り合わせると選手がどのような軌跡・スピードで移動したかが一目で分かる．さらに，上方から眺めたときの選手・ボールの位置変化を表示することができる．図-6はゴールが成功したときの選手とボールの位置，および0.5秒後に到達できる位置を選手の周りに円として表示している．この図では，選手の位置取りのよさを確認することができる．

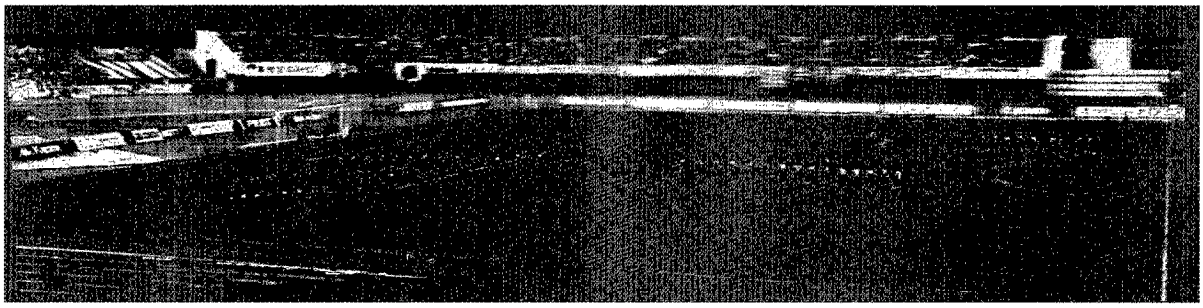


図-5　ビデオモザイキングによる移動軌跡表示

これらの表示機能を使い，ユーザの興味をさらに深く誘導し，検索・表示・データ解釈のサイクルを進めることが可能となる．

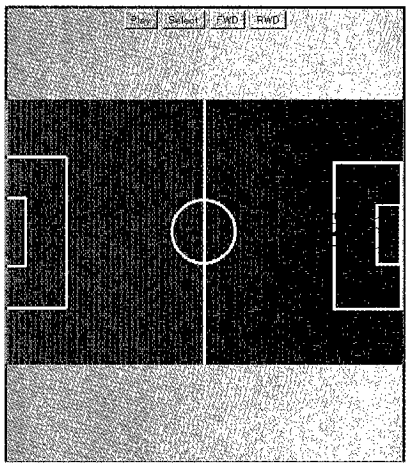


図-6　オブジェクトの位置表示

■ 今後の動向 ——

2000年のBSディジタルテレビ放送に続き，地上波ディジタル放送が2003年に始まろうとしているが，そのときには現在のアナログ放送とは異なった視聴方法が考えられる．アナログ放送では，ユーザは実況中継を視聴する，または特定の時間にセットしたビデオデッキに映像を録画し，録画開始時間の頭出ししかできなかった．しかし，ディジタル放送では，映像だけでなくデータを同時に送信できることから，番組のマルチメディア化，特に双方向通信が大きな長所となり，さまざまなアプリケーションが現在検討されている．ディジタル放送は放送と通信を融合する架け橋となるメディアで，次世代メディアの核となり得る重要な技術である．

米国では，低価格で大容量のハードディスクが受像機に組み込まれたPersonal Video Recoder（PVR）が普及し始め，ユーザの好みによる映像が，ユーザの過去の選択事例に従い録画する機能を持っている．PVRは電子番組表と連動し，ユーザが欲する可能性の番組を長時間録画するが，ユーザの視聴時間は，いつの時代でも大きく変わらないので，効率よく映像を検索し，要約して提示することが重要となる．そこでデータ放送が検索，要約のための注釈などの情報提供になるとともに，発信者側ではユーザの嗜好を個人単位で収集でき，双方にとって利点がある．"personalization"は今後のキーワードであり，発信者の提供する注釈と，個人の嗜好を組み合わせた次世代PVRは，ユーザを退屈させない，映像を含めたデータサーバとなるであろう．

従来，放送はマスメディアであったが，ディジタル放送によって，パーソナルメディアとしての特徴を持つようになる．そのとき，重要なコンテンツは映像だけでなく，映像に付随するデータも含めて視聴者の嗜好を反映することが可能で，コンテンツ提供者の意図をより強く反映したシナリオを描くことができるであろう．

参考文献

1) Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D. and Yanker, P.: Query by Image and Video Content: The QBIC System, IEEE Computer, Vol.28, No.9, pp.23-32 (1995).
2) Wactlar, H., Kanade, T., Smith, M. and Stevens, S.: Intelligent Access to Digital Video: The Informedia Project, IEEE Computer, Vol.29, No.5 (1996).
3) 柴田正啓: 放送と情報処理: コンテンツ記述の標準化 MPEG-7, 情報処理, Vol.41, No.2, pp.176-182 (Feb. 2000).
4) Echigo, T., Radke, R., Ramadge, P., Miyamori, H. and Iisaku, S.: Ghost Error Elimination and Superimposition of Moving Objects in Video Mosaicing, IEEE ICIP-99, 28AO3.5 (1999).
5) Tan, Y. P.: Digital Video Analysis and Manipulation, Ph.D. Thesis, Princeton University (1997).
6) Miyamori, H., Echigo, T. and Iisaku, S.: Proposal of Query by Short-time Action Descriptions in a Scene, IAPR Workshop on Machine Vision Applications, 3-18, pp.111-114 (1998).
7) Kurokawa, M., Echigo, T., Tomita, A., Maeda, J., Miyamori, H. and Iisaku, S.: Representation and Retrieval of Video Scene by using Object Actions and Their Spatio-temporal Relationship, IEEE ICIP-99, 26AO2.1 (1999).

（平成12年2月1日受付）