10. カメラ移動に基づく認識

本章ではこれまでの２次元処理や予測で得た MPEG ベースの ET を意味のレベルから評価し，改善することでシーンクラスを判別し，その結果として ET を GT に近づけることを考える．まず，ブロック画素単位のイベントの集合からシーンの三次元構造を求めるために因子分解法を導入する．次にカメラ軌道に応じた三次元復元の度合いを考察する．さらに，意味を扱う次元に拡張した場合の因子分解法について実験する．

Keyword　カメラ軌道，因子分解法，クラスタ化，一般化因子分解

10.1　はじめに

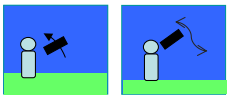
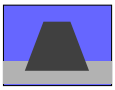
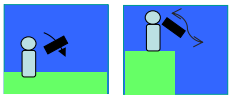
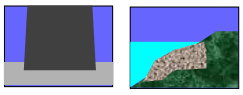
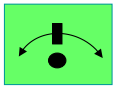
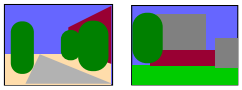
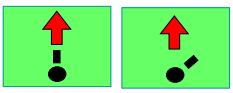
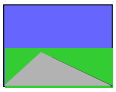
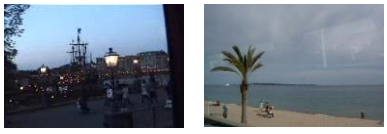
風景シーンにおいては明らかに特定の制約が存在し，人間がリアルタイムでビデオカメラを使用する場合には特にその傾向が強い．ネガティブな制約はフレーム長と計算量が小さいことである．ポジティブな制約は以下のとおりである．本章の目的はシーンクラスを自動的に決定することであるので，再構成された３D シーン構造について，システムが高解像度で計算する必要はない．また，事前に定義されたシーンクラスはたかだか 5 個程度を想定する．

10.2　カメラ軌道の分類

人が風景シーンを撮影する場合，車載カメラや空撮とは異なり，いくつかのパターンに限定されることが多い．経験的にそれらの典型的なパターンをまとめた結果を表 10.1 に示す．ここで動きのタイプ｛UP,DOWN,PAN,THROUGH｝のそれぞれにおいて典型的な動きのモデルが考えられ，この

表 10.1　風景撮影における主要なカメラ軌道

Tab.10.1 Primary camera motions for the landscape scenes.

Type	Motion Model	View Model	View Example
UP			
DOWN			
AROUND (PAN)			
THRO-UGH			

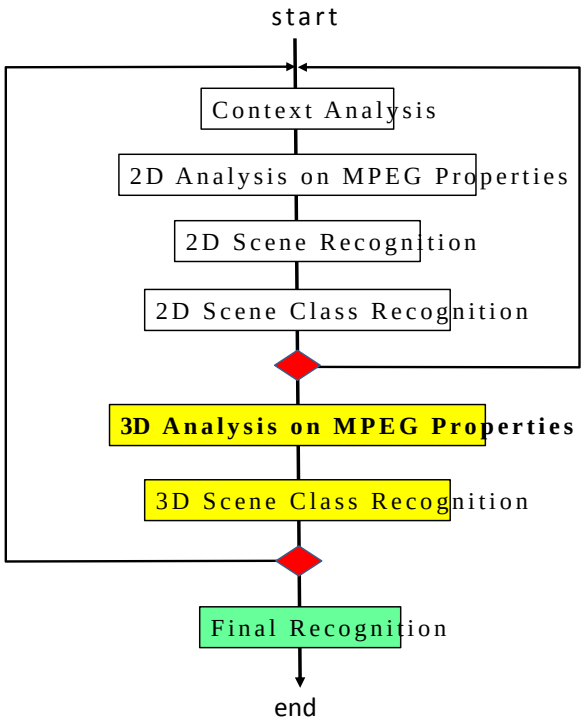


図 10.2 システムフロー
Fig. 10.2. System flow.

選択に関する好みや体格などに起因する微妙なカメラ軌道の違いはユーザプロファイルで記述される．そして動きのモデルは 3 次元のカメラ軌道を限定するため，カメラ特性に基づいて 2 次元の ViewModel が限定される．このような限定をもとに実際に撮影された映像（その例は ViewModel に示した）からカメラ軌道を算出する際の制約条件を設定することができ，後述の因子分解法などの三次元復元の計算を簡略化することができると考えられる．

10.3　２次元処理から３次元処理への流れ

従来，単眼カメラで得られる２次元画像の時系列 I[t1:t2]からもとのシーンの３次元形状 S を復元する問題は純粋に幾何学的に解かれることが多かった．移動ステレオや因子分解法，あるいは非線形最小 2 乗法などがそれにあたる．他方で，透視投影が仮定されるときには３D 再構成メカニズムが難しくなるという問題もあった．Kanade らが 90 年代に提案したように．因子分解法はビデオ画像から 3 次元再構成を行う際に非常にシンプルである．しかしながら，その計算においては，特徴点の抽出は画素レベルであることが要求され，シンプルに実現する際には正射影が基本的には導入されている．この正射影は MAP (計量アフィン射影：Metric Affine Projection)の最も簡単なクラスであることが知れられている[D3].

だが，それらをもってしてもシーンクラスの認識は後段の 3 次元形状や 3 次元運動に関する判定規則に委ねられ，意味へのマッピングは容易ではなかった．

しかし，図 10.1 のようなカメラ運動の制約を仮定すると，カメラ運動に対して支配的であるユーザの時空間軌道 u~とシーンクラスの時系列 SC~に応じて，画像面上の 2 次元パターンの時系列

P(I~)と対応するシーンの 3 次元構造との間の関係を限定できる．この u~と SC~がシーンのコンテキストを形成し，逆にこれを知ることでその後の 2 次元処理を適応化できるであろう．

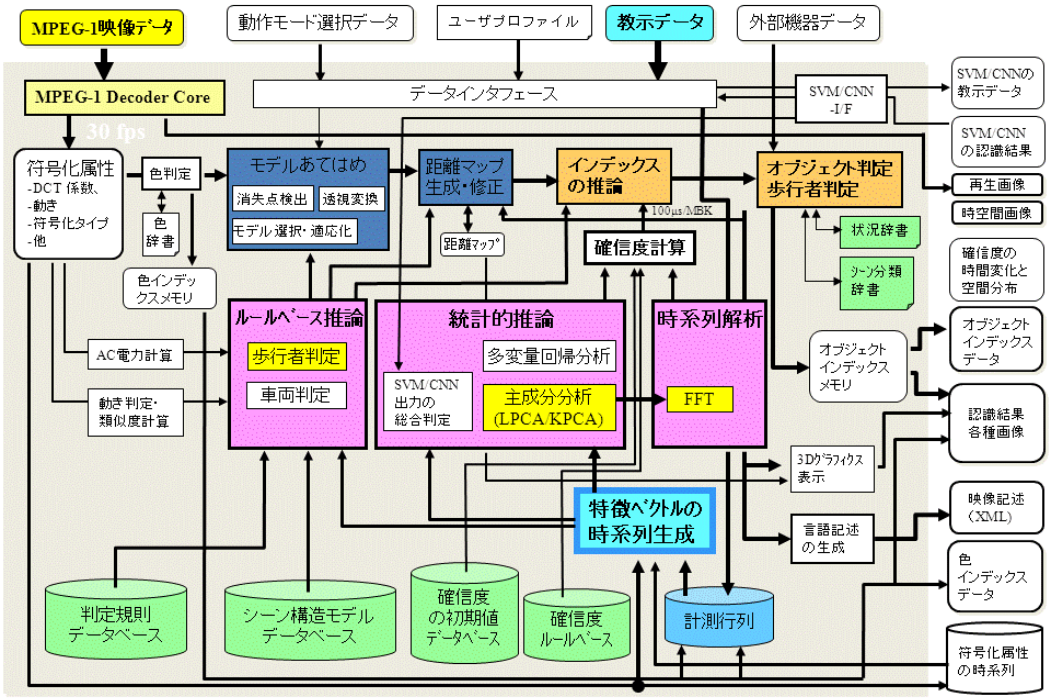


図 10.3 システム構成

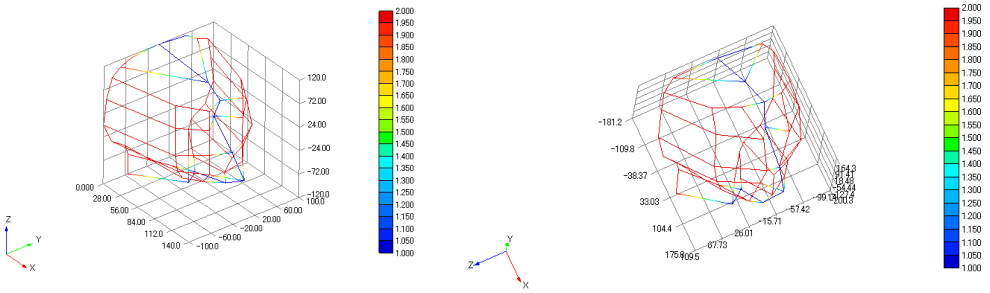
Fig.10.3. System architecture.

前章までの議論と上記の考察から風景認識のフローをまとめると図 10.2 のようになる．Context Analysis ではカメラの現在位置やユーザプロファイルに書かれたユーザの行動内容および嗜好をもとに現在の状況を限定する．限定された状況下でシーンクラスの候補値を選択し，「MPEG 属性の 2 次元解析」における主要戦略を決定し，語彙の限定や確信度のバイアス値(開始時点では初期値)が与えられる．これが 3S モデルであり，歩行者認識では図 7.9 の適応化機構や図 7.10 の状態遷移図による歩行者挙動のモデリングがそれに相当する．その後，2 次元シーン認識では 6 章から 9 章で述べた認識が行われ．その結果を踏まえて 2D Scene Class Recognition においてシーンクラスが更新される．その後、認識結果が所定の評価関数を満たすならば次の 3D Scene Class Recognition に向かい，満たしていないならば再度 Context Analysis から繰り返す．この評価関数は学習段階でなければ正解との比較に基づくものではないことが多い．仮に正解を用いるとすると，キャリブレーション用のテストパターンか基準シーン（正解があらかじめ用意されている）の場合である．通常は正解を知らずに認識動作を行うので，11 章で述べる集合知あるいはモデル関数を用いたメカニズムとなる．このことは 3D Scene Class Recognition の後段の分岐でも同様である．個々の演算の関係を見やすくするために，9 章までの 2 次元処理に基づく MBV のシステム構成を図 10.3 にまとめた．

10.4 因子分解法

本章では，シーン構造の 3 次元推定を行い，シーンクラスの認識率を改善することを目的とする．そこで，因子分解法を導入し，ブロック画素単位の特徴点を動きベクトルでトラッキングした結果として得られる計測行列に適用する．（因子分解法の詳細は 3 章と APPENDIX を参照．）

ブロック画素単位の因子分解を実行する前に，本稿で用いる因子分解法の基本性能を確認する．図 10.4 にその動作内容を示す．図 10.4(a)の原データ（人間の顔の 3 次元モデル）を一方向に回転させることで近距離を周回するカメラ軌道を模擬し，軌道上のフレーム時刻に対応した離散点を原点とするカメラ座標系で，オブジェクト（顔の 3 次元モデル）の特徴点の透視変換を行った．その結果得られる特徴点の 2 次元座標の時系列から計測行列を作成した．この計測行列を特異値分解すると（計測行列）＝（運動行列）×（形状行列）となり，この形状行列の上位 3 ランクの



(a) Original shape. (b) Reconstructed shape.

図 10.4 因子分解法の基本性能

Fig.10.4. Basic perfoemnace of factorization method.

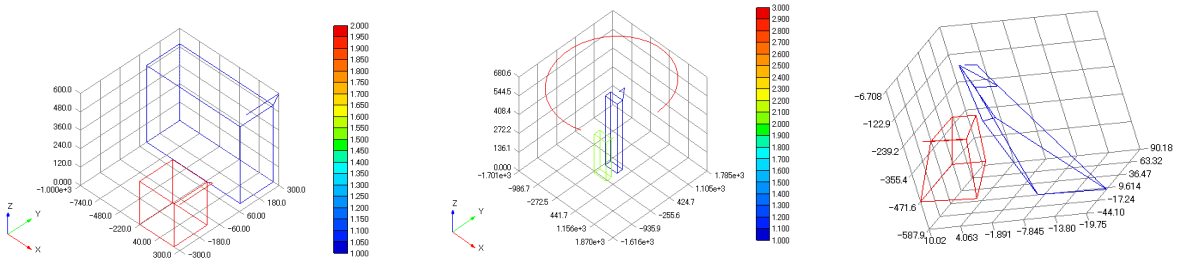
成分を取ることにより立体構造が復元される（図 10.4(b)）．ただし，ここでは計量拘束により座標回転の自由度を抑制することで復元結果の補正を行うこと（APPENDIX 参照）は省略した．なお，3 次元復元の精度は，原物体の形状，観測するカメラの軌道，特徴点の数と位置精度，観測するフレーム数などに影響を受けることが実験でわかっている．次節ではそのうち，カメラ軌道の影響について簡単に考察する．

10.5 カメラ軌道の影響

ここでは 3 種類の軌道｛円周，直進，パニング｝について三次元復元実験を行った．

（１）軌道１：円周

図 10.5(a)を原形状とするオブジェクトについて，その周囲をオブジェクトの上部に近い高さで円周状に周回し，その間，常にカメラの姿勢は原点を向くように設定した．



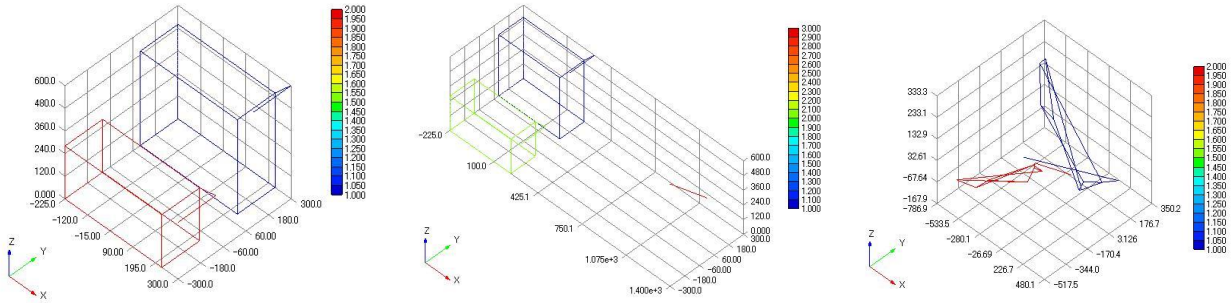
(1) Original shape. (b) Camera trajectory No.1 (c) Reconstructed shape.

図 10.5 軌道 1 に対する復元結果

Fig. 10.5. Reconstructed results using trajectory No.1.

（２）軌道２：直進

図 10.6 は R=1500,φ=70 の位置から 50/frame でターゲットオブジェクトに接近し，30 フレームについて観測した結果を計測行列に格納し，因子分解法で復元した結果である．

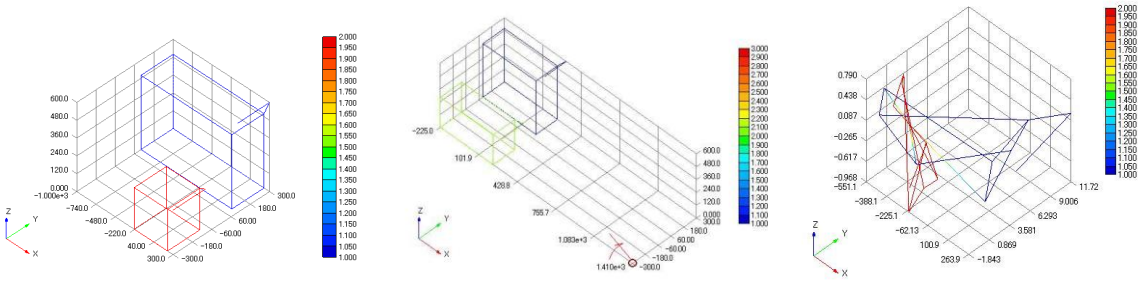


(a) Original shape. (b) Camera trajectory No.2 (c) Reconstructed shape.

図 10.6 軌道 2 に対する復元結果
Fig. 10.6. Reconstructed results using trajectory No.2.

（３）軌道３：パニング

同様のシーン構造に対して遠方からパニングした場合の復元結果を図 10.7 に示す．カメラ位置が固定のため軌道は描かれていないが，オブジェクト中心に対して水平角 Phai=-10~20[deg]の範囲を 1deg/frame で 30 フレームにわたってパニングした結果として計測行列を獲得し，因子分解法で復元した結果が(c)である．主観評価では軌道 1，軌道 2 に比べて復元精度が悪く，復元形状のねじれも見られる．

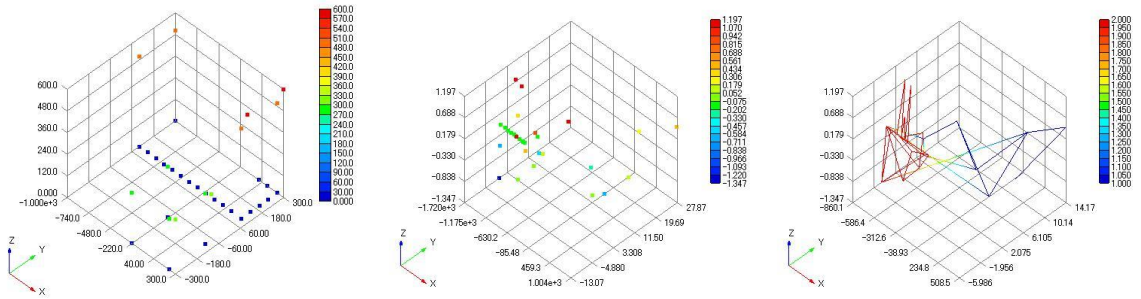


(1) Original shape. (b) Camera trajectory No.3 (c) Reconstructed shape.

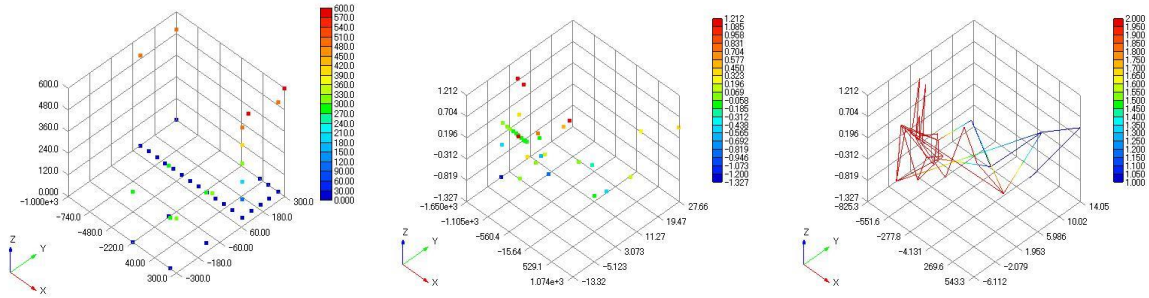
図 10.7 軌道 3 に対する復元結果
Fig. 10.7. Reconstructed results using trajectory No.3.

10.6 特徴点数の影響

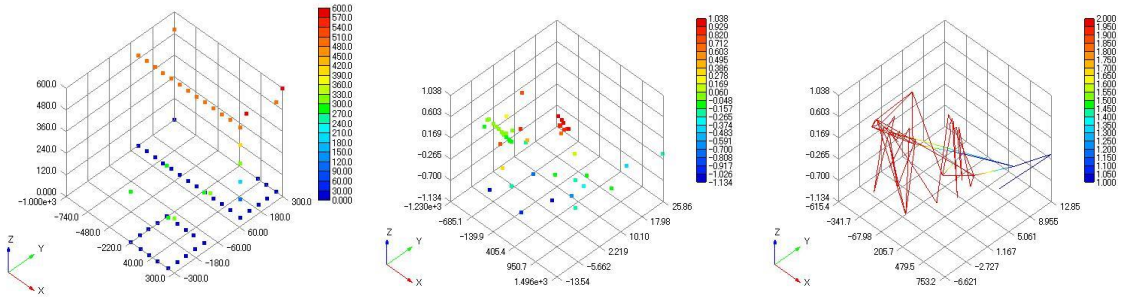
復元形状のねじれに関しては同じシーン構造に対して特徴点の数を増やすことで改善される可能性もある．そこで，段階的にモデル構造での特徴点を増やし，復元形状の変化を考察した（図 10.8）．この結果，特徴点数の増加に際しても特別な改善は認められず，全体の復元形状はあまり変化しなかった．軌道 1（円周軌道）では特徴点数増加に伴い，逆に復元結果にノイズが目立つケースも見られた．軌道上の移動に伴って変化するはずの個々の特徴点位置が軌道の条件によってはあまり変化しなかったり，一直線上に密集したりするケースもあり，これらは復元計算においてノイズをもたらす原因となっていることが推察される．



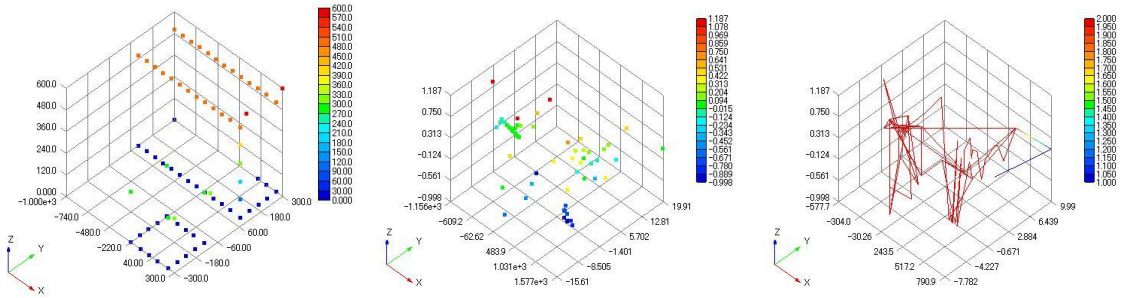
(1) Level-4 of Original, Reconstructed. Reconstructed wireframe, respectively.



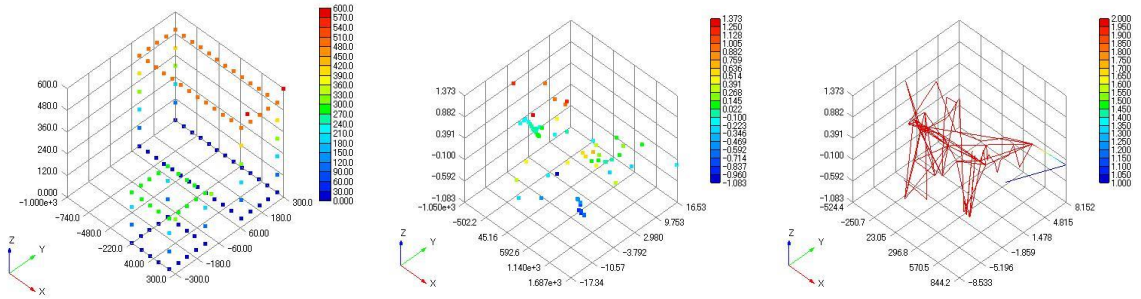
(2) Level-5 Original, Reconstructed. Reconstructed wireframe, respectively.



(3) Level-7 Original, Reconstructed. Reconstructed wireframe, respectively.



(4) Level-8 Original, Reconstructed. Reconstructed wireframe, respectively.



(5) Level-11 Original, Reconstructed. Reconstructed wireframe, respectively.

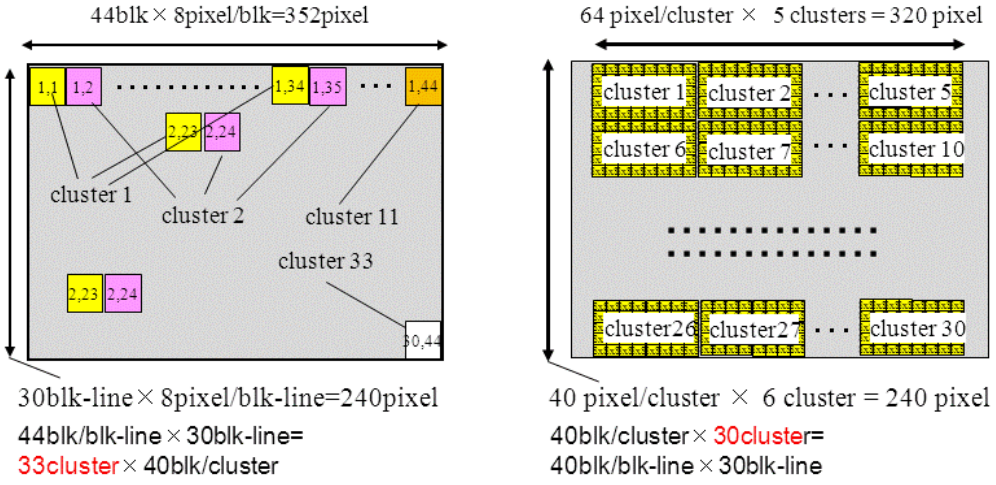
図 10.8 軌道 3 に対する復元結果
Fig. 10.8. Reconstructed results using trajectory No.3.

10.7　ブロックベースの因子分解法

上記は因子分解法を画素単位で適用したものであるが，MPEG ベースでシーン構造を自動認識する際には，より簡単で高速に動画像処理できるための工夫が必要である．

まず，特徴点に関しては MBV の観点から，ブロック画素ベースの計算で抽出した動きベクトルを適用する．これにより各フレームの特徴点数は一定数となり，それぞれの 2 次元位置も固定できる．ただし，個々の特徴点はフレーム間で規定のブロックアドレス上を移動する．しかし，この前提条件によって計測行列の時間方向の深さが十分な値をとることができなくなり，ランク問題が発生する．それは計測行列の時間的深さが十分な値をとれないためである．（本稿ではその時間深さを 20 に設定した．これは 2 秒間の MPEG シーンにおいて，前向き動きベクトルを保持する予測フレームの数に等しい．そして 2 秒の理由はこの提案するメカニズムがハンディビデオ装置にインストールされる際に，実際の観測時間に帰着される．もう一つの理由は，ハンディビデオ装置のユーザは通常，パンとチルトを用い，撮影されたシーンが同じ視界をさほど長く保てないからである．）さらには，ブロックの数は MPEG-1 規格の場合，1 画面あたり 1320 である．したがって，これらすべてを考慮すると，ランク問題のためにその計測行列はしばしば SVD にとって不適当となる．そこで，MPEG-1 の 1 フレームの画像サイズ（352×240）に対してブロック画素（8×8）単位で特徴点を定義すると，水平方向に 44 個，垂直方向に 30 個並ぶため，合計で 44×30＝1320 個の特徴点が存在する．この 1320 個の特徴点の時系列を 1 回の因子分解で処理するには計測行列のランクの制約から，少なくとも 660 フレームの時系列が必要になる．P フレームならば 10 フレーム/秒なので 66 秒の観測を行わないとシーンの 3 次元構造が復元できないことになる．これは風景撮影に応用する際には，処理時間と記憶サイズの観点から非実用的である．たいていの風景映像は定点観測でない限り，1 分以上同じ視野を撮影し続けることはまれであるからである．

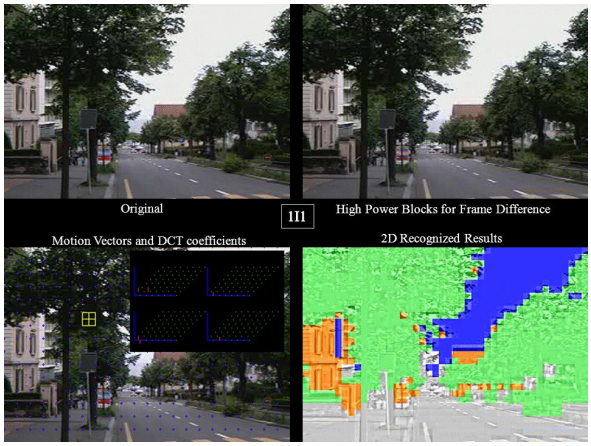
一方，因子分解法の定義式によれば，計測行列の特異値分解に必要なランクさえ満たしていれ



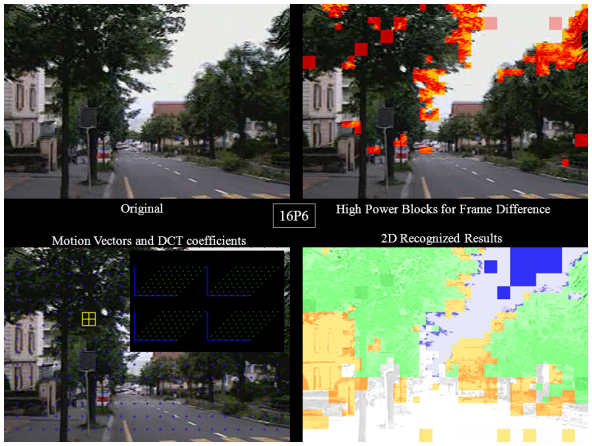
(a) Type-1　　(b) Type-2

図 10.9 ブロックベースの画像分割

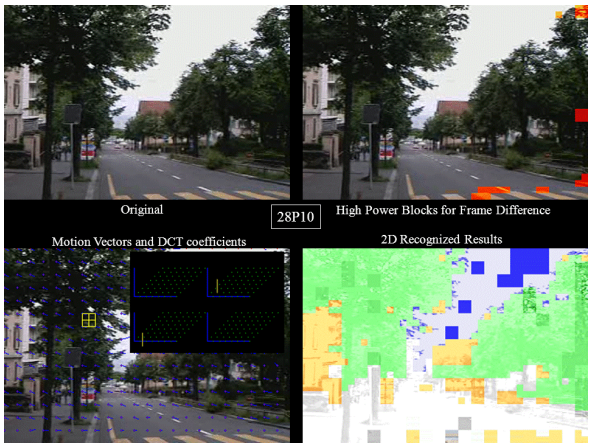
Fig.10.9. Block based image decomposition



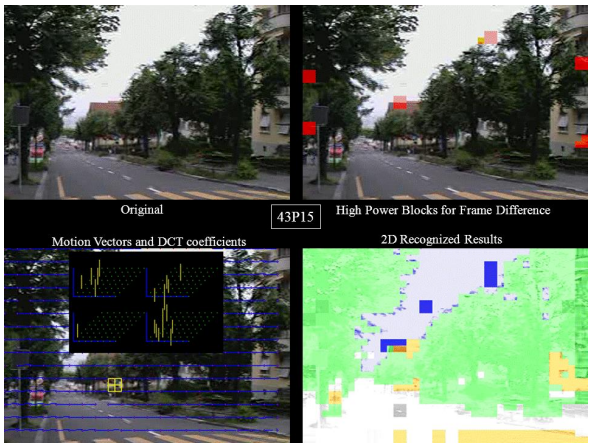
(a)2 次元処理結果（1I1）



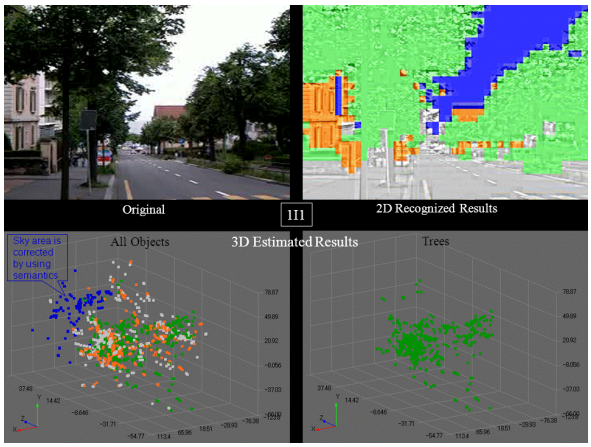
(b) 2 次元処理結果（16P16）



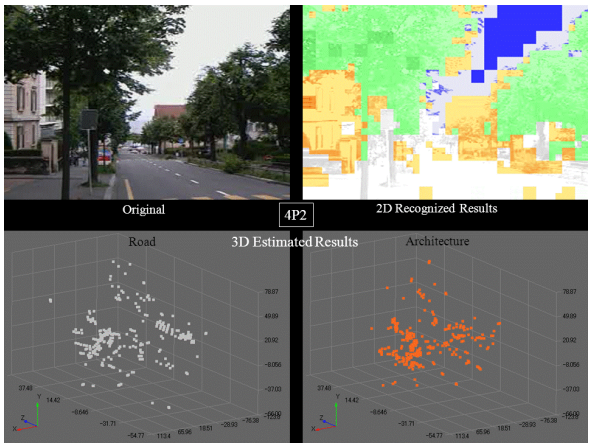
(c)2 次元処理結果（28P10）



(d) 2 次元処理結果（43P15）



(e)因子分解の結果（1I1）



(f) 因子分解の結果（4P2）

図 10.10 実験シーン No.1 （Lausanne）

Fig. 10.10. Experimental Scene No.1 (Lausanne).

ば観測する特徴点の数には制約はない．そこで 1 画面のすべてを 1 度に処理するのではなく，いくつかのクラスターに分割してそれぞれのクラスター内で 3 次元復元を行うことを考える．図 10.9 はクラスター化の一例である．8×8 画素の BLK にひとつの動きベクトルを対応付ける

（MBK の動きベクトルを転写し，時間方向の偏移を BLK 単位で管理する）と仮定し，水平方向に 8 BLK，垂直方向に 5 BLK で合計 40BLK で構成される小画像について因子分解を適用する．この場合，1 回の因子分解に必要な計測行列は 2 K×40 で K≧20 であるから，最低 20 フレームの P ピクチャがあればよいことになる．

クラスターはブロック形状タイプに限定されているわけではない．各クラスターの定義は初期画像平面をブロック画素のグループに分割する方法のみによって規定される．この分割は，計測行列を特異値分解する際の制約条件から，各クラスター中のブロック数（すなわち，クラスター中の特徴点数）が NK を超えないという条件のもとで達成される．ここで，N は観測される属性の数であり，K は観測するフレームの数である．これは，各クラスターの時系列が対応する計測行列を形成し，その行サイズは NK（この場合，水平座標と垂直座標の 2 つが対応する属性なので，N=2）である．なお，各クラスターのトラッキング操作は動きベクトルによって達成される．計算上の観点からは，観測時間が短い場合でも各特徴点に関してフレーム間の生成消滅を管理することは若干複雑な問題となる．

前節までに行ったように，3 次元 CG による合成画像をテスト画像とする場合はオブジェクトモデリングの際に特徴点を 3 次元座標で定義できるため，それらを画面上に透視変換した結果として 2 次元座標の時系列を精度よく管理できる．ところが実画像では

動きから特徴点時系列による計測行列 X を復元する際，動きが画素単位で検出されていれば各特徴点の 2 次元座標は画素単位で検出できる．しかし，MPEG ブロックベースの特徴点はブロック格子上に制約された動きとなる．

10.8　実験

（1） シーン 1 ：Lausanne

上記図 10.9 のタイプのクラスター化に基づき，9 章で用いたシーン Lausanne に上記のクラスター別因子分解を適用した結果を図 10.10 に示す．Lausannne は小高い位置にある鉄道駅からレマン湖に向けてゆるやかな下り坂の直線道路左側の歩道で立ち止まり，レマン湖側から鉄道駅側に向けて右に 180 度パンした映像である．9 章で示したように，Lausanne では道路右手のアパートに対する認識率がシーン中ほどで著しく劣化する．これに対してクラスター別因子分解が出力する 3 次元シーン構造がこの劣化をどの程度補正できるかが実験の焦点となる．

因子分解で出力された各ブロックの 3 次元情報と 2 次元認識で出した ET（図 10.10(a)(b)(c)(d)）を組み合わせ，因子分解の結果がどの程度妥当であるのかを定量的にチェックする．

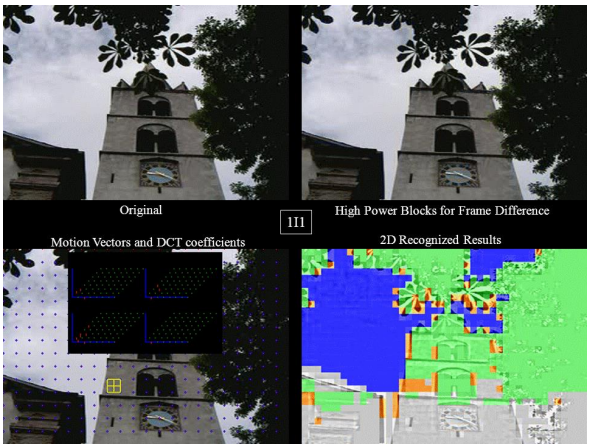
オブジェクト分類するために 3 次元表示した各ブロックの色を ET における 2 次元インデックス画像の色に合わせて｛青：空，緑：樹木，橙：建物，灰：道路｝として標示した（図 10.10(e)(f)）．ただし，空の部分は奥行きが正確に復元されないため，初期段階で奥行き（Z 軸）を最大値に設定した．もとの映像データは正確な 3 次元情報を持たない通常のビデオ映像であるため，その他のオブジェクトについては次の判定規範を考える：

1） 復元した 3 次元配置について，ブロック間の相対位置は適切か

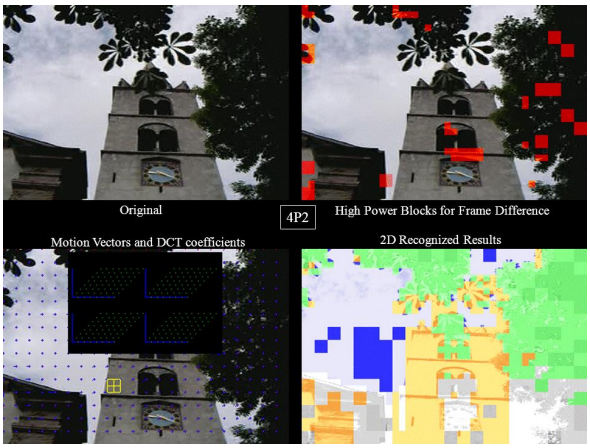
2） 復元した 3 次元配置について，オブジェクトとしてのクラスターを形成するか

3） 復元した 3 次元位置について，オブジェクト間の位置関係は適切か

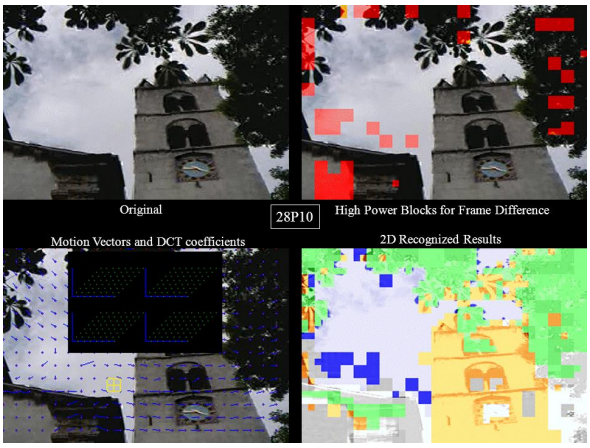
これらはいずれも自動的に判定することは困難であり，定量化に際しては人間の主観評価に頼ら



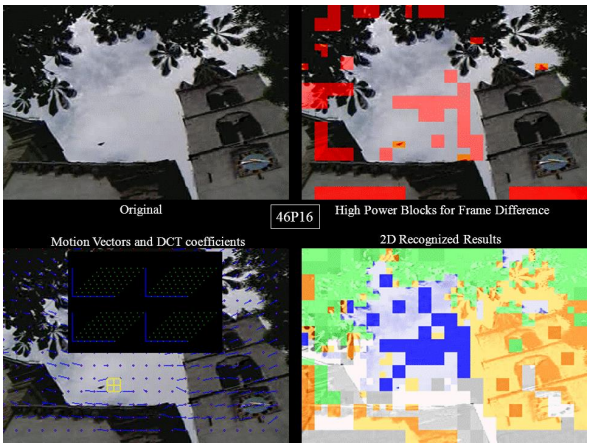
(a)2 次元処理結果（1I1）



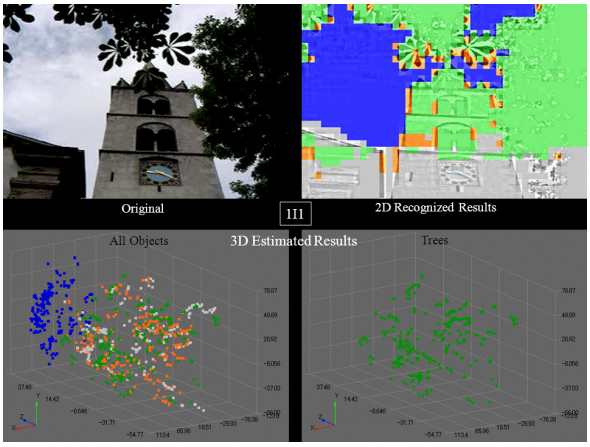
(b) 2 次元処理結果（4P2）



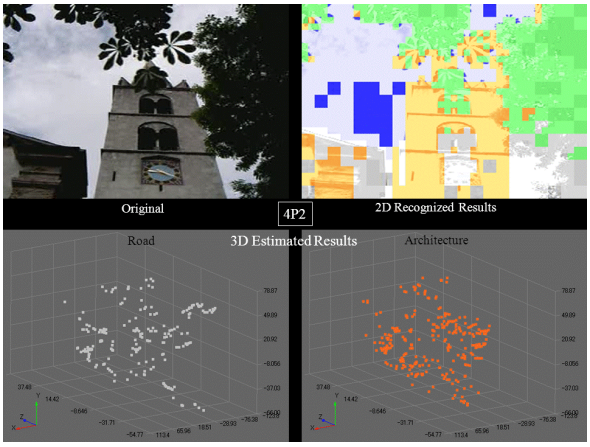
(c)2 次元処理結果（28P10）



(d) 2 次元処理結果（43P15）



(e)因子分解の結果（1I1）



(f) 因子分解の結果（4P2）

図 10.11　実験シーン No.2（Martigny）

Fig. 10.11　　Experimental Scene No.2 (Martigny).

ざるを得ない．２）のクラスターの形成度合いについては２次元インデックスのクラスターを参考にしてある程度の参考指標を算出することはできる．しかし，トータルとしては１）～３）のすべてを行うことが必要であるため今回は主観評価を採用した．この結果，「おおむねオブジェクトの配置を再現してはいるものの，まだ改善の余地がある」という結論に達した．改善のポイン

トは

A) カメラ軌道（Lausanne ではパニング）に応じた復元性能の改善

B) 初期ブロック（1I1 で定義したブロック画素群）の視界からの消失に対する対処

C) ブロック画素単位の 2 次元位置精度と，それを生成する動きベクトルの精度改善

などである．これらは今後の課題としたい．

次に，Martigny のシーンについての実験結果を図 10.11 に示す．これは教会の上部を地上から見上げた状態から左にパンし，徐々に視線を地上に下げるというシーンである．ここでは最初の教会上部と空および樹木の間で誤認識が多く発生しており，これを 3 次元シーン構造の復元でどの程度補正できるかが最終的な目標となる．Lausanne と同様の主観評価により，ここでも「おおむねオブジェクトの配置を再現してはいるものの，まだ改善の余地がある」という結論に達した．改善のポイントも Lausanne と同様になる．

10.9　クラスター概念の拡張

さらに一般化すると，クラスターへの“分割”はブロック画素の識別子（ID）からクラスター ID への“マッピング”に置き換えることができる．この段階ではあるクラスターに入れるべきブロックを選ぶ際に，そのマッピングの規則だけが定義されていればよく，それ以外には特に時空間的な制約は必要ない．したがって，各ブロック画素について，異なる複数のクラスターID への重複したマッピングもまた可能となる．

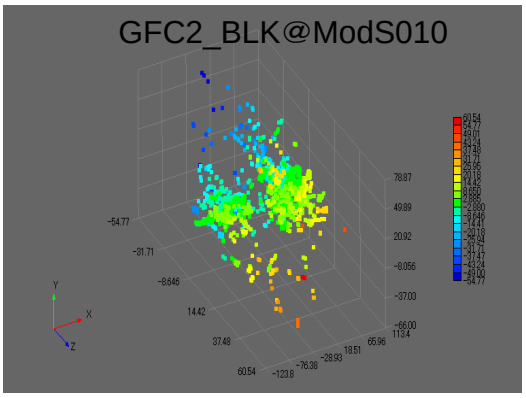
一方でこのようなマッピングを定義すると，それはラベリングの上位概念であることが思い起こされる．すなわち，画像理解プロセスの開始段階ではルールによるマッピング，その次の初期段階では信号処理段的規則によるマッピング，さらに理解プロセスが進展するとマッピングそのものがそれに伴って進化・変遷し，最終的なマッピングはオブジェクトラベリング（ひとつの画素領域に複数のラベルが可能）に収束すると解釈することができる．

この途中段階のマッピングにおいて，ラベリングの重複を許容する冗長度のあるマッピングは，それぞれの部分的に再構成された 3 D 情報を一つの復元解に統合する際に有効となる．これは個々のマッピングそのものの精度が必ずしも十分ではなく，復元誤差を含むからである．

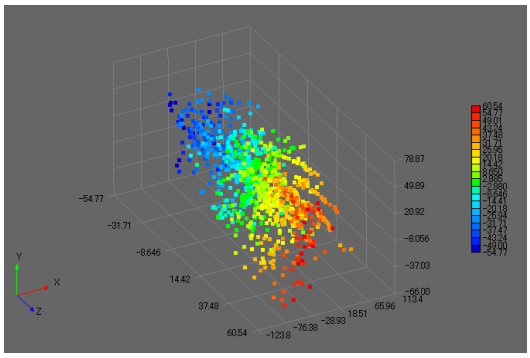
10.10　意味情報の自動補正について

ここまでは，画像面上の位置という 2 次元特徴量を P 個の特徴点（本稿ではブロック画素）について K フレームの時系列として集めた計測行列を特異値分解することで 3 次元情報であるカメラ軌道と形状を同時にもとめることを行った．これを 3 章で示したように，N 次元特徴量に拡張すれば 3 次元情報以外の新たな属性群を求めることができるはずである．ただし，認識のターゲットとする属性が出力行列（因子分解された結果の行列）においてどの次元にあるかを一意的に定める方法については未定である．だが，APPENDIX で示したように，左辺の計測行列 W において位置とそれ以外の属性をブロック行列として分離しておくことで，右辺に出力される拡張運動行列と拡張形状行列の双方において三次元位置情報とそれ以外の属性（今，これを意味情報と呼ぶことにする）は，双方に干渉がなければ分離されて出てくるはずである．だが，計測行列を特異値分解した結果，右辺では 3 次元位置情報が他の属性に優先して高い特異値に対応付けられるという保証はない．また，意味情報についてもそれに含まれる個々の次元の属性がいったい何

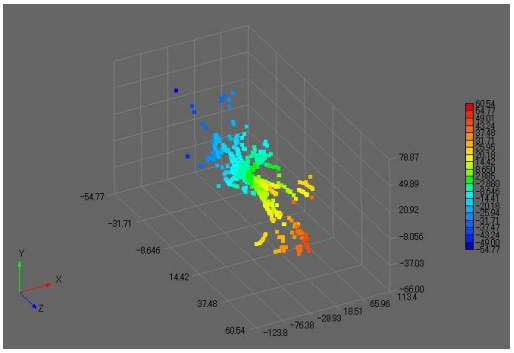
(a)原画像



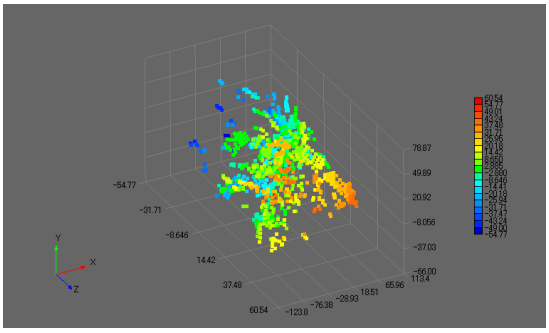
(b) GFC2



(c) GFC4



(d) GFC12



(e) GFC19

図 10.12 実験シーン No.1（Lausanne）

Fig. 10.12 Experimental Scene No.1 (Lausanne).

であるのかは何らかの制約条件がない限り一意的には定められない．そこで，APPENDIX で示したようにあらかじめ多変量線形回帰式で属性次元を定め，意味情報に関する導出を制約することが考えられる．これらについては，より実験条件を絞り込んだ上で別途チャレンジすることとし，今回は上位 3 次元すなわち 3 次元位置の精度が他の付加的な属性次元によって改善されるかどうかについての検証を試みる．

図 10.12 は以下の場合についての因子分解の実験結果である．ここで，GFC'N'は N 次元の因子分解を意味する．N は最大 19 であり，各次元の内訳は以下のとおりである：

n=1,2：マクロブロック番号（0,..,329）とブロック番号(0,..,3)

n=3 ：多変量回帰分析の切片で常に 1.0

n=4,5,6：YUV に関する DCT 係数の DC 成分

n=7,8：ブロック画素の水平座標（1,...,22）と垂直座標（1,...,15）

n=9：AC 成分のアクティビティ（BLK 全体）
n=10：AC 成分の水平方向のアクティビティ
n=11：AC 成分の垂直方向のアクティビティ
n=12：動きベクトルの水平成分
n=13：動きベクトルの垂直成分
n=14～16：符号化属性タイプ
n=17～19：1 フレーム前の DCT 係数の DC 成分（YUV）

この 19 次元の属性から次元を n=1 から順に N 個サンプルして係数行列を形成し，N=2,4,12,19 について因子分解（GF"N"と略記）を行い，上位 3 次元を視覚化した結果を図 10.12 に示す．GF2 は前節の 2 次元位置情報についてのみの因子分解の結果と同じになっている．予想では付加的な次元の効果により N が大きくなるほど復元形状は実際映像の 3 次元構造に近づくと考えられたが，現時点ではそうなっていないことがわかる．付加次元の構成の再検討も含めて今後の課題とする．

10.11　シーンクラスの認識に向けて

本章ではブロック画素で構成するクラスターベースの因子分解を提案した．この計算そのものは良好に行われ，画素単位であるいはブロック画素単位で画面全体の特徴点を 1 つの計測行列に格納する場合に比べて非常に高速である．ただし，復元精度は劣化するため，性能改善が望まれる．本来，静止画でも人間はそのシーンクラスと奥行きの概略をしることができる．だが，それを可能にするのは膨大な画像記憶とオブジェクトとの対応付けに関するデータベースやルールベースを人間が有するためであり，それによって日々確率的な解釈を可能にしている．MBV ではこのメカニズムに習い，画像処理のみで完結させるのではなく，不完全な三次元復元を完全な三次元復元に補正することを上述の概念マッピングの最適化によって行おうとするものである．ここでは，画像認識をアシストする先験的知識が組み込まれ，例えば，統計的な人の出現なども考慮される．したがって，概念マッピングの最適化プロセスは単体の認識システムで完結するものではなく，複数の認識エージェントの情報交換により達成されるものであることがわかる．だが，常にフルタイムですべてのエージェントが活動しているわけではないので，その情報交換はたえずリアルタイムに行われるとは限らず，分散する共有知識によって時空間的に繋がれることになる．その一形態がシーンクラスであり，セマンティック Web と同様に大局的には準静的なデータベースであるが，局所的には準リアルタイムに変化をとげる．このメカニズムでは理解は予測推定とほぼ同様であり，最新のスパースなセンシングデータを同化していくプロセスとなる．

11. 複数の認識器の協調

本章では EM アルゴリズムを導入することで複数の認識器を協調させ，シーン中のオブジェクトインデックスの認識率を向上させる．これにより推定解（ET）の学習のみで正解（GT）に進化していく集合知メカニズムを考える．具体的には GLAD 手法を取り入れる．

Keyword GLAD, 集合知, EM, GT, ET

11.1 集合知の応用

複数の認識器の統合を考える際，AdaBoost は協調的学習におけるもっともよく知られた方法の一つである．しかし，通常，AdaBoost は認識対象となるオブジェクトの数値的性質をアルゴリズム内部に持つことはない．一方，3 章で述べた GLAD(Generative model of Labels, Abilities, and Difficulties)[H1][H2]に基づく集合知の手法は認識器の認識能力とオブジェクトの判別容易性という２つの主要特徴を内包している．また，推定解（ET）を巡回的に学習し，正解（GT）を初期学習以外では一切用いる必要がない．これは AdaBoost とは決定的に異なるユニークな性質である．問題に特有のパラメータを認識対象に応じて自動獲得できるため，判別容易性の概念を多次元に拡張することができる（図 3.9）．他方，Graphical Model（図 3.8）の構造も問題に応じて変えていくことが考えられ，Collapsed Gibbs Sampling を用いたシーン予測手法[C19]などが報告されている．

本章では複数の認識器を協調させて認識率を改善する手法の可能性を探る．実験条件は以下のように設定した：

（１）対象物を表す語彙の数：3
（２）対象画像の数：10 フレーム（映像名：Lausanne）
（３）認識器の数：4（No.1～4）
（４）入力情報：
　①原認識器で認識した結果 ET0
　②4 個の認識器で構成する認識器群 6 セット．各認識器は ET0 から擬似的に生成．
　③認識率を算出するための正解値（GT：Ground Truth）
（５）実行内容：3 章で示した EM アルゴリズム（反復回数 10 回）
（６）出力情報
　オブジェクトの先験的存在確率
　各認識器の単独の認識率
　集合知の認識率（EM アルゴリズムの最終結果）
　認識能力に関するベクトル $\boldsymbol{\alpha}$
　判別容易性に関するベクトル $\boldsymbol{\beta}$
　EM アルゴリズムの反復回数に応じて段階的に変化する認識結果の画像出力

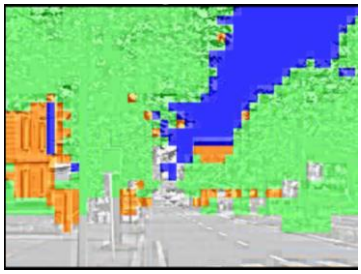
11.2 認識解の探索と収束

図 11(a)は GT と ET の間の関係を示すグラフィカルモデルである．基本的な確率構造は次式で表される：

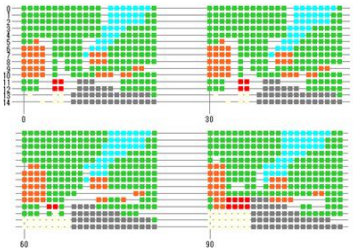
$$p(l_{ik} = z_k | \alpha_i, \beta_k) = \frac{1}{1+e^{-\alpha_i \beta_k}} \tag{11.1}$$

EM アルゴリズムで対数尤度を最大化する際，E-step は次式で表現される：

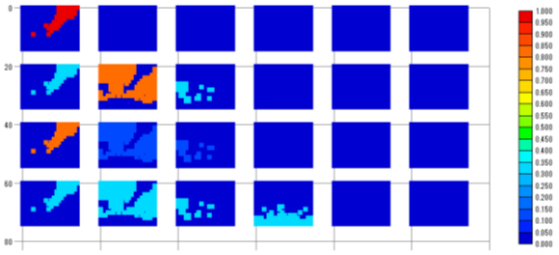
(a) Original image.



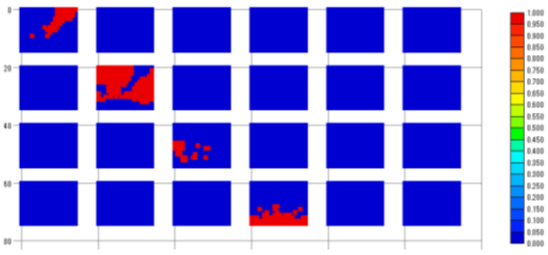
(b) Recognized results (ET) by MBV.



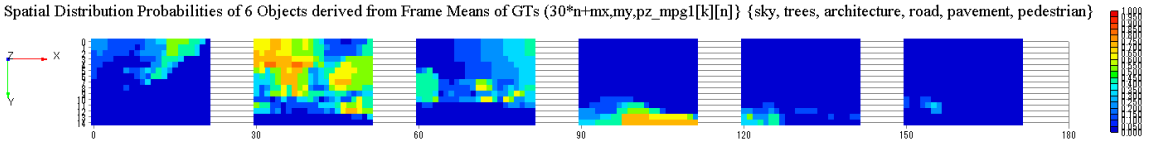
(c) GTs of original images.



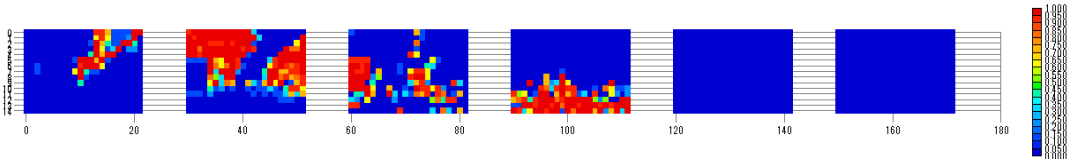
(d) 4 Pseudo ETs for 6 objects (Set-1).



(e) 4 Pseudo ETs for 6 objects (Set-2).



(f) Object existence probabilities according to the GTs(c).



(g) Pseudo object existence probabilities derived from ETs(b).

図 11.1 ET と GT による実験データセット
Fig. 11.1 Experimental data set of ETs and GTs.

$$p(z_k|\boldsymbol{l},\boldsymbol{\alpha},\boldsymbol{\beta}) = p(z_k|\boldsymbol{l}_k,\boldsymbol{\alpha},\beta_k) \propto p(z_k|\boldsymbol{\alpha},\beta_k)p(\boldsymbol{l}_k|z_k,\boldsymbol{\alpha},\beta_k)$$
$$\propto p(z_k)\prod_i p(l_{ik}|z_k,\alpha_i,\beta_k) \tag{11.2}$$

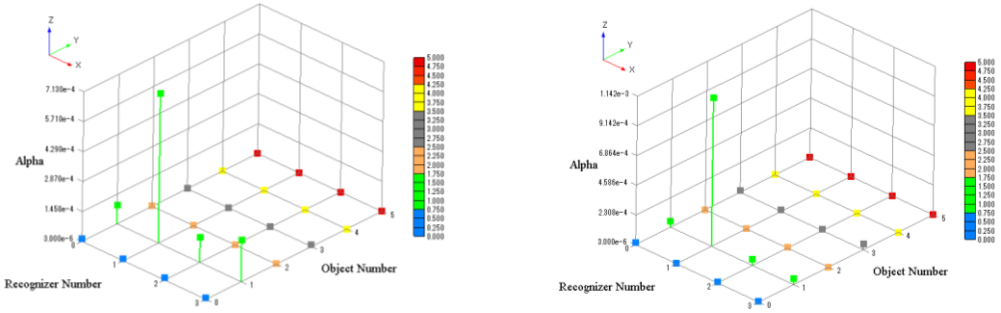
また，M-step は次式で表される：

$$Q(\boldsymbol{\alpha},\boldsymbol{\beta}) = E\left[ln\{p(\boldsymbol{l},\boldsymbol{Z}|\boldsymbol{\alpha},\boldsymbol{\beta})\}\right] = E\left[ln\{\prod_k (p(z_k)\prod_i p(l_{ik}|z_k,\alpha_i,\beta_k))\}\right]$$
$$= \sum_k E\left[ln\{p(z_k)\}\right] + \sum_{ik} E\left[ln\{p(l_{ik}|z_k,\alpha_i,\beta_k)\}\right] \tag{11.3}$$

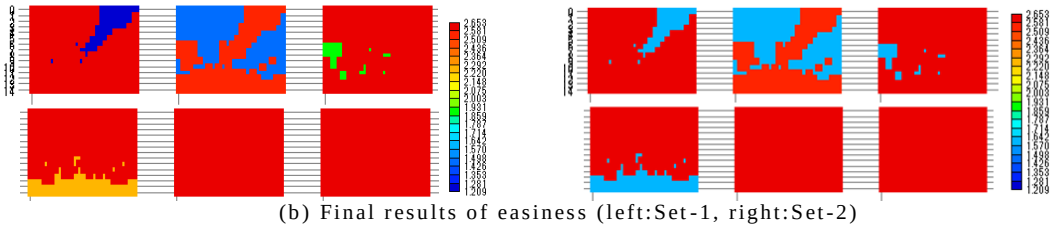
(1)式の確率モデルに従って，実際的な計算式は以下のようになる：

$$Q(\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_k \sum_{g=0}^{1} p^g \, ln\{p(z_k = g)\} + \sum_{ik} \sum_{g=0}^{1} p^g \, ln\{p(l_{ik}|z_k = g,\alpha_i,\beta_k)\} \tag{11.4}$$

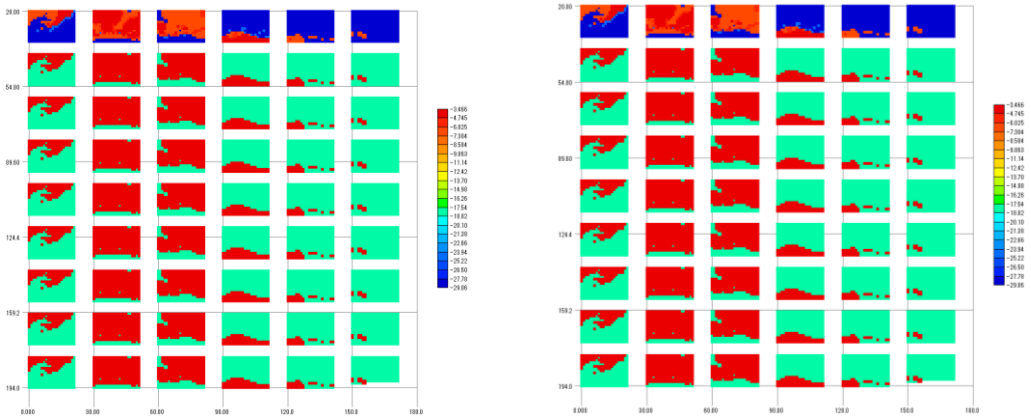
ただし，



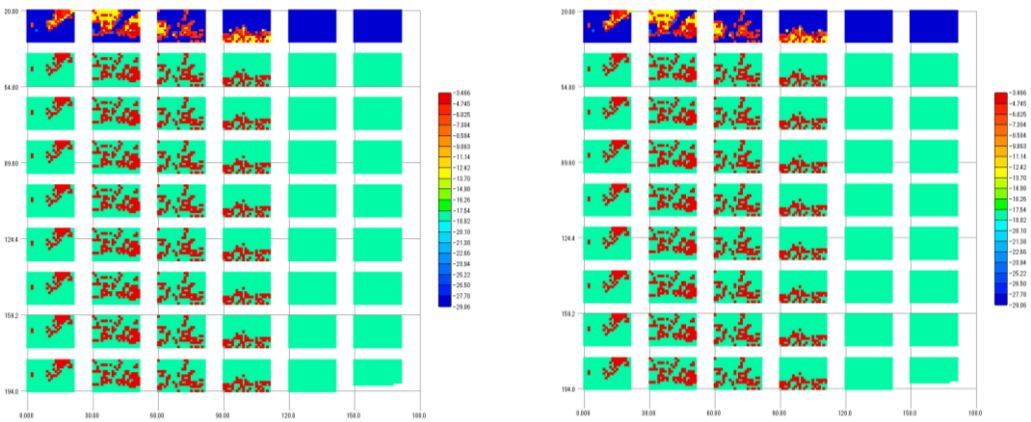
(a) Final results of accuracy (left:Set-1, right:Set-2).



(b) Final results of easiness (left:Set-1, right:Set-2)



(c) EM evolution of log-likelihoods with GT based a priori probabilities (left:Set-1, right:Set-2).



(d) EM evolution of log-likelihoods with ET based a priori probabilities (left:Set-1, right:Set-2).

Figure 11.2: Experimental results.

$$p^g = p\left(z_k = g \middle| \boldsymbol{l}, \boldsymbol{\alpha}^{old}, \boldsymbol{\beta}^{old}\right). \tag{11.5}$$

である．よって，次式の偏微分方程式を得る．

$$\begin{cases} \dfrac{\partial Q}{\partial \alpha_i} = \sum_k \left(p^1 l_{ik} + (1-p^1)(1-l_{ik}) - \sigma \right) \beta_k \\ \dfrac{\partial Q}{\partial \beta_k} = \sum_i \left(p^1 l_{ik} + (1-p^1)(1-l_{ik}) - \sigma \right) \alpha_i \end{cases} \tag{11.6}$$

ここで，次式で定義するロジスティック関数を用いた：

$$\sigma = \frac{1}{1+e^{-\alpha_i \beta_k}} \tag{11.7}$$

以上により，各サイクルの M-step において最急勾配法を適用でき，対数尤度の評価関数 Q について最大値探索を行うために Q を直接計算する必要はなくなる．

Figure 11.1 (a) はテスト画像の第 1 フレームである．Figure 11.1(b) は 9 章で示した MBV の認識結果である．Figure 11.1(c) は第 1 フレームに関する MBK 単位の GT であり，22×15 のインデックスに相当する．マクロブロック(16×16 pixels)上の各カラーインデックスは対応するオブジェクトカテゴリを示す．今回，{空，樹木，建物，道路，歩道，歩行者} の 6 カテゴリを用いた．Figure 11.1(d) は 22×15MBK に対する 4 個の認識器の空間的認識特性を示す．各 MBK の色は実験用に設定した認識精度を示す．そして，4 個の認識器の各ブロックラインは大局的進化における実験用 CV を表している．あるいは，車載器における局所的進化を想定するならば車 1 台に搭載された複数の認識器と考えることもできる．Figure 11.1(f) はオブジェクトの存在確率であり，10 フレームの GT に関するフレーム平均として計算した．Figure 11(g) は ET から算出した擬似オブジェクト存在確率である．

初期学習や定期的保守および認識特性の調整においては GT を用いることが望ましい．通常，その作業は非常に高いコストを要する．ターゲットオブジェクトを最も信頼度の高い，たとえば交通標識などを選べばコストに見合う可能性はある．しかし ET を学習して対応するという厳しい観点からみれば，ET ベースの先験確率に焦点を当てるべきである．

認識タスクという観点からは，少なくとも 4 つのケースがある．第 1 のケース（Case-1）は先験確率も集合処理も行わず，各認識器が独立して動作し，何ら事前知識を用いない場合である．第 2 のケース（Case-2）は先験的確率を用いない場合の集合処理（GLAD ベースの EM 進化）である．第 3 のケース（Case-3）は先験的確率を用いるが集合処理は行わない場合である．最後のケース（Case-4)は先験的確率も集合処理も用いる場合である．

Figure 11.2 では Case-4 に関して Set-1 認識器群と Set-2 認識器群の 2 つを適用した場合の結果を視覚的に表示した．Figure 11.2(c)(d)では Set-1 と Set-2 の間で視覚的および空間的差異は認められないように見えるが，明らかに微小ながら数値上の差異が存在する．同様に，時間的変化に関しても，第 2 回以降の EM 反復においても各 Set に関して小さい規模の数値的収束を確認した．

Table 11.1 は Figure8(d)(e)と Figure 9 に示した特性を持つ 10 セットの認識器群に関する数値的な実験結果を要約したものである．Case-1 と Case-3 に関しては各数値は一つの認識器群における平均確率を意味する．Case-0 は初期の ET0 であり，MBV の出力から与えらえる．

実際応用の観点からいうと，{Case-1, Case-2, Case-3, Case-4, followed by some feedbacks to update appropriate properties and cooperation members (recognizers), ...} が最もありそうな進化列の一つである．これは Table 1 におおまかにその傾向が出ている．

数値性能を綿密に考察すると，Case-4 は必ずしも Case-3 を上回ってはいないことがわかる．こ

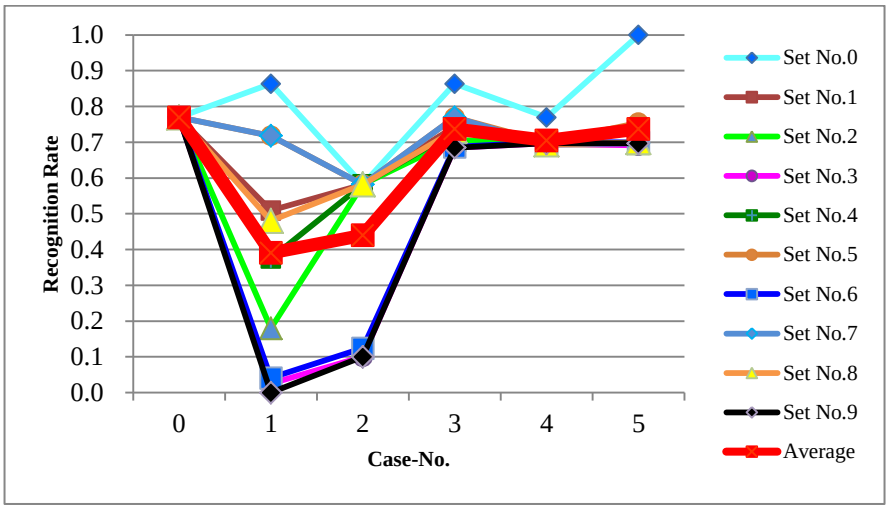


Fig. 11.3 Case based changes of recognition rates.

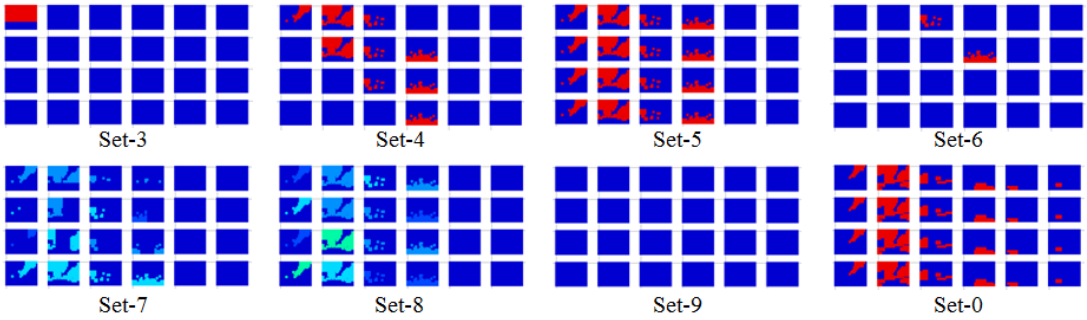


Fig. 11.4 Configuration of recognizers.

の結果に関しては以下のような点が考察される：

1) 各擬似認識器群の構成:

- 認識能力の空間分布は基本的には MBV ベースの ET0（Figure 7(b)）から作られたものであり，その認識率は 0.76979 である．
- ４つの認識器の選択と組み合わせによる影響が大きい

2) 先験確率を導入する際のアルゴリズム的戦略:

本章では Case-3 における戦略として，“各 MBK に関して，認識能力の数値（確信度に相当）が未定義またはゼロに等しいならば，あるいは事前に設定した閾値を下回るならば，事前確率に置き換える”を用いた．ここで，事前確率とは先験確率と同義だが，GT と比較して得た絶対的な（100%の信頼度という意味）数値ではなく，あくまでも各認識器の能力に基づいた推定値としての認識結果（確信度）であり，ET と同等である．しかしながら，上記の戦略は GT-less の集合知を考える際には実際上不可能である．それは各認識器に関する先験情報がシステムにおいて正しく得られる保証がないからである．

3) 事前確率の信頼性と確信度:

本システムでは事前確率そのものは過去に基づく集合知であると考える．図 11.3 では ET0 ベースの事前確率（Case-0）がかなり強力かつ支配的である．このような状況ではシステムは，ある MBK について ET0 が正しくないような場合でも新しい更新を棄却してしまう．そこで，事前確率 $\boldsymbol{c}_A$ を認識器の出力する確信度 $\boldsymbol{c}_{RR}$ と ET0 ベースの確信度 $\boldsymbol{c}_{ET0}$ との荷重線形和として計算

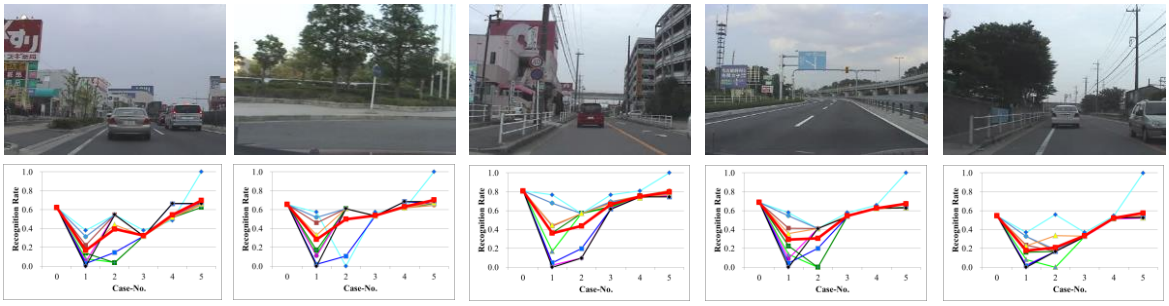


Figure 11.5: Evolution patterns extracted from other scenes.

する Case-5 を新規に追加する：

$$\boldsymbol{c}_A = w_0 \boldsymbol{c}_{RR} + (1 - w_0)\boldsymbol{c}_{ET0} \tag{11.8}$$

今回，w_0=0.5 とした．図 11.5 で示されるように Case-5 は認識率を改善し，ローカルの進化過程においてはすべてのシーケンスについて同じパターンを確認することができた．.

4) 集合知の意味:

クラウド中で最高性能の認識器の出力を ET_{best} とすると，クラウドで想定する集合結果 ETC は ET_{best} の値を向上させる貢献ものではないかもしれない．しかしながら，最低性能のメンバーやそれに類似するレベルのメンバーにとっては ETC は間違いなく性能を向上させる．この効果は各メンバーの性能や優先順位が事前によく知らされていない場合に非常に重要である．それは GT が獲得されないことを想定しているからである．

結論として，より際立った効果を GT-less 集合知に関して検証するにはより多くの数のテスト認識器群と ET を用意する必要がある．

プライバシーの問題はさておき，もしフルスケールの人の流れに相当する有限個の人の数の制約が保存則などにより良好にインストールされれば，マルチトラッキングのタスクにおける解空間は組み合わせ爆発を回避できるであろう．そして ET の探索プロセスは良好にトレーニングされて進化した推論能力のもとではかなり削減されるであろう．

通常，もしターゲットエリアの交通量が田舎や郊外の居住地域のようにきわめて低いならば，同時に発生するイベントについてさほど多くの数の車載認識器を期待できない．そこで，シーンの周辺部について過去の周期的記憶を活用する案が浮上してくる．このアイデアは静的環境や周期的な人の流れについて有効である．だが，特殊な周期的性質を持たないような移動体については持たない過去の情報を収集することが困難となる．そこで，Neuro-ITS アーキテクチャ[A16]における“予測からセンシング制御へ”に焦点があてられる．そのもっとも簡単な実現形態は，単体の車載のセンシングデータストレージを活用した過去との集合知である．これはカーナビやドライブレコーダにおいて容易に導入できる．この“過去”の概念をフレーム間の時間スケールに拡張すれば，もう一つのミクロスケールの local evolution が車載の複数の認識器において考えられる．これはビデオトラッキングにおいて効果を発するとみられる．

12. おわりに

本論文では MPEG で符号化された画像に含まれる特徴量を利用してブロック画素単位でインデキシングを行うシステムをめざして基本理論を構成し，交通映像と景観映像に関する実験結果と考察を述べた．MPEG 映像はすでにディジタル放送，DVD，インターネット，PC などで膨大な使用実績があると同時に多種多様なアーカイブが構成されている．携帯やスマフォなどモバイル通信とも親和性が高く，付与されるメタデータも近年，G-空間技術の進歩とともにさらにリッチになっている．この結果，複数視点による集合知の形成も可能になり，本稿の後半で述べたような集合知ベースの画像理解も今後進むと見られる．MPEG 画像特徴量はブロック画像単位に得られるので，これまで歩行者認識などで研究されてきた画素単位の特徴量に比べて認識への応用価値は低いと思われてきた．しかし，近年普及が急速に進むハイビジョン映像では本稿の実験で用いた MPEG-1 画像の 10 倍以上のブロック数を扱っており，ブロック画像単位で構成する認識結果はそれだけ空間分解能が上がっている．したがって，MPEG Based Vision の認識結果は MPEG の空間スケーラビリティによる高解像度化とともに従来の VGA クラスの画素単位の認識結果に近づいている．一方で，画素単位に見せることもブロック画像単位の結果の後処理として可能であるため，景観認識などの応用では実用上の差異はほぼなくなるであろう．

一方で，移動体認識や 100%近い認識精度を要求される交通標識の認識を無記憶の単独の認識器で行うには専用の画像特徴量と認識器とを用いた画素単位のパターン認識が必要である．しかし，冒頭で述べたように，人間の目では，MPEG 復号画像を見て細かい判別を行っているため，MPEG 特徴量にも可能性は残されている．むしろ，過去の認識結果をシーンクラスなどで体系化するとともに GLAD のような手法で集合知化し，高精度の認識結果を進化させていくメカニズムをシステムに持たせることのほうが無記憶の画像認識器を発展させることよりも労力は少ない．蓄積と通信に優れ，世の中に湯水のごとくゆきわたった MPEG 映像をあえてビットマップに戻して新たに無記憶の高精度認識器に適した特徴量に変換するということがペイするケースとペイしないケースがあることを考えるべきであろう．

さらに，本論文後半で述べたように認識対象の出現予測と連携させることにより，全体としての認識能力は飛躍的に向上すると見られ，ここでも単なる信号処理ではなく状況や予測に適した知能処理を画像理解の本質ととらえることが重要とみられる．

本論文では景観画像が主として MBV の対象となったが，MBV の応用範囲はグラフィクス画像，アニメーション医用画像，顕微鏡画像，天体観測画像，エネルギー分布など物理;・化学分野での特性把握，人の移動・環境・気象・社会事象・経済現象を含む空間情報科学分野の時空間特性の認識においても有用とみられる．特に実時間事象については蓄積・通信に適した性能を活かし，携帯・スマフォ，プローブカー，ロボットなどとの連携により，データ同化を組み入れることでさらに MBV の活用範囲は広がると考えられる．